Distribuição Normal Assimétrica para Dados de Expressão Gênica

Priscila da Silva Gomes

Distribuição Normal Assimétrica para Dados de Expressão Gênica

Priscila da Silva Gomes

Orientadores: Dra. Vera L. D. Tomazzela

Dr. Francisco Louzada-Neto.

Trabalho apresentado ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

UFSCar - São Carlos

Maio/2009

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

G633dn	Gomes, Priscila da Silva. Distribuição normal assimétrica para dados de expressão gênica / Priscila da Silva Gomes. -- São Carlos : UFSCar, 2009. 62 f. Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2009. 1. Estatística matemática. 2. Expressão gênica. 3. Distribuição normal assimétrica. I. Título. CDD: 519.5 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística
Via Washington Luís, Km 235 - C.P.676 - CGC 45358058/0001-40
FONE: (016) 260-8292/260-8241 - FAX: (016) 260-8243
13565-905 - SÃO CARLOS-SP-BRASIL

ATA DO EXAME DE DISSERTAÇÃO DE MESTRADO DA CANDIDATA:
Priscila da Silva Gomes

Aos vinte e sete dias do mês de fevereiro do ano de dois mil e nove, às nove horas, na Sala de Reuniões do Departamento de Estatística, reuniu-se a Comissão Examinadora nas formas e termos do Artigo 25º do Regimento Interno do Programa de Pós-Graduação em Estatística da UFSCar, composta pelos membros: Profa. Dra. Vera Lucia Damasceno Tomazella (DEs-UFSCar, Orientadora), Prof. Dr. Francisco Louzada Neto (DEs-UFSCar, Orientador), Profa. Dra. Júlia Maria Pavan Soler (IME-USP) e Prof. Dr. Luis A. Milan (DEs-UFSCar), para Exame de Dissertação de Mestrado da candidata Priscila da Silva Gomes, sob o título "Distribuição Normal Assimétrica para Dados de Expressão Gênica". A sessão foi aberta pela Profa. Dra. Vera Lucia Damasceno Tomazella (Presidente), iniciando-se pela apresentação da dissertação. Em seguida, foi feita a argüição da candidata pelos membros da Comissão Examinadora. A Comissão Examinadora considerou o tema relevante para Estatística e julgou a exposição feita pela candidata clara e objetiva. A candidata respondeu satisfatoriamente as questões formuladas. Pelo apresentado acima, a comissão atribuiu as seguintes avaliações: Profa. Dra. Vera Lucia Damasceno Tomazella, nível A; Prof. Dr. Francisco Louzada Neto, nível A; Profa. Dra. Júlia Maria Pavan Soler e Prof. Dr. Luis A. Milan, nível A. De acordo com o parágrafo 5º do Artigo 25º, a candidata foi considerada **aprovada**. Encerrada a sessão secreta, o Presidente informou o resultado da defesa. Nada mais havendo a tratar, eu, Maria Isabel Rinaldo Pessôa de Araujo, Secretária deste Programa, lavrei a presente ata, que assino juntamente com os membros da Banca Examinadora.

Profa. Dra. Vera Lucia Damasceno Tomazella

Prof. Dr. Francisco Louzada Neto

Profa. Dra. Júlia Maria Pavan Soler

Prof. Dr. Luis A. Milan

Maria Isabel R. P. Araujo

Agradecimentos

Gostaria de agradecer, aos meus pais pela paciência e por sempre me apoiarem em todas as escolhas feitas no decorrer da minha vida, certamente sem eles não estaria aqui.

Aos meus amigos, que sempre estiveram presentes me ajudando a superar a pressão e vencer os obstáculos que surgiram no meio do caminho.

Especialmente ao André, pelo amor, amizade e total apoio, compartilhados nas pacientes horas de trabalho, e por estar sempre ao meu lado me dando força e me incentivando a não desistir.

Aos meus orientadores, Vera e Neto, pela orientação e incentivo na elaboração e condução do trabalho.

Ao professor Luis Milan, membro da banca do exame de qualificação, pelas diversas conversas e pelas sugestões feitas.

Finalmente, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio concedido para a realização deste trabalho.

Resumo

Os *microarrays* são ferramentas utilizadas para medir os níveis de expressão de uma grande quantidade de genes ou fragmentos de genes simultaneamente em situações variadas. Com esta ferramenta é possível determinar possíveis genes causadores de doenças de origem genética. Uma abordagem estatística comumente utilizada para determinar se um gene g apresenta evidências para níveis de expressão diferentes consiste no teste t, que exige a suposição de normalidade aos dados (Saraiva, 2006; Baldi & Long, 2001). No entanto, esta suposição pode não condizer com a natureza dos dados analisados. Neste trabalho, será utilizada a distribuição normal assimétrica descrita formalmente por Azzalini (1985), que tem a distribuição normal como caso particular, com o intuito de flexibilizar a suposição de normalidade. Considerando a abordagem clássica, é realizado um estudo de simulação para detectar diferenças entre os níveis de expressão gênica em situações de controle e tratamento através do teste t, também é considerado um estudo de simulação para analisar o poder do teste t quando é assumido um modelo assimétrico para o conjunto de dados. Também é realizado o teste da razão de verossimilhança, para verificar se o ajuste de um modelo assimétrico aos dados é adequado.

Palavras-chave: Distribuição Normal Assimétrica, Teste t, Teste da Razão de Verossimilhança, *Microarray*, Expressão Gênica.

Abstract

Microarrays technologies are used to measure the expression levels of a large amount of genes or fragments of genes simultaneously in different situations. This technology is useful to determine genes that are responsible for genetic diseases. A common statistical methodology used to determine whether a gene g has evidences to different expression levels is the t-test which requires the assumption of normality for the data (Saraiva, 2006; Baldi & Long, 2001). However this assumption sometimes does not agree with the nature of the analyzed data. In this work we use the skew-normal distribution described formally by Azzalini (1985), which has the normal distribution as a particular case, in order to relax the assumption of normality. Considering a frequentist approach we made a simulation study to detect differences between the gene expression levels in situations of control and treatment through the t-test. Another simulation was made to examine the power of the t-test when we assume an asymmetrical model for the data. Also we used the likelihood ratio test to verify the adequability of an asymmetrical model for the data.

Keywords: Skew-normal distribution, t-test, likelihood ratio test, microarray, genic expression.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Nos últimos anos, uma enorme massa de dados tem ficado disponível para que biólogos moleculares, bioquímicos e outros pesquisadores possam analisá-los. O desenvolvimento de computadores capazes de armazenar e gerenciar um grande número de informações e, as descobertas no campo da biologia molecular tornaram possível a criação de bancos de dados genéticos e algoritmos de mineração de dados para analisar proteínas, genes e coleções completas de DNA (ácido desoxirribonucleico) em um nível genômico, tendo como objetivo principal a produção de conhecimentos para a compreensão do código da vida (DNA).

O rápido desenvolvimento de tecnologias biológicas na última década, fez com que pesquisadores e laboratórios trabalhassem em conjunto na descoberta do sequenciamento genético de diversos organismos vivos, além de disponibilizarem estas informações em bancos de dados públicos, permitindo assim, o estudo de processos biológicos relacionados ao genoma.

O genoma humano é a impressão digital de todas as estruturas e atividades celulares no corpo humano. Ele é composto por 23 pares de cromossomos, no qual cada par possui um cromossomo proveniente da mãe e outro do pai. Cada cromossomo é formado por uma longa cadeia de duas fitas de DNA chamada de hélice dupla (Schwender *et al.*, 2006).

Os genes são segmentos de DNA e carregam a informação necessária para a síntese de proteínas, e estas, atuam de forma vital no sistema biológico de um ser vivo, por isso

é importante entender como os genes são traduzidos em proteínas.

Através do processo de pareamento de bases complementares, o DNA é transcrito em RNA mensageiro (mRNA). O mRNA deixa o núcleo da célula e é traduzido em proteína. O processo de conversão de uma sequência de DNA em proteína é conhecida como *expressão gênica* (Schwender *et al.*, 2006; Lee, 2004). O estudo da expressão gênica torna possível a compreensão de um estado alterado do gene que pode causar por exemplo, câncer e doenças genéticas.

Uma das principais ferramentas para o estudo da expressão gênica são os arranjos de DNA, ou simplesmente, *microarrays*. A tecnologia dos *microarrays* permite a identificação e quantificação da expressão gênica, simultaneamente, para todos os genes no organismo analisado. O objetivo principal desta técnica é medir a concentração de mRNA em células ou tecidos de interesse.

O princípio dos arranjos de DNA começou a ser abordado em meados 1991, sendo Lennon e Lehrach alguns dos pioneiros (Lennon & Lehrach, 1991). Porém, esta técnica passa a ser disseminada anos à frente, em trabalhos como os de Nguyen & Smart Jr (1995) e Zhao (1995).

Os *microarrays* são matrizes compostas por pequenas placas de vidro ou nylon, nas quais seus elementos (*spots*) são moléculas de DNA. O *microarray* tem como função explorar a habilidade que uma dada molécula de cDNA (versão transcrita do mRNA) tem de se hibridizar a uma sequência complementar no *array*. Através do uso dos *microarrays* os cientistas podem medir, em um único experimento, os níveis de expressão de centenas ou milhares de genes dentro de uma célula, tecido ou organismo, medindo a quantidade de cDNA que se hibridizou a cada *spot* do *array* (Causton *et al.*, 2003).

Os dados numéricos (relacionados às medidas dos níveis de expressão dos genes) obtidos são positivos e possuem grande variabilidade. Para que a análise e intrepretação dos dados seja feita é necessário o uso de métodos estatísticos. Os primeiros métodos estatísticos descritos para a análise de dados de expressão gênica foram discutidos por Schena *et al.* (1995), Schena *et al.* (1996) e DeRisi *et al.* (1996).

Atualmente, é possível encontrar diversos artigos que utilizam diferentes métodos estatísticos para a análise de dados de expressão gênica, entre eles, Baldi & Long (2001)

com a utilização do teste t e uma abordagem bayesiana e Efron *et al.* (2001) através de uma abordagem bayesiana empírica. Estes trabalhos possuem uma característica em comum, que é a suposição de normalidade aos dados transformados.

O objetivo deste trabalho é flexibilizar essa suposição utilizando a distribuição normal assimétrica (Azzalini, 1985), que possui como caso particular a distribuição normal.

Para análise da expressão gênica, serão considerados as situações controle (por exemplo, medidas de níveis de expressão das sondas de cDNA's provenientes de células normais) e tratamento (medidas de níveis de expressão das sondas de cDNA's em qualquer situação diferente do controle). Porém, na literatura é possivel encontrar a análise de dados de expressão gênica para mais de dois grupos, ver por exemplo Kerr & Churchill (2001).

Este trabalho está organizado da seguinte forma: no Capítulo 2 será detalhado o processo para obtenção da expressão gênica de um determinado gene; o Capítulo 3 apresenta a distribuição normal assimétrica e algumas de suas propriedades probabilísticas; no Capítulo 4, é realizado um estudo de simulação utilizando o teste t para detectar diferenças entre os níveis de expressão gênica em situações de controle e tratamento, quando se supõe que um dos mesmos segue uma distribuição normal assimétrica. Neste capítulo também é verificado o poder do teste t, e é realizado o teste t com um conjunto de dados reais. O Capítulo 5 apresenta o TRV (Teste da Razão de Verossimilhança), onde é verificada a adequabilidade de um modelo assimétrico aos dados; no Capítulo 6 são apresentadas as considerações finais a respeito do que foi estudado e também as propostas futuras.

Capítulo 2

Expressão Gênica

A célula é a menor unidade de vida presente em um organismo e, existe um grande número de transformações químicas específicas, que não apenas fornecem a energia necessária à célula mas também, coordena todos os eventos e atividades dentro dela. O processo da vida envolve uma coleção de moléculas desde a água à compostos orgânicos (por exemplo, gorduras e açúcares), e macromoléculas (DNA - ácido desoxirribonucleico, proteínas, e polisacarídeos) que definem a estrutura das células. As macromoléculas controlam e governam a maioria das atividades presentes em um ser vivo. As moléculas de DNA armazenam informações sobre a estrutura das macromoléculas, permitindo que elas sejam feitas, precisamente, de acordo com as especificações e necessidades da célula (Lee, 2004; Griffiths *et al.*, 2004).

As células somáticas (responsáveis pela formação do corpo de um organismo) de quase todas as plantas e animais contêm duas cópias de seus respectivos genoma e, os organismos com essa característica são conhecidos como diplóides. O genoma é formado por uma ou mais moléculas de DNA extremamente longas, que são organizadas em cromossomos. Isto é, cada cromossomo é composto por duas longas fitas de DNA formando uma hélice dupla. Em organismos diplóides, cada cromossomo está presente em dobro. Por exemplo, as células somáticas humanas contêm dois conjuntos compostos por 23 cromossomos, resultando num total de 46 cromossomos. Dois cromossomos com a mesma coleção de genes são chamados de homólogos.

As fitas de DNA são compostas por sequências de nucleotídeos, onde cada nu-

cleotídeo é composto por um grupo fosfato, açúcares e uma das quatro bases nitrogenadas - adenina (A), timina (T), citosina (C) e guanina (G). Para descrever o DNA, basta saber a sequência de um das fitas, a razão para este fato é o pareamento de bases complementares: a base A em uma fita está sempre ligada (via ligações de hidrogênio) à base T na fita oposta, enquanto C está sempre ligada à G.

As moléculas de DNA são responsáveis pela conversão da informação necessária para a produção de proteínas ou moléculas de RNA (ácido ribonucleico) encontradas em um organismo. Esse processo, pode ser descrito nos passos a seguir (Crick, 1970):

- **Replicação:** a informação contida no DNA é duplicada, conservando a informação genética.
- **Transcrição:** permite a passagem da informação contida no DNA para o RNA menssageiro (mRNA), através do emparelhamento de bases complementares, com exceção de uma base pois, no RNA a base timina (T) é substituída pela base uracila (U).
- **Tradução:** possibilita a síntese de proteínas mediante a informação contida no mRNA.

As proteínas, por sua vez, atuam de forma vital para o funcionamento das células. Porém, no processo de tradução, apenas algumas partes dos genes - os *exons* - são necessárias. É por isso, que no processo de transcrição, os *introns* - partes não codificantes dos genes - são removidos através do *RNA-splicing* (junção do RNA). No entanto, os mesmos *exons* nem sempre são conservados. Então, diferentes combinações dos *exons* de um gene, podem ser unidas para produzir diferentes mRNAs, que podem levar a síntese de diferentes proteínas. Acredita-se que cerca de 60 % dos genes são afetados por essa união do RNA (para mais detalhes ver, Schwender *et al.*, 2006; Griffiths *et al.*, 2004; Zhang, 2006).

A *expressão gênica* pode ser descrita como o processo pelo qual o mRNA e, eventualmente a proteína são sintetizados a partir da sequência de DNA referente ao gene (Lee, 2004). O entendimento de quais genes são expressos e sob quais circunstâncias, fornece informações valiosas sobre os processos biológicos que acontecem nas células. A

partir disso, é possível estudar o que acontece em um estado alterado do gene, como por exemplo, câncer.

2.1 Medindo Expressão Gênica

Uma das principais ferramentas para o estudo da expressão gênica são os arranjos de DNA (*DNA array*) ou *microarrays* (microarranjos). Os *microarrays* tornam possível a quantificação da expressão gênica, medindo a hibridização ou combinação do DNA fixado em uma pequena lâmina de vidro, plástico ou mebranda de nylon, ao mRNA referente a amostra sob estudo. Dessa forma, a tecnologia de *microarray* permite que o nível de expressão de milhares de genes contidos no genoma possa ser medido simultaneamente (Müller & Nicolau, 2004; Lee, 2004). Tal habilidade para medir simultaneamente grandes proporções de genes no genoma, abre as portas para a investigação de iterações entre os genes numa grande escala, permitindo a descoberta do papel desempenhado pelos genes cujas funções ainda são desconhecidas.

Existem diferentes tecnologias de *microarrays*, entre elas estão os arranjos de cDNA e os arranjos de oligonucleotídeos. Embora o objetivo de ambas as tecnologias seja a hibridização, elas diferem no modo em que as sequências de DNA são colocadas no *array* e no tamanho dessas sequências. Sobek *et al.* (2006) e Hoheisel (2006) descrevem em detalhes as atuais tecnologias de *microarrays* e suas diferentes aplicações nos dias atuais.

Nos *microarrays* de cDNA, o mRNA de duas amostras biológicas diferentes (por exemplo, controle e tratamento) é transcrito de modo reverso em cDNA, marcado com diferentes fluorescências (neste caso, verde e vermelho) e então, colocado na lâmina de vidro onde é hibridizado as sequências de DNA presentes na mesma. Após a hibridização, a lâmina é escaneada medindo a intensidade da fluorescência de cada cor, uma maior fluorescência indica uma quantidade maior de cDNA hibridizado, ou seja, indica uma expressão gênica maior na amostra. Uma descrição mais detalhada sobre os *microarrays* de cDNA pode ser encontrada em Schena *et al.* (1995) e DeRisi *et al.* (1996).

As próximas subseções apresentam as etapas de preparação dos *microarrays* de cDNA, alguns tipos de transformações aplicadas aos dados e uma breve descrição de

técnicas estatísticas na análise dos mesmos.

2.1.1 Etapas da Preparação dos *Microarrays* de cDNA

O processo para a fabricação do *microarray* de cDNA e a obtenção dos dados para a análise pode ser descrito por:

1. **Fabricação do arranjo:** roboticamente é feita a deposição de sequências de cDNA conhecidas em posições específicas de uma lâmina de vidro;
2. Preparação das amostras biológicas a serem estudadas, uma correspondente à situação padrão ou controle e a outra correspondente à situação que se deseja estudar (tratamento).
3. Extração do mRNA de cada amostra.
4. **Marcação do mRNA:** através de transcrição reversa a amostra controle é marcada com cianina 3 (Cy3) - verde - fluorescente e a amostra do tratamento é marcada com cianina 5 (Cy5) - vermelho - fluorescente;
5. Ambas as amostras são colocadas no *microarray*, resultando na hibridização das mesmas com as sequências de cDNA contidas no arranjo.
6. É feito o escaneamento do *microarray* hibridizado.

Cada sonda de cDNA hibridizará apenas com um dos mRNA recolhidos das amostras controle ou tratamento. A intensidade e a cor final da hibridização das sondas depende do nível de concentração do mRNA de cada amostra. O resultado do experimento são duas imagens com pontos iluminados, uma com pontos verdes e outra com pontos vermelhos. A imagem dos pontos é salva e processada, onde cada *spot* do arranjo recebe valores numéricos referentes à intensidade de sua cor. Este processo pode ser visualizado na Figura 2.1.

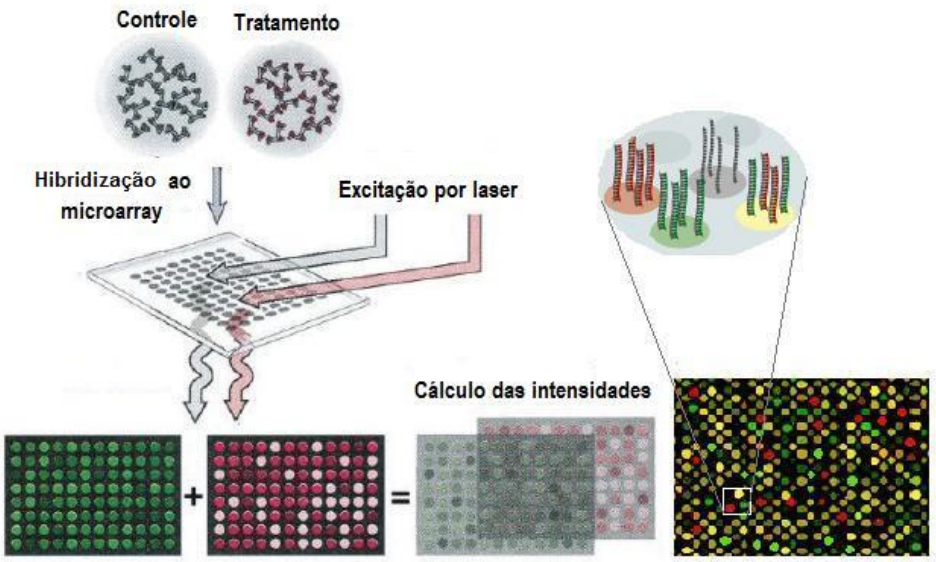

FIGURA 2.1: Experimento com arranjos de DNA.

Os dados obtidos através de experimentos de *microarray* possuem duas características básicas, o significado biológico e o significado estatístico. O significado biológico é o que interessa aos pesquisadores e mostra o quanto a expressão de um gene é influenciada pela condição sob estudo. Já o significado estatístico mostra o quão confiável é o significado biológico. Devido as grandes fontes de variação nos experimentos de *microarray* (Kerr & Churchill, 2001; Zhang, 2006), a análise estatística é crucial para o sucesso na interpretação do fenômeno biológico sob estudo. Por essa razão, é necessária a compreensão do formato dos dados de *microarray*, ou seja, saber interpretá-los para que uma análise estatística significativa seja feita.

O estudo de dados de expressão gênica podem envolver as seguintes etapas:

- A identificação de genes diferencialmente expressos, isto é, genes cujo o nível de expressão difere fortemente entre grupos (por exemplo, câncer *versus* não câncer).
- Encontrar genes cujos níveis de expressão variam juntos (grupos de genes - *clustering genes*).
- Construir uma regra de classificação baseada nos dados de expressão gênica, para associar novas observações a uma determinada classe, por exemplo, câncer ou não câncer.

Antes de iniciar qualquer uma dessas análises é necessário processar os dados gerados no experimento de *microarray*. Os dados de expressão gênica obtidos por arranjos de cDNA ou arranjos de oligonucleotídeos são similares em formato. Existem diversos *softwares* (comerciais e livres) que têm disponível pacotes para a análise das imagens e quantificação dos níveis de expressão. De modo geral, os resultados obtidos são compostos por matrizes que incluem a localização dos *spots* (pontos) no *microarray*, identidade dos genes, média e mediana da intensidade dos pixels dentro de cada *spot* e intensidades locais dos *backgrounds* (regiões entre pontos).

A hibridização simultânea de duas amostras diferentes rotuladas com marcadores Cy3 (verde) e Cy5 (vermelho) requer uma análise especial. Os dois marcadores têm diferentes propriedades e sensividades à luz. A emissão de fluorescência do marcador Cy5 tem menor intensidade, por isso os sinais de fluorescência dos dois marcadores têm que ser "normalizados" (Causton *et al.*, 2003; Lee, 2004).

O propósito da normalização, é minimizar a variação nas medidas dos níveis de expressão das amostras de mRNA hibridizadas, de tal modo que, as diferenças biológicas (diferenças na expressão gênica) possam ser melhor distinguidas. Para maiores detalhes sobre o processo de normalização dos dados de expressão gênica ver, Lee (2004); Zhang (2006); Causton *et al.* (2003) e Kerr & Churchill (2001).

2.1.2 Transformação dos Dados

Mesmo após a normalização, as medidas dos níveis de expressão não possuem necessariamente propriedades estatísticas desejadas, como por exemplo, variância constante e normalidade. Por essa razão, a tranformação dos dados é importante. Nesta subseção, serão abordadas algumas transformações que têm se mostrado úteis na análise de dados de *microarray* (para mais detalhes, Lee, 2004).

Alguns Métodos de Transformação dos Dados

- **Transformação logarítmica:**

  É a transformação mais comum aplicada aos dados de expressão gênica,

$$y_{ij} = \log x_{ij},$$

onde $i = 1, 2, \ldots, G$ é referente ao gene em questão e $j = 1, 2, \ldots, n$ é o número de observações para cada gene. O logaritmo utilizado pode ser na base 2, 10 ou na base e (logaritmo natural).

- **Transformação da raiz quadrada:**
  Normalmente é utilizada para estabilizar a variância dos dados, isto é, a variância tenderá a ser uma constante
  $$y_{ij} = \sqrt{x_{ij}},$$
  onde $i = 1, 2, \ldots, G$ é referente ao gene em questão e $j = 1, 2, \ldots, n$ é o número de observações para cada gene.

- **Transformação de Box-Cox:**
  As transformações anteriores são dois membros da família de transformações Box-Cox, que é definida por:
  $$y_{ij} = \frac{x_{ij}^d - 1}{d},$$
  onde $i = 1, 2, \ldots, G$ é referente ao gene em questão, $j = 1, 2, \ldots, n$ é o número de observações para cada gene e $d = \frac{1}{2}$, correspondendo a transformação da raiz quadrada. O limite da função y_{ij}, quando $d \to 0$ corresponde a transformação logarítmica. Quando $d = 1$, nenhuma transformação é adotada, exceto por um pequeno deslocamento.

2.1.3 Técnicas Estatísticas Para a Análise de Dados de Expressão Gênica

Atualmente, é possível encontrar diversos artigos que utilizam diferentes métodos estatísticos para a análise de dados de expressão gênica, entre eles:

- **Teste** t: utiliza-se o teste t para verificar se as médias dos valores de intensidade obtidos para o controle e para o tratamento de um determinado gene g são iguais, isto é para verificar se há diferença no nível de expressão obtido para o controle e o tratamento para cada gene. Artigos como os de Baldi & Long (2001) e de Menezes *et al.* (2004) utilizam o teste t e/ou variações do mesmo, para verificar se existe diferença entre os níveis de expressão dos genes;

- **Abordagem Bayesiana**: uma alternativa aos métodos clássicos, é possível encontrar na literatura diversos artigos que utilizam abordagem Bayesiana para analisar dados de expressões gênica, como Efron *et al.* (2001), que trabalha com uma abordagem bayesiana empírica, e Saraiva (2006), que em seu trabalho utiliza diversos métodos estatísticos, entre eles uma abordagem bayesiana paramétrica e não paramétrica.

- **Análise de Cluster**: técnica de classificação de objetos ou indivíduos em diferentes grupos. Mais precisamente, é uma técnica que particiona um conjunto de dados em subconjuntos ou grupos, de tal forma que os elementos em cada cluster tenham uma ou mais características em comum, que estejam próximos segundo uma medida de distância definida. Essa técnica aplicada a dados de expressão gênica pode ser vista em Kuriakose *et al.* (2004).

Outras técnicas envolvem análise de componentes principais e análise fatorial.

2.2 Considerações Finais

Neste capítulo foi descrito o processo para a obtenção dos dados de *microarray*, assim como técnicas para a análise dos mesmos. Também foram descritos alguns métodos estatísticos para a análise dos dados de expressão gênica. O Capítulo 3 apresenta a distribuição normal e suas propriedades. No final do capítulo é feita uma aplicação utilizando o método *bootstrap*.

Capítulo 3

A Distribuição Normal Assimétrica

A análise estatística para o estudo de dados contínuos tem sido desenvolvida, em grande parte, com base no modelo normal. No entanto, a suposição de simetria para os dados pode fazer com que sejam feitas inferências pouco apropriadas sobre os parâmetros de interesse. Uma alternativa para essa situação seria flexibilizar a suposição de normalidade, considerando uma classe de distribuições assimétricas que possui como caso particular a distribuição normal.

Embora a distribuição normal seja o modelo de probabilidade mais utilizado em estatística, existem muitos fenômenos que não podem ser descritos pela distribuição normal e nem por distribuições simétricas. Aplicações de distribuições assimétricas são comumente encontradas em diversas áreas, tais como: ciências atuárias, demografia, economia, engenharia, ciência ambiental, finanças e ciências médicas (ver, Johnson *et al.*, 1994, 1995; Seshadri, 1999).

A distribuição normal assimétrica (*skew-normal*) foi formalmente introduzida por Azzalini (1985). A partir de então, as propriedades probabilísticas dessa distribuição foram amplamente estudadas. Em Arellano-Valle & Azzalini (2006) é apresentada a parametrização centrada dessa distribuição, corrigindo o problema de singularidade da matriz de informação de Fisher quando o parâmetro de assimetria é nulo. O artigo de Arellano-Valle & Azzalini (2006) mostra a correlação (unificação) das diversas propostas originadas no decorrer dos anos, para o caso multivariado da distribuição normal assimétrica.

Devido à sua contínua variação de normalidade para não normalidade, e por ser matematicamente tratável, esta distribuição tem sido aplicada a diversas áreas (ver por exemplo, Pourahmadi, 2007; Ghosh *et al.*, 2006).

Neste capítulo, é apresentada uma revisão da distribuição normal assimétrica considerando duas representações, uma formalmente introduzida por Azzalini (1985) e outra por Henze (1986), o qual obtém a densidade proposta por Azzalini através de transfomação de variáveis, onde a variável em questão é obtida como combinação linear de variáveis aleatórias independentes. Serão apresentadas também algumas propriedades e métodos de inferência para os parâmetros dessa distribuição.

3.1 Definição e Propriedades

Nesta Seção serão descritas a distribuição normal assimétrica e suas propriedades.

Definição: Uma variável aleatória Z tem distribuição normal assimétrica padrão com parâmetro de assimetria λ, denotada por $SN(\lambda)$, se sua densidade é dada por

$$f(z) = 2\phi(z)\Phi(\lambda z), \quad z \in \Re, \tag{3.1}$$

onde λ é um valor definido em $\Re$, e $\phi(.)$ e $\Phi(.)$ denotam respectivamente, a função densidade de probabilidade e a função de distribuição acumulada da distribuição normal padrão.

A Figura 3.1 mostra o comportamento de $f(z)$ para diferentes valores de λ, e é fácil notar que a distribuição normal padrão está incluída como caso particular da normal assimétrica padrão quando $\lambda = 0$.

A função de distribuição acumulada (fda) da normal assimétrica padrão denotada por $F(z;\lambda)$ é dada por

$$F(z;\lambda) = 2\int_{-\infty}^{z}\int_{-\infty}^{\lambda t}\phi(t)\Phi(u)dudt. \tag{3.2}$$

A seguir serão descritas algumas propriedades (Azzalini, 1985) da distribuição normal assimétrica padrão.

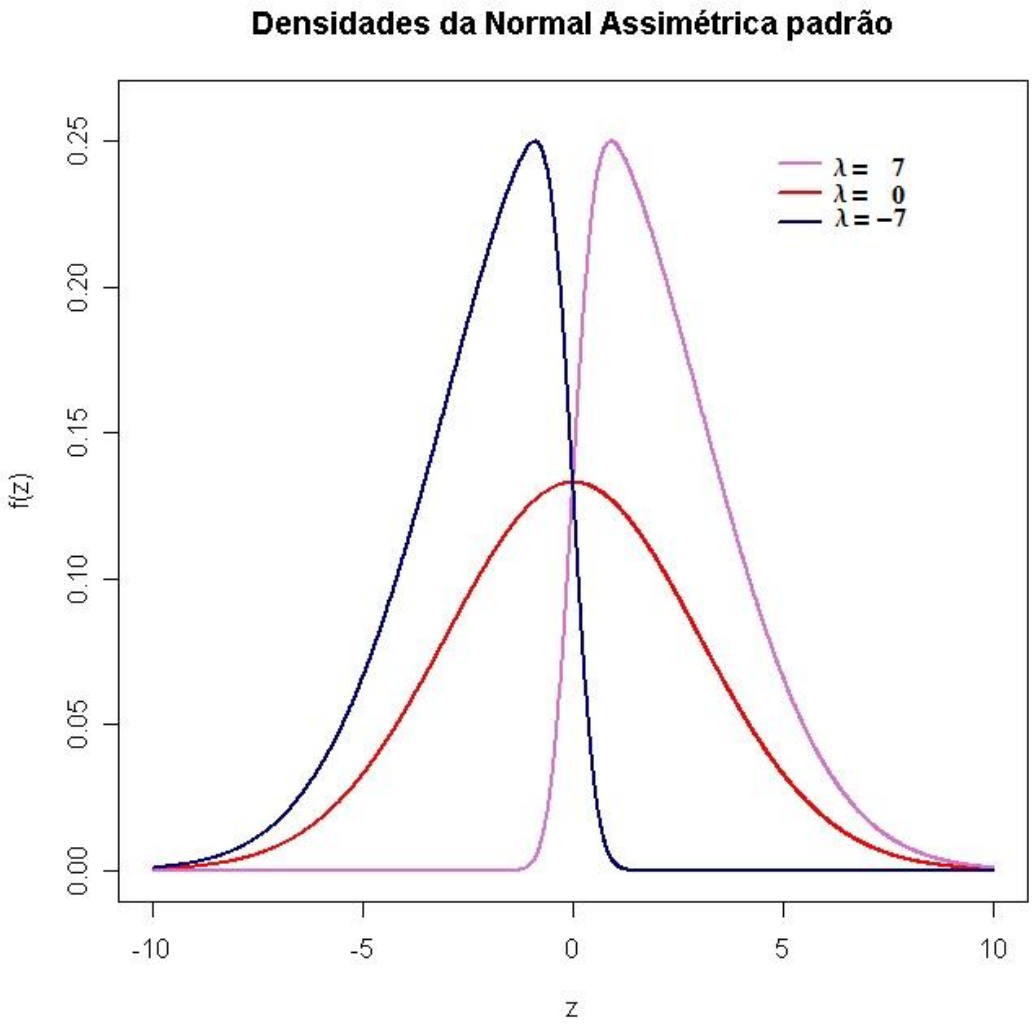

FIGURA 3.1: Gráfico da densidade normal assimétrica padrão para valores diferentes de λ.

Propriedades

1. $SN(0)$ é igual à $N(0,1)$.
2. Se $Z \sim SN(\lambda)$ então $Y = |Z| \sim HN(0,1)$, onde $HN(0,1)$ representa a distribuição denominada *half-normal* com densidade dada por $f(y) = 2\phi(y)I_{[y>0]}(y)$.
3. Se $Z \sim SN(\lambda)$ então quando $\lambda \to \infty$, $Z \xrightarrow{D} HN(0,1)$.
4. Se $Z \sim SN(\lambda)$ então $-Z \sim SN(-\lambda)$.
5. Se $Z \sim SN(\lambda)$ então a fdp de Z é unimodal e $logf(z)$ é uma função côncava.
6. $1 - F(-z, \lambda) = F(z, -\lambda)$.

7. $F(z,1) = \{\Phi(z)\}^2$.

8. Se $Z \sim SN(\lambda)$ então $Z^2 \sim \chi^2_1$.

3.2 Distribuição Normal Assimétrica com parâmetros de locação e escala

Definição: Uma variável aleatória Y tem distribuição normal assimétrica ($Y \sim SN(\mu, \sigma, \lambda)$) com parâmetro de assimetria λ, parâmetro de locação μ e parâmetro de escala σ ($\sigma > 0$) se sua densidade é dada por

$$f(y) = 2\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\Phi\left(\lambda\left(\frac{y-\mu}{\sigma}\right)\right), y \in \Re \tag{3.3}$$

onde $\phi(.)$ e $\Phi(.)$ denotam respectivamente, a função densidade de probabilidade e a função de distribuição acumulada da distribuição normal padrão.

Note que se $Z \sim SN(\lambda)$ e $Y = \mu + \sigma Z$, então $Y \sim SN(\mu, \sigma, \lambda)$. Ou seja, qualquer combinação linear de uma variável aleatória normal assimétrica padrão também terá distribuição normal assimétrica.

Propriedades

9. Se $Y \sim SN(\mu, \sigma, \lambda)$ então $X = a + bY \sim SN(a + b\mu, b\sigma, \lambda)$,$a, b \in \Re$.

10. A função geradora de momentos de Y é dada por

$$M_y(t) = 2\exp\left(\frac{(t-\mu)^2}{2\sigma^2}\right)\Phi\left(\delta\left(\frac{y-\mu}{\sigma}\right)\right), \tag{3.4}$$

onde $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$.

11. A média de Y é dada por

$$E(Y) = \mu + \sigma\delta\sqrt{\frac{2}{\pi}} \tag{3.5}$$

12. A variância de Y é dada por

$$Var(Y) = \sigma^2\left[1 - \frac{2}{\pi}\delta^2\right] \tag{3.6}$$

13. O terceiro momento e o coeficiente de assimetria de Y são dados, respectivamente, por

$$E(Y^3) = \mu^3 + 3\mu^3\sigma\delta\sqrt{\frac{2}{\pi}} + 3\mu\sigma^2 + 3\sigma^3\delta\sqrt{\frac{2}{\pi}} - \sigma^3\delta^3\sqrt{\frac{2}{\pi}},$$

$$\gamma_1 = \sqrt{\frac{2}{\pi}}\delta^3\left[\frac{4}{\pi} - 1\right]\left[1 - \frac{2}{\pi}\delta^2\right]^{-3/2},$$

onde γ_1 assume valores no intervalo (-0.99527 , 0.99527).

3.3 Geração de uma Distribuição Normal Assimétrica

Existem diversas formas de se obter a classe de distribuições normais assimétricas (ver, Freitas, 2006): construção por condicionamento, representação de Henze (1986), estatísticas de ordem, a versão de Cartinhour (1990) e a versão de Arellano-Valle & Azzalini (2006). Nos três primeiros casos, a construção é feita a partir de uma distribuição normal bivariada, enquanto Cartinhour (1990) parte de uma distribuição normal multivariada, na qual todas as componentes do vetor aleatório são truncadas em um intervalo fechado. A seguir, é descrita a representação de Henze.

A representação de Henze consiste em construir a distribuição normal assimétrica como uma combinação linear de variáveis aleatórias independentes, e este método é muito eficiente em simulações, pois a partir dessa representação pode-se implementar facilmente algoritmos computacionais para a geração de amostras.

Sejam $X_0 \sim N(0,1)$ e $X_1 \sim N(0,1)$ variáveis aleatórias independentes, $\delta \in (-1,1)$ e $Z = \delta|X_0| + \sqrt{1-\delta^2}X_1$. Então, $Z \sim SN(\lambda)$, onde $\lambda = \frac{\delta}{\sqrt{1-\delta^2}}$.

3.4 Inferência Sobre os Parâmetros da Distribuição Normal Assimétrica

Esta Seção apresenta dois métodos clássicos de inferência para os parâmetros do modelo normal assimétrico: o método dos momentos e o método de máxima verossimilhança.

3.4.1 Método dos Momentos

O método dos momentos consiste em estimar os parâmetros populacionais através de sistemas de equações que envolvem os momentos amostrais. Como visto na equação (3.4), a função geradora de momentos para uma variável $Y \sim SN(\mu, \sigma, \lambda)$ é dada por

$$M_y(t) = 2\exp\left(\frac{(t-\mu)^2}{2\sigma^2}\right)\Phi\left(\delta\left(\frac{y-\mu}{\sigma}\right)\right).$$

Considere $X_1, X_2, \ldots, X_n$ uma amostra de tamanho n da variável aleatória Y padronizada

$$X = \frac{Y-\bar{y}}{s}.$$

onde $\bar{y} = \sum_{i=1}^{n} y_i/n$ e $s^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2/n-1$ são, respectivamente, a média e a variância amostral de Y.

Então, $X \sim SN(\mu_X, \sigma_X, \lambda)$, onde

$$\begin{aligned} \mu_X &= \frac{\mu-\bar{y}}{s}, \\ \sigma_X &= \frac{\sigma}{s}. \end{aligned}$$

Tem-se que $\bar{x} = 0, s_X = 1$ e $m_3 = \frac{1}{n}\sum_{i=1}^{n} x_i^3$. Igualando os momentos amostrais aos momentos populacionais é obtido o seguinte sistema de equações

$$\begin{cases} \mu_X + \sigma_X\delta\sqrt{\frac{2}{\pi}} = 0 \\ \sigma_X^2[1-\frac{2}{\pi}\delta^2] = 1 \\ \mu_X^3 + 3\mu_X^3\sigma_X\delta\sqrt{\frac{2}{\pi}} + 3\mu_X\sigma_X^2 + 3\sigma_X^3\delta\sqrt{\frac{2}{\pi}} - \sigma_X^3\delta^3\sqrt{\frac{2}{\pi}} = m_3. \end{cases}$$

Deste sistema,

$$\hat{\mu}_X = -\frac{1}{s}\left(\frac{2}{4-\pi}m_3\right)^{1/3},$$
$$\hat{\sigma}_X = \sqrt{1+\hat{\mu}_X^2}$$

$$\text{e } \hat{\delta} = -\frac{\hat{\mu}_X}{\hat{\sigma}_X}\sqrt{\frac{2}{\pi}}.$$

Portanto os estimadores de momentos de μ, σ e λ são dados por

$$\hat{\mu} = \bar{y} + s\hat{\mu}_X,$$
$$\hat{\sigma} = \sqrt{s^2(1+\hat{\mu}_X^2)}$$
$$\text{e } \hat{\lambda} = \frac{\hat{\delta}}{\sqrt{1-\hat{\delta}}},$$

sendo este último sob a condição $|\hat{\delta}| < 1$, caso contrário $\hat{\lambda}$ não está definido.

Quando $\delta = 0$, verifica-se que os estimadores encontrados não se comportam bem com o caso particular (distribuição normal), pois o estimador de momentos $\hat{\sigma}$ superestima o parâmetro σ e $\hat{\mu}$ sub ou superestima μ dependendo do sinal de $\hat{\mu}_X$. Para mais detalhes, ver Pewsey (2000), onde é proposta a utilização da parametrização centrada de Azzalini (1985) para diminuir os efeitos de super ou subestimação dos parâmetros.

3.4.2 Método de Máxima Verossimilhança

Considere $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ uma amostra aleatória de tamanho n de uma variável aleatória $Z \sim SN(\lambda)$. A função de verossimilhança é dada por

$$L(\lambda, \mathbf{z}) = \prod_{i=1}^{n} 2\phi(z_i)\Phi(\lambda z_i).$$

Se λ é positivo e suficientemente grande, a probabilidade de $z_i > 0$ é razoavelmente grande, para todo $i = 1, 2, \ldots, n$. Neste caso, $L(\lambda, \mathbf{z})$ é uma função monótona crescente em λ e, portanto, o estimador que maximiza a função de verossimilhança será infinito, o que levaria a inferir que $Z \sim HN(0, 1)$. Esse mesmo raciocínio segue quando λ é suficientemente distante, à esquerda de 0.

Considere agora somente amostras onde o estimador de máxima verossimilhança (emv) é finito. Para obter o emv para λ é preciso maximizar o logaritmo da função de verossimilhança, que é dado por

$$\log L(\lambda; \mathbf{z}) = \sum_{i=1}^{n} [\log(2\phi(z_i)) + \log \Phi(\lambda z_i)].$$

A primeira derivada é dada por

$$\frac{\partial \log L(\lambda; \mathbf{z})}{\partial \lambda} = \sum_{i=1}^{n} \frac{z_i \phi(\lambda z_i)}{\Phi(\lambda z_i)},$$

onde as raízes dessa derivada podem ser obtidas numericamente.

Sabendo que a informação de Fisher é dada por

$$I(\lambda) = nE_Z \left[\frac{Z^2 \phi^2(\lambda Z)}{\Phi(\lambda Z)} \right],$$

obtém-se

$$I(\lambda) = n \int 2z^2 \phi(z) \frac{\phi^2(\lambda z)}{\Phi(\lambda z)} dz.$$

Seja $X = \mu + \sigma Z$, isto é, $X \sim SN(\mu, \sigma, \lambda)$. Considerando uma amostra aleatória de tamanho n, a sua função de verossimilhança é

$$L(\mu, \sigma, \lambda | \mathbf{x}) = \left(\frac{2}{\sigma}\right)^n \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \exp\left\{ -\frac{1}{2} \left[\frac{\sum_{i=1}^{n} (\mu - x_i)^2}{\sigma^2} \right] \right\} \prod_{i=1}^{n} \Phi\left(\lambda \left(\frac{x_i - \mu}{\sigma} \right) \right), \quad (3.7)$$

e a sua matriz de informação de Fisher é dada por

$$I(\mu, \sigma, \lambda) = \begin{bmatrix} \frac{1+\lambda^2 a_0}{\sigma^2} & \frac{E(Z)\frac{1+2\lambda^2}{1+\lambda^2} + \lambda^2 a_1}{\sigma^2} & \frac{\sqrt{2/\pi}(1+\lambda^2)^{3/2} - \lambda a_1}{\sigma} \\ \frac{E(Z)\frac{1+2\lambda^2}{1+\lambda^2} + \lambda^2 a_1}{\sigma^2} & \frac{2+\lambda^2 a_2}{\sigma^2} & -\frac{\lambda a_2}{\sigma} \\ \frac{\sqrt{2/\pi}(1+\lambda^2)^{3/2} - \lambda a_1}{\sigma} & -\frac{\lambda a_2}{\sigma} & a_2 \end{bmatrix}$$

onde $a_k = E_Z \left[\frac{Z^k \phi(\lambda Z)}{\Phi(\lambda Z)} \right], k = 0, 1, 2.$

Um dos problemas encontrados nessa matriz é o fato dela ser singular quando $\lambda = 0$, o que pode impedir que seja avaliada a existência de assimetria ao utilizar testes baseados na informação de Fisher, como por exemplo, o teste de Wald. A maior dificuldade encontrada com o método de máxima verossimilhança (ver Freitas, 2006) é que estudos de simulação têm mostrado que o emv de λ pode ser infinito, ainda que o verdadeiro

valor desse parâmetro seja finito. Uma alternativa é usar o algoritmo EM que requer a obtenção de uma variável latente. Outra alternativa é utilizar a parametrização centrada (Azzalini, 1985), da forma como foi proposto em Pewsey (2000).

A Seção 2.5 apresenta um estudo de simulação, no qual são feitas inferências sobre os parâmetros de interesse da distribuição normal assimétrica através dos métodos *bootstrap* paramétrico e *bootstrap* não-paramétrico.

3.5 Método de Simulação *Bootstrap*

O *bootstrap* é um método computacional para avaliar a precisão de estimativas e testes, sem necessidade de muitas suposições ou desenvolvimentos analíticos complicados. O método foi proposto por Efron (1979) e tem sido amplamente aplicado na solução dos mais diversos problemas estatísticos.

A técnica *bootstrap* consiste na reamostragem dos dados, permitindo dessa forma aproximar a distribuição de uma função das observações, pela distribuição empírica dos dados, baseada em uma amostra de tamanho finito. A amostragem (com reposição) é feita da distribuição da qual os dados são obtidos quando esta é conhecida (*bootstrap* paramétrico) ou da amostra original (*bootstrap* não-paramétrico). Neste último caso, supõe-se que as observações são obtidas da função de distribuição empírica $F(x)$, que designa uma massa de probabilidade igual a $\frac{1}{n}$ para cada ponto amostral.

O *bootstrap* aborda o cálculo do intervalo de confiança dos parâmetros e dos p-valores, em circunstâncias em que outras técnicas não são aplicáveis, em particular, no caso em que o tamanho amostral é reduzido. A técnica *bootstrap* trata a amostra original como se esta representasse exatamente toda a população (conjunto de experiências, realizações). A sua grande virtude consiste em apresentar solução para casos em que a dedução da precisão da estimativa e do vício aparenta ser impossível ou mesmo demasiado complexa.

No *bootstrap* paramétrico, as estimativas de máxima verossimilhança (emv) são obtidas por meio do modelo ajustado, isto é, gera-se dados do modelo ajustado com os valores dos parâmetros fixados nas emv obtidas da amostra original, e no *bootstrap* não-paramétrico as emv são baseadas em B reamostras com reposição obtidas da amostra

original. O *bootstrap* não-paramétrico é mais robusto contra suposições distribucionais, ao passo que o *bootstrap* paramétrico é esperado ser mais eficiente quando as suposições paramétricas são verdadeiras. Maiores detalhes sobre a técnica podem ser obtidos em Efron & Gong (1983) e Davison & Hinkley (1997).

O procedimento para obtenção da amostra *bootstrap* não-paramétrico

O *bootstrap* não-paramétrico consiste na reamostragem da amostra em questão e pode ser descrito no algoritmo abaixo:

- Considere $X = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória contendo n observações.
- Gerar B amostras de tamanho n, com reposição, de X: $X^*_{(1)}, X^*_{(2)}, \ldots, X^*_{(B)}$.
- Calcular as estimativas dos parâmetros para cada amostra gerada.
- Calcular os resumos inferenciais (média, mediana e desvio padrão) das estimativas dos parâmetros.
- Encontrar o intervalo de confiança $(1 - \alpha)100\%$ para os parâmetros de interesse.

O procedimento para obtenção da amostra *bootstrap* paramétrico

O *bootstrap* paramétrico consiste em estimar os parâmetros através da amostra sob a suposição de uma determinada distribuição, e utilizar essas estimativas para gerar novas amostras. O algoritmo é descrito abaixo:

- Considere $X = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória contendo n observações.
- Calcular as estimativas dos parâmetros da amostra observada.
- Gerar B amostras, de tamanho n, com a mesma distribuição de X, utilizando as estimativas dos parâmetros encontradas no segundo passo do algoritmo.
- Calcular os resumos inferenciais (média, mediana e desvio padrão) das estimativas dos parâmetros.
- Encontrar o intervalo de confiança $(1 - \alpha)100\%$ para os parâmetros de interesse.

3.5.1 Intervalo de Confiança via Método *Bootstrap*

Em inferência estatística, o interesse está na quantificação do erro cometido ao se estimar um parâmetro de interesse θ através de $\hat{\theta}$. Uma estratégia usual para a busca de medidas de incerteza, que expressem este erro, é a estimação do erro padrão de $\hat{\theta}$. Entretanto, métodos analíticos para a obtenção destas medidas nem sempre estão disponíveis, ou constituem processos altamente complexos, enquanto métodos assintóticos, nos quais a construção de intervalo de confiança é baseada, dependem de aproximações nem sempre alcançadas. Neste contexto, o método *bootstrap* constitui uma eficiente alternativa, fornecendo estimativas do erro padrão de $\hat{\theta}$ livres de complexidades algébricas e possibilitando a obtenção de intervalos de confiança sem necessidade de pressupostos sobre a distribuição do estimador.

Desta forma, o método *bootstrap* é utilizado para a obtenção de estimativas intervalares empíricas para os estimadores dos parâmetros de interesse, através da reamostragem do conjunto de dados original.

Seja θ o parâmetro de interesse. Para cada amostra calcula-se a emv para θ. Ao todo serão B reamostragens, cujas emv's associadas são dispostas de maneita ordenada: $\hat{\theta}_1, \hat{\theta}_2, \ldots \hat{\theta}_B$. Utiliza-se, então $\hat{\theta}_{1(B+1)}(\frac{\alpha}{2})$ e $\hat{\theta}_{2(B+1)}(\frac{1-\alpha}{2})$ como sendo os limites inferiores e superiores do intervalo $100(1-\alpha)\%$ de confiança para θ.

Em geral, o número de reamostragens B é fixado em 1000. Dessa forma, intervalos de confiança podem ser obtidos pelo $100(1-\alpha)\%$ percentil *bootstrap* para o parâmetro de interesse. Segundo Efron & Gong (1983), os intervalos *bootstrap*, embora aproximados, oferecem melhores aproximações que os intervalos de confiança padrão.

3.5.2 Aplicação

Neste trabalho é feito um estudo de simulação, onde foram geradas 10 amostras de tamanho 10 de uma distribuição normal assimétrica, através da representação de Henze (Henze, 1986), com parâmetro de locação $\mu = 0$, escala $\sigma = 1$ e assimetria $\lambda = 1$. A partir de cada amostra, foram feitas 2000 reamostragens através dos métodos *bootstrap* paramétrico e não-paramétrico (totalizando 4000 reamostras para cada amostra) e, para cada reamostra foram estimados os valores dos parâmetros através de máxima

verossimilhaça. Os intervalos de confiança foram construídos de maneira empírica e foi considerado um nível de confiança igual a 95%.

De acordo com a teoria apresentada na Subseção 3.4.2, os emv's podem ser encontrados pela maximização da função de verossimilhança ou, equivalentemente, pela maximização do logaritmo da função de verossimilhança. A variância dos emv's pode ser estimada pela matriz de informação de Fisher.

TABELA 3.1: Comparação entre médias e desvio padrão.

	Bootstrap NP		Bootstrap P	
Parâmetros	Média	Desvio Padrão	Média	Desvio Padrão
λ	1.07851	0.56873	0.80906	1.44669
μ	0.01677	0.26862	0.25116	0.43559
σ	0.99692	0.29484	0.92805	0.33963

TABELA 3.2: Comparação entre medianas.

Parâmetros	Bootstrap NP	Bootstrap P
λ	1.25205	0.61649
μ	-0.03257	0.25310
σ	1.11204	0.84605

TABELA 3.3: Comparação entre intervalos 95%.

Parâmetros	Bootstrap NP	Bootstrap P
λ	(0.00116, 1.69402)	(-1.62616, 3.34179)
μ	(-0.2169, 0.65225)	(-0.42943, 1.33860)
σ	(0.50145, 1.41607)	(0.47979, 1.73919)

As Tabelas 3.1, 3.2 e 3.3 comparam as estimativas dos parâmetros para ambos os métodos sendo, respectivamente, as médias e desvios padrão, medianas e intervalos de confiança.

O *bootstrap* não-paramétrico apresentou estimativas cujas esperanças estão mais próximas dos valores reais dos parâmetros. Ao comparar os intervalos de confiança, é

possível verificar que ambos os métodos criaram intervalos que contêm os verdadeiros valores dos três parâmetros, porém, os intervalos gerados pelo *bootstrap* não-paramétrico possuem comprimento muito menor que os mesmos para o caso paramétrico, o que indica que o *bootstrap* não-paramétrico seja uma melhor escolha nessa situação, ou seja, quando $n = 10, \mu = 0, \sigma = 1$ e $\lambda = 1$.

3.6 Considerações Finais

Neste Capítulo foi apresentada a distribuição normal assimétrica e suas propriedades. Também foi apresentado um método de geração da mesma através da representação de Henze (1986), além de dois métodos clássicos de inferência sobre os parâmetros de interesse (método dos momentos e método de máxima verossimilhança). O método de simulação *bootstrap* foi utilizado para verificar a precisão das estimativas dos parâmetros de locação, escala e assimetria do modelo normal assimétrico, através do *bootstrap* paramétrico e não-paramétrico. Verificou-se que o *bootstrap* não-paramétrico, na situação considerada, apresentou melhores resultados com relação ao *bootstrap* paramétrico, tanto na proximidade do valor real dos parâmetros quanto ao erro-padrão das estimativas.

O Capítulo 4 mostra o teste t para dados de expressão gênica, onde é feita uma simulação para verificar a robustez do teste. Neste capítulo, também é feita uma aplicação a um conjunto de dados reais.

Capítulo 4

O Teste t para Dados de Expressão Gênica

Para cada gene encontrado no experimento de *microarray*, é desejado desempenhar um teste estatístico para determinar se esse gene é diferencialmente expresso para um grau de significância no grupo tratamento comparado com o grupo controle. O teste t é um teste adequado para verificar se as médias de duas populações normais são iguais.

Na análise de dados de expressão gênica é comum considerar que o logaritmo das observações tanto do controle como do tratamento, para um determinado gene g, possui uma distribuição normal (Baldi & Long, 2001; Arfin *et al.*, 1995; Saraiva, 2006). Porém, é possível que exista uma certa assimetria nessas observações, fazendo com que a suposição de normalidade seja pouco apropriada. Uma alternativa para essa situação é o uso da distribuição normal assimétrica para este tipo de conjunto de dados.

Dessa forma, será considerado que o logaritmo das observações de controle segue uma distribuição normal assimétrica (ver Capítulo 2), com parâmetros de locação μ_{gc}, escala σ_{gc} e assimetria λ_{gc},

$$X_c = (x^c_{g1}, x^c_{g2}, \ldots, x^c_{gn_c})' \sim SN(\mu_{gc}, \sigma_{gc}, \lambda_{gc})$$

e o logaritmo das observações de tratamento segue uma distribuição normal assimétrica, com parâmetros de locação μ_{gt}, escala σ_{gt} e assimetria λ_{gt},

$$X_t = (x^t_{g1}, x^t_{g2}, \ldots, x^t_{gn_t})' \sim SN(\mu_{gt}, \sigma_{gt}, \lambda_{gt})$$

onde n_c e n_t correspondem, respectivamente, ao número de observações dos níveis de expressão nas situações de controle e de tratamento.

Para determinar se o gene g apresenta ou não evidências para níveis de expressão diferentes, considere o teste de hipóteses sob a forma

$$I\!H_0 : E[X_c] = E[X_t] \; versus \; I\!H_1 : E[X_c] \neq E[X_t]. \tag{4.1}$$

Assim, o teste t é utilizado para cada gene g e baseado nesse teste, as médias e variâncias amostrais são utilizadas para determinar se o gene g apresenta evidências para diferença, sob a forma

$$t_g = \frac{\bar{x}_{gc} - \bar{x}_{gt}}{\sqrt{\frac{s^2_{gc}}{n_c} + \frac{s^2_{gt}}{n_t}}},$$

com t_g seguindo uma distribuição t-Student, com p graus de liberdade,

$$p = \frac{\left[\frac{s^2_{gc}}{n_c} + \frac{s^2_{gt}}{n_t}\right]^2}{\frac{\left(\frac{s^2_{gc}}{n_c}\right)^2}{n_c - 1} + \frac{\left(\frac{s^2_{gt}}{n_t}\right)^2}{n_t - 1}}, \tag{4.2}$$

onde $\bar{x}_{gc} = \sum_{i=1}^{n_c} x^c_{gi}/n_c$ e $\bar{x}_{gt} = \sum_{i=1}^{n_t} x^t_{gi}/n_t$ são, respectivamente, as médias amostrais das observações de controle e tratamento para o gene g, enquanto que $s^2_{gc} = \sum_{i=1}^{n_c}(x^c_{gi} - \bar{x}_{gc})^2/n_c - 1$ e $s^2_{gt} = \sum_{i=1}^{n_t}(x^t_{gi} - \bar{x}_{gt})^2/n_t - 1$ são, respectivamente, as variâncias amostrais das observações de controle e tratamento para o gene g.

Fixado um nível de significância α, se $|t_g|$ é maior que o valor de referência $t_{1-\frac{\alpha}{2},p}$ (quantil $1 - \frac{\alpha}{2}$ da distribuição t-Student com p graus de liberdade), então há evidência de que o gene g apresenta níveis de expressão significativamente diferentes, quando são comparadas as situações de tratamento e controle.

O maior problema encontrado na aplicação do teste t aos dados de expressão gênica, é o fato dos tamanhos amostrais (n_c e n_t), para um determinado gene g, serem pequenos devido ao custo do experimento (Saraiva, 2006). A Seção 4.1 apresenta um estudo de simulação onde a sensibilidade do teste t é verificada ao serem consideradas variações nos parâmetros de locação, escala e assimetria.

4.1 Estudo de Simulação

Nesta Seção é realizado um estudo de simulação, onde o logaritmo dos níveis de expressão da situação controle são gerados de uma distribuição normal assimétrica $SN(\mu_{gc}, \sigma_{gc}, \lambda_{gc})$, com $\mu_{gc} = -0.036$, $\sigma_{gc} = 0.2$ e $\lambda_{gc} = 0$ (ou seja, $N(-0.036, 0.04)$), baseado nas observações do experimento realizado com as células da bactéria *Escherichia Coli* (Arfin *et al.*, 1995), onde é verificado os níveis de expressão com realação aos padrões IHF^+ e IHF^-.

Foram utilizados 3 tamanhos amostrais para as medidas dos níveis de expressão para cada gene, nas situações controle e tratamento, onde $n_c = n_t = (5, 10, 30)$. No entanto, nesta Seção serão apresentadas apenas as Tabelas (com o número de genes diferencialmente expressos) para $n_c = n_t = 5$, as demais tabelas podem ser vistas no Apêndice C.

Para a situação tratamento, os dados foram gerados de uma distribuição normal assimétrica

$$SN(\mu_{gt} = \mu_{gc} + \psi, \sigma_{gt} = \gamma\sigma_{gc}, \lambda_{gt} = \lambda_{gc} + \eta),$$

onde os deslocamentos para os parâmetros de locação, escala e assimetria foram, respectivamente, $\psi = (-1, -0.8, -0.5, 0, 0.5, 0.8, 1), \gamma = (0.5, 1, 2, 3, 4)$ e $\eta = (-1, -0.5, 0, 0.5, 1)$. Para cada variação de ψ, γ e η foi aplicado o teste t e foi verificado quantos genes apresentaram evidências para níveis de expressão diferentes.

O algoritmo utilizado para a realização do teste t, com um nível de confiança igual a 95% é descrito a seguir:

- Foram simulados 1000 genes, $g = 1, 2, \ldots, 1000$, para a situação controle e para cada gene g foram geradas 5 observações ($n_{cg} = 5$) com distribuição normal assimétrica $SN(\mu_{gc}, \sigma_{gc}, \lambda_{gc})$.
- Foram simulados 1000 genes, $g = 1, 2, \ldots, 1000$, para a situação tratamento e para cada gene g foram geradas 5 observações ($n_{tg} = 5$) com distribuição normal assimétrica $SN(\mu_{gt}, \sigma_{gt}, \lambda_{gt})$.
- Para cada gene g, $g = 1, \ldots, 1000$, foi aplicado o teste t com $\alpha = 0.05$.

As Tabelas 4.1 a 4.5 mostram a quantidade de genes (em 1000) detectados com evidências para níveis de expressão diferentes pelo teste t, quando λ_{gt} é fixado e são consideradas diferentes variações nos parâmetros de locação e escala do grupo tratamento.

TABELA 4.1: Número de genes diferencialmente expressos para $\lambda_{gt} = -1$.

	μ_{gt}						
σ_{gt}	-1.036	-0.836	-0.536	-0.036	0.464	0.764	0.964
0.1	1000	1000	989	73	950	1000	1000
0.2	1000	1000	993	109	789	998	1000
0.4	1000	1000	939	168	290	798	956
0.6	991	960	813	202	89	375	605
0.8	938	899	683	193	50	173	336

TABELA 4.2: Número de genes diferencialmente expressos para $\lambda_{gt} = -0.5$.

	μ_{gt}						
σ_{gt}	-1.036	-0.836	-0.536	-0.036	0.464	0.764	0.964
0.1	1000	1000	990	65	962	1000	1000
0.2	1000	1000	970	78	854	999	1000
0.4	998	977	820	114	358	795	949
0.6	949	829	577	100	140	426	658
0.8	812	692	434	118	75	217	378

TABELA 4.3: Número de genes diferencialmente expressos para $\lambda_{gt} = 0$.

	μ_{gt}						
σ_{gt}	-1.036	-0.836	-0.536	-0.036	0.464	0.764	0.964
0.1	1000	1000	966	50	974	1000	1000
0.2	1000	1000	920	48	919	1000	1000
0.4	975	890	537	44	545	889	984
0.6	786	597	313	50	281	622	790
0.8	521	395	196	46	192	414	570

TABELA 4.4: Número de genes diferencialmente expressos para $\lambda_{gt} = 0.5$.

	μ_{gt}						
σ_{gt}	-1.036	-0.836	-0.536	-0.036	0.464	0.764	0.964
0.1	1000	1000	964	63	988	1000	1000
0.2	1000	1000	840	66	977	1000	1000
0.4	948	818	342	87	800	984	999
0.6	679	426	143	117	606	847	946
0.8	404	239	86	100	431	667	822

TABELA 4.5: Número de genes diferencialmente expressos para $\lambda_{gt} = 1$.

	μ_{gt}						
σ_{gt}	-1.036	-0.836	-0.536	-0.036	0.464	0.764	0.964
0.1	1000	1000	950	84	996	1000	1000
0.2	1000	997	803	113	990	1000	1000
0.4	947	784	284	181	931	999	1000
0.6	626	393	80	209	791	977	998
0.8	297	202	61	224	713	881	944

É possível verificar que o teste t é muito sensível a diferenças entre médias, pois fixado o parâmetro de escala (σ_{gt}) o número de genes detectados com evidências para diferença aumenta, em geral, conforme os parâmetros de locação dos grupos controle e tratamento se distanciam (ou seja, quando o deslocamento de coluna μ_{gt} varia). Porém, quando o parâmetro de locação é fixado e é variado o parâmetro de escala do tratamento (σ_{gt}), a sensibilidade do teste passa a ser menor, detectando menos genes com evidências para diferença. Por exemplo, na Tabela 4.1 onde $\mu_{gt} = -1.036$ e $\sigma_{gt} = 0.8$ foram detectados 938 genes diferencialmente expressos, um número menor que nos casos onde $\mu_{gt} = -1.036$ e $\sigma_{gt} < 0.8$.

O valor do parâmetro de assimetria também influencia a detecção de genes diferencialmente expressos pelo teste t, como é possível verificar nas Tabelas 4.1 a 4.5. Na Tabela 4.1, fixado o parâmetro de escala (σ_{gt}) e variando o parâmetro de locação, o parâmetro de assimetria ($\lambda_{gt} = -1$) faz com que mais genes com diferença sejam detectados quando o parâmetro de locação é negativo, por exemplo, onde $\mu_{gt} = -0.836$

e $\sigma_{gt} = 0.8$ (899 genes). Na Tabela 4.5, fixado o parâmetro de escala (σ_{gt}) e variando o parâmetro de locação (colunas), o parâmetro de assimetria ($\lambda_{gt} = 1$) faz com que mais genes com diferença sejam detectados quando o parâmetro de locação é positivo, por exemplo, onde $\mu_{gt} = -1.036$ e $\sigma_{gt} = 0.6$, tem-se 626 genes com diferença.

A Tabela 4.3 considera o parâmetro de assimetria igual a zero. Neste caso, as observações de controle e as observações de tratamento seguem uma distribuição normal. Quando $\mu_{gt} = -0.036$ e $\sigma_{gt} = 0.2$, a distribuição do tratamento é igual ao do controle, e como está sendo considerado um nível de confiança de 95% o número de genes detectados com evidências para diferença é esperado (4.8%).

Em todas as tabelas é possível encontrar um número de genes próximo ao nível de significância adotado (5%), por exemplo, na Tabela 4.1 quando $\sigma_{gt} = 0.8$ e $\mu_{gt} = 0.464$ (50 genes), na Tabela 4.2 quando $\sigma_{gt} = 0.1$ e $\mu_{gt} = -0.036$ (65 genes), na Tabela 4.4 quando $\sigma_{gt} = 0.2$ e $\mu_{gt} = -0.036$ (66 genes), e na Tabela 4.5 quando $\sigma_{gt} = 0.8$ e $\mu_{gt} = -0.536$ (61 genes). Este fato ocorre devido a dependência dos três parâmetros pela média do tratamento (ver equação 3.5), indicando nestes casos, a proximidade das populações controle e tratamento.

4.1.1 Avaliação do Poder do Teste t

Nesta Subseção, é conduzido um estudo de simulação com o propósito de avaliar o poder do teste t sob diferentes instâncias. Para o cálculo do poder do teste, valores pseudo-aleatórios são gerados sob a hipótese alternativa.

O poder de um teste estatístico é definido como a probabilidade de rejeitar a hipótese nula dado que a mesma é falsa. Na prática, é importante que se tenham testes com nível de significância próximo do nível nominal e, que o poder do teste seja alto, mesmo em situações onde as amostras são pequenas.

Neste estudo de simulação, foram utilizadas duas situações considerando diferentes tamanhos amostrais ($n_{gc} = n_{gt} = n = 5, 10, 30, 50, 100$ e 300) para calcular o poder do teste t. Em ambas as situações, é considerado fixo o parâmetro de escala, em uma $\sigma_{gt} = \sigma_{gc}$ e na outra $\sigma_{gt} = 4\sigma_{gc}$, esses valores foram escolhidos para que fosse possível visualizar o que acontece com o poder do teste, quando se tem um deslocamento "grande"

no parâmetro de escala em conjunto com as variações nos demais parâmetros.

Para ambas as situações serão considerados, nesta subseção, dois tamanhos amostrais $n = 5$ e $n = 50$ para que sejam feitas as comparações entre os gráficos do poder do teste t, no entanto os gráficos do poder para os tamanhos amostrais restantes podem ser vistos no Apêndice A e as Tabelas referentes aos mesmos poderão ser visualizadas no Apêndice B.

As Figuras 4.1 e 4.2 mostram os gráficos do poder considerando $\sigma_{gt} = \sigma_{gc}$ e $n = 5$ e $n = 50$, respectivamente, enquanto as Figuras 4.3 e 4.4 mostram os gráficos do poder considerando $\sigma_{gt} = 4\sigma_{gc}$ e $n = 5$ e $n = 50$, respectivamente.

A Figura 4.1 mostra o poder do teste t, quando são deslocados os parâmetros de assimetria e de locação, considerando o parâmetro de escala fixo. Quando a diferença entre os parâmetros de locação é nula ($\psi = 0$), ou seja, quando se está sob $\mathcal{H}_0$, a variação do parâmetro de assimetria faz com que o poder do teste aumente além do esperado, isto é, o teste realizado em situações onde uma das populações é assimétrica rejeita mais do que nominalmente deveria (5% das vezes).

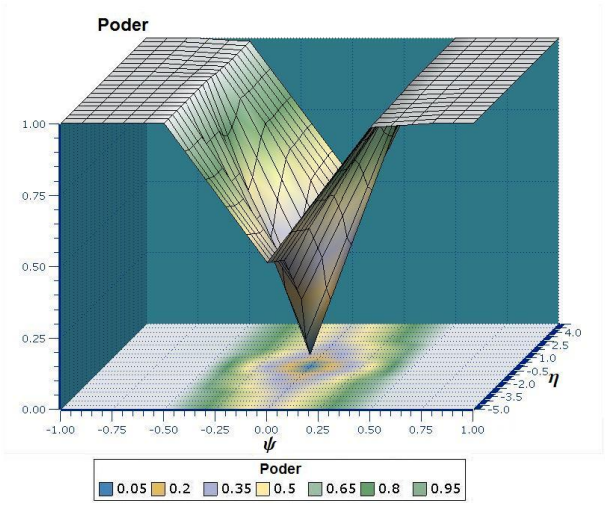

FIGURA 4.1: Gráfico do Poder do Teste t ($\sigma_{gt} = \sigma_{gc}$ e $n = 5$).

A Figura 4.2 mostra o poder do teste t para $n = 50$ e $\sigma_{gt} = \sigma_{gc}$. Quando as Figuras 4.1 e 4.2 são comparadas, é possível verificar que com o aumento do tamanho amostral o poder do teste t também aumenta, fazendo com que o gráfico da Figura 4.2 tenha um aspecto mais afunilado. Também é possível perceber que quando $\psi = 0$ e η (incremento na assimetria do tratamento) é variado, o teste t é menos poderoso.

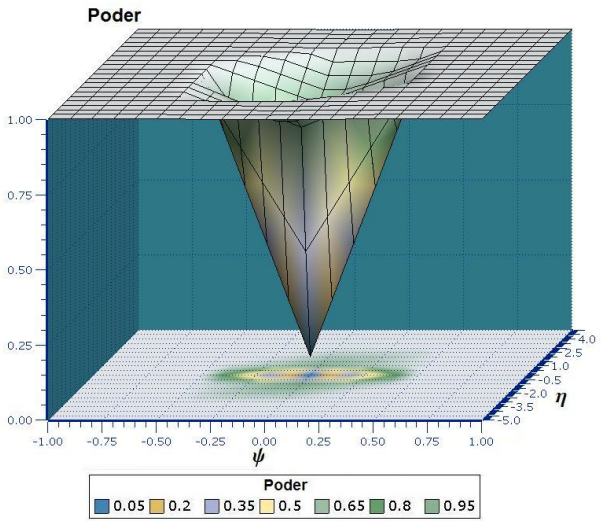

FIGURA 4.2: Gráfico do Poder do Teste t ($\sigma_{gt} = \sigma_{gc}$ e $n = 50$).

A Figura 4.3 mostra o poder do teste t, $n = 5$, quando são deslocados os parâmetros de assimetria e de locação, considerando o parâmetro de escala do tratamento como sendo 4 vezes maior do que o parâmentro de escala do controle. Em situações onde existe diferença entre as populações, em particular, onde $\psi = 0.5$ e o parâmetro de assimetria varia entre -2 a -5, a proporção de rejeições é de cerca de 5%, assim como no caso em que os parâmetros de locação e assimetria são iguais a zero.

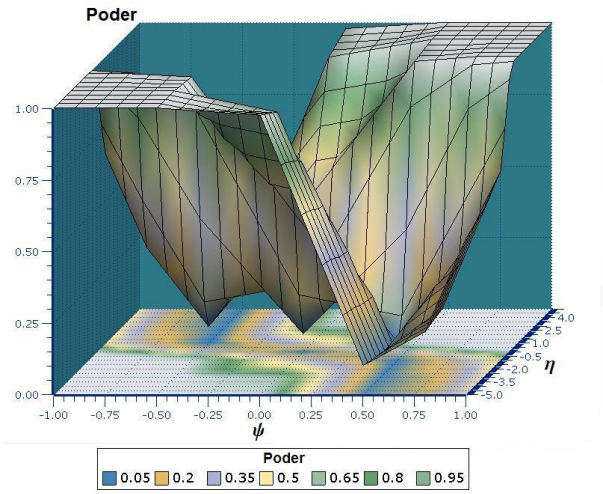

FIGURA 4.3: Gráfico do Poder do Teste t ($\sigma_{gt} = 4\sigma_{gc}$ e $n = 5$).

A Figura 4.4 mostra o poder do teste t para $n = 50$ e $\sigma_{gt} = 4\sigma_{gc}$. Ao comparar as Figuras 4.3 e 4.4, as pontas das quedas representam conjuntos de parâmetros que caracterizam a distribuição cujo centro de massa (esperança) se aproxima do mesmo para a distribuição do grupo controle, ou seja, onde a proporção de rejeições é de cerca de 5%.

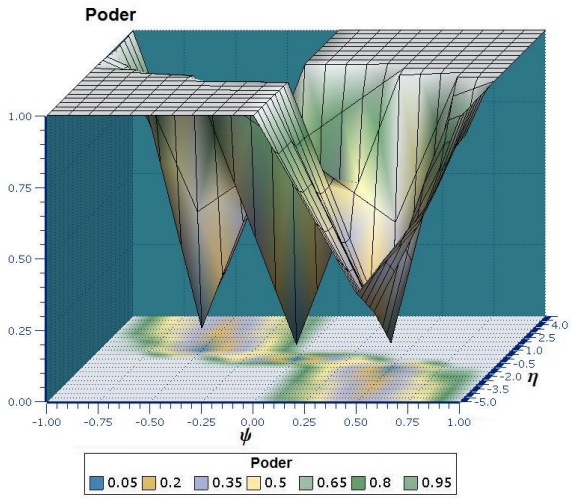

FIGURA 4.4: Gráfico do Poder do Teste t ($\sigma_{gt} = 4\sigma_{gc}$ e $n = 50$).

4.2 Exemplo com Dados Reais

Nesta Seção, o teste t é aplicado a um conjunto de dados, cujos níveis de expressão dos tecidos normais e cancerígenos são obtidos a partir de 22 pacientes com tumores epiteliais na cabeça e no pescoço, *HNSCC - head and neck squamous cell carcinoma* (para uma descrição detalhada do experimento, Kuriakose *et al.*, 2004). Os dados obtidos estavam na sua forma original, ou seja, apenas normalizados representando a medida de intensidade de fluorescência para cada característica de interesse (controle e tratamento). Para ser efetuada a análise dos mesmos, foi necessário fazer um ajuste (transformação) dos dados utilizando neste caso, o logaritmo na base 2 (para mais detalhes, Causton *et al.*, 2003).

Inicialmente, são considerados 12642 genes, onde cada gene possui 22 observações de controle (tecido normal) e 22 observações de tratamento (tecido doente). Após a aplicação do logaritmo aos dados, foi utilizado um teste não-paramétrico para selecionar apenas os genes cujas amostras de tratamento e controle não possuiam evidências de normalidade.

O teste de aderência não-paramétrico utilizado foi o de Shapiro-Wilk (Govindarajulu, 2007). Este teste consiste em verificar se a amostra de interesse é proveniente de

uma população normal ou não. As hipóteses são apresentadas da seguinte forma:

$$\mathbb{H}_0: \text{ A amostra provêm de uma população normal}$$

$$\mathbb{H}_1: \text{ A amostra não provêm de uma população normal.}$$

A estatística de teste é dada por,

$$W = \frac{1}{D}\left[\sum_{i=1}^{n} a_i \left(x^{(n-i+1)} - x^{(i)}\right)\right]^2, \tag{4.3}$$

onde $D = \sum_{i=1}^{n}(x_i - \bar{x})^2$, $x^{(*)}$ é a observação ordenada e a_i é o coeficiente de Shapiro-Wilk obtido na tabela do teste.

O teste foi realizado com um nível de significância de 5%, e foram selecionados os genes onde pelo menos uma das amostras (controle ou tratamento) não apresentava evidências de normalidade, totalizando 4882 genes.

Para verificar quais genes eram diferencialmente expressos, foi realizado o teste t, considerando um nível de significância $\alpha = 0.05$, com cada um dos 4882 genes selecionados. Ao todo foram detectados 1063 genes diferencialmente expressos.

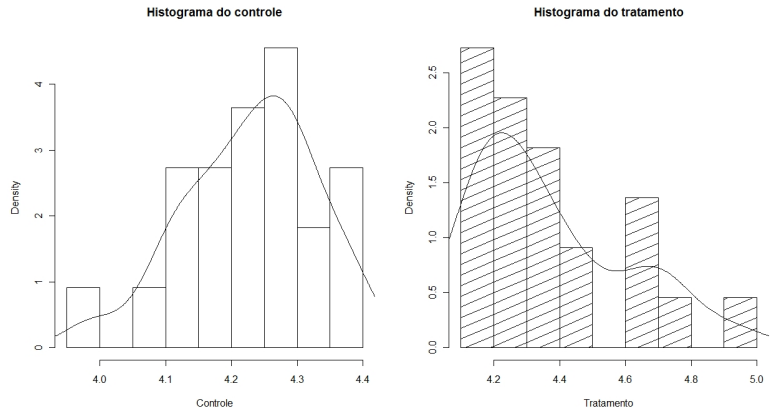

FIGURA 4.5: Histograma do grupo controle e tratamento para o gene HIPK3.

A Figura 4.5 mostra o histograma do controle e do tratamento, para um dos genes detectado como diferencialmente expresso. Nela, é possível observar que tanto controle como tratamento possuem distribuições assimétricas, sendo a assimetria negativa para o controle e positiva para o tratamento (ver Tabela 4.6).

A Tabela 4.6 mostra 10 genes detectados como diferencialmente expressos. Nela é possível observar o p-valor fornecido pelo teste t, as médias e o coeficiente de assimetria

para os grupos controle e tratamento.

TABELA 4.6: Genes detectados como diferencialmente expressos pelo teste t.

Gene	Média Cont.	Média Trat.	p-valor teste t	Coef. Ass. Cont.	Coef. Ass. Trat.
HIPK3	4.22806	4.37193	0.01254	-0.36133	0.63305
IK	7.32629	7.02849	0.03036	0.58595	0.359619
U19495	6.76365	5.81531	0.00360	0.29754	0.35259
FCER1A	6.05201	5.11335	0.00079	0.20792	0.63955
TRAPPC6A	7.95359	7.80221	0.04852	0.27262	0.07119
ST5	12.84925	13.00788	0.04776	-0.52476	-0.13311
MT1X	10.79427	9.80674	0.00061	0.55396	-0.40430
SEPT9	10.24643	10.47866	0.00574	0.05229	0.62812
CDC37	8.81576	9.05266	0.00203	0.10033	0.62812
PKIA	4.18937	3.97271	0.04178	0.39098	0.26130

A Figura 4.6 mostra as médias e variâncias dos grupos controle e tratamento observados, onde os pontos vermelhos representam os genes que foram detectados como diferencialmente expressos. Nela é possível verificar que os genes diferencialmente expressos, encontrados pelo teste t, possuem médias mais distantes da reta (média controle = média tratamento). Esse fato ocorre devido as variâncias dos grupos que, em sua maioria, acabam sendo menores que 1.

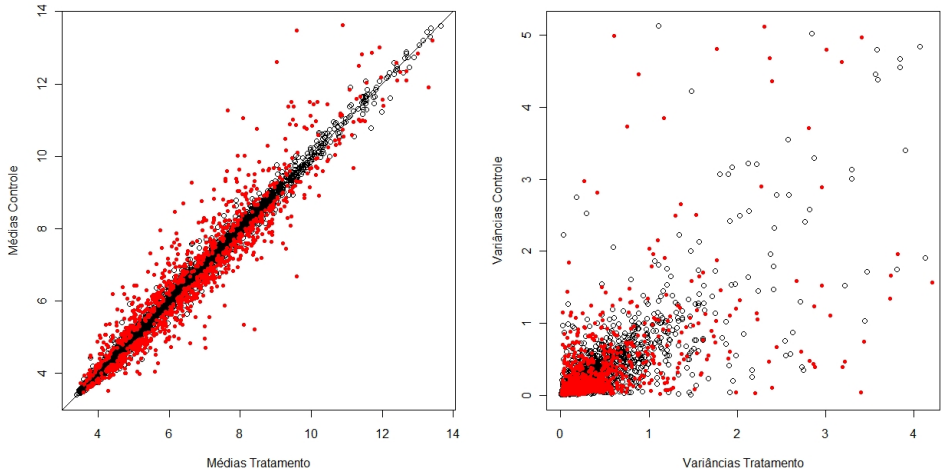

FIGURA 4.6: Médias e variâncias do controle e tratamento.

4.3 Considerações Finais

Este capítulo abordou o teste t para dados de expressão gênica e foi verificado, no estudo de simulação, que o teste t é mais sensível a variações na média quando o parâmetro de locação do grupo tratamento se distancia do parâmetro do grupo controle, do que quando são variadas as distâncias entre os parâmetros de escala dos grupos. Também foi feita a avaliação do poder do teste t e, como esperado, foi possível ver que conforme o tamanho amostral aumenta o teste se torna mais poderoso.

Na aplicação ao conjunto de dados reais, entre 4882 genes sob teste, foi possível encontrar 1063 genes diferencialmente expressos. Também foi possível verificar que embora a suposição de normalidade tenha sido violada, o teste t foi robusto o bastante para identificar como diferencialmente expressos todos os genes que estavam mais distantes da reta média controle = média tratamento (ver Figura 4.6). No entanto, é preciso tomar cuidado com o uso do teste t quando a suposição de normalidade é violada, pois os resultados podem levar a falsos positivos, ou seja, encontrar como diferencialmente expressos genes que na verdade não são.

No Capítulo 5, é utilizado o teste da razão de verossimilhança (TRV) para verificar se o ajuste de um modelo assimétrico aos dados é adequado, para isso é feito um estudo de simulação, onde é verificado o tamanho e poder do teste. O TRV também é aplicado ao gene HIPK3, encontrado como diferencialmente expresso pelo teste t.

Capítulo 5

Teste da Razão de Verossimilhança

Existem situações onde um trabalho de pesquisa é forçado a considerar o problema fundamental de escolher um modelo que descreva o melhor possível os dados em estudo (Hoel *et al.*, 1971). Isto é, a questão não é necessariamente se um certo parâmetro assume um determinado valor, mas sim, saber se os dados são ajustados adequadamente por um entre dois ou mais modelos paramétricos. Um procedimento formal usado nestas situações envolve o uso de testes de hipóteses para seleção de modelos. Muitos testes de hipóteses são orientados aos parâmetros: um modelo de distribuição básico referente aos dados é fornecido e a questão de interesse é se certo parâmetro ou vetor de parâmetros assumido para descrever o modelo de distribuição toma um certo valor hipotético.

O objetivo dos testes de hipóteses estatísticos é validar ou refutar uma hipótese através dos resultados da amostra, expressos através do valor de uma estatística. Dessa forma é possível testar,

$$I\!H_0 : \theta \in \Theta_0 \ \times \ I\!H_1 : \theta \in \Theta_1.$$

A estatística da razão de verossimilhanças (*likelihood ratio test*), para testar as hipóteses acima, é uma ferramenta padrão de seleção de e atua em várias áreas, com a finalidade de solucionar os problemas de seleção de modelos. Este teste é uma generalização natural do lema de Neymann-Pearson (Casella & Berger, 2002) quando as duas hipóteses são simples.

Seja $x_1, x_2, \ldots, x_n$ uma amostra aleatória de uma distribuição com função den-

sidade de probabilidade $f(x;\tilde{\theta})$, onde $\tilde{\theta} = (\theta_1, \ldots, \theta_k)'$ é um vetor de parâmetros desconhecidos assumindo valores no espaço amostral Ω. A função de verossimilhança para θ é dada por

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i;\tilde{\theta}).$$

O valor de θ que maximiza essa função é chamado de estimativa de máxima verossimilhança. Dentro da estrutura de um modelo, toda informação que as observações fornecem está contida na razão de verossimlhnaças de dois modelos: um que está sujeito à hipótese nula, $I\!H_0$, e outro que está sujeito à hipótese alternativa, $I\!H_1$. A razão de verossimilhanças pode ser interpretada como o grau em que as observações disponíveis apoiam uma hipótese contra a outra.

Suponha então dois modelos, um modelo completo cujo vetor de parâmetros é constituído por todos os parâmetros e o modelo restrito que é obtido pelas restrições impostas ao modelo completo. O teste é construído usando a razão da função de verossimilhança que maximiza os espaços Θ_0 e Θ_1. Assim estatística do teste é dada por:

$$\Lambda = \frac{\sup_{\theta\in\Theta_0} L(\theta|x)}{\sup_{\theta\in\Theta} L(\theta|x)}, \tag{5.1}$$

onde Θ é todo o espaço paramétrico e sup é o valor máximo atingido pela função de verossimilhança dentro do espaço amostral.

Deste modo, serão comparados o valor máximo atingido pela função de verossimilhança quando $\theta \in \Theta_0$ com o valor máximo atingido quando $\theta \in \Theta_1$.

Porém, na maioria dos casos, a distribuição exata da razão de verossimlhanças correspondentes às hipóteses é muito difícil de se determinar. Um resultado conveniente entretanto, diz que assintoticamente o TRV segue distribuição qui-quadrado com k graus de liberdade,

$$-2\log(\Lambda) \sim \chi^2_k,$$

onde k é a diferença entre o número de parâmetros definidos sob $I\!H_0$ e sob $I\!H_1$.

Uma característica importante em dados de expressão gênica são os tamanhos amostrais, que normalmente são pequenos, por isso tem-se o interesse em verificar a

robustez do TRV e o poder do teste quando são utilizadas amostras com tamanhos pequenos. O estudo de simulação apresentado na Seção 5.1 visa verificar se o ajuste do modelo normal assimétrico é adequado aos dados e tem o intuito de delinear o tamanho (taxa de rejeição de $I\!H_0$ quando é verdadeira) e o poder do teste, considerando diferentes tamanhos amostrais.

5.1 Aplicação

Nesta Seção são realizadas duas aplicações do TRV, onde uma é o estudo de simulação e a outra uma aplicação do TRV ao gene HIPK3 (Kuriakose *et al.*, 2004) encontrado como diferencialmente expresso no Capítulo 4.

O estudo de simulação nesta Seção visa verificar se o modelo possui distribuição normal ou se tem distribuição normal assimétrica. Dessa forma, são testadas as seguintes hipóteses,

$$I\!H_0 : \lambda = 0 \ \times \ I\!H_1 : \lambda \neq 0,$$

com o intuito de verificar o poder do teste, isto é, verificar a probabilidade da amostra pertencer à região crítica dado o valor do parâmetro. Para isso são consideradas duas abordagens, o procedimento padrão (teoria assintótica) e o procedimento *bootstrap*.

Devido aos tamanhos amostrais pequenos, que serão considerados neste estudo de simulação, para a realização do teste é considerada a seguinte correção de continuidade:

$$-2\log(\Lambda) \sim \frac{1}{2} + \frac{1}{2}\chi^2_1,$$

onde $k = 1$, pois apenas λ está sob teste.

Também é considerado o logaritmo da medida dos níveis de expressão gênica, onde os dados utilizados são os mesmos apresentados para a situação controle do Capítulo 4, isto é, os dados têm uma distribuição normal assimétrica $SN(-0.036, 0.04, 0) \equiv N(-0,036, 0.04)$. Foram considerados 6 tamanhos amostrais ($n = 5, 10, 30, 50, 100$ e 300) e 11 valores de assimetria ($\lambda = -3, -2, -1.5, -1, 0, 1, 1.5, 2, 3$).

Procedimento padrão: Para cada tamanho amostral e para cada valor de λ, foram consideradas 1000 replicações do teste. Os resultados obtidos podem ser vistos na Tabela 5.1. Se nesta tabela, for fixada a linha (tamanho amostral) e forem variadas as colunas (valores de λ) é possível ver o poder do teste. No entanto, se a coluna for fixada e forem variadas as linhas é possível ver o tamanho do teste para cada valor de λ.

TABELA 5.1: Tamanho e poder do teste (procedimento padrão)

	λ										
n	-3	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	3
5	0.316	0.292	0.243	0.233	0.233	0.126	0.226	0.227	0.265	0.299	0.337
10	0.421	0.351	0.272	0.240	0.221	0.122	0.234	0.219	0.294	0.359	0.437
30	0.527	0.321	0.210	0.150	0.092	0.056	0.107	0.156	0.217	0.349	0.535
50	0.664	0.414	0.232	0.125	0.078	0.060	0.100	0.125	0.254	0.396	0.677
100	0.888	0.599	0.354	0.151	0.083	0.047	0.088	0.170	0.362	0.616	0.901
300	0.999	0.947	0.684	0.272	0.068	0.070	0.077	0.287	0.689	0.944	1

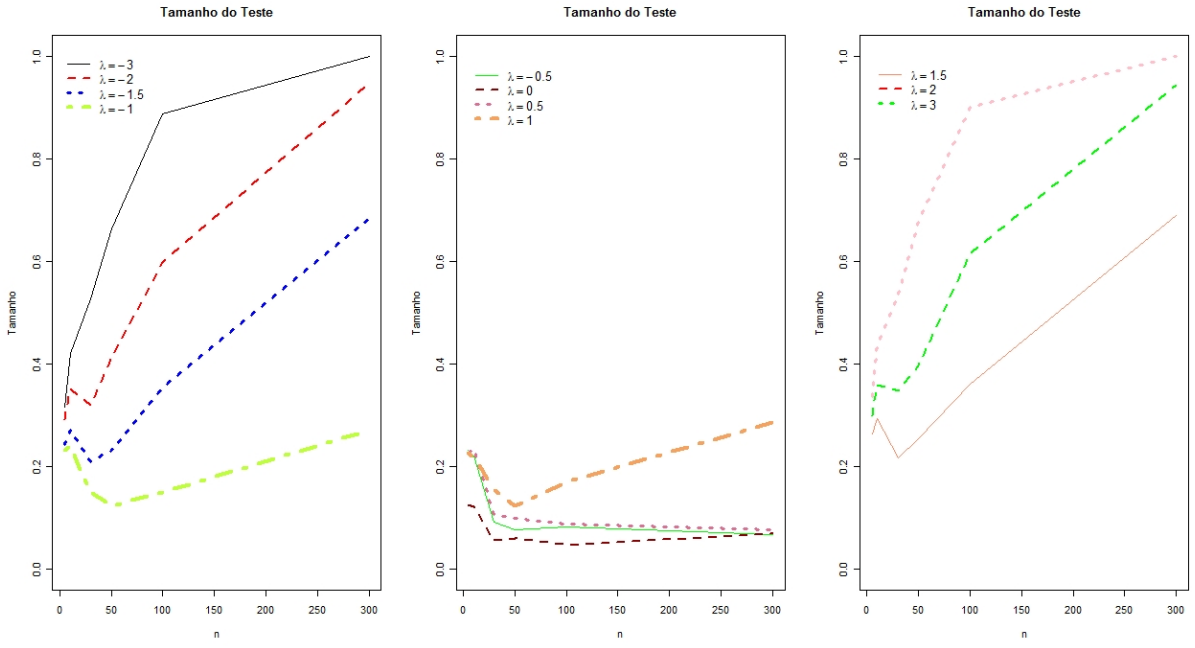

FIGURA 5.1: Tamanho do teste (procedimento padrão).

Os gráficos do tamanho do teste podem ser vistos na Figura 5.1. O poder do teste pode ser visto na Figura 5.2. É possível notar que o poder do teste aumenta conforme λ se distancia de zero e, para amostras grandes, por exemplo, 300 e 100 o poder "é mais sensível", pois tem um aumento considerável ao menor distanciamento de $\lambda = 0$.

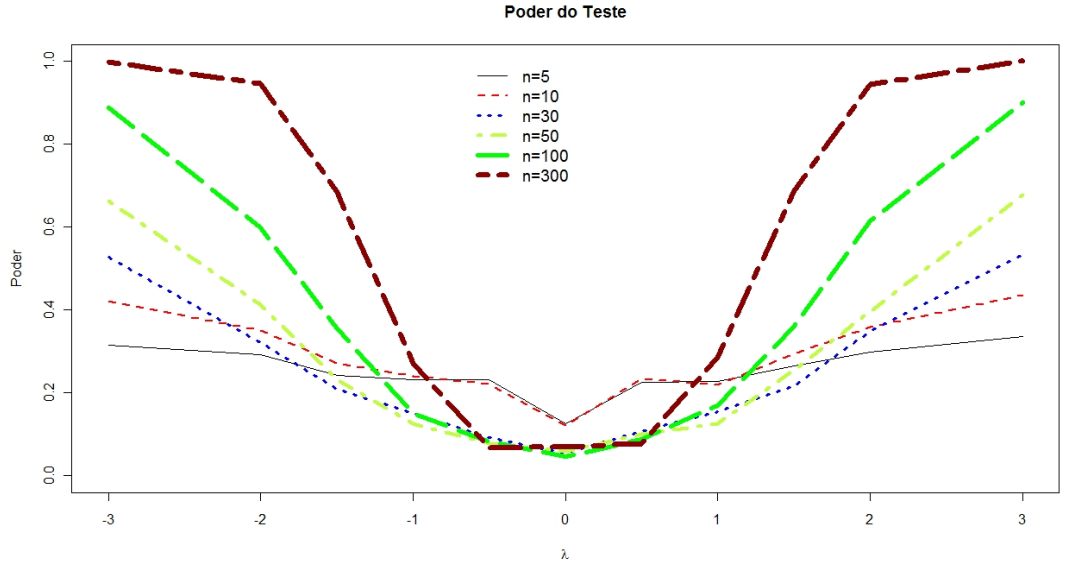

FIGURA 5.2: Poder do teste (procedimento padrão).

Procedimento *bootstrap*: O método utilizado foi o *bootstrap* não-paramétrico (ver Seção 2.5). Foram geradas 100 amostras para cada tamanho amostral e para cada amostra foram utilizadas 1000 reamostras. A Tabela 5.2 mostra o tamanho e poder do teste obtidos, nela é possível verificar que os seus valores são bem parecidos com os da Tabela 5.1, porém, o teste via *bootstrap* apresenta maiores valores tanto no tamanho quanto no poder do teste. Isto é, o teste de hipótese via *bootstrap* superestima o tamanho do teste em até 10%, esse fato também pode ser visualizado nas Figuras 5.3 e 5.4.

TABELA 5.2: Tamanho e poder do teste (*bootstrap*)

	λ										
n	-3	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	3
5	0.435	0.382	0.370	0.334	0.281	0.198	0.281	0.323	0.366	0.414	0.448
10	0.408	0.349	0.229	0.176	0.126	0.084	0.148	0.189	0.293	0.314	0.422
30	0.489	0.369	0.290	0.208	0.155	0.112	0.157	0.206	0.299	0.360	0.523
50	0.599	0.441	0.349	0.181	0.180	0.168	0.164	0.224	0.275	0.425	0.598
100	0.807	0.559	0.427	0.263	0.164	0.141	0.164	0.241	0.399	0.596	0.784
300	0.995	0.901	0.634	0.351	0.174	0.157	0.198	0.356	0.616	0.894	0.988

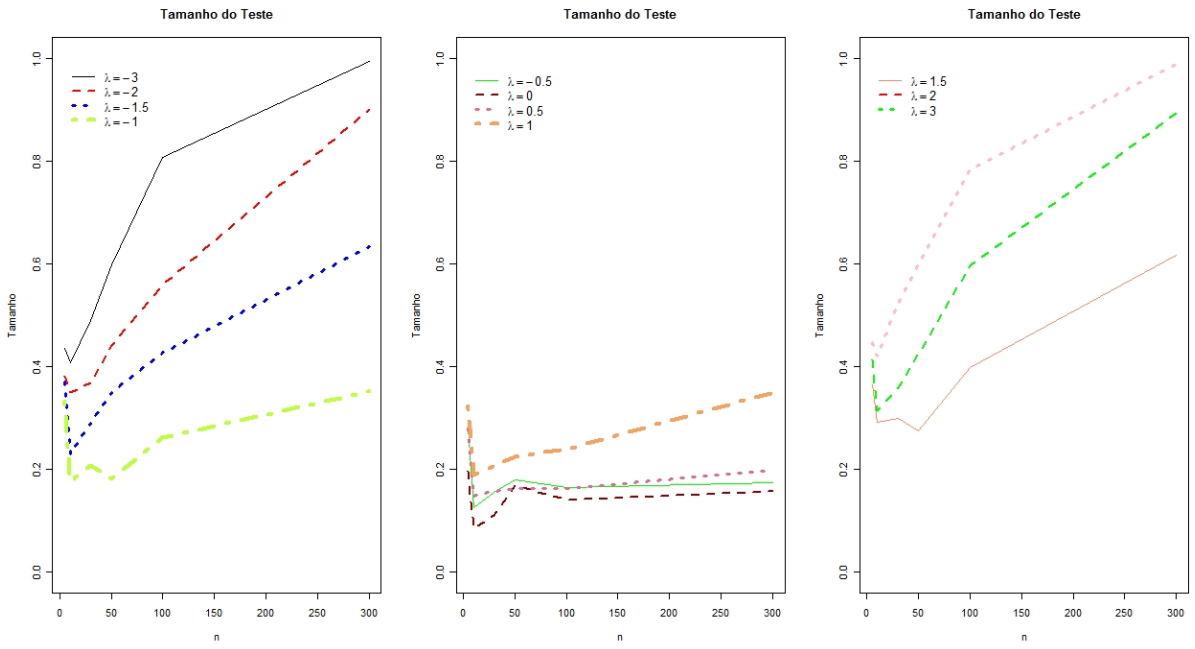

FIGURA 5.3: Tamanho do teste (*bootstrap*).

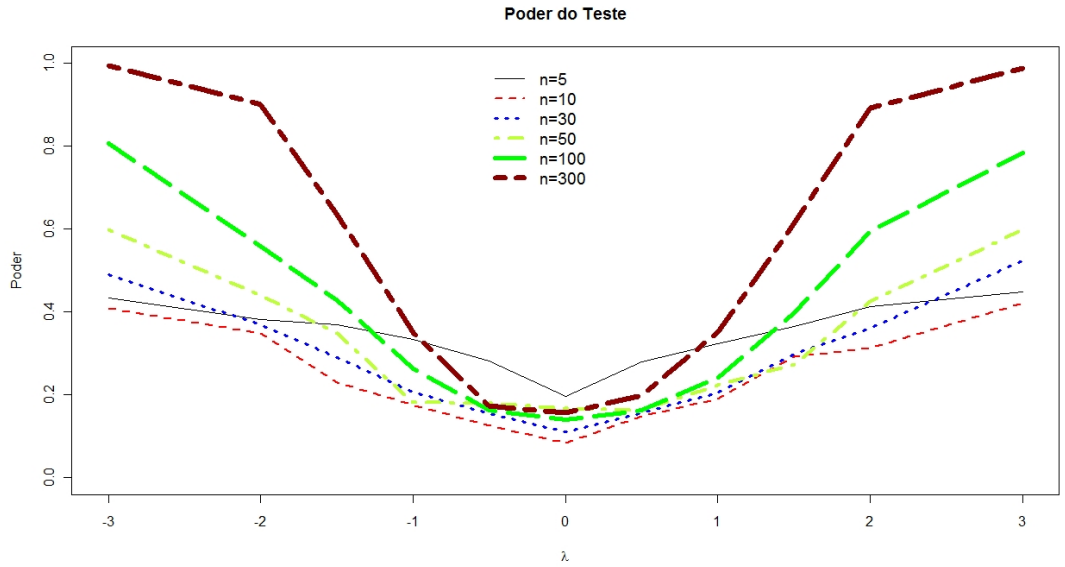

FIGURA 5.4: Poder do teste (*bootstrap*).

Considerando o conjunto de dados reais utilizado no Capítulo 4 (ver, Kuriakose *et al.*, 2004), foi realizado o TRV, para verificar se o ajuste do modelo assimétrico aos dados é adequado. O teste foi realizado para o gene HIPK3 (Kuriakose *et al.*, 2004) encontrado como diferencialmente expresso no Capítulo 4. Para cada situação (controle e tratamento) foi realizado um teste, onde as hipóteses para o grupo controle são:

$$I\!H_0 : \lambda_c = 0 \ \times \ I\!H_1 : \lambda_c \neq 0,$$

e para o grupo tratamento:

$$I\!H_0 : \lambda_t = 0 \ \times \ I\!H_1 : \lambda_t \neq 0.$$

O resultado do TRV para o controle foi 1.138105 indicando a aceitação da hipótese nula e para o grupo tratamento 11.55242 indicando a rejeição da hipótese nula. Embora a Figura 4.5 mostre uma leve assimetria nos dados do controle, o teste mostra que o ajuste do modelo normal assimétrico aos dados não é adequado. O mesmo não ocorre para o grupo tratamento.

5.2 Considerações Finais

O Capítulo 5 mostrou o teste da razão de verossimilhança, que pode ser aplicado aos dados antes de ser feito o teste t, onde foi verficada a adequabilidade de um modelo assimétrico aos dados, ou seja, verificar se a suposição de assimetria é significativa, para isso, foi calculado o poder do teste utilizando o procedimento padrão (teoria assintótica) e o procedimento *bootstrap*. O procedimento padrão mostrou-se mais sensível do que o procedimento *bootstrap* ao poder do teste. O TRV também foi aplicado ao gene HIPK3 e foi possível ver que o ajuste de um modelo assimétrico ao grupo controle não é adequado. Mas mesmo a assimetria sendo significativa, é possivel aplicar o teste t se realmente couber aos dados uma distribuição normal assimétrica, visto que o teste t se mostrou robusto quando a mesma é considerada (ver Capítulo 4).

No Capítulo 6 serão apresentadas as conclusões obtidas acerca do desenvolvimento deste trabalho.

Capítulo 6

Conclusões e Propostas Futuras

O processo de desenvolvimento dos experimentos de *microarray* para situações de tratamento e controle foi brevemente descrito. Neste trabalho também foi apresentado um resumo sobre a distribuição normal assimétrica e algumas de suas propriedades probabilísticas. Também foi realizado um estudo de simulação, utilizando os métodos *bootstrap* não-paramétrico e *bootstrap* paramétrico para fazer inferências sobre os parâmetros dessa distribuição.

Para a aplicação do teste t utilizado por Baldi & Long (2001) foi feito um estudo de simulação, no qual as observações dos genes da situação tratamento foram gerados de uma distribuição normal assimétrica. Pode-se verificar que, embora houvesse assimetria, o teste t foi capaz de detectar, em geral, mais genes com diferença do que quando é realizado o teste considerando a assimetria igual a zero, ou seja, controle e tratamento com distribuição normal, indicando dessa forma a robustez do mesmo. No entanto, é preciso tomar cuidado com o uso do teste t quando a suposição de normalidade é violada, pois os resultados podem levar a falsos positivos, ou seja, encontrar como diferencialmente expressos genes que na verdade não são.

O teste t também foi aplicado a um conjunto de dados reais (Kuriakose *et al.*, 2004) (onde foi verificado através do teste de Shapiro - Wilk a ausência de normalidade). Embora os dados possam ser considerados assimétricos, foi possível identificar um grande número de genes diferencialmente expressos (1063), no entanto este número poderia ter sido maior se o teste t fosse mais sensível a grandes diferenças nas variâncias.

Também foi realizado o TRV, para verificar a adequabilidade do modelo assimétrico aos dados. Ao aplicar o TRV às observações do gene HIPK3, foi possível verificar que embora o histograma (Figura 4.5) indicasse assimetria para as observações do controle, o TRV aceitou a hipótese nula ($I\!H_0 : \lambda_c = 0$). Este fato pode ter ocorrido pela assimetria do controle ser menos acentuada que a assimetria do tratamento (ver coeficiente de assimetria na Tabela 4.6). Mas mesmo a assimetria sendo significativa, é possivel aplicar o teste t se realmente couber aos dados uma distribuição normal assimétrica, visto que o teste t se mostrou robusto quando a mesma é considerada (ver Capítulo 4).

Referências Bibliográficas

Arellano-Valle, R. B. & Azzalini, A. (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics*, **33**(3), 561–574.

Arfin, S. M., Long, A. D., Ito, E. T., Torelli, L., Riehle, M. M., Paegle, E. & Hatfield, G. W. (1995). Global gene expression profiling in Escherichia Coli K12. *Journal of Biological Chemistry*, **51**(4), 1579–1580.

Azzalini, A. (1985). A class of distribution which includes the normal ones. *Scandinavian Journal of Statistics*, **12**, 171–178.

Baldi, P. & Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized test t and statistical inferences of gene changes. *Bioinformatics*, **17**(6), 509–519.

Cartinhour, J. (1990). One dimensional marginal density function of a truncated multivariate normal density function. *Comm. Stat. - Theory and Methods*, **19**, 197–203.

Casella, G. & Berger, R. L. (2002). *Statistical Inference*. Duxbury, United States.

Causton, H. C., Quackenbush, J. & Brazma, A. (2003). *Microarray gene expression data analysis*. Blackwell Publishing, United Kingdom.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, **227**(5258), 561–563.

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press, Cambridge.

de Menezes, R. X., M., B. J. & Houwelingen, H. C. (2004). Microarray data analysis: a hierarchical t-test to handle heteroscedasticity. *Applied Bioinformatics*, **3**(4), 229–235.

DeRisi, J. L., Penland, L., Brown, P. O., Bitter, M. L., Meltezer, P. S., Ray, M., Chen, Y., Su, Y. A. & Trent, J. M. (1996). Use of a cDNA microarray to analyze gene expression patterns in human cancer. *Nature Genetics*, **14**, 457–460.

Efron, B. (1979). Bootstrap methods: another look at the jacknife. *Annals of Statistics*, (7), 1–26.

Efron, B. & Gong, G. (1983). A leisurely look at the bootstrap,the jacknife and cross-validation. *The American Statistician*, **37**, 36–48.

Efron, B., Tibishirani, R., Storey, J. D. & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Association*, **96**, 1151–1160.

Freitas, L. A. (2006). *Modelo de regressão com erros normais assimétricos: uma abordagem bayesiana*. Tese de mestrado, DEs – UFSCar, São Carlos.

Ghosh, P., Branco, M. D. & Chakraborty, H. (2006). Bivariate random effect model using skew-normal distribution with application to HIV-RNA. *Statistics in Medicine*, **26**(6), 1255–1267.

Govindarajulu, Z. (2007). *Nonparametric Inference*. World Scientific, New Jersey.

Griffiths, A. J. F., Wessler, S. R., Lewontin, R. C., Gelbart, W. M., Suzuki, D. T. & Miller, J. H. (2004). *Introduction to genetic analysis*. W. H. Freeman, United States.

Henze, N. (1986). A probabilistic representation of the skew-normal distribution. *Scand J. Statistics*, **13**, 271–275.

Hoel, G. P., Port, C. S. & Stone, J. C. (1971). *Introduction to statistical theory*, volume 2. Houghton Mifflin Company, Boston.

Hoheisel, J. D. (2006). Microarray technology: beyond transcript profiling and genotype analysis. *Nature*, (7), 200–210.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1994). *Continuous univariate distribution*, volume 1. Wiley, New York.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995). *Continuous univariate distribution*, volume 2. Wiley, New York.

Kerr, M. K. & Churchill, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics*, **2**(2), 183–201.

Kuriakose, M. A., Chen, W. T., He, Z. M., Sikora, A. G., Zhang, P., Zhang, Z. Y., Qiu, W. L., Hsu, D. F., McMunn-Coffran, C., Brown, S. M., Elango, E. M., Delacure, M. D. & Chen, F. A. (2004). Selection and validation of differentially expressed genes in head and neck cancer. *Cellular and Molecular Life Sciences*, **61**, 1372–1383.

Lee, M.-L. T. (2004). *Analysis of microarray gene expression data*. Kluwer Academic Publishers, United States.

Lennon, G. G. & Lehrach, H. (1991). Hybridization analyses of arrayed cdna libraries. *Trends Genetics*, (7), 314–317.

Müller, U. R. & Nicolau, D. (2004). *Microarray technology and its applications*. Springer, New York.

Nguyen, K. B. & Smart Jr, G. C. (1995). Morphometrics of infective juveniles of steinernema spp. and heterorhabditis bacteriophora (Nemata: Rhabditida). *Journal of Nematology*, **27**(2), 206–212.

Pewsey, A. (2000). Problems of inference for Azzalini's skew-normal distribution. *Journal of Applied Statistics*, **27**(7), 859–870.

Pourahmadi, M. (2007). Skew-normal ARMA models with nonlinear heteroscedastic predictors. *Communications in Statistics: Theory and Methods*, **36**(9), 1803–1819.

Saraiva, E. F. (2006). *Métodos estatísticos aplicados à análise da expressão gênica*. Tese de mestrado, DEs – UFSCar, São Carlos.

Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O. & Davis, R. W. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceeding of the National Academy of Sciences*, **93**, 10614–10619.

Schwender, H., Rabstein, S. & Ickstadt, K. (2006). Do you speak Genomish? *Chance*, **19**(3), 3–10.

Seshadri, V. (1999). *Inverse gaussian distribution, statistical theory and applications*. Springer, New York.

Sobek, J., Bartcherer, K., Jacob, A., Hoheisel, J. D. & Angenendt, P. (2006). Microarray technology as a universal tool for high-throughput analysis of biological systems. *Combinatorial Chemistry & High Throughput Screening*, **9**(5), 365–380.

Zhang, A. (2006). *Advanced analysis of gene expression microarray data*. World Scientific, United States.

Zhao, E. M. (1995). Infraspecific classification of some chinese snakes. *Sichuan Journal of Zoology*, (14), 107–112.

Apêndice A

Figuras do Poder do Teste t

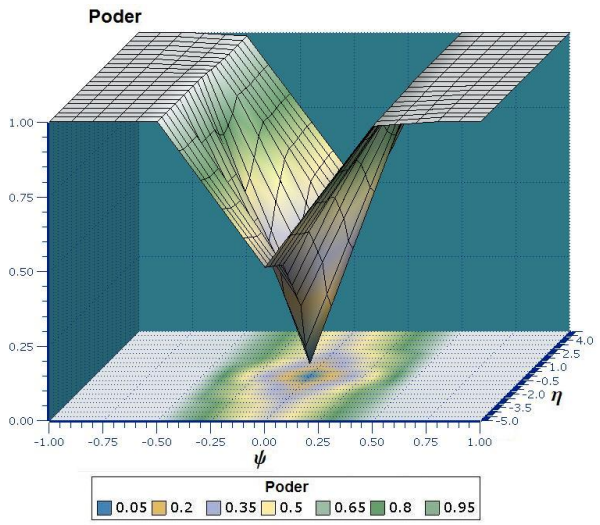

FIGURA A.1: Gráfico do Poder do Teste t ($\sigma_{gt} = \sigma_{gc}$ e $n = 10$).

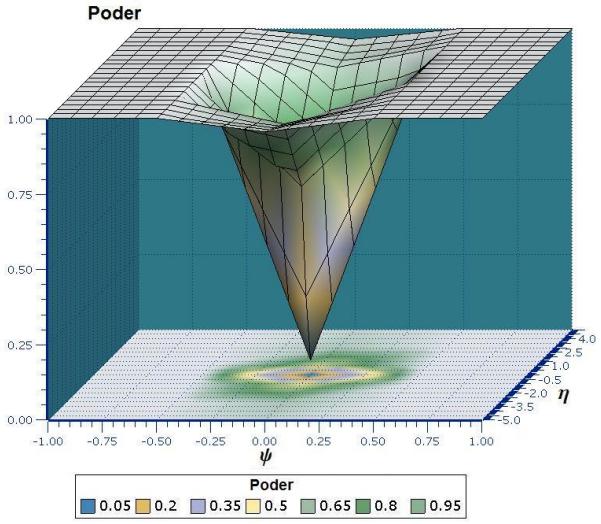

FIGURA A.2: Gráfico do Poder do Teste t ($\sigma_{gt} = \sigma_{gc}$ e $n = 30$).

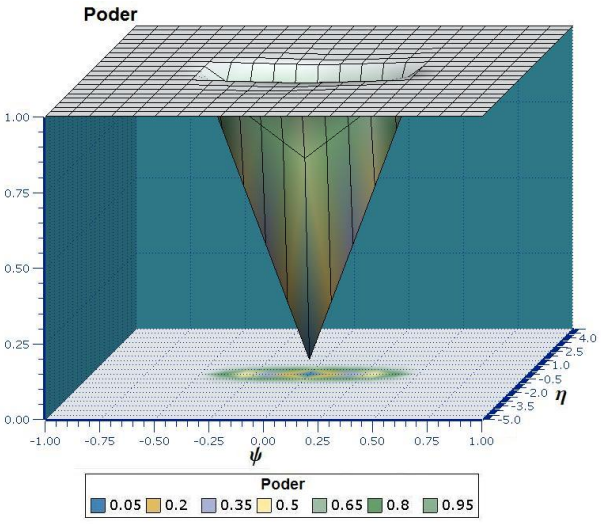

FIGURA A.3: Gráfico do Poder do Teste t ($\sigma_{gt} = \sigma_{gc}$ e $n = 100$).

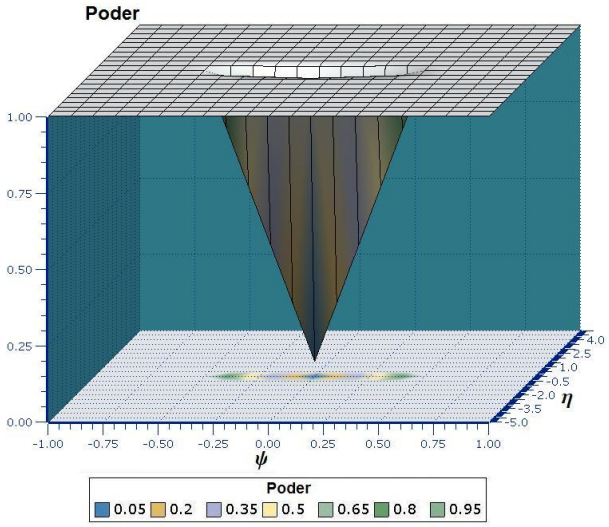

FIGURA A.4: Gráfico do Poder do Teste t ($\sigma_{gt} = \sigma_{gc}$ e $n = 300$).

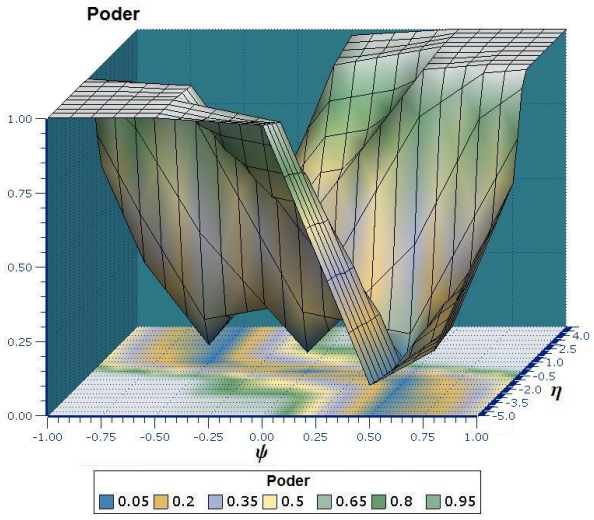

FIGURA A.5: Gráfico do Poder do Teste t ($\sigma_{gt} = 16\sigma_{gc}$ e $n = 10$).

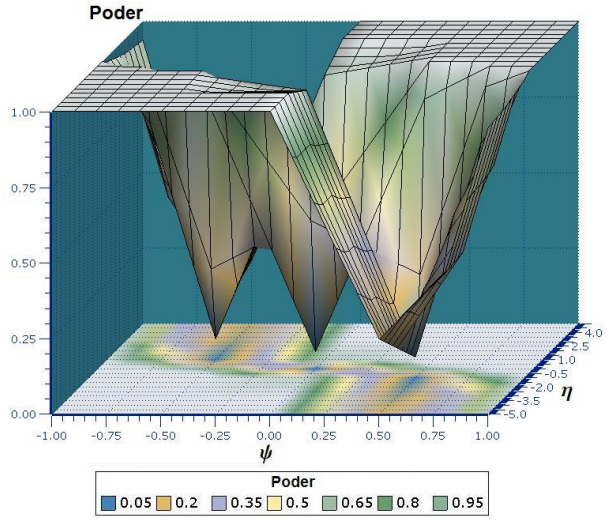

FIGURA A.6: Gráfico do Poder do Teste t ($\sigma_{gt} = 16\sigma_{gc}$ e $n = 30$).

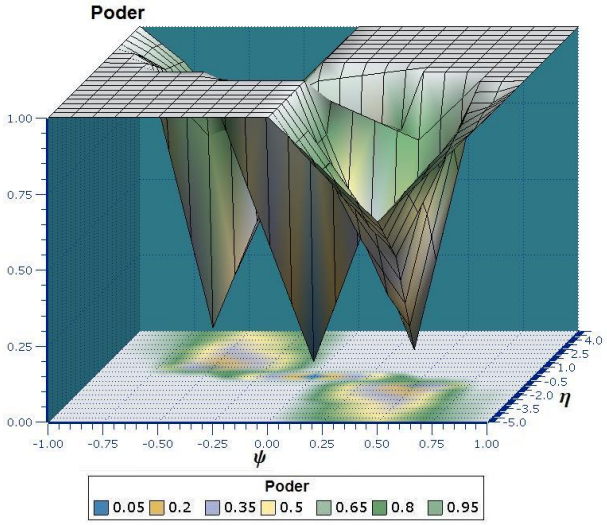

FIGURA A.7: Gráfico do Poder do Teste t ($\sigma_{gt} = 16\sigma_{gc}$ e $n = 100$).

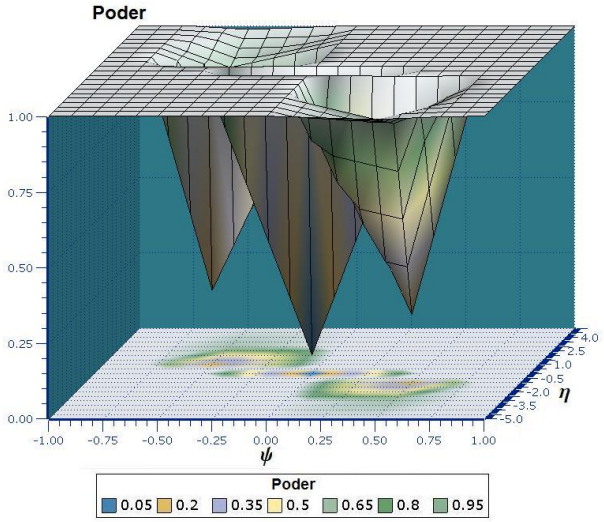

FIGURA A.8: Gráfico do Poder do Teste t ($\sigma_{gt} = 16\sigma_{gc}$ e $n = 300$).

Apêndice B

Tabelas Poder do Teste t

TABELA B.1: Poder do teste $\sigma_{gt} = \sigma_{gc}$ e $n = 5$

	η								
ψ	-5	-3	-1	-0.5	0	0.5	1	3	5
-1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
-0.8	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.998
-0.5	1.000	0.999	0.994	0.980	0.925	0.892	0.831	0.789	0.808
0	0.244	0.248	0.118	0.089	0.050	0.073	0.129	0.238	0.246
0.5	0.821	0.817	0.826	0.873	0.922	0.983	0.998	1.000	1.000
0.8	0.999	1.000	0.999	0.998	1.000	1.000	1.000	1.000	1.000
1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

TABELA B.2: Poder do teste $\sigma_{gt} = 16\sigma_{gc}$ e $n = 5$

	η								
ψ	-5	-3	-1	-0.5	0	0.5	1	3	5
-1	1.000	1.000	0.987	0.895	0.680	0.476	0.401	0.329	0.351
-0.8	1.000	1.000	0.925	0.801	0.500	0.290	0.181	0.163	0.154
-0.5	0.998	0.994	0.792	0.528	0.258	0.105	0.083	0.065	0.073
0	0.668	0.581	0.242	0.125	0.067	0.159	0.281	0.561	0.680
0.5	0.056	0.048	0.074	0.111	0.246	0.542	0.761	0.992	0.998
0.8	0.159	0.162	0.216	0.326	0.491	0.754	0.945	0.999	1.000
1	0.368	0.346	0.390	0.474	0.657	0.887	0.982	1.000	1.000

TABELA B.3: Poder do teste $\sigma_{gt} = \sigma_{gc}$ e $n = 10$

	η								
ψ	-5	-3	-1	-0.5	0	0.5	1	3	5
-1	1.000	1.00	1.000	1.000	1.000	1.000	1.000	1.000	1.000
-0.8	1.000	1.00	1.000	1.000	1.000	1.000	1.000	1.000	1.000
-0.5	1.000	1.00	1.000	1.000	0.999	0.998	0.994	0.986	0.987
0	0.512	0.46	0.262	0.109	0.040	0.119	0.257	0.484	0.507
0.5	0.985	0.99	0.996	1.000	0.999	1.000	1.000	1.000	1.000
0.8	1.000	1.00	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1	1.000	1.00	1.000	1.000	1.000	1.000	1.000	1.000	1.000

TABELA B.4: Poder do teste $\sigma_{gt} = 16\sigma_{gc}$ e $n = 10$

	η								
ψ	-5	-3	-1	-0.5	0	0.5	1	3	5
-1	1.000	1.000	1.000	0.996	0.951	0.780	0.659	0.577	0.587
-0.8	1.000	1.000	1.000	0.987	0.817	0.513	0.333	0.223	0.211
-0.5	1.000	1.000	0.989	0.857	0.464	0.157	0.056	0.066	0.075
0	0.975	0.926	0.510	0.212	0.061	0.220	0.515	0.938	0.978
0.5	0.100	0.080	0.062	0.138	0.424	0.838	0.983	1.000	1.000
0.8	0.218	0.239	0.367	0.534	0.846	0.986	1.000	1.000	1.000
1	0.547	0.579	0.661	0.794	0.949	0.998	1.000	1.000	1.000

TABELA B.5: Poder do teste $\sigma_{gt} = \sigma_{gc}$ e $n = 30$

	η								
ψ	-5	-3	-1	-0.5	0	0.5	1	3	5
-1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.00	1.000
-0.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.00	1.000
-0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.00	1.000
0	0.955	0.922	0.655	0.272	0.046	0.286	0.679	0.93	0.963
0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.00	1.000
0.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.00	1.000
1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.00	1.000

TABELA B.6: Poder do teste $\sigma_{gt} = 16\sigma_{gc}$ e $n = 30$

	η								
ψ	-5	-3	-1	-0.5	0	0.5	1	3	5
-1	1.000	1.000	1.000	1.000	1.000	1.000	0.983	0.951	0.936
-0.8	1.000	1.000	1.000	1.000	0.997	0.944	0.779	0.475	0.439
-0.5	1.000	1.000	1.000	1.000	0.893	0.313	0.067	0.164	0.208
0	1.000	1.000	0.950	0.503	0.054	0.498	0.940	1.000	1.000
0.5	0.247	0.157	0.067	0.325	0.891	1.000	1.000	1.000	1.000
0.8	0.463	0.473	0.774	0.946	1.000	1.000	1.000	1.000	1.000
1	0.946	0.957	0.985	0.999	1.000	1.000	1.000	1.000	1.000

TABELA B.7: Poder do teste $\sigma_{gt} = \sigma_{gc}$ e $n = 50$

	η								
ψ	-5	-3	-1	-0.5	0	0.5	1	3	5
-1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
-0.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
-0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0	0.997	0.998	0.852	0.424	0.059	0.477	0.871	0.996	0.996
0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

TABELA B.8: Poder do teste $\sigma_{gt} = 16\sigma_{gc}$ e $n = 50$

	η								
ψ	-5	-3	-1	-0.5	0	0.5	1	3	5
-1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.990	0.995
-0.8	1.000	1.000	1.000	1.000	1.000	0.994	0.936	0.673	0.605
-0.5	1.000	1.000	1.000	1.000	0.989	0.498	0.075	0.243	0.359
0	1.000	1.000	0.996	0.748	0.048	0.745	0.999	1.000	1.000
0.5	0.376	0.253	0.084	0.491	0.991	1.000	1.000	1.000	1.000
0.8	0.616	0.676	0.947	0.998	1.000	1.000	1.000	1.000	1.000
1	0.995	0.996	1.000	1.000	1.000	1.000	1.000	1.000	1.000

TABELA B.9: Poder do teste $\sigma_{gt} = \sigma_{gc}$ e $n = 100$

	η								
ψ	-5	-3	-1	-0.5	0	0.5	1	3	5
-1	1.000	1.000	1.00	1.000	1.000	1.000	1.00	1.000	1.000
-0.8	1.000	1.000	1.00	1.000	1.000	1.000	1.00	1.000	1.000
-0.5	1.000	1.000	1.00	1.000	1.000	1.000	1.00	1.000	1.000
0	1.000	1.000	0.99	0.726	0.046	0.744	0.99	1.000	1.000
0.5	1.000	1.000	1.00	1.000	1.000	1.000	1.00	1.000	1.000
0.8	1.000	1.000	1.00	1.000	1.000	1.000	1.00	1.000	1.000
1	1.000	1.000	1.00	1.000	1.000	1.000	1.00	1.000	1.000

TABELA B.10: Poder do teste $\sigma_{gt} = 16\sigma_{gc}$ e $n = 100$

	η								
ψ	-5	-3	-1	-0.5	0	0.5	1	3	5
-1	1.000	1.000	1.000	1.00	1.000	1.000	1.000	1.000	1.000
-0.8	1.000	1.000	1.000	1.00	1.000	1.000	0.998	0.918	0.884
-0.5	1.000	1.000	1.000	1.00	1.000	0.776	0.128	0.451	0.652
0	1.000	1.000	1.000	0.95	0.046	0.955	1.000	1.000	1.000
0.5	0.656	0.481	0.116	0.79	1.000	1.000	1.000	1.000	1.000
0.8	0.863	0.917	0.998	1.00	1.000	1.000	1.000	1.000	1.000
1	1.000	1.000	1.000	1.00	1.000	1.000	1.000	1.000	1.000

TABELA B.11: Poder do teste $\sigma_{gt} = \sigma_{gc}$ e $n = 300$

	η								
ψ	-5	-3	-1	-0.5	0	0.5	1	3	5
-1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
-0.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
-0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0	1.000	1.000	1.000	0.989	0.047	0.998	1.000	1.000	1.000
0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

TABELA B.12: Poder do teste $\sigma_{gt} = 16\sigma_{gc}$ e $n = 300$

	η								
ψ	-5	-3	-1	-0.5	0	0.5	1	3	5
-1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
-0.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
-0.5	1.000	1.000	1.000	1.000	1.000	0.998	0.243	0.921	0.987
0	1.000	1.000	1.000	1.000	0.059	1.000	1.000	1.000	1.000
0.5	0.989	0.915	0.223	0.996	1.000	1.000	1.000	1.000	1.000
0.8	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Apêndice C

Teste t para Diferentes Tamanhos Amostrais

As Tabelas abaixo mostram o número de genes diferencialmente expressos para os seguintes tamanhos amostrais $n_{gt} = n_{gc} = n = 10$ e 30 e para os seguintes valores de $\lambda_{gt} = (-1, -0.5, 0, 0.5, 1)$.

TABELA C.1: Número de genes diferencialmente expressos para $\lambda_{gt} = -1$ e $n = 10$.

	μ_{gt}						
σ_{gt}	-1.036	-0.836	-0.536	-0.036	0.464	0.764	0.964
0.05	1000	1000	1000	106	1000	1000	1000
0.2	1000	1000	1000	269	995	1000	1000
0.8	1000	1000	1000	379	556	985	998
1.8	1000	1000	999	449	145	687	939
3.2	1000	999	982	490	51	291	618

TABELA C.2: Número de genes diferencialmente expressos para $\lambda_{gt} = -0.5$ e $n = 10$.

	μ_{gt}						
σ_{gt}	-1.036	-0.836	-0.536	-0.036	0.464	0.764	0.964
0.05	1000	1000	1000	66	1000	1000	1000
0.2	1000	1000	1000	125	996	1000	1000
0.8	1000	1000	995	173	698	994	1000
1.8	1000	1000	925	172	292	812	972
3.2	998	985	795	149	120	482	757

TABELA C.3: Número de genes diferencialmente expressos para $\lambda_{gt} = 0$ e $n = 10$.

	μ_{gt}						
σ_{gt}	-1.036	-0.836	-0.536	-0.036	0.464	0.764	0.964
0.05	1000	1000	1000	45	1000	1000	1000
0.2	1000	1000	999	51	999	1000	1000
0.8	1000	999	912	52	917	1000	1000
1.8	996	951	641	41	630	952	996
3.2	929	795	431	47	393	793	923

TABELA C.4: Número de genes diferencialmente expressos para $\lambda_{gt} = 0.5$ e $n = 10$.

	μ_{gt}						
σ_{gt}	-1.036	-0.836	-0.536	-0.036	0.464	0.764	0.964
0.05	1000	1000	1000	67	1000	1000	1000
0.2	1000	1000	996	104	1000	1000	1000
0.8	1000	997	695	161	997	1000	1000
1.8	964	813	318	181	939	999	1000
3.2	756	510	117	184	831	984	996

TABELA C.5: Número de genes diferencialmente expressos para $\lambda_{gt} = 1$ e $n = 10$.

	μ_{gt}						
σ_{gt}	-1.036	-0.836	-0.536	-0.036	0.464	0.764	0.964
0.05	1000	1000	1000	101	1000	1000	1000
0.2	1000	1000	992	244	1000	1000	1000
0.8	1000	995	553	408	1000	1000	1000
1.8	942	691	163	477	995	1000	1000
3.2	620	304	67	453	980	1000	1000

TABELA C.6: Número de genes diferencialmente expressos para $\lambda_{gt} = -1$ e $n = 30$.

	μ_{gt}						
σ_{gt}	-1.036	-0.836	-0.536	-0.036	0.464	0.764	0.964
0.05	1000	1000	1000	261	1000	1000	1000
0.2	1000	1000	1000	647	1000	1000	1000
0.8	1000	1000	1000	886	962	1000	1000
1.8	1000	1000	1000	920	380	994	1000
3.2	1000	1000	1000	934	66	776	989

TABELA C.7: Número de genes diferencialmente expressos para $\lambda_{gt} = -0.5$ e $n = 30$.

	μ_{gt}						
σ_{gt}	-1.036	-0.836	-0.536	-0.036	0.464	0.764	0.964
0.05	1000	1000	1000	136	1000	1000	1000
0.2	1000	1000	1000	279	1000	1000	1000
0.8	1000	1000	1000	449	989	1000	1000
1.8	1000	1000	1000	475	741	999	1000
3.2	1000	1000	999	493	324	943	998

TABELA C.8: Número de genes diferencialmente expressos para $\lambda_{gt} = 0$ e $n = 30$.

	μ_{gt}						
σ_{gt}	-1.036	-0.836	-0.536	-0.036	0.464	0.764	0.964
0.05	1000	1000	1000	54	1000	1000	1000
0.2	1000	1000	1000	51	1000	1000	1000
0.8	1000	1000	1000	38	1000	1000	1000
1.8	1000	1000	993	55	984	1000	1000
3.2	1000	999	895	48	887	1000	1000

TABELA C.9: Número de genes diferencialmente expressos para $\lambda_{gt} = 0.5$ e $n = 30$.

	μ_{gt}						
σ_{gt}	-1.036	-0.836	-0.536	-0.036	0.464	0.764	0.964
0.05	1000	1000	1000	124	1000	1000	1000
0.2	1000	1000	1000	271	1000	1000	1000
0.8	1000	1000	995	455	1000	1000	1000
1.8	1000	1000	709	469	1000	1000	1000
3.2	997	939	315	520	1000	1000	1000

TABELA C.10: Número de genes diferencialmente expressos para $\lambda_{gt} = 1$ e $n = 30$.

	μ_{gt}						
σ_{gt}	-1.036	-0.836	-0.536	-0.036	0.464	0.764	0.964
0.05	1000	1000	1000	270	1000	1000	1000
0.2	1000	1000	1000	625	1000	1000	1000
0.8	1000	1000	970	899	1000	1000	1000
1.8	1000	998	410	916	1000	1000	1000
3.2	981	762	75	922	1000	1000	1000