
TESTES DE PERMUTAÇÃO E BOOTSTRAP EM ANÁLISE ESTATÍSTICA DE
FORMAS: APLICAÇÕES À ZOOLOGIA

FÁBIO VERÍSSIMO JAQUES DA SILVEIRA

Orientador: Prof. Dr. Getúlio José Amorim do Amaral

Co-orientador: Prof. Dr. Diego Astúa de Moraes

Área de Concentração: Estatística Aplicada

Dissertação submetida como requerimento parcial para obtenção do grau
de Mestre em Estatística pela Universidade Federal de Pernambuco

Recife, fevereiro de 2008

Silveira, Fábio Veríssimo Jaques da

Testes de permutação e bootstrap em análise estatística de formas: aplicações à zoologia / Fábio Veríssimo Jaques da Silveira. – Recife: O Autor, 2008.

x, 49 folhas : il., fig., tab.

Dissertação (mestrado) – Universidade Federal de Pernambuco. CCEN. Departamento de Estatística, 2008.

Inclui bibliografia.

1. Análise multivariada. 2. Bootstrap. I. Título.

519.535

CDD (22.ed.)

MEI2008-026

Universidade Federal de Pernambuco

Pós-Graduação em Estatística

28 de fevereiro de 2008

(data)

Nós recomendamos que a dissertação de mestrado de autoria de

Fábio Veríssimo Jaques da Silveira

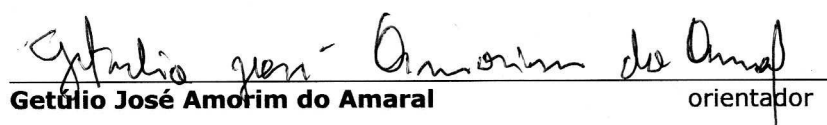
intitulada

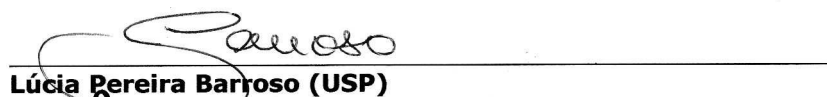
"Testes de Permutação e Bootstrap em Análise Estatística de Formas: Aplicações à Zoologia"

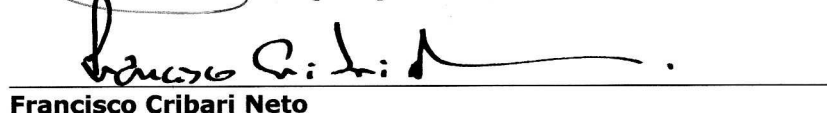
seja aceita como cumprimento parcial dos requerimentos para o grau de Mestre em Estatística.


Coordenador da Pós-Graduação em Estatística

Banca Examinadora:


Getúlio José Amorim do Amaral orientador


Lúcia Pereira Barroso (USP)


Francisco Cribari Neto

Este documento será anexado à versão final da dissertação.

*Até aqui me tem ajudado O Senhor.
Unicamente a Ele dedico este trabalho.*

Agradecimentos

A Deus.

A Mirna Jaques, minha mãe.

A Mário Silveira Filho, meu pai, e a Breno Silveira, meu irmão.

Aos meus familiares, em especial, meus tios Jadson, Suelí, Edilene, Marise, Leopoldo e Fernando, e aos meus primos Jafson e Thamires.

Aos meus irmãos em Cristo, Thiago Simões, Luiz Santana, Poliane Fernanda, Priscila Raquel, Priscila Batista, Abdias, Isaías, Fabi, Jefferson, Roberval, Rosângela, Bruno Trajano, Plínio, Pr. Arthur Enéias, Carlos Domingues e Flávio Souza, pelas orações e pela valiosa amizade de cada um deles.

Aos meus colegas de mestrado, Raphael, Abraão, Lilian, Edwin, Allan, Andréa Prudente, Silvia, Cácio, Wallace, Andrea Santos, Lídia, Marcelo, Juliana, Geraldo, Hemílio, Rejane, Cecílio, Larissa, Olga, Alice, Wagner e Alessandro pela amizade e apoio.

Aos meus amigos da Matemática, Eduardo Jorge, Artur Ribeiro, Ruy de Deus, Gleidson e Tarciana.

Ao professor Getúlio Amaral, pela orientação.

Ao professor Diego Astúa, pela co-orientação.

Aos professores integrantes da banca examinadora, pelas valiosas sugestões.

Aos professores Antonio Carlos R. Monteiro, Jacira Guiró Marino, Henrique Araújo, Francisco Cysneiros e Helena Maia.

À CAPES, pelo apoio financeiro.

Resumo

O objetivo desta dissertação é aplicar métodos de análise estatística de formas a um problema real da zoologia. Quatro testes foram considerados: T^2 de Hotelling, de Goodall, o teste com a estatística de James e o teste com a estatística λ . Estes dois últimos testes não são usados com frequência na literatura estatística.

Também foram considerados testes com bootstrap e testes de permutação. Tais métodos são não-paramétricos e não impõem suposições sobre a distribuição dos dados. Versões dos testes de permutação e bootstrap são comparadas com os testes originais.

São apresentadas também conclusões sobre a aplicação à zoologia. Foram realizados testes para detectar a presença de dimorfismo sexual em seis espécies de marsupiais, baseando-se na mandíbula desses animais.

Palavras-chave: análise de formas; bootstrap; dimorfismo sexual; testes de permutação; zoologia.

Abstract

The aim of this dissertation is to apply methods of statistical shape analysis to a real problem of zoology.

Four test are considered: Goodall, Hotelling, James and λ ; the latter two are seldom used in the statistical literature.

Methods based on bootstrap and permutation tests are also considered. These methods are nonparametric and do not entail assumptions about the distribution of the data. Bootstrap and permutation versions of the tests are compared to the original tests.

Conclusions about the zoology application are presented. Tests were performed to check the presence of sexual dimorphism in six species of marsupials, studying its jaw.

Keywords: shape analysis; bootstrap; sexual dimorphism; permutation tests; zoology.

Lista de Figuras	viii
Lista de Tabelas	ix
1 Introdução	1
1.1 Introdução	1
1.2 Revisão da Literatura	3
1.3 Organização da Dissertação	5
1.4 Plataforma Computacional	6
2 Análise Estatística de Formas	7
2.1 Representação Matemática de Formas	7
2.2 Sistemas de Coordenadas	11
2.2.1 Coordenadas Procrustes Completas	11
2.2.2 Coordenadas Tangentes	14
2.2.3 Inversa Generalizada de Moore-Penrose	15
2.2.4 Coordenadas Polares de Kent	16
2.3 Algumas Distribuições de Probabilidade em Análise de Formas	17
2.3.1 Distribuição Uniforme	17

2.3.2	Distribuição de Bingham Complexa	18
2.3.3	Relação da Distribuição Normal Complexa com a Bingham Complexa . . .	19
2.4	Testes para Um Grupo de Objetos	20
2.4.1	Teste T^2 de Hotelling para uma Forma Média Especificada	20
2.4.2	Teste de Goodall para uma Forma Média Especificada	21
2.5	Testes para Duas Populações de Objetos	22
2.5.1	Teste T^2 de Hotelling e Teste com a Estatística de James para Comparar Formas Médias de Duas Populações	23
2.5.2	Teste de Goodall para Comparar Formas Médias de Duas Populações . . .	24
2.5.3	Teste de Hipótese Utilizando a Estatística λ para Comparar Formas Médias de Duas Populações	25
3	Testes de Permutação e Testes Bootstrap	27
3.1	Testes de Permutação para Duas Populações	27
3.2	Testes Bootstrap para Duas Populações	30
3.3	Teste de Permutação e Teste Bootstrap Utilizando a Estatística λ	31
3.3.1	Teste de Permutação com a Estatística λ	31
3.3.2	Teste Bootstrap com a estatística λ	32
4	Análise de Dados de Zoologia	36
4.1	Descrição dos Dados	38
4.2	Resultados	40
4.2.1	Uso dos Testes por Zoólogos	42
5	Conclusão	45
	Referências Bibliográficas	46

Lista de Figuras

4.1	<i>Didelphis albiventris</i>	37
4.2	Mandíbula do <i>D. albiventris</i>	37
4.3	<i>Didelphis marsupialis</i>	37
4.4	<i>Didelphis aurita</i>	37
4.5	Marcos anatômicos na mandíbula do <i>Lutreolina crassicaudata</i>	38
4.6	<i>Lutreolina crassicaudata</i>	38

Lista de Tabelas

4.1	Descrição dos marcos anatômicos.	39
4.2	Tamanho das amostras por grupo.	40
4.3	<i>P</i> -valores dos testes para verificar dimorfismo sexual	44

1.1 Introdução

O estudo das formas de objetos tem recebido crescente atenção em várias disciplinas. Avanços na tecnologia de computadores tem tornado mais fácil a captura e manipulação de imagens de objetos. Essa informação pode ser usada para responder questões relevantes em muitas disciplinas incluindo biologia, medicina, arqueologia e computação visual. Alguns exemplos de objetos que têm sido estudados são vértebras de ratos, crânios de gorilas e exames de ressonância magnética em cérebros.

O conceito de forma de um objeto tem papel central nesse estudo. Análise estatística de formas está relacionada com características e comparações de formas de objetos.

Algumas etapas devem ser realizadas com o intuito de representar a forma de um objeto de maneira matematicamente conveniente. Uma aproximação conveniente é colocar *landmarks* (marcos anatômicos) no objeto, que são pontos para identificar posições especiais no objeto. As coordenadas numéricas dos marcos anatômicos são então usadas para representar um objeto. Essas coordenadas pertencem a um espaço que é chamado espaço de marcos anatômicos. A informação sobre a forma de um objeto é o que resta após se levar em conta os efeitos de translação, escala e rotação [Kendall (1984)].

Um novo conjunto de coordenadas de um objeto, que será chamado coordenadas de pré-

formas, pode ser obtido a partir das coordenadas de um objeto no espaço de marcos anatômicos. Transformações apropriadas são usadas para remover os efeitos de escala e translação. O novo sistema de coordenadas também representa um mapeamento do espaço de marcos anatômicos no novo espaço. Este recebe o nome de espaço das pré-formas.

Primeiramente, vamos nos concentrar em formas de objetos com duas dimensões, isto é, formas planas.

Duas importantes informações de uma amostra aleatória de objetos, a forma média e a matriz produto (ou *SSP*), podem ser calculadas usando coordenadas pré-forma. A matriz produto representa a variação das coordenadas pré-forma e a forma média é definida como o autovetor associado ao maior autovalor dessa matriz.

A forma é finalmente obtida removendo a informação de rotação das coordenadas pré-forma do objeto. A informação de rotação é eliminada rotacionando um objeto para que ele fique tão próximo quanto possível de um molde. O novo conjunto de coordenadas do objeto está dentro de um novo espaço, que é chamado espaço de formas.

O espaço de pré-formas e o espaço de formas são espaços não-euclidianos. É, portanto, difícil desenvolver uma análise estatística padrão nesses espaços. Para evitar as dificuldades de espaços não-euclidianos, é possível definir uma aproximação linear ao espaço. Um espaço tangente é uma aproximação linear local ao espaço em um ponto particular. Para uma dada amostra aleatória de objetos, as coordenadas pré-formas desses objetos podem ser projetadas no espaço tangente da forma média amostral. As novas coordenadas são chamadas coordenadas tangentes.

Métodos inferenciais em análise de formas são freqüentemente realizados no espaço tangente. Tais métodos funcionam melhor quando os dados estão muito concentrados. No espaço tangente, dispõem-se de muitos procedimentos comumente usados em análise multivariada linear padrão. Por exemplo, a variabilidade da forma pode ser estudada aplicando análise de componentes principais às coordenadas tangentes.

Há outras possíveis aproximações à análise estatística de formas que não serão consideradas nesta dissertação. Por exemplo, análise de tamanho-e-forma, análise de forma de reflexão e análise de tamanho-e-forma de reflexão. Na análise estatística de tamanho-e-forma de objetos,

a informação sobre tamanho é retida, e a informação sobre rotação e locação é descartada.

Se o objetivo é desenvolver um estudo sobre a forma da reflexão de objetos, a informação sobre reflexão deve ser removida das formas desses objetos. Semelhantemente, se o objetivo é desenvolver um estudo sobre tamanho e forma de reflexão de objetos, a informação sobre reflexão deve ser removida do tamanho e forma desses objetos [veja Dryden & Mardia (1998, p. 57)].

1.2 Revisão da Literatura

O primeiro trabalho em análise estatística de formas foi feito por Kendall (1977). Em outro artigo, Kendall (1984) dá uma descrição mais completa do campo de pesquisa. Muitos conceitos importantes, incluindo espaços de formas, variedades de formas, análise de Procrustes e densidades de formas, são apresentados e discutidos com detalhes. Ele também esclarece as diferenças entre análise estatística de formas e teoria das formas, que é estudada por topólogos.

Em Kendall (1984) um sistema de coordenadas é também introduzido; mais tarde, chamaremos isso de sistema de coordenadas de Kendall. Um fato interessante sobre esse sistema é que a locação é removida usando uma matriz especial, a matriz de Helmert. Uma contribuição importante de Kendall (1984) foi a definição matemática de forma, onde ele define um espaço matemático para representar a forma de um conjunto rotulado de k pontos em m dimensões.

Por outro lado, em Bookstein (1984) e Bookstein (1986) é apresentada uma base matemática para o estudo da morfometria. Nesse caso, os objetos considerados estão relacionados a disciplinas tais como biologia e medicina e têm marcos anatômicos escolhidos segundo características biológicas ou médicas. Ele também introduz o que se conhece por sistema de coordenadas de Bookstein, que remove os efeitos de translação, rotação e escala manipulando dois dos marcos anatômicos de tal maneira que eles estejam em posições fixas.

Quando convidado para comentar o artigo de Bookstein (1986), Kendall estabeleceu a conexão entre suas duas teorias. O conjunto rotulado de k pontos em m dimensões de Kendall corresponde aos marcos anatômicos de Bookstein. Embora eles usem maneiras diferentes de calcular tamanho e sistemas de coordenadas diferentes, suas idéias são muito similares no sentido de representar a forma de um objeto como um ponto em uma variedade.

Análise de Procrustes pode ser considerada como uma metodologia para estimar, para um conjunto particular de objetos, as transformações “ótimas” de escala, rotação e translação. O tópico de análise de Procrustes foi estudado por Goodall (1991), que definiu a forma média em termos de análise de Procrustes. Se a soma dos quadrados das distâncias entre um ponto e a pré-forma é mínima, diz-se que este ponto é a forma média.

Um modelo gaussiano para os marcos anatômicos foi introduzido por Goodall (1991). Este modelo tem um parâmetro para cada transformação: escala, rotação e translação. Goodall (1991) também apresentou algoritmos para análise de Procrustes, incluindo um algoritmo para análise de Procrustes ordinários, que minimiza a soma de quadrados das distâncias entre duas observações, e um método mais geral baseado em mínimos quadrados ponderados. Ele também apresentou um algoritmo iterativo para estimar as transformações com várias observações. Esse segundo algoritmo é chamado de análise de Procrustes generalizada.

Depois de aplicar as transformações às pré-formas, as coordenadas ajustadas de Procrustes são obtidas. A forma média também pode ser obtida como a média dessas coordenadas.

Goodall também definiu testes para formas nos casos de uma e duas populações. Esses testes foram baseados nas estatísticas do tipo razão F e T^2 de Hotelling. O teste da razão F é chamado na literatura de teste de Goodall.

Mardia & Walder (1994) consideraram testes para dados com marcos anatômicos emparelhados. Eles usaram um modelo gaussiano para os marcos, onde para cada objeto há duas observações. O caso de dois raios-x para o mesmo objeto foi dado como um exemplo. Eles propuseram uma densidade de forma emparelhada e usaram esta densidade para desenvolver inferência. Eles estimaram os parâmetros desta distribuição por máxima verossimilhança e derivaram a estatística da razão de verossimilhanças, que pode ser usada para testar hipóteses e construir regiões de confiança.

Um modelo probabilístico importante para análise estatística de formas foi apresentado por Kent (1994). Este modelo é a distribuição Bingham complexa, uma versão complexa da distribuição Bingham real. Uma propriedade importante da distribuição Bingham complexa é a simetria complexa. Esta simetria complexa significa que um vetor e qualquer versão rotacionada

desse vetor terão a mesma distribuição. Esta propriedade é útil porque a análise de formas pode ser desenvolvida enquanto se trabalha com pré-formas.

A distribuição Watson complexa, que é um caso especial da distribuição Bingham complexa, foi discutida por Mardia & Dryden (1999). Foram considerados estimação por máxima verossimilhança e procedimentos de teste de hipótese. Eles também ilustraram o uso dessa distribuição em análise de formas.

Kent introduziu um método para calcular a forma média que resiste a outliers para dados de marcos anatômicos em duas dimensões. Seu modelo usa uma distribuição gaussiana angular central para as pré-formas. A forma média é calculada por estimação por máxima verossimilhança usando o algoritmo EM.

A geometria do espaço forma foi estudada por Kendall (1984), Le & Kendall (1993) e Kendall et al. (1999). Veja também Dryden & Mardia (1998, cap. 5 e 7).

1.3 Organização da Dissertação

Esta dissertação está dividida em cinco capítulos. Além deste capítulo introdutório, há mais dois capítulos que apresentam alguns fundamentos teóricos, um capítulo dedicado à discussão da aplicação dos testes e a conclusão. No capítulo 2, serão revisados conceitos relacionados à representação matemática de formas, tais como matrizes de configuração e espaço de formas, sistema de coordenadas de Kendall e análise de Procrustes; algumas distribuições de probabilidade comumente utilizadas em análise de formas serão vistas nesse capítulo; serão também revisados os testes T^2 de Hotelling, de Goodall, o teste com a estatística de James e o teste com a estatística λ na sua forma exata (forma tabulada). No capítulo 3, serão revistos os conceitos de testes de permutação e testes bootstrap para duas populações; em seguida, estes métodos serão aplicados para se obter versões de permutação e bootstrap do teste com a estatística λ para duas populações. No capítulo 4, será feita a análise dos dados dos marsupiais, comentando-se os resultados dos testes e a presença ou não de dimorfismo sexual nas espécies consideradas. O capítulo 5 conclui a dissertação.

1.4 Plataforma Computacional

Os resultados numéricos apresentados nesta dissertação foram obtidos utilizando o ambiente de programação e análise de dados R em sua versão 2.6.0 para sistema operacional Microsoft Windows. O R se encontra disponível gratuitamente através do site <http://www.R-project.org>. Detalhes sobre esta linguagem de programação podem ser encontrados em Cribari-Neto & Zarkos (1999).

A presente dissertação de mestrado foi digitada utilizando o sistema de tipografia L^AT_EX, que consiste em uma série de macros ou rotinas do sistema T_EX que facilitam o desenvolvimento da edição do texto. Detalhes sobre o sistema de tipografia L^AT_EX podem ser encontrados em De Castro Korgi (2003).

2.1 Representação Matemática de Formas

Seja Y uma matriz $k \times m$ de coordenadas cartesianas de k marcos anatômicos em m dimensões, dada por

$$Y = \begin{pmatrix} y_{1,1} & \cdots & y_{1,m} \\ \vdots & \ddots & \vdots \\ y_{k,1} & \cdots & y_{k,m} \end{pmatrix}. \quad (2.1)$$

Uma configuração é um conjunto de marcos anatômicos em um determinado objeto e a matriz Y é comumente chamada *matriz de configuração*.

A forma de uma matriz de configuração é obtida removendo-se a informação sobre escala, locação e rotação. O espaço de formas é o conjunto de todas as possíveis formas. A dimensão do espaço de formas associado a objetos com k marcos anatômicos em m dimensões é:

$$km - m - 1 - \frac{m(m-1)}{2}.$$

O termo km é a dimensão total da matriz de configuração Y e se subtrai m , 1 e $m(m-1)/2$ como consequência da remoção de locação, escala e rotação respectivamente [Dryden & Mardia (1998, p. 56)]. A notação Σ_m^k designa um espaço de formas de k marcos anatômicos e m dimensões.

O espaço de marcos anatômicos é um espaço real $\mathbb{R}^m$ onde são representadas as coordenadas

cartesianas de cada marco. Por exemplo, para objetos bidimensionais ($m = 2$), o espaço de marcos anatômicos é $\mathbb{R}^2$. Nesta dissertação, só será tratado o caso $m = 2$.

Algumas transformações precisam ser desenvolvidas na matriz Y com o intuito de remover os efeitos de locação, escala e rotação. Quando $m = 2$, a matriz de configuração pode ser escrita como um vetor complexo. Defina um vetor complexo $k \times 1$

$$z^0 = (y_{1,1} + iy_{1,2}, \dots, y_{k,1} + iy_{k,2})^T = (z_{(1)}^0, \dots, z_{(k)}^0)^T, \quad (2.2)$$

cujos elementos correspondem às coordenadas complexas dos marcos anatômicos. O superescrito 0 é usado para indicar que a configuração conserva os efeitos de locação, escala e rotação. Os detalhes de cada transformação para o caso $m = 2$ serão dados a seguir.

O primeiro passo é remover a locação. Isso pode ser feito de várias maneiras dependendo do sistema de coordenadas. Aqui serão usadas as coordenadas de Kendall. Para o sistema de coordenadas de Kendall, serão necessários detalhes sobre a matriz de Helmert e sobre a sub-matriz de Helmert. A sub-matriz de Helmert fornece uma transformação linear particular que remove a locação pré-multiplicando z^0 [veja Small (1996, p. 130) e Dryden & Mardia (1998, p.34)].

A matriz de Helmert completa H^F é uma matriz ortogonal $k \times k$, cuja primeira linha tem todos os elementos iguais a $1/\sqrt{k}$, e tem a $(j + 1)$ -ésima linha, para $j \geq 1$ dada por

$$(h_j, \dots, h_j, -jh_j, 0, \dots, 0), \quad h_j = -\{j(j + 1)\}^{-1/2},$$

com $j = 1, \dots, k - 1$, onde o número de zeros na $(j + 1)$ -ésima linha é igual a $k - j - 1$. Por exemplo, se o número de marcos anatômicos é 5, a matriz de Helmert completa é dada por

$$H^F = \begin{pmatrix} 1/\sqrt{5} & 1/\sqrt{5} & 1/\sqrt{5} & 1/\sqrt{5} & 1/\sqrt{5} \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 & 0 & 0 \\ -1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} & 0 & 0 \\ -1/\sqrt{12} & -1/\sqrt{12} & -1/\sqrt{12} & 3/\sqrt{12} & 0 \\ -1/\sqrt{20} & -1/\sqrt{20} & -1/\sqrt{20} & -1/\sqrt{20} & 4/\sqrt{20} \end{pmatrix}.$$

Pode ser mostrado por cálculo direto que a matriz de Helmert H^F é uma matriz ortogonal. A locação da configuração complexa z^0 é removida multiplicando-a pela sub-matriz de Helmert de dimensão $(k-1) \times k$, que é a matriz de Helmert H^F com a primeira linha removida. A sub-matriz de Helmert será denotada por H . A configuração helmertizada é dada por

$$\omega = Hz^0. \quad (2.3)$$

Diz-se que uma configuração é centrada se $1_k^T z^0 = 0$, onde 1_k é um vetor de uns de dimensão $k \times 1$. Configurações helmertizadas estão conectadas a configurações centradas pela seguinte propriedade da matriz de Helmert [veja Dryden & Mardia (1998, p. 54)]:

$$H^T H = I_k - \frac{1}{k} 1_k 1_k^T,$$

onde I_k é a matriz identidade de ordem $k \times k$. Além disso, uma vez que H^F é ortogonal, tem-se $HH^T = I_{k-1}$. Desse modo, sendo o vetor $z^0 = (z_{(1)}^0, \dots, z_{(k)}^0)^T$ uma configuração complexa, então

$$(I_k - \frac{1}{k} 1_k 1_k^T) z^0 = z^0 - \bar{z}^0 1_k,$$

onde $\bar{z}^0 = k^{-1} \sum_{i=1}^k z_{(i)}^0$. Portanto, uma vez que $z^0 - \bar{z}^0 1_k$ é uma configuração centrada, tem-se que as configurações centradas são iguais às configurações helmertizadas multiplicadas por H^T . Então, é sempre possível obter configurações helmertizadas a partir de configurações centradas e vice-versa.

A escala pode ser removida da configuração helmertizada ω usando

$$z = \frac{\omega}{\sqrt{\omega^* \omega}} = \frac{Hz^0}{\sqrt{(Hz^0)^* Hz^0}}, \quad (2.4)$$

onde ω^* é o transposto conjugado complexo de ω . O vetor z é chamado de pré-forma da configuração complexa z^0 . Este nome foi usado por Kendall (1984). Note que uma pré-forma é uma forma com a informação de rotação conservada.

Será dado o conceito de espaço de pré-formas, pois tem um papel muito importante [veja Dryden & Mardia (1998, p.59) e Small (1996, p.9)]. O espaço de pré-formas é o espaço de todos os possíveis vetores complexos $(k-1) \times 1$ que não possuem informação de translação nem escala.

Desse modo, o espaço das pré-formas consiste em uma hipersfera unitária complexa em $(k - 1)$ dimensões complexas; isto é,

$$CS^{k-1} = \{z \in \mathbb{C}^{k-1} : z^* z = 1\}, \quad (2.5)$$

onde $\mathbb{C}^{k-1}$ é o espaço complexo $(k - 1)$ -dimensional.

O espaço de formas pode ser pensado como o espaço de pré-formas com a informação de rotação removida. A informação de rotação no vetor pré-forma z pode ser eliminada definindo-se a seguinte classe de equivalência:

$$[z] = \{e^{i\theta} z : \theta \in [0, 2\pi)\}, \quad (2.6)$$

onde por $[z]$ se identifica qualquer uma de suas versões rotacionadas. Kendall (1984) ressalta que o espaço de formas quando $m = 2$ é o espaço complexo projetado $\mathbb{C}P^{k-2}$, o espaço de linhas complexas passando pela origem.

Um problema importante na análise de formas é estimar a forma média de uma amostra aleatória de configurações. Seja $z_1^0, \dots, z_n^0$ uma amostra aleatória de configurações complexas oriunda de uma população de objetos Π , onde cada z_i^0 está definido em (2.2).

Sejam $z_1, \dots, z_n$ as pré-formas de $z_1^0, \dots, z_n^0$, onde z_i está definido em (2.4) e $z_i \in CS^{k-1}$. A forma média Procrustes completa $\hat{\mu}$ é o autovetor correspondente ao maior autovalor da matriz complexa de somas de quadrados e produtos (*SSP*), definida por [veja Kent (1994)]

$$\hat{S} = \sum_{i=1}^n z_i z_i^*.$$

Uma vez que a matriz complexa $\hat{S}$ satisfaz a condição $\hat{S} = \hat{S}^*$, esta matriz é hermitiana [Axler (1997, p. 128)]. Dado que a distribuição das pré-formas tem uma densidade com respeito à distribuição uniforme na esfera das pré-formas e $n \geq k - 1$, então $\hat{S}$ tem posto completo com probabilidade 1. Aplicando o teorema da decomposição espectral para matrizes hermitianas, dado em Mirsky (1955, p. 388), $\hat{S}$ é escrito como

$$\hat{S} = \sum_{j=1}^{k-1} \hat{\lambda}_j \hat{\mu}_j \hat{\mu}_j^*, \quad (2.7)$$

onde $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_{k-1} \geq 0$ são os autovalores e $\hat{\mu}_1, \dots, \hat{\mu}_{k-1}$ são os autovetores correspondentes de $\hat{S}$.

Dado que $\hat{\lambda}_1 > \hat{\lambda}_2 > \dots$, que será comumente o caso na prática, a forma média $\hat{\mu}$ está bem definida e é dada por

$$\hat{\mu} = \hat{\mu}_1. \quad (2.8)$$

2.2 Sistemas de Coordenadas

Em análise estatística de formas, existem vários sistemas de coordenadas para serem usados. Cada sistema de coordenadas é útil para alguns aspectos da análise. Três sistemas de coordenadas serão considerados aqui: coordenadas Procrustes completas, coordenadas tangentes e coordenadas polares de Kent.

2.2.1 Coordenadas Procrustes Completas

Análise de Procrustes é uma técnica usada para *casar* dois objetos. Quando dois ou mais objetos são considerados, eles podem ter diferentes rotações, translações e escalas. Então a técnica de análise de Procrustes é usada para *casar* um objeto com outro. Isto é feito usando as pré-formas desses objetos, uma vez que as pré-formas têm a mesma translação e escala.

Iremos explicar como ocorre o ajustamento de Procrustes dadas duas configurações complexas $y = (y_1, \dots, y_k)^T$ e $w = (w_1, \dots, w_k)^T$, ambas pertencentes a $\mathbb{C}^k$, com $y^* 1_k = 0 = w^* 1_k$, onde y^* denota o transposto do complexo conjugado de y .

Matematicamente, um procedimento adequado para ajustar duas configurações y e w é representado pelo modelo de regressão complexa

$$\begin{aligned} y &= (a + ib)1_k + \beta e^{i\theta} w + \epsilon \\ &= [1_k, w]A + \epsilon \\ &= X_D A + \epsilon, \end{aligned} \quad (2.9)$$

onde $A = (A_1, A_2)^T = (a + ib, \beta e^{i\theta})$ são parâmetros complexos 2×1 com translação $a + ib$, escala $\beta > 0$ e rotação $0 \leq \theta < 2\pi$; ϵ é um vetor complexo de erros; e $X_D = [1_k, w]$ é a matriz de

planejamento $k \times 2$.

Para obter as estimativas dos parâmetros desta regressão complexa, é necessário minimizar a soma de quadrados dos erros que é dada por

$$D^2(y, w) = \epsilon^* \epsilon = (y - X_D A)^* (y - X_D A).$$

A superimposição Procrustes completa de w em y é obtida por estimar A com $\hat{A}$, onde

$$\hat{A} = (\hat{a} + i\hat{b}, \hat{\beta}e^{i\hat{\theta}})^T = \operatorname{arginf} \epsilon^* \epsilon = \operatorname{arginf} (y - X_D A)^* (y - X_D A).$$

O ajuste Procrustes completo (superimposição) de w em y é dado por

$$w^P = X_D \hat{A} = (\hat{a} + i\hat{b})1_k + \hat{\beta}e^{i\hat{\theta}}w,$$

onde o vetor (β, θ, a, b) é escolhido de tal maneira que minimize

$$D^2(y, w) = \|y - w\beta e^{i\theta} - (a + ib)1_k\|^2.$$

A fim de que a soma de quadrados $D^2(y, w)$ seja minimizada, deve-se escolher os seguintes valores para os parâmetros:

$$\hat{a} + i\hat{b} = 0, \tag{2.10}$$

$$\hat{\theta} = \arg(w^* y) = -\arg(y^* w), \tag{2.11}$$

$$\hat{\beta} = \frac{(w^* y y^* w)^{1/2}}{(w^* w)}. \tag{2.12}$$

A prova destes resultados será apresentada agora e pode ser encontrada em Dryden & Mardia (1998). Deseja-se minimizar a seguinte expressão, para os parâmetros β, θ, a, b :

$$\begin{aligned} D^2 &= \epsilon^* \epsilon \\ &= \|y - w\beta e^{i\theta} - (a + ib)1_k\|^2 \\ &= y^* y + \beta^2 w^* w - y^* w \beta e^{i\theta} - w^* y \beta e^{-i\theta} + k(a^2 + b^2), \end{aligned} \tag{2.13}$$

em que y e w são centrados. De imediato, vem que $a = b = 0$. Seja $y^* w = \gamma e^{i\phi}$ ($\gamma \geq 0$); então

$$\beta(y^* w e^{i\theta} + w^* y e^{-i\theta}) = \beta(\gamma e^{i(\theta+\phi)} \gamma e^{-i(\theta+\phi)}) = 2\beta\gamma \cos(\theta + \phi).$$

Ora, para minimizar $\|y - \beta e^{i\theta} w\|^2$ sobre θ , precisa-se maximizar $2\beta\gamma \cos(\theta + \phi)$. Uma solução para θ é $\hat{\theta} = -w = -\arg(y^*w)$. Para encontrar o valor de β que minimiza D^2 , faz-se

$$\frac{\partial D^2}{\partial \beta} = 0 = 2\beta w^*w - 2\gamma,$$

em que $\gamma = |y^*w|$. Logo,

$$\hat{\beta} = \frac{|y^*w|}{(w^*w)}.$$

Esta é a solução de mínimos quadrados ordinários, mas com variáveis complexas. A solução pode ser escrita da seguinte maneira:

$$\hat{A} = (\hat{A}_1, \hat{A}_2)^T = (X_D^* X_D)^{-1} X_D^* y \Rightarrow \hat{A}_1 = 0, \quad \hat{A}_2 = \frac{w^* y}{(w^* w)}. \quad (2.14)$$

Note que o ajuste Procrustes completo de w em y é dado explicitamente por

$$w^P = X_D \hat{A} = \hat{\beta} e^{i\hat{\theta}} w = \frac{w^* y w}{(w^* w)}. \quad (2.15)$$

O vetor de resíduos $r = y - X_D \hat{A}$ é dado por

$$r = [I_k - X_D (X_D^* X_D)^{-1} X_D^*] y = (I_k - H_c) y,$$

onde H_c é a *matriz chapéu* de X_D . Isto é,

$$H_c = X_D (X_D^* X_D)^{-1} X_D^*.$$

O valor minimizado da função D^2 é

$$D^2(r, 0) = r^* r = y^* y - \frac{(y^* w w^* y)}{(w^* w)}. \quad (2.16)$$

A expressão (2.16) não é simétrica em y e w a menos que $y^* y = w^* w$. Uma padronização conveniente é tomar as configurações como unitárias, ou seja,

$$\sqrt{y^* y} = \sqrt{w^* w} = 1.$$

Se a padronização for considerada, então temos uma medida adequada de distância entre formas.

Agora definiremos *distância Procrustes completa* entre duas configurações complexas w e y :

$$\begin{aligned} d_F(w, y) &= \inf_{\beta, \theta, a, b} \left\| \frac{y}{\|y\|} - \frac{w}{\|w\|} \beta e^{i\theta} - a - ib \right\| \\ &= \left\{ 1 - \frac{y^* w w^* y}{w^* w y^* y} \right\}^{1/2}. \end{aligned} \quad (2.17)$$

Esta distância é utilizada no teste de Goodall, seção 2.4.2.

Para uma dada amostra de pré-formas, a análise de Procrustes é desenvolvida ajustando a pré-forma de cada objeto à forma média. As novas coordenadas são chamadas *ajustes Procrustes* ou coordenadas Procrustes e serão definidas abaixo.

Seja $z_1, \dots, z_n$ uma amostra aleatória de pré-formas e sejam $\omega_1, \dots, \omega_n$ as correspondentes configurações helmertizadas.

As configurações têm uma rotação arbitrária [veja Dryden & Mardia (1998, pp. 44-45)]. Assim, antes de proceder com a análise estatística de formas, é necessário rotacionar todas as configurações de tal maneira que estejam o mais próximo possível da forma média amostral. O cálculo é o mesmo de (2.15):

$$\omega_i^P = \frac{\omega_i^* \hat{\mu} \omega_i}{\omega_i^* \omega_i}, \quad i = 1, \dots, n, \quad (2.18)$$

onde $\omega_1^P, \dots, \omega_n^P$ são os ajustes Procrustes completos ou coordenadas Procrustes completas.

Uma vez que as pré-formas podem ser escritas como $z_i = \omega_i / \|\omega_i\|$, onde cada z_i está definida em (2.4) e $\|\omega_i\| = \sqrt{\omega_i^* \omega_i}$, as coordenadas Procrustes também podem ser calculadas de

$$\omega_i^P = z_i^* \hat{\mu} z_i, \quad i = 1, \dots, n.$$

2.2.2 Coordenadas Tangentes

Outro sistema de coordenadas útil é o sistema de coordenadas tangentes. Os conceitos de vetores tangentes e espaço tangente precisam ser apresentados antes da definição de coordenadas tangentes [veja Small (1996, pp. 42-46)]. O espaço tangente do espaço de formas $\mathbb{C}P^{k-2}$ no ponto z é o espaço vetorial de todos os vetores tangentes a $\mathbb{C}P^{k-2}$ no ponto z . Quando fazemos inferência no espaço tangente, é comumente usado o espaço tangente da pré-forma média amostral.

A variabilidade da análise de forma pode ser tratada no espaço tangente. Esse espaço é uma versão linearizada do espaço de formas. Uma das maiores vantagens do espaço tangente é que técnicas padrão de análise multivariada podem ser usadas diretamente.

Existem diferentes tipos de coordenadas no espaço tangente. Aqui, usaremos as coordenadas tangentes Procrustes parciais, que são dadas por

$$t_i = e^{i\hat{\theta}}[I_{k-1} - \hat{\mu}\hat{\mu}^*]z_i, \quad i = 1, \dots, n, \quad (2.19)$$

onde z_i é um vetor pré-forma definido em (2.4) e $\hat{\theta}$ minimiza $\|\hat{\mu} - ze^{i\hat{\theta}}\|^2$ e $\|z\| = \sqrt{z^*z}$. Suponha que $z_1, \dots, z_n$ é uma amostra aleatória de pré-formas e $t_1, \dots, t_n$ são suas coordenadas tangentes, onde cada z_i e t_i são calculados usando (2.4) e (2.19), respectivamente. Seja v_i um vetor de tamanho $(2k-2) \times 1$ obtido empilhando as partes real e imaginária das coordenadas de cada t_i . Se $t_i = x_i + iy_i$, essa operação é representada por $cvec$ onde

$$v_i = cvec(t_i) = (x_i^T, y_i^T)^T, \quad (2.20)$$

onde $x_i = \Re(t_i)$ é a parte real de t_i e $y_i = \Im(t_i)$ é a parte imaginária de t_i . Se o número de marcos anatômicos é k , um vetor pré-forma z_i tem dimensão $(k-1)$ e seu correspondente vetor de coordenadas tangentes v_i , onde v_i é dado em (2.20), tem dimensão $(2k-2)$.

Métodos padrão de análise multivariada podem ser aplicados a coordenadas tangentes reais v_i . Quando os dados estão altamente concentrados, métodos baseados na distribuição normal multivariada podem ser aplicados às coordenadas reais tangentes v_i [veja Dryden & Mardia (1998, p. 151)]. Alguns desses métodos serão considerados nas próximas seções.

Nos métodos citados acima, para obter as estatísticas de teste utilizadas, é necessário calcular a inversa generalizada de alguma matriz em questão. Por isso, será dada agora uma breve descrição da inversa generalizada de Moore-Penrose.

2.2.3 Inversa Generalizada de Moore-Penrose

Seja U uma matriz qualquer. Diz-se que G é uma matriz inversa generalizada da matriz U se G satisfaz:

$$UGU = U.$$

De fato, a matriz U tem infinitas inversas generalizadas e existem algoritmos para obtê-las. Veja, por exemplo, Searle (1971, p. 4), um algoritmo para calcular uma inversa generalizada de U baseando-se em seu posto.

Apesar de apresentar infinitas inversas generalizadas, para qualquer matriz U , existe uma matriz K que é única e satisfaz as seguintes condições:

$$\begin{aligned} UKU &= U \\ KUK &= K \\ (KU)^T &= KU \\ (UK)^T &= UK. \end{aligned}$$

Estas são as chamadas condições de Penrose [Searle (1971, p. 16)] e K é denominada *inversa generalizada de Moore-Penrose*.

Neste trabalho, a notação utilizada será U^- para designar a inversa generalizada de Moore-Penrose de U .

2.2.4 Coordenadas Polares de Kent

O seguinte sistema de coordenadas polares foi proposto por Kent (1994). Dado um ponto $(z_1, \dots, z_{k-1})^T \in CS^{k-2}$, este é transformado em $(s_1, \dots, s_{k-2}, \theta_1, \dots, \theta_{k-1})$, onde

$$\Re(z_j) = s_j^{1/2} \cos(\theta_j), \quad \Im(z_j) = s_j^{1/2} \sin(\theta_j), \quad (2.21)$$

para $j = 1, \dots, k-1$, $s_j \geq 0$, $0 \leq \theta_j < 2\pi$ e $s_{k-1} = 1 - s_1 - \dots - s_{k-2}$. Mais detalhes sobre este sistema de coordenadas podem ser encontrados em Shelupsky (1962). As coordenadas $s_1, \dots, s_{k-2}$ pertencem ao *simplex* unitário de dimensão $k-2$, S_{k-2} . Podemos identificar a esfera de pré-formas com o produto cartesiano $S_{k-2} \times [0, 2\pi)^{k-1}$, obtendo a medida de volume de CS^{k-2} por

$$2^{2-k} ds_1 \dots ds_{k-2} d\theta_1 \dots d\theta_{k-1}. \quad (2.22)$$

O volume total é

$$\frac{2\pi^{k-1}}{(k-2)!},$$

uma vez que o volume do j -ésimo *simplex* é $1/j!$, para $j = 1, 2, \dots$

O conjunto de coordenadas de Kent tem a vantagem de que a densidade uniforme na esfera de pré-formas é uniforme nessas coordenadas.

Coordenadas de forma podem ser obtidas rotacionando z a um eixo fixado. Considere a informação de rotação da figura original em θ_{k-1} ; então, as coordenadas de forma (de dimensão $2k - 4$) são

$$(s_1, \dots, s_{k-2}, \phi_1, \dots, \phi_{k-2}), \quad (2.23)$$

onde $\phi_j = \theta_j - \theta_{k-1}$, $j = 1, \dots, k - 2$. Então a medida de volume no espaço de formas é dada por

$$2^{2-k} ds_1 \dots ds_{k-2} d\phi_1 \dots d\phi_{k-2}. \quad (2.24)$$

E o volume total é

$$\frac{\pi^{k-2}}{(k-2)!}.$$

2.3 Algumas Distribuições de Probabilidade em Análise de Formas

Nesta seção, serão vistas três importantes distribuições de probabilidade no espaço de formas, que fornecem modelos para análise de formas. Trataremos da distribuição uniforme, da distribuição de Bingham complexa e da relação desta com a distribuição normal complexa.

2.3.1 Distribuição Uniforme

Foi mostrada uma medida de volume no espaço de pré-formas na equação (2.22) e uma medida de volume no espaço de formas na equação (2.24), usando coordenadas polares de Kent. Ao normalizar a medida de volume no espaço de formas, obtém-se a medida uniforme $d\gamma$ neste mesmo espaço. Considere um ponto $z = (z_1, \dots, z_{k-1})^T$; as coordenadas polares de Kent são obtidas por $\Re(z_j) = s_j^{1/2} \cos(\theta_j)$, $\Im(z_j) = s_j^{1/2} \sin(\theta_j)$ para $j = 1, \dots, k - 1$, $s_j \geq 0$, $0 \leq \theta_j < 2\pi$ (2.21). A *medida de forma uniforme* é dada por

$$d\gamma = \frac{(k-2)!}{(2\pi)^{k-2}} ds_1 \dots ds_{k-2} d\phi_1 \dots d\phi_{k-2},$$

com $\int d\gamma = 1$.

2.3.2 Distribuição de Bingham Complexa

Vamos considerar o caso de uma distribuição de probabilidade na esfera de pré-formas CS^{k-1} , onde CS^{k-1} é a esfera unitária complexa em $k-1$ dimensões definida em (2.5). No caso de análise de formas, considere k marcos anatômicos em $m = 2$ dimensões com coordenadas complexas escritas como z^0 (2.2), isto é, vetores complexos $k \times 1$. Pré-multiplicando z^0 pela sub-matriz de Helmert (de dimensão $(k-1) \times k$), obtêm-se $k-1$ configurações helmertizadas ω (2.3). Normalizando por $\|\omega\|$, segue que a pré-forma é dada por

$$z = (z_1, \dots, z_{k-1})^T = \frac{\omega}{\|\omega\|} \in CS^{k-1}. \quad (2.25)$$

A distribuição de Bingham complexa em CS^{k-1} (a notação é $CB_{k-1}(A)$) tem função densidade de probabilidade

$$f(z) = c(A)^{-1} \exp(z^* A z), \quad z \in CS^{k-1}, \quad (2.26)$$

onde z^* denota o transposto conjugado complexo de z , A é uma matriz hermitiana ($A = A^*$) de tamanho $(k-1) \times (k-1)$ e $c(A)$ é uma constante normalizadora dada por:

$$c(A) = 2\pi^{k-1} \sum_{j=1}^{k-1} a_j \exp(\lambda_j), \quad a_j^{-1} = \prod_{i \neq j} (\lambda_j - \lambda_i),$$

em que $\lambda_1 < \lambda_2 < \dots < \lambda_{k-1} = 0$ representam os autovalores de A . Perceba que $c(A) = c(\Lambda)$ depende apenas dos autovalores de A e $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_{k-1})$. A prova deste resultado pode ser encontrada em Dryden & Mardia (1998, p. 113).

A distribuição tem a seguinte propriedade:

$$f(e^{i\theta} z) = f(z)$$

e, portanto, é invariante a rotações da pré-forma z . Então, se um objeto é rotacionado, ele tem a mesma densidade e tanto este (o rotacionado) quanto o objeto original participarão identicamente da inferência. Esta propriedade faz da Bingham complexa, uma distribuição adequada para análise de formas (locação e escala foram previamente removidos, pois z pertence à esfera de pré-formas).

Uma vez que $z^*z = 1$ para $z \in CS^{k-1}$, pode-se ver que as matrizes parâmetro A e $A + \alpha I$ definem a mesma distribuição de Bingham complexa, com $c(A + \alpha I) = c(A) \exp(\alpha)$, onde α é um número complexo.

Para mais detalhes sobre as propriedades da distribuição de Bingham complexa, veja Kent (1994).

2.3.3 Relação da Distribuição Normal Complexa com a Bingham Complexa

Seja $z_j = x_j + iy_j$ uma variável aleatória com distribuição conjunta normal complexa com média $\xi_j = \mu_j + i\nu_j$, $j = 1, \dots, p$ e $\Sigma = \Sigma_1 + i\Sigma_2$ é uma matriz de covariância $p \times p$ hermitiana. Se $x = (x_1, \dots, x_p, y_1, \dots, y_p)^T$ e $\mu = (\mu_1, \dots, \mu_p, \nu_1, \dots, \nu_p)^T$, tem-se que

$$x \sim N_{2p} \left(\mu, \frac{1}{2} \begin{bmatrix} \Sigma_1 & \Sigma_2 \\ -\Sigma_2 & \Sigma_1 \end{bmatrix} \right), \quad (2.27)$$

onde $\Sigma_2 = -\Sigma_2^T$ é anti-simétrica e Σ_1 é simétrica positiva definida. Em particular, $\text{var}(x_j) = \text{var}(y_j)$ e $\text{cov}(x_j, y_j) = 0$, e em cada ponto a estrutura de covariância é isotrópica. Dados $z = (z_1, \dots, z_p)^T$ e $\xi = (\xi_1, \dots, \xi_p)$, a função densidade de probabilidade da distribuição normal complexa é

$$f(z) = \frac{1}{\pi^p |\Sigma|} e^{-(z-\xi)^* \Sigma^{-1} (z-\xi)}. \quad (2.28)$$

A notação $z \sim CN(\xi, \Sigma)$ denota que z tem distribuição normal complexa com vetor de médias ξ e matriz de covariância Σ .

A distribuição de Bingham complexa pode ser obtida condicionando uma distribuição normal complexa multivariada com média zero a ter norma 1. Ou seja, se $w \sim CN_{k-1}(0, \Sigma)$, tem-se

$$w | \{\|w\| = 1\} \sim CB_{k-1}(-\Sigma^{-1}).$$

Então, uma interpretação da matriz hermitiana $-\Sigma$ é que esta é a inversa da matriz de covariância de uma variável aleatória normal complexa com média zero, que é condicionada a ter norma 1 para se obter a distribuição de Bingham complexa. Ora, a distribuição de Bingham complexa é um exemplo de aproximação condicionada.

2.4 Testes para Um Grupo de Objetos

Nós consideramos dois métodos para testar se a forma média é igual a um valor particular. Um é o teste T^2 de Hotelling para uma amostra e o outro é o teste de Goodall para uma amostra. O primeiro é menos restrito que o segundo, porém mais complexo. O teste de Goodall supõe que a distribuição conjunta no espaço de marcos anatômicos é normal complexa e isotrópica [veja Dryden & Mardia (1998, p. 160)], o que significa que a variância para cada marco é a mesma. Por outro lado, o teste T^2 de Hotelling supõe normalidade para as observações no espaço tangente e a isotropia não é assumida.

2.4.1 Teste T^2 de Hotelling para uma Forma Média Especificada

Considere as suposições do teste T^2 de Hotelling para uma amostra. Seja $z_1^0, \dots, z_n^0$ uma amostra aleatória de configurações complexas, sejam $z_1, \dots, z_n$ as pré-formas dessas configurações, onde z_i é calculado a partir de (2.4), e seja $\hat{\mu}$ a forma média dessa amostra, calculada usando (2.8). Sejam $v_1, \dots, v_n$ as coordenadas tangentes Procrustes parciais dessas pré-formas, onde v_i é obtida de (2.20).

Suponha que v_i , definido em (2.20), tem distribuição normal multivariada. Tem-se que a média tangente amostral é dada por

$$\bar{v} = \frac{1}{n} \sum_{i=1}^n v_i \quad (2.29)$$

e a matriz de covariância tangente amostral é

$$S_v = \frac{1}{n} \sum_{i=1}^n (v_i - \bar{v})(v_i - \bar{v})^T. \quad (2.30)$$

O teste T^2 de Hotelling objetiva avaliar as hipóteses

$$H_0 : [\mu] = [\mu_0] \quad vs. \quad H_1 : [\mu] \neq [\mu_0],$$

onde $[\mu_0]$ é um valor pré-especificado para a forma média. Aqui, $[\mu]$ pode ser pensado como uma classe de equivalência de pré-formas. As coordenadas tangentes parciais γ_0 para a pré-forma média μ_0 são dadas por

$$\gamma_0 = (I_{2k-2} - cvec(\hat{\mu})cvec(\hat{\mu}^T))cvec\left(\frac{\mu_0^P}{\|\mu_0^P\|}\right), \quad (2.31)$$

onde $cvec(\cdot)$ está definido em (2.20) e μ_0^P é o ajuste Procrustes de μ_0 , calculado usando (2.18).

A estatística usada para este teste é dada por

$$F = \frac{(n - M)}{M} (\bar{v} - \gamma_0)^T S_v^- (\bar{v} - \gamma_0), \quad (2.32)$$

onde γ_0 é dado em (2.31), S_v^- é a inversa generalizada de Moore-Penrose de S_v (seção 2.2.3) e $M = 2k - 4$ é a dimensão do espaço tangente.

Esta estatística tem distribuição $F_{M, n-M}$ sob H_0 . A hipótese H_0 é rejeitada ao nível $\alpha \in (1, 0)$ se $F \geq F(M, n - M, \alpha)$, onde $F(M, n - M, \alpha)$ é o quantil $1 - \alpha$ da distribuição F com graus de liberdade M e $n - M$.

2.4.2 Teste de Goodall para uma Forma Média Especificada

Neste caso, supõe-se isotropia. Seja $z_1, \dots, z_n$ uma amostra aleatória de pré-formas, onde cada z_i é dado em (2.4). Considere as coordenadas tangentes $v_1, \dots, v_n$ dessas pré-formas, onde v_i está definido em (2.20).

Para o teste de Goodall, é necessário supor que as coordenadas tangentes seguem um modelo normal isotrópico. Então v_i tem uma distribuição normal multivariada com vetor de médias μ e matriz de covariância $\Sigma = \sigma^2 I_{2k}$, onde σ^2 é uma constante positiva e I_{2k} é a matriz identidade de dimensão $2k \times 2k$ [veja Goodall (1991, p. 314) e Dryden & Mardia (1998, p. 160)].

Assim como no teste T^2 de Hotelling, as hipóteses consideradas são

$$H_0 : [\mu] = [\mu_0] \quad vs. \quad H_1 : [\mu] \neq [\mu_0].$$

O teste de Goodall é baseado em distâncias Procrustes ao quadrado (2.17). Para as pré-formas z_i e z_j , definidas em (2.4), essa distância é dada por

$$d_F^2(z_i, z_j) = 1 - z_i^* z_j z_j^* z_i, \quad (2.33)$$

para $i = 1, \dots, n$.

Se $\hat{\mu}$, o estimador de μ , estiver próximo de μ , e σ for pequeno, a análise de variância (ANOVA) aproximada é dada por

$$\sum_{i=1}^n d_F^2(z_i, \mu) \approx \sum_{i=1}^n d_F^2(z_i, \hat{\mu}) + n d_F^2(\mu, \hat{\mu}),$$

[veja Dryden & Mardia (1998, p. 160)].

Sob a hipótese H_0 , as distribuições das distâncias Procrustes ao quadrado são, aproximadamente, distribuições qui-quadrado,

$$d_F^2(z_i, \mu_0) \sim \tau_0^2 \chi_M^2,$$

onde $\tau_0 = \sigma / \|\mu_0\|$ e $M = 2k - 4$. A prova desse resultado envolve expansão em série de Taylor [veja Dryden & Mardia (1998, p. 161)].

Usando este resultado e a propriedade aditiva da independência de distribuições qui-quadrado, obtém-se:

$$\sum_{i=1}^n d_F^2(z_i, \mu_0) \sim \tau_0^2 \chi_{(n-1)M}^2.$$

A distância Procrustes $d_F^2(\mu_0, \hat{\mu})$ é aproximadamente independente de $\sum d_F^2(z_i, \hat{\mu})$. Usando mais uma vez a propriedade aditiva da independência de distribuições qui-quadrado, tem-se, aproximadamente,

$$n d_F^2(\mu_0, \hat{\mu}) \sim \tau_0^2 \chi_M^2.$$

Sob a hipótese H_0 , tem-se o seguinte resultado aproximado:

$$F = (n-1)n \frac{d_F^2(\mu_0, \hat{\mu})}{\sum_{i=1}^n d_F^2(z_i, \hat{\mu})} \sim F_{M, (n-1)M}. \quad (2.34)$$

Este resultado vale se σ for pequeno e μ_0 for próximo de $\hat{\mu}$. Rejeita-se H_0 para valores grandes dessa estatística.

2.5 Testes para Duas Populações de Objetos

Quatro testes para comparar as formas médias de duas populações são considerados nesta seção. Primeiro, serão vistos os testes T^2 de Hotelling e o teste com a estatística de James, que tem semelhanças com o T^2 de Hotelling, exceto a suposição de que as estruturas de dispersão entre as populações são diferentes. Em seguida será apresentado o teste de Goodall para duas populações que, juntamente com o T^2 de Hotelling, é uma versão estendida dos testes apresentados na seção 2.4. Por fim, veremos o teste baseado na estatística λ .

2.5.1 Teste T^2 de Hotelling e Teste com a Estatística de James para Comparar Formas Médias de Duas Populações

Estes testes são usados para comparar as médias de duas populações no espaço de pré-formas. No entanto, as quantidades a serem usadas estão no espaço tangente. Esse aspecto será esclarecido após as definições dessas quantidades.

Considere uma amostra aleatória $z_{1j}^0, \dots, z_{n_jj}^0$ de configurações complexas oriunda da população Π_j , onde $j = 1, 2$ (com Π_1 e Π_2 independentes). Sejam $z_{1j}, \dots, z_{n_jj}$ as pré-formas de $z_{1j}^0, \dots, z_{n_jj}^0$, onde z_{lj} é calculada de z_{lj}^0 por (2.4).

Seja $[\mu_j]$ a forma média da população Π_j . Para testar

$$H_0 : [\mu_1] = [\mu_2] \quad vs. \quad H_1 : [\mu_1] \neq [\mu_2],$$

pode-se realizar um teste T^2 de Hotelling para duas amostras independentes no espaço Procrustes tangente. Este é o espaço tangente da pré-forma média amostral $\hat{\mu}$ calculada a partir das duas amostras combinadas. Isto é, calcula-se $\hat{\mu}$ como em (2.8), só que agora com base na amostra $z_{11}, \dots, z_{n_11}, z_{12}, \dots, z_{n_22}$, com $n_1 + n_2$ observações.

Sejam $v_1, \dots, v_{n_1}$ e $w_1, \dots, w_{n_2}$, respectivamente, as coordenadas tangentes Procrustes parciais de $z_{11}, \dots, z_{n_11}$ e $z_{12}, \dots, z_{n_22}$, calculadas por (2.20), no espaço tangente de $\hat{\mu}$. É proposto um modelo normal multivariado no espaço tangente, onde

$$v_i \sim N(\xi_1, \Sigma), \quad w_l \sim N(\xi_2, \Sigma), \quad i = 1, \dots, n_1, \quad l = 1, \dots, n_2, \quad (2.35)$$

v_i e w_l sendo mutuamente independentes. Supõe-se que as matrizes de covariância sejam iguais para v_i e w_l . O quadrado da distância de Mahalanobis entre $\bar{v}$ e $\bar{w}$, calculada por (2.29), é

$$D^2 = (\bar{v} - \bar{w})^T S_u^- (\bar{v} - \bar{w}), \quad (2.36)$$

onde $S_u = (n_1 S_v + n_2 S_w) / (n_1 + n_2 - 2)$, S_v e S_w são calculados por (2.30) e S_u^- é a inversa generalizada de Moore-Penrose de S_u . Sob a hipótese H_0 , tem-se $\xi_1 = \xi_2$. A estatística de teste é

$$F = \frac{n_1 n_2 (n_1 + n_2 - M - 1)}{(n_1 + n_2)(n_1 + n_2 - 2)M} D^2, \quad (2.37)$$

onde M é a dimensão do espaço tangente. Esta estatística tem distribuição F_{M, n_1+n_2-M-1} sob a hipótese nula. Portanto, rejeita-se H_0 para valores grandes de F .

Perceba que para realizar o teste de hipótese utilizando a estatística T^2 de Hotelling, deve-se supor a igualdade entre as matrizes de covariância de v_i e w_i (2.35). Considere que esta suposição não possa ser feita. James (1954) propôs uma estatística de teste que leva em consideração a possível diferença entre as matrizes de covariância. A estatística de James é dada por

$$J = (\bar{v} - \bar{w})^T \left\{ \frac{1}{n_1} S_v + \frac{1}{n_2} S_w \right\}^{-1} (\bar{v} - \bar{w})$$

e tem distribuição assintótica χ_M^2 .

2.5.2 Teste de Goodall para Comparar Formas Médias de Duas Populações

Considere uma amostra aleatória $z_{1j}^0, \dots, z_{n_j j}^0$ de configurações complexas oriunda da população Π_j , $j = 1, 2$ (com Π_1 e Π_2 independentes). Sejam $z_{1j}, \dots, z_{n_j j}$ as pré-formas de $z_{1j}^0, \dots, z_{n_j j}^0$, onde z_{lj} é calculada de z_{lj}^0 por (2.4).

Supõe-se que ambas as populações têm variância comum σ^2 para cada coordenada (isotropia). Seja $[\mu_j]$ a forma média da população Π_j . Deseja-se testar as seguintes hipóteses:

$$H_0 : [\mu_1] = [\mu_2] (= [\mu_0]) \quad vs. \quad H_1 : [\mu_1] \neq [\mu_2].$$

Sejam $[\hat{\mu}_1]$ e $[\hat{\mu}_2]$ as médias Procrustes completas de cada amostra, e $\hat{\mu}_1$ e $\hat{\mu}_2$ as respectivas pré-formas médias amostrais. Sob H_0 , com σ pequeno, as distâncias Procrustes são aproximadamente distribuídas por

$$\begin{aligned} \sum_{i=1}^{n_1} d_F^2(z_{i1}, \hat{\mu}_1) &\sim \tau_0^2 \chi_{(n_1-1)M}^2, \\ \sum_{i=1}^{n_2} d_F^2(z_{i2}, \hat{\mu}_2) &\sim \tau_0^2 \chi_{(n_2-1)M}^2, \\ d_F^2(\hat{\mu}_1, \hat{\mu}_2) &\sim \tau_0^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \chi_M^2, \end{aligned}$$

onde $\tau_0 = \sigma / \|\mu_0\|$. Novamente, as provas dos resultados podem ser obtidas usando expansão em série de Taylor (veja Dryden e Mardia, 1998, p. 162). Em adição, tais estatísticas são aproximadamente mutuamente independentes (exatamente no caso das duas primeiras expressões).

Então, sob H_0 , tem-se o seguinte resultado aproximado:

$$F = \frac{n_1 + n_2 - 2}{n_1^{-1} + n_2^{-1}} \frac{d_F^2(\hat{\mu}_1, \hat{\mu}_2)}{\sum_{i=1}^{n_1} d_F^2(z_{i1}, \hat{\mu}_1) + \sum_{i=1}^{n_2} d_F^2(z_{i2}, \hat{\mu}_2)} \sim F_{M, (n_1 + n_2 - 2)M}. \quad (2.38)$$

Este resultado é válido para σ pequeno. Rejeita-se H_0 para valores grandes de F .

2.5.3 Teste de Hipótese Utilizando a Estatística λ para Comparar Formas Médias de Duas Populações

Seja $z_1, \dots, z_n$ uma amostra aleatória de pré-formas, onde cada z_i é dado em (2.4). Para $i = 1, \dots, n$, z_i pertence à hipersfera unitária CS^d , com $d = k - 1$, definida em (2.5). Seja $\hat{\mu}$ a forma média amostral (2.8). Escolha uma matriz $\widehat{M}$ de tal maneira que suas linhas sejam vetores unitários mutuamente ortogonais e $\widehat{M}\hat{\mu} = \mathbf{0}_{d-1}$, onde $\mathbf{0}_{d-1}$ é um vetor de zeros com tamanho $(d - 1) \times 1$. $\widehat{M}$ representa uma projeção no espaço tangente de $\hat{\mu}$.

Considere a estatística

$$T(\mu) = 2n\mu^* \widehat{M}^* \widehat{G}^{-1} \widehat{M} \mu, \quad (2.39)$$

onde μ é um vetor unitário complexo ($\mu^* \mu = 1$) de tamanho $d \times 1$; $\widehat{M}_{(d-1) \times d}$ é uma matriz complexa tal que $\widehat{M}\hat{\mu} = \mathbf{0}_{d-1}$, $\widehat{M}\widehat{M}^* = I_{d-1}$ e $\widehat{M}^* \widehat{M} = I_d - \mu\mu^*$; e $\widehat{G}$ é uma matriz complexa positiva definida hermitiana, que é um estimador consistente da matriz de covariância assintótica de $n^{1/2}\widehat{M}\mu_0$, sob a hipótese $H_0 : \mu = e^{i\theta}\mu_0$, onde θ está definido em (2.6) e μ_0 é um valor especificado.

Para escolher o $\widehat{G}$, considere que se construa uma matriz $\widehat{S}$ como em (2.7), baseando-se na amostra de pré-formas $z_1, \dots, z_n$ dada. Esta matriz tem d autovalores e d autovetores, cujo autovetor associado ao maior autovalor foi definido em (2.8) como a forma média amostral. Dados os autovalores $\hat{\lambda}_1 < \hat{\lambda}_2 < \dots < \hat{\lambda}_d$, exclua $\hat{\lambda}_d$ e tome os autovetores correspondentes aos autovalores remanescentes, isto é, $\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_{d-1}$. Escreva $\widehat{M} = [\hat{\mu}_1, \dots, \hat{\mu}_{d-1}]^*$. Designe $\widehat{\mathcal{G}}_{uv}$ como o elemento da u -ésima linha e v -ésima coluna de $\widehat{G}$, para $u, v = 1, \dots, d - 1$. Então,

$$\widehat{\mathcal{G}}_{uv} = \frac{1}{n} (\hat{\lambda}_d - \hat{\lambda}_u)^{-1} (\hat{\lambda}_d - \hat{\lambda}_v)^{-1} \times \sum_{j=1}^n (\hat{\mu}_u^* z_j) (z_j^* \hat{\mu}_v) (z_j^* \hat{\mu}) (\hat{\mu}^* z_j), \quad (2.40)$$

em que $\hat{\mu}$ é a forma média amostral dada em (2.8). Esta construção de $\widehat{G}$ (2.40) foi proposta em Bhattacharya & Patrangenaru (2003) e usada em Amaral et al. (2007), p. 3.

Considere agora duas amostras aleatórias $z_{11}, \dots, z_{n_1 1}$ (oriunda de Π_1) e $z_{12}, \dots, z_{n_2 2}$ (oriunda de Π_2), onde $z_{ij} \in CS^d$ para $i = 1, \dots, n_j$ e $j = 1, 2$, e seja $\hat{\mu}_j$ o estimador de μ_0 baseado na amostra j . Tem-se também que $n_j = n_j(n)$ denota uma função de n , o índice do tamanho da amostra. Suponha que $n_j(n) = n\psi_j(n)$, onde

$$\liminf_{n \rightarrow \infty} \min_{j=1,2} \psi_j(n) > 0, \quad (2.41)$$

$$\limsup_{n \rightarrow \infty} \max_{j=1,2} \psi_j(n) < \infty. \quad (2.42)$$

Para mais detalhes, veja Amaral et al. (2007).

Para $j = 1, 2$, é requerido que

$$n^{1/2} \widehat{M} \mu_0 \xrightarrow{\mathcal{D}} CN_{d-1}(0, G_j), \quad (2.43)$$

onde G_j tem posto completo e $CN_{d-1}(0, G_j)$ designa a distribuição normal complexa $(d-1)$ -variada (veja a seção 2.3.3). Seja $\widehat{G}_j$ um estimador consistente para G_j , em que $\widehat{G}_j$ é calculada da mesma maneira que $\widehat{G}$, isto é, usando (2.40) baseando-se na amostra oriunda da população Π_j para $j = 1, 2$. Defina

$$\widehat{\mathbf{A}}_0 = n \sum_{j=1}^2 \widehat{M}_j^* \widehat{G}_j^{-1} \widehat{M}_j \quad \text{e} \quad T_0(\mu) = 2\mu^* \widehat{\mathbf{A}}_0 \mu. \quad (2.44)$$

A estatística λ é dada por

$$\lambda \equiv \min_{\mu: \|\mu\|=1} T_0(\mu), \quad (2.45)$$

onde $T_0(\mu)$ é dado em (2.44).

Em Amaral et al. (2007) é apresentado o seguinte teorema que fornece a distribuição assintótica da estatística λ :

Teorema 2.5.1. *Suponha que (2.43) vale para $j = 1, 2$. Suponha também que $\widehat{G}_j$ é um estimador consistente de G_j e que (2.41) e (2.42) são verdade. Tem-se que, sob H_0 ,*

$$\lambda \xrightarrow{\mathcal{D}} \chi_{2(d-1)}^2 \quad \text{quando } n \rightarrow \infty.$$

Testes de Permutação e Testes Bootstrap

Neste capítulo, serão apresentados dois métodos úteis para a realização de testes de hipótese. Primeiro, será apresentado o conceito de testes de permutação com duas amostras oriundas de populações diferentes. Em seguida, será abordado o teste de hipótese com reamostragem de bootstrap. Depois de apresentadas as definições desses dois tipos de teste, veremos como eles se aplicam à análise estatística de formas para comparar formas médias de dois grupos, tomando como exemplo os testes de permutação e bootstrap da estatística λ , descrita na seção 2.5.3.

3.1 Testes de Permutação para Duas Populações

Esta é a principal aplicação de testes de permutação. Sejam $\mathbf{x} = (x_1, \dots, x_n)$ e $\mathbf{y} = (y_1, \dots, y_m)$ duas amostras aleatórias independentes em que $\mathbf{x}$ consiste em n realizações independentes de uma variável aleatória X com distribuição de probabilidade F_X e $\mathbf{y}$ representa m realizações independentes de uma variável aleatória Y distribuída segundo F_Y . Deseja-se testar:

$$H_0 : F_X = F_Y \quad \text{vs.} \quad H_1 : F_X \neq F_Y, \quad (3.1)$$

onde por $F_X = F_Y$, entende-se que $Prob_X\{A\} = Prob_Y\{A\}$, para qualquer subconjunto A do espaço amostral comum de X e Y . Se H_0 é verdadeira, não existe diferença entre os comportamentos probabilísticos das variáveis aleatórias X e Y .

Para se testar hipóteses, é necessário dispor de uma estatística de teste. Suponha que tenhamos à disposição a estatística de teste $T(\mathbf{x}, \mathbf{y}) = T$, que é uma função das amostras dadas. Sem perda de generalidade, considere que H_0 em (3.1) é rejeitada para valores grandes de T . Uma vez observada T , defina o *nível descritivo* ou *nível de significância alcançado* (ASL), a menor probabilidade possível acima da qual H_0 não é rejeitada. Em outras palavras:

$$ASL = Prob_{H_0}\{T^* \geq T\}. \quad (3.2)$$

O termo T em (3.2) é fixado como um valor observado e T^* designa uma variável aleatória distribuída sob H_0 , isto é, a distribuição de T se H_0 é verdade. Quanto menor for o ASL , mais forte é a evidência contra a veracidade de H_0 .

Testar a hipótese H_0 consiste em calcular o ASL , verificar se este é pequeno (ou grande) e, baseado nesta informação, tomar uma decisão. Por exemplo, se adotarmos o nível de significância $\alpha = 0.1$ para o teste e calcularmos um $ASL < 0.1$, rejeitamos H_0 .

O teste de permutação de Fisher é uma maneira hábil de se calcular um ASL para uma hipótese nula como em (3.1). O procedimento do teste é o seguinte: combinar as duas amostras $\mathbf{x}$ e $\mathbf{y}$ obtendo uma amostra de tamanho $n + m$; a partir da amostra combinada, tomar uma amostra, sem reposição, de tamanho n para designar o primeiro grupo e os m elementos restantes para designar o segundo grupo; calcula-se a estatística T para as duas novas amostras, fazendo isso um grande número de vezes. Se o valor da estatística T aplicada às amostras $\mathbf{x}$ e $\mathbf{y}$ originais não figurar entre os 95% valores calculados centrais, rejeita-se H_0 ao nível 5%.

Formalizando, seja $\mathbf{u}_0 = (\mathbf{x}, \mathbf{y})$ uma amostra formada combinando as amostras $\mathbf{x}$ e $\mathbf{y}$. Por conveniência, ordene os elementos de $\mathbf{u}_0$ e obtenha $\mathbf{u} = (u_1, u_2, \dots, u_r)$, em que $r = n + m$. Uma vez ordenados os elementos de $\mathbf{u}_0$, defina um vetor $\mathbf{p} = (p_1, p_2, \dots, p_r)$ que designa a qual população o i -ésimo elemento de $\mathbf{u}$ pertence. Isto é, se o elemento u_i pertencente a $\mathbf{u}$ foi obtido da amostra $\mathbf{x}$, tem-se que $p_i = x$, caso contrário, $p_i = y$. Dessa forma, temos que o vetor $\mathbf{p}$ é formado por n x 's e m y 's. Por exemplo, se $n = 2$ e $m = 3$, uma possível observação do vetor $\mathbf{p}$ seria $\mathbf{p} = (x, y, y, x, y)$. Existem 10 vetores diferentes $\mathbf{p}$ formados por dois x 's e três y 's. Assim como, no caso geral, existem $r!/(n!m!) = P_r^{n,m}$ (permutação de r elementos com repetição de n

e m elementos) maneiras de se obter o vetor $\mathbf{p}$.

O seguinte lema é de suma importância para realização de testes de permutação [Efron & Tibshirani (1993)].

Lema 3.1.1 (Lema de Permutação). *Sob $H_0 : F_X = F_Y$ (3.1), qualquer uma das observações do vetor $\mathbf{p}$ tem probabilidade de ocorrência $1/P_r^{n,m}$, onde $P_r^{n,m} = \binom{r}{n}$.*

Considere agora uma estatística T que possa ser escrita como função dos vetores $\mathbf{u}$ e $\mathbf{p}$, digamos $T = f(\mathbf{u}, \mathbf{p})$, e de tal maneira que, para valores grandes de T , rejeite-se a hipótese nula.

Por exemplo, se estivéssemos interessados em avaliar se as médias de duas populações são iguais (como é o nosso caso), poderíamos pensar na estatística T como a diferença entre as médias das duas populações, isto é, $T = \bar{x} - \bar{y}$. Escrevendo T como função de $\mathbf{u}$ e $\mathbf{p}$, teríamos:

$$T = \frac{1}{n} \sum_{p_i=x} u_i - \frac{1}{m} \sum_{p_i=y} u_i,$$

em que $\sum_{p_i=x} u_i$ indica a soma de todos os u_i 's para os quais $p_i = x$.

Seja $\mathbf{p}^*$ um vetor indicando qualquer uma das possíveis $P_r^{n,m}$ permutações dos elementos do vetor $\mathbf{p}$ e defina *réplica de permutação* de T como

$$T^* = T(\mathbf{p}^*) = f(\mathbf{p}^*, \mathbf{u}).$$

Existem $P_r^{n,m}$ réplicas de permutação, cada uma com a mesma probabilidade de ocorrência. Isto decorre do Lema 3.1.1.

Define-se *nível descritivo de permutação* (ASL_{perm}) como a probabilidade de T^* exceder T , ou seja,

$$ASL_{\text{perm}} = \text{Prob}\{T^* \geq T\} = \frac{\#\{T^* \geq T\}}{P_r^{n,m}}, \quad (3.3)$$

onde $\#\{T^* \geq T\}$ indica o número de vezes em que $T^* \geq T$.

Há casos em que é mais conveniente aproximar o ASL_{perm} via simulação de Monte Carlo, devido à grande quantidade de indivíduos de determinados grupos em estudo. Veja, por exemplo, os tamanhos amostrais dos grupos macho e fêmea da espécie *Caluromys philander* na tabela 4.2. Se fôssemos calcular o ASL_{perm} pelo modo direto, seria necessário calcular $(50 + 52)! = 102!$.

Em Efron & Tibshirani (1993), encontra-se o seguinte algoritmo para aproximar o ASL_{perm} .

Algoritmo 3.1.1. *Cálculo do $\widehat{ASL}_{\text{perm}}$.*

1. *Selecione aleatoriamente B permutações entre as $P_r^{n,m}$ permutações do vetor $\mathbf{p}$, digamos $\mathbf{p}^*(1), \mathbf{p}^*(2), \dots, \mathbf{p}^*(B)$.*
2. *Calcule as réplicas de permutação de T em cada um dos B vetores obtidos no passo 1:*

$$T^*(b) = f(\mathbf{p}^*(b), \mathbf{u}), \quad b = 1, \dots, B.$$

3. *Estima-se o ASL_{perm} por:*

$$\widehat{ASL}_{\text{perm}} = \frac{\#\{T^*(b) \geq T\}}{B}.$$

3.2 Testes Bootstrap para Duas Populações

Para o teste bootstrap com duas populações, considere suposições similares àquelas referentes ao teste de permutação, isto é, deseja-se testar as mesmas hipóteses que em (3.1) e se dispõe das amostras aleatórias independentes $\mathbf{x} = (x_1, \dots, x_n)$ e $\mathbf{y} = (y_1, \dots, y_m)$, como descritas na seção 3.1. Segundo Efron & Tibshirani (1993), os testes bootstrap e os testes de permutação apresentam resultados similares, porém isto pode não ser verdade em situações complexas em análise estatística de formas.

Um teste de hipótese bootstrap, assim como um teste de permutação, baseia-se numa estatística de teste. Podemos pensar novamente em uma estatística T designando a diferença entre as médias de duas populações em estudo. Semelhante ao caso de permutação, defina $T = T_{\text{obs}} = \bar{x} - \bar{y}$. O nível de significância alcançado é definido como em (3.2).

Defina $\mathbf{u} = (\mathbf{x}, \mathbf{y})$ uma amostra combinada de $\mathbf{x}$ e $\mathbf{y}$, mas dessa vez, não é necessário que seus elementos estejam ordenados. Para calcular o $\widehat{ASL}_{\text{boot}}$, segue o algoritmo.

Algoritmo 3.2.1. *Cálculo do $\widehat{ASL}_{\text{boot}}$.*

1. *Retire B amostras de tamanho $r = n + m$ com reposição de $\mathbf{u}$: $\mathbf{u}^*(1), \mathbf{u}^*(2), \dots, \mathbf{u}^*(B)$.
Seja $\mathbf{x}^*(b)$ o vetor formado pelos primeiros n elementos de $\mathbf{u}^*(b)$ e $\mathbf{y}^*(b)$ o vetor formado pelos demais m elementos de $\mathbf{u}^*(b)$, com $b = 1, 2, \dots, B$.*

2. Calcule T em cada amostra,

$$T(\mathbf{u}^*(b)) = \bar{\mathbf{x}}^*(b) - \bar{\mathbf{y}}^*(b), \quad b = 1, 2, \dots, B.$$

3. O $\widehat{ASL}_{\text{boot}}$ é dado por

$$\widehat{ASL}_{\text{boot}} = \frac{\#\{T(\mathbf{u}^*(b)) \geq T_{\text{obs}}\}}{B}, \quad (3.4)$$

onde T_{obs} é o valor observado da estatística T .

Quando se faz o teste de permutação, as duas amostras aleatórias retiradas da amostra combinada (seção 3.1) são feitas sem reposição. Neste caso, podemos interpretar o ASL exato como a probabilidade do evento $\{T^* \geq T\}$ ocorrer, calculada em todo o espaço amostral (3.3). A cardinalidade do espaço amostral é finita e igual a $P_r^{n,m}$. Por ser difícil de calcular, este ASL é estimado.

No caso do teste bootstrap, o procedimento equivale a retirar duas amostras aleatórias independentes (uma com n e outra com m elementos) de $\mathbf{u}$, mas agora com reposição. O $\widehat{ASL}_{\text{boot}}$ é uma aproximação para o ASL_{boot} , mas não tem interpretação como uma probabilidade exata.

3.3 Teste de Permutação e Teste Bootstrap Utilizando a Estatística λ

Agora, serão apresentados os passos para se realizar testes de permutação e bootstrap utilizando a estatística λ descrita na seção 2.5.3. Para o teste bootstrap será necessário *transformar* a hipótese nula, uma vez que, tipicamente, os dados não a satisfarão [ver Amaral et al. (2007)].

3.3.1 Teste de Permutação com a Estatística λ

Dadas as amostras aleatórias de pré-formas $z_{11}, \dots, z_{n11}$ (oriunda de Π_1) e $z_{12}, \dots, z_{n22}$ (oriunda de Π_2), deseja-se testar:

$$H_0 : [\mu_1] = [\mu_2] \quad vs. \quad H_1 : [\mu_1] \neq [\mu_2], \quad (3.5)$$

onde $[\mu_1]$ e $[\mu_2]$ são as formas médias das populações Π_1 e Π_2 respectivamente.

Será apresentado a seguir um algoritmo para realizar um teste de permutação utilizando a estatística λ para verificar a igualdade entre as formas médias das duas populações.

Algoritmo 3.3.1. *Teste de Permutação com a estatística λ .*

1. Para ambas amostras ($j = 1, 2$), obtenha $\widehat{\mu}_j$, $\widehat{M}_j$, $\widehat{G}_j$.
2. Determine o menor autovalor, λ , e o autovetor correspondente, $\widehat{\mu}_0$, da seguinte matriz hermitiana:

$$n \sum_{j=1}^2 \widehat{M}_j^* \widehat{G}_j^{-1} \widehat{M}_j. \quad (3.6)$$

3. Para cada ciclo de permutação b ($b = 1, \dots, B$), amostre aleatoriamente a partir da amostra combinada $z_{11}, \dots, z_{n_1 1}, z_{12}, \dots, z_{n_2 2}$, sem reposição, para formar duas novas amostras de tamanhos n_1 e n_2 ; calcule $\widehat{\mu}_j^{(b)}$, $\widehat{M}_j^{(b)}$ e $\widehat{G}_j^{(b)}$; determine o menor autovalor, $\lambda^{(b)}$, da matriz

$$n \sum_{j=1}^2 (\widehat{M}_j^{(b)})^* (\widehat{G}_j^{(b)})^{-1} \widehat{M}_j^{(b)},$$

isto é, uma matriz como em (3.6), substituindo $\widehat{M}_j$ e $\widehat{G}_j$ por $\widehat{M}_j^{(b)}$ e $\widehat{G}_j^{(b)}$.

4. O p -valor de permutação é dado por

$$\widehat{ASL}_{\text{perm}} = \frac{(1 + \#\{b : \lambda^{(b)} \geq \lambda\})}{(B + 1)},$$

onde λ é obtido no passo 2.

Note que no passo 4, seguiu-se Davison & Hinkley (1997, seção 4.3, p. 158), adicionando 1 ao numerador e ao denominador na determinação do $\widehat{ASL}_{\text{perm}}$.

3.3.2 Teste Bootstrap com a estatística λ

Um ponto importante no teste de hipótese bootstrap é que a reamostragem deve ser feita sob a hipótese nula. Tipicamente, os dados não satisfarão tal hipótese exatamente, por isso é preciso que seja feita uma transformação preliminar nos dados. Vejamos como é feita esta transformação, antes de apresentarmos o algoritmo para realização do teste bootstrap.

Suponha que η e γ sejam vetores unitários pertencentes a CS^d (2.5). Deseja-se escolher uma pré-forma $\tilde{\eta}$ da forma $[\eta]$ de η , de tal maneira que $\tilde{\eta}$ mova-se ao longo de uma geodésica *horizontal*

no espaço de pré-formas, que corresponde a uma geodésica no espaço de formas [ver Kendall et al. (1999, cap. 6)]. Na prática, a escolha de $\tilde{\eta}$ é feita da seguinte maneira:

$$\tilde{\eta} = \frac{\eta(\eta^*\gamma)}{|\eta^*\gamma|}, \quad (3.7)$$

em que $\tilde{\eta}$ tem a mesma norma de η e $\tilde{\eta}^*\gamma \in \mathbb{R}$.

Defina

$$\begin{aligned} \tilde{c} &= \frac{\tilde{\eta} - \gamma(\gamma^*\tilde{\eta})}{\|\tilde{\eta} - \gamma(\gamma^*\tilde{\eta})\|}, \\ \tilde{\alpha} &= \arccos(\gamma^*\tilde{\eta}) \in [0, \pi), \\ \tilde{A} &= \gamma\tilde{c}^* - \tilde{c}\gamma^*. \end{aligned}$$

Lema 3.3.1. *Suponha que γ e η pertencem a CS^d (2.5), isto é $\|\gamma\|^2 = \gamma^*\gamma = 1$ e $\|\eta\|^2 = \eta^*\eta = 1$, e suponha que $|\gamma^*\eta| < 1$. Dados $\tilde{\alpha}$, $\tilde{A}$ e $\tilde{c}$ definidos acima, a matriz*

$$\mathbf{U} = \exp(\tilde{\alpha}\tilde{A}) = I_d + \sum_{j=1}^{\infty} \frac{\tilde{\alpha}^j}{j!} \tilde{A}^j$$

tem as seguintes propriedades:

(i) $\mathbf{U}$ é uma matriz unitária $d \times d$;

(ii) $\mathbf{U}$ pode ser escrita como

$$\mathbf{U} = I_d + \sin(\tilde{\alpha})\tilde{A} + \{\cos(\tilde{\alpha}) - 1\}(\gamma\gamma^* + \tilde{c}\tilde{c}^*);$$

(iii) $\mathbf{U}\tilde{\eta} = \gamma$;

(iv) $\forall z \in \mathbb{C}^d$ tal que $\gamma^*z = 0$ e $\eta^*z = 0$, tem-se $\mathbf{U}z = z$.

A prova deste resultado para dados direcionais pode ser encontrada em Amaral et al. (2007).

Para o caso de formas, a prova é análoga.

Quando este resultado é aplicado na prática, toma-se $\eta = \hat{\mu}_i$ e $\gamma = \hat{\mu}_0$, define-se $\tilde{\eta}$ como em (3.7), e se toma $\mathbf{U}_i$ como uma matriz unitária dada pelo Lema 3.3.1.

No primeiro passo do algoritmo bootstrap, é necessário calcular a matriz $\hat{G}_j$, para $j = 1, 2$. Amaral et al. (2007) notam que a escolha de $\hat{G}_j$ pode depender de se é suposta estrutura de

dispersão comum entre as populações. Supondo que não haja estrutura de dispersão comum, calcula-se $\widehat{G}_j$ com base na amostra j , $j = 1, 2$, como em (2.40). Caso contrário, se é considerada esta estrutura comum, uma versão combinada de (2.40) pode ser usada para obter $\widehat{G}$ e definir $\widehat{G}_j = \rho_j \widehat{G}$, em que ρ_j é um multiplicador escalar (um “peso”) que leva em conta os diferentes tamanhos amostrais. Ainda em Amaral et al. (2007), são feitas considerações sobre a escolha de $\widehat{G}$ baseada na existência ou não de simetria rotacional.

Dadas as amostras aleatórias de pré-formas $z_{11}, \dots, z_{n_1 1}$ (oriunda de Π_1) e $z_{12}, \dots, z_{n_2 2}$ (oriunda de Π_2), deseja-se testar:

$$H_0 : [\mu_1] = [\mu_2] \quad vs. \quad H_1 : [\mu_1] \neq [\mu_2], \quad (3.8)$$

onde $[\mu_1]$ e $[\mu_2]$ são as formas médias das populações Π_1 e Π_2 respectivamente.

Será apresentado a seguir um algoritmo para realizar um teste bootstrap utilizando a estatística λ , para verificar a igualdade entre as formas médias de Π_1 e Π_2 .

Algoritmo 3.3.2. *Teste Bootstrap com a estatística λ .*

1. Para ambas amostras ($j = 1, 2$), obtenha $\widehat{\mu}_j$, $\widehat{M}_j$, $\widehat{G}_j$.
2. Determine o menor autovalor, λ , e o autovetor correspondente, $\widehat{\mu}_0$, da seguinte matriz hermitiana:

$$n \sum_{j=1}^2 \widehat{M}_j^* \widehat{G}_j^{-1} \widehat{M}_j. \quad (3.9)$$

3. Transforme ambas amostras ($j = 1, 2$) usando a matriz unitária $\mathbf{U}_j$ obtida no Lema 3.3.1, que move $\widehat{\mu}_j$ a $\widehat{\mu}_0$ ($j = 1, 2$).
4. Para cada ciclo bootstrap b ($b = 1, \dots, B$), a partir da amostra j transformada usando $\mathbf{U}_j$ (Lema 3.3.1), retire uma nova amostra, com reposição, de tamanho n_j ($j = 1, 2$); calcule $\widehat{\mu}_j^{(b)}$, $\widehat{M}_j^{(b)}$ e $\widehat{G}_j^{(b)}$; determine o menor autovalor, $\lambda^{(b)}$, da matriz

$$n \sum_{j=1}^2 (\widehat{M}_j^{(b)})^* (\widehat{G}_j^{(b)})^{-1} \widehat{M}_j^{(b)},$$

isto é, uma matriz como em (3.9), substituindo $\widehat{M}_j$ e $\widehat{G}_j$ por $\widehat{M}_j^{(b)}$ e $\widehat{G}_j^{(b)}$.

5. O *p*-valor bootstrap é dado por

$$\widehat{ASL}_{\text{boot}} = \frac{(1 + \#\{b : \lambda^{(b)} \geq \lambda\})}{(B + 1)},$$

onde λ é obtido no passo 2.

Note que no passo 5 seguiu-se Davison & Hinkley (1997, seção 4.4, p. 161), adicionando 1 ao numerador e ao denominador na determinação do $\widehat{ASL}_{\text{boot}}$.

Análise de Dados de Zoologia

A análise estatística de formas de objetos, ou morfometria geométrica, é uma ferramenta muito útil em algumas áreas da ciência, como, por exemplo, na Zoologia. Neste capítulo, discutiremos uma de suas aplicações ao estudo de algumas espécies de marsupiais americanos.

Tradicionalmente visto como um grupo morfologicamente conservador, estes animais apresentam, no entanto, diversas modificações morfológicas relacionadas às suas características ecológicas. Através de técnicas de análise estatística de formas, pode-se avaliar e quantificar a diversidade e diferenciação morfológica existente nos marsupiais americanos atuais, no intuito de poder relacionar esta diversidade aos seus hábitos alimentares, locomotores e seu relacionamento filogenético [Astúa de Moraes (2004)].

O objeto de estudo aqui são as mandíbulas dos marsupiais. Segundo Astúa de Moraes (2004), a forma do crânio e, sobretudo, da mandíbula, apresenta correlação com os hábitos alimentares dos marsupiais americanos, indicando que a diferenciação sutil na especialização alimentar entre espécies de dietas qualitativamente similares se reflete em modificações sutis de um plano morfológico básico, que se mantêm ao longo da filogenia. Filogenia é o termo comumente utilizado para hipóteses de relações evolutivas (ou seja, relações filogenéticas) de um grupo de organismos, isto é, determinar as relações ancestrais entre espécies conhecidas (ambas as que vivem e as extintas) [veja Amorim (1997)].



Figura 4.1: *Didelphis albiventris*



Figura 4.2: Mandíbula do *D. albiventris*

O objetivo principal desta análise é identificar dimorfismo sexual, baseando-se nas formas médias das mandíbulas de machos e fêmeas de cada espécie. A análise de dimorfismo sexual significa quantificar as diferenças no tamanho e na forma de alguma parte do corpo (como crânio, mandíbula, escápula, etc.) entre macho e fêmea de determinada espécie. Oliveira et al. (2005), por exemplo, analisou as diferenças no tamanho e forma dos crânios de machos e fêmeas de 346 espécimes adultos de *Arctocephalus australis* (lobo-marinho) utilizando morfometria geométrica. Foi verificada presença de substancial dimorfismo sexual no tamanho e na forma dos crânios de *A. australis*. As diferenças na forma e no tamanho do crânio entre machos e fêmeas desta espécie podem estar associadas à seleção sexual, a sistemas reprodutivos e à história de vida.



Figura 4.3: *Didelphis marsupialis*



Figura 4.4: *Didelphis aurita*

As seis espécies de marsupiais americanos consideradas neste estudo pertencem à família

Didelphidae. São elas: *Didelphis albiventris*, *Didelphis aurita*, *Didelphis marsupialis*, *Caluromys philander*, *Caluromys derbianus* e *Marmosops incanus*.

4.1 Descrição dos Dados

O banco de dados utilizado neste trabalho foi cedido pelo Professor Doutor Diego Astúa de Moraes, do Departamento de Zoologia da Universidade Federal de Pernambuco. Os dados correspondem a coordenadas cartesianas bidimensionais referentes a mandíbulas de 418 indivíduos (marsupiais da família *Didelphidae*). Para cada indivíduo, dispõe-se de um conjunto de coordenadas referentes a 19 marcos anatômicos na mandíbula do marsupial. Os dados foram tratados para ficarem dispostos na forma de matrizes de configuração, como descrito em (2.1). Tem-se, portanto, 418 matrizes de dimensão 19×2 , onde cada linha de uma matriz corresponde à coordenada no plano de um marco anatômico e cada matriz representa o conjunto de coordenadas de marcos na mandíbula de um indivíduo.

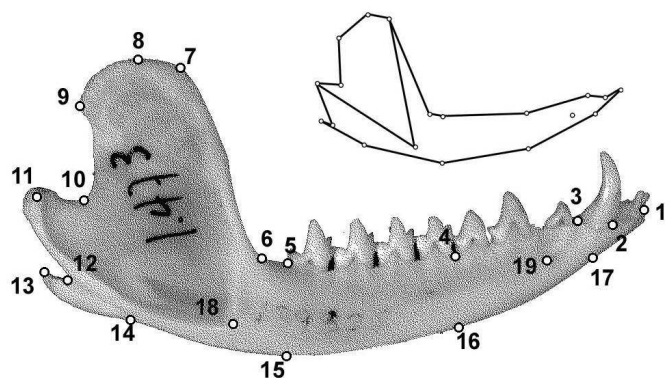


Figura 4.5: Marcos anatômicos na mandíbula do *Lutreolina crassicaudata*



Figura 4.6: *Lutreolina crassicaudata*

Para obter esses dados, foram tiradas fotografias das mandíbulas dos espécimes, alinhando o plano do processo coronóide paralelo à base e à lente da câmera (lado labial direcionado para a lente). Quando os dois ramos mandibulares encontravam-se fundidos, a mandíbula foi apoiada em uma caixa furada, de modo que um ramo ficasse paralelo à base e à lente da câmera e o outro suspenso, no interior da caixa. Os marcos anatômicos foram digitalizados nas imagens com o software TPSDig. Este software foi desenvolvido por F. James

Rohlf da State University of New York at Stone Brook e pode ser obtido através do site <http://life.bio.sunysb.edu/morph/soft-dataacq.html>. Para mais detalhes técnicos relacionados à aquisição das imagens, veja Astúa de Moraes (2004, pp. 56-59).

Em cada exemplar de mandíbula foram determinados 19 marcos anatômicos. A figura 4.5 ilustra os 19 marcos pontuados na mandíbula de um marsupial da espécie *Lutreolina crassicaudata*, pertencente à mesma família que os marsupiais em estudo, portanto a escolha dos marcos é equivalente para todas as espécies mencionadas. A descrição dos marcos anatômicos pode ser vista na tabela 4.1.

Marcos anatômicos	Descrição
(1)	Base do primeiro incisivo inferior.
(2)	Base do quarto incisivo inferior.
(3)	Extremidade posterior do alvéolo do canino inferior.
(4)	Base do terceiro premolar e do primeiro molar inferior direito, no ponto onde os dentes se encontram com o dentário.
(5)	Base posterior do quarto molar inferior direito, no ponto onde o dente se encontra com o dentário.
(6)	Ponto de encontro entre o ramo horizontal da mandíbula e o início do processo coronóide (ponto de maior inflexão e base do processo coronóide).
(7)	Ponto de maior inflexão (extremidade dorsal) da borda anterior do processo coronóide.
(8)	Ponto mais alto (mais dorsal) do processo coronóide.
(9)	Extremidade posterior do processo coronóide (início da borda posterior do processo coronóide).
(10)	Ponto de maior inflexão da curvatura entre o processo articular e a borda posterior do processo coronóide (base da borda posterior).
(11)	Extremidade lateral ou labial do côndilo articular.
(12)	Base posterior do processo angular.
(13)	Extremidade do processo angular.
(14)	Base anterior do processo angular.
(15)	Projeção do marco anatômico 5 na borda ventral da mandíbula, perpendicularmente à linha formada pelos marcos 4 e 5.
(16)	Projeção do marco anatômico 4 na borda ventral da mandíbula, perpendicularmente à linha formada pelos marcos 4 e 5.
(17)	Projeção do marco anatômico 3 na borda ventral da mandíbula, perpendicularmente à linha formada pelos marcos 3 e 4.
(18)	Extremidade antero-ventral da fossa massetérica.
(19)	Forâmen mental anterior.

Tabela 4.1: Descrição dos marcos anatômicos.

4.2 Resultados

Foram realizados testes de hipótese para verificar se as formas médias das mandíbulas de machos e fêmeas são iguais. Os testes foram T^2 de Hotelling, de Goodall, teste com a estatística de James e o teste com a estatística λ , obtendo-se os p -valores usuais apresentados na tabela 4.3. Com base em cada uma das quatro estatísticas envolvidas nos testes mencionados foram ainda realizados testes de permutação e testes bootstrap (com 200 réplicas de permutação e 200 réplicas bootstrap para cada teste). Todos os testes foram realizados utilizando a função `resampletest`, presente no pacote `shapes` do R, versão 2.6.0. Esta função recebe os dados, que devem estar dispostos como matrizes de tamanho $k \times m$ (onde k é o número de marcos anatômicos e m é a dimensão em que a forma está sendo estudada), e retorna os p -valores referentes aos quatro testes citados. Para mais detalhes sobre o pacote `shapes`, consulte <http://www.maths.nott.ac.uk/~ild/shapes>.

A primeira coluna da tabela 4.3 contém abreviaturas dos nomes científicos das espécies estudadas; são elas: ALB (*D. albiventris*), AUR (*D. aurita*), MAR (*D. marsupialis*), PHI (*C. philander*), DER (*C. derbianus*) e INC (*M. incanus*). Na tabela 4.2, N_m e N_f remetem ao número de indivíduos dos grupos macho e fêmea, respectivamente, para cada espécie; N_T é o número de indivíduos em cada amostra de espécie. Note que somente para as espécies *D. albiventris* e *M. incanus*, $N_T = N_m + N_f$. Isto porque no banco de dados não havia informação do sexo de alguns indivíduos; por exemplo, 6 dos 68 indivíduos do grupo *C. derbianus* foram excluídos das amostras para os testes de dimorfismo sexual.

Espécie	N_m	N_f	N_T
<i>D. albiventris</i>	29	31	60
<i>D. aurita</i>	28	27	57
<i>D. marsupialis</i>	27	33	65
<i>C. philander</i>	52	50	104
<i>C. derbianus</i>	30	32	68
<i>M. incanus</i>	34	30	64

Tabela 4.2: Tamanho das amostras por grupo.

Pode ser visto na tabela 4.3 que os p -valores dos testes T^2 de Hotelling e James com bootstrap

foram muito diferentes daqueles respectivos p -valores resultantes dos testes feitos da maneira usual e com permutação. Por exemplo, o teste de James na forma usual forneceu p -valor 0 para a espécie *D. aurita*, enquanto o teste com bootstrap da mesma estatística forneceu p -valor 0.8407; esta grande diferença deve-se ao fato de os tamanhos amostrais terem sido muito pequenos, $N_m = 28$ e $N_f = 27$ (veja a tabela 4.2). Por outro lado, os testes de permutação e os testes na forma usual, para estas duas estatísticas, apresentaram p -valores próximos, como é o caso da espécie *C. derbianus* que, no teste T^2 de Hotelling, apresentou p -valor usual 0.0945 e p -valor de permutação 0.1094.

Observou-se que os testes de Goodall de permutação e com bootstrap apresentaram p -valores muito próximos. E, em alguns casos, iguais até a quarta casa decimal; a espécie *M. incanus*, por exemplo, apresentou p -valores de permutação e bootstrap iguais a 0.0049. Os p -valores usuais da estatística do teste de Goodall foram muito diferentes dos p -valores de permutação e bootstrap no caso das espécies *C. philander* e *C. derbianus*.

Os testes com a estatística λ foram os que apresentaram os menores p -valores. Se adotássemos um nível de significância 0.01, os testes na forma usual rejeitariam a hipótese de igualdade entre as formas médias das mandíbulas de macho e fêmea para todas as espécies. Comparando os testes de permutação e bootstrap com a estatística λ , vê-se que os p -valores diferiram para algumas espécies, como, por exemplo, *C. derbianus*, para a qual observou-se p -valores de permutação e de bootstrap iguais a 0.0597 e 0.1293, respectivamente; no entanto, para a espécie *D. aurita*, tais p -valores foram iguais até a quarta casa decimal (0.0049).

É bastante notável na tabela 4.3 a diferença entre os p -valores associados aos testes relacionados à espécie *C. philander* e aqueles referentes às demais espécies. Os p -valores, principalmente os de permutação e bootstrap, referentes aos testes com *C. philander* distaram razoavelmente pouco entre si e bastante se comparados à maior parte dos p -valores associados a testes com outras espécies. Isto pode ser devido ao tamanho das amostras dos grupos macho e fêmea da espécie *C. philander* (52 e 50, respectivamente) serem bem maiores que os referentes às demais espécies; por exemplo, para *D. aurita*, $N_m = 28$ e $N_f = 27$ (tabela 4.2). O teste bootstrap funciona melhor para tamanhos de amostras grandes; os resultados são mais confiáveis à medida

que o tamanho da amostra aumenta.

A estatística λ é assintoticamente pivotal, portanto, espera-se uma redução no erro das estimativas com a aplicação do método bootstrap.

De certo modo, os resultados dos testes refletem o fato que dados reais estão sendo utilizados. Comparações destes testes devem ser feitas com estudos de simulação. Ao se trabalhar com dados reais, algumas suposições impostas em casos de simulação não podem ser fixadas. Por exemplo, no caso de simulação, podemos produzir um conjunto de dados com baixa concentração, porém, para dados reais, conjuntos de dados com tais características não são freqüentemente observados.

4.2.1 Uso dos Testes por Zoólogos

Um profissional interessado no uso destes testes pode perguntar: Qual é o melhor teste? Que teste devo usar para que situação? Estas perguntas podem ser respondidas a partir de estudos de simulação. Em Amaral et al. (2007), por exemplo, encontram-se estudos de simulação referentes aos mesmos testes T^2 de Hotelling, de Goodall, James e o teste com a estatística λ , cada um feito da maneira usual, com permutação e com bootstrap. Nesse artigo, foram simuladas pré-formas (vetores complexos unitários) com quatro marcos anatômicos, diretamente da distribuição de Bingham complexa (seção 2.3.2). Também foram feitos estudos gerando dados diretamente no espaço de marcos anatômicos, simulando vetores distribuídos segundo uma normal multivariada.

Os testes de Goodall e T^2 de Hotelling são baseados na normalidade das coordenadas tangentes. Esta propriedade só se verifica na prática quando a variabilidade das formas em relação à forma média é pequena. Por outro lado, os testes de permutação e de bootstrap tentam solucionar este problema. Deve-se notar que o aspecto da variabilidade pequena é relativo e deve ser interpretado como tal.

As estatísticas λ e de James são mais gerais do que aquelas de Hotelling e Goodall. A de Goodall apresenta a restrição de mesmas variâncias dos marcos anatômicos e a T^2 de Hotelling é obtida assumindo-se que os grupos apresentam a mesma matriz de covariâncias.

Testes de hipótese em análise estatística de formas podem ser úteis não somente na detecção de dimorfismo sexual, mas também na definição de novas espécies. Alguns critérios de identi-

ificação de espécies na prática são válidos, mas um teste indicando a igualdade de formas entre espécies, pode invalidar tais critérios. Por exemplo, as espécies de peixe-agulha *Hemiramphus brasiliensis* e *Hemiramphus balao* [veja Araújo (2008)].

Espécie	Usual				Permutação				Bootstrap			
	H	G	J	λ	H	G	J	λ	H	G	J	λ
ALB	0.0105	0	0	1.66×10^{-15}	0.0099	0.0049	0.0099	0.0049	0.7661	0.0049	0.7661	0.0248
AUR	0.0028	0	0	0	0.0099	0.0049	0.0099	0.0049	0.8407	0.0049	0.8407	0.0049
MAR	0.0170	0.0064	0	0	0.0149	0.1144	0.0248	0.0049	0.6616	0.0746	0.6865	0.0049
PHI	0.2194	0.4123	0.088	0.0003	0.2388	0.3333	0.2288	0.0447	0.5273	0.3681	0.5223	0.0895
DER	0.0945	0.1037	0	2.14×10^{-5}	0.1094	0.2189	0.1094	0.0597	0.8457	0.2835	0.8457	0.1293
INC	3.03×10^{-7}	0	0	0	0.0049	0.0049	0.0049	0.0049	0.1791	0.0049	0.1791	0.0049

Tabela 4.3: *P*-valores dos testes para verificar dimorfismo sexual

Este trabalho interdisciplinar proporcionou benefícios para zoólogos e estatísticos. Este tipo de trabalho desperta nos estatísticos a motivação para desenvolver novos métodos. Além disso, pesquisas interdisciplinares fazem com que os zoólogos passem a conhecer novos modelos. Alguns até bem conhecidos pelos estatísticos, mas pouco utilizados por profissionais de zoologia, biologia e medicina. Por exemplo, os testes com a estatística λ e com a estatística de James, que requerem suposições mais fracas que os outros testes trabalhados (Hotelling e Goodall).

Pelas suposições dos testes T^2 de Hotelling e de Goodall, percebe-se que estes são bem restritivos, uma vez que o primeiro exige a igualdade entre as matrizes de covariância dos grupos que estão sendo comparados e o segundo restringe que os marcos anatômicos tenham a mesma variância. Outros testes, como o de James e o teste com a estatística λ , são menos restritivos.

Uma possível solução para o bom funcionamento dos testes é trabalhar com as versões de permutação e bootstrap dos mesmos. Testes de permutação e testes com bootstrap são não paramétricos e têm a vantagem de não impor nenhuma suposição sobre a distribuição dos dados.

Dos testes realizados com os dados dos marsupiais, em geral, os testes de permutação sugerem a presença de dimorfismo sexual nas espécies estudadas que pertencem ao gênero *Didelphis* e na espécie *Marmosops incanus*. O mesmo ocorreu nas versões bootstrap dos testes de Goodall e com λ referentes a estas mesmas espécies. Já com os marsupiais estudados do gênero *Caluromys*,

apenas o teste de permutação com a estatística λ rejeitou a hipótese de igualdade entre as formas das mandíbulas.

Referências Bibliográficas

- [1] Amaral, G., Dryden, I. & Wood, A. (2007), ‘Pivotal bootstrap methods for k -sample problems in directional statistics and shape analysis’, *Journal of the American Statistical Association* **102**, 695-707.
- [2] Amorim, D. (1997), *Elementos Básicos de Sistemática Filogenética*, Holos Editora & Sociedade Brasileira de Entomologia, Ribeirão Preto.
- [3] Araújo, L. (2008), Agrupamento em análise de formas, Dissertação de mestrado, Universidade Federal Rural de Pernambuco.
- [4] Astúa de Moraes, D. (2004), Evolução morfológica do crânio e elementos pós-cranianos dos marsupiais americanos (Didelphimorphia, Paucituberculata, Microbiotheria), Tese de doutorado, Universidade de São Paulo.
- [5] Axler, S. (1997), *Linear Algebra Done Right*, second ed., Springer-Verlag, New York.
- [6] Bhattacharya, R. & Patrangenaru, V. (2003), ‘Large-sample theory of intrinsic and extrinsic sample means on manifolds, I’, *Annals of Statistics* **31**, 1-29.
- [7] Bookstein, F.L. (1984), ‘A statistical method for biological shape comparisons’, *Journal of Theoretical Biology* **107**, 475-520.

- [8] Bookstein, F.L. (1986), ‘Size and shape spaces for landmark data in two dimensions (with discussion)’, *Statistical Science* **1**, 181-242.
- [9] Cribari-Neto, F. & Zarkos, S. (1999), ‘R: Yet another econometric programming environment.’, *Journal of Applied Econometrics* **14**, 319-329.
- [10] Davison, A., & Hinkley, D. (1997), *Bootstrap Methods and Their Application*, Cambridge University Press, New York.
- [11] De Castro Korgi, R. (2003), *El Universo L^AT_EX*, Universidad Nacional de Colombia.
- [12] Dryden, I., & Mardia, K. (1998), *Statistical Shape Analysis*, Wiley and Sons, Chichester.
- [13] Efron, B. & Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Chapman and Hall/CRC.
- [14] Goodall, C. R. (1991), ‘Procrustes methods in the statistical analysis of shape (with discussion)’, *Journal of the Royal Statistical Society, Series B* **53**, 285-339.
- [15] James, G. (1954), ‘Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown.’, *Biometrika* **41**, 19-43.
- [16] Kendall, D. G. (1977), ‘The diffusion of shape’, *Advances in Applied Probability* **9**, 428-430.
- [17] Kendall, D. G. (1984), ‘Shape manifolds, procrustean metric and complex projective spaces’, *Bulletin of the London Mathematical Society* **16**, 81-121.
- [18] Kendall, D. G., Barden, D., Carne, T. K. & Le, H. (1999), *Shape and Shape Theory*, John Wiley & Sons, London.
- [19] Kent, J. T. (1994), ‘The complex Bingham distribution and shape analysis’, *Journal of the Royal Statistical Society, Series B* **56**, 285-299.
- [20] Le, H., & Kendall, D. G. (1993), ‘The Rimannian structure of Euclidean shape spaces: a novel environment for statistics’, *Annals of Statistics* **21**, 1225-1271.

- [21] Mardia, K. V. & Dryden, I. L. (1999), ‘The complex Watson distribution and shape analysis’, *Journal of the Royal Statistical Society, Series B* **61**, 913-926.
- [22] Mardia, K. V. & Walder, A. N. (1994), ‘Shape analysis of paired landmark data’, *Biometrika* **81**, 185-196.
- [23] Mirsky, L. (1955), *An Introduction to Linear Algebra*, Oxford University Press, London.
- [24] Oliveira, L., Hingst-Zaher, E. & Morgante, J. (2005), ‘Size and shape sexual dimorphism in the skull of the South American fur seal, *Arctocephalus australis* (Zimmermann, 1783) (Carnivora: Otariidae)’, *The Latin American Journal of Aquatic Mammals* **4**, 27-40.
- [25] Searle, S. (1971), *Linear Models*, John Wiley & Sons, New York.
- [26] Shelupsky, D. (1962), ‘An introduction to spherical coordinates’, *American Mathematical Monthly* **69**, 644-646.
- [27] Small, C. G. (1996), *The Statistical Theory of Shape*, Springer-Verlag, New York.