UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS NATURAIS E EXATAS
CURSO DE PÓS-GRADUAÇÃO EM ESTATÍSTICA E MODELAGEM QUANTITATIVA

COMPARAÇÃO DE ROTAS DE COLETA DE LEITE, COM BASE EM VARIÁVEIS FÍSICO-QUÍMICAS, UTILIZANDO ANÁLISE DE VARIÂNCIA MULTIVARIADA NÃO-PARAMÉTRICA

MONOGRAFIA DE ESPECIALIZAÇÃO

Enio Júnior Seidel

Santa Maria, RS, Brasil
2010

COMPARAÇÃO DE ROTAS DE COLETA DE LEITE, COM BASE EM VARIÁVEIS FÍSICO-QUÍMICAS, UTILIZANDO ANÁLISE DE VARIÂNCIA MULTIVARIADA NÃO-PARAMÉTRICA

por

Enio Júnior Seidel

Monografia apresentada ao Curso de Especialização em Estatística e Modelagem Quantitativa, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Especialista em Estatística e Modelagem Quantitativa**

Orientador: Prof. Dr. Luis Felipe Dias Lopes

Santa Maria, RS, Brasil

2010

Seidel, Enio Júnior

Comparação de rotas de coleta de leite, com base em variáveis físico-químicas, utilizando análise de variância multivariada não-paramétrica / por Enio Júnior Seidel; orientador Luis Felipe Dias Lopes – Santa Maria, 2010.

50 p.

Monografia (especialização) – Universidade Federal de Santa Maria, Centro de Ciências Naturais e Exatas, Curso de Pós-Graduação em Estatística e Modelagem Quantitativa, RS, 2010.

1. Comparação de grupos 2. Variáveis físico-químicas 3. Métodos não-paramétricos 4. Análise de variância multivariada 5. Estatística e Modelagem Quantitativa. I. Lopes, Luis Felipe Dias, orient. II. Título.

CDU

Universidade Federal de Santa Maria
Centro de Ciências Naturais e Exatas
Curso de Pós-Graduação em Estatística e Modelagem Quantitativa

A Comissão Examinadora, abaixo assinada,
aprova a Monografia de Especialização

COMPARAÇÃO DE ROTAS DE COLETA DE LEITE, COM BASE EM VARIÁVEIS FÍSICO-QUÍMICAS, UTILIZANDO ANÁLISE DE VARIÂNCIA MULTIVARIADA NÃO-PARAMÉTRICA

elaborada por
Enio Júnior Seidel

como requisito parcial para obtenção do grau de
Especialista em Estatística e Modelagem Quantitativa

COMISSÃO EXAMINADORA:

Luis Felipe Dias Lopes, Dr.
(Presidente/Orientador)

Angela Pellegrin Ansuj, Dra. (UFSM)

Dario Trevisan de Almeida, Msc. (UFSM)

Santa Maria, 10 de março de 2010.

AGRADECIMENTOS

À Universidade Federal de Santa Maria, pela oportunidade de cursar a especialização.

Ao professor Dr. Luis Felipe Dias Lopes, meu orientador, pelo incentivo, atenção e contribuições para o desenvolvimento do trabalho.

Aos professores do Curso de Pós-graduação em Estatística e Modelagem Quantitativa, pelos conhecimentos transmitidos.

À professora Dra. Angela Pellegrin Ansuj e aos professores Msc. Dario Trevisan de Almeida e Dr. Ivanor Müller, membros da banca examinadora, pelas contribuições para o aprimoramento do trabalho.

À minha família, pelo apoio e confiança dedicados.

Aos colegas e amigos.

A Deus por tudo.

RESUMO

Monografia de Especialização
Curso de Pós-Graduação em Estatística e Modelagem Quantitativa
Universidade Federal de Santa Maria, RS, Brasil

COMPARAÇÃO DE ROTAS DE COLETA DE LEITE, COM BASE EM VARIÁVEIS FÍSICO-QUÍMICAS, UTILIZANDO ANÁLISE DE VARIÂNCIA MULTIVARIADA NÃO-PARAMÉTRICA

Autor: Enio Júnior Seidel
Orientador: Dr. Luis Felipe Dias Lopes
Data e Local de Defesa: Santa Maria, 10 de março de 2010.

O objetivo desta pesquisa foi desenvolver um estudo sobre Análise de Variância Multivariada Não-Paramétrica para comparação de rotas de coleta de leite, com base em variáveis físico-químicas. Além disso, se teve o objetivo de elaborar um material didático para a utilização dos métodos não-paramétricos de comparação de grupos no *software* R. Foram consideradas 81 observações coletadas no período de outubro a dezembro de 2007, em três rotas de coleta do leite realizadas por uma usina de laticínios, denominadas de rota 1, rota 2 e rota 3. Foram considerados 13 fornecedores na rota 1, 34 fornecedores na rota 2 e 34 na rota 3. As variáveis consideradas na análise foram as seguintes: Água Excedente (%); Acidez (ºD); Gordura (%); Densidade (g/mL); Lactose (%) e Proteínas (%). Para realizar os procedimentos de análise, primeiramente, foram comparadas as rotas de coleta do leite utilizando métodos não-paramétricos univariados. Após, foram utilizados procedimentos não-paramétricos multivariados para realizar as comparações. Para a aplicação das técnicas e desenvolvimento do estudo utilizou-se o *software* R. Considerando a comparação das rotas de forma univariada verificou-se diferença significativa entre as rotas para apenas uma das variáveis físico-químicas consideradas, porém quando se fez a comparação das rotas considerando todas as variáveis simultaneamente, verificou-se que não ocorreram diferenças significativas. Em relação a utilização do software R, foi possível mostrar os comandos a serem seguidos para realizar os procedimentos de análise realizados nesta pesquisa.

Palavras-chave: Comparação de grupos; Variáveis físico-químicas; Métodos não-paramétricos; Análise de variância multivariada.

ABSTRACT

Monografia de especialização
Curso de Pós-Graduação em Estatística e Modelagem Quantitativa
Universidade Federal de Santa Maria, RS, Brasil

COMPARAÇÃO DE ROTAS DE COLETA DE LEITE, COM BASE EM VARIÁVEIS FÍSICO-QUÍMICAS, UTILIZANDO ANÁLISE DE VARIÂNCIA MULTIVARIADA NÃO-PARAMÉTRICA

(COMPARISON OF MILK COLLECTION ROUTES BASED ON PHYSICOCHEMICAL VARIABLES USING NONPARAMETRIC MULTIVARIATE ANALYSIS OF VARIANCE)

Author: Enio Júnior Seidel
Adviser: Dr. Luis Felipe Dias Lopes
Date and Place of Defense: Santa Maria, March, 10, 2010.

The objective of this research was to develop a study on Nonparametric Multivariate Analysis of Variance for the comparison of milk collection routes based on physicochemical variables. Also, it was aimed to develop a teaching material for the use of nonparametric methods to compare groups in the R software. Eighty-one observations collected from October to December 2007 were considered in three milk collection routes made by a dairy plant, called Route 1, Route 2 and Route 3. In addition, 13 suppliers on route 1, 34 on route 2 and 34 on route 3 were also considered. The variables considered in the analysis were the following: Water Surplus (%), Acidity (° D), Fat (%), Density (g / mL), Lactose (%) and Protein (%). To perform the testing procedures, the milk collection routes were compared using univariate nonparametric methods. After, multivariate nonparametric procedures were used to perform comparisons. For the application of techniques and development of the study, the R software was used. Considering the comparison of routes using univariate analysis, a significant difference between routes for only one of physicochemical variables studied was observed; however, when the routes were compared considering all variables simultaneously, it was found that there were no significant differences. In relation to the use of the R software, it was possible to show the commands to be followed to perform the testing procedures carried out in this research.

Key words: Comparison of groups, physicochemical variables, Nonparametric methods, Multivariate analysis of variance.

LISTA DE TABELAS

Tabela 1 – Padrões mínimos de composição química do leite de acordo com a Instrução Normativa 51 19

Tabela 2 – Padrões físicos normais do leite de acordo com a Instrução Normativa 51 19

Tabela 3 – Composição nutricional do leite 20

Tabela 4 – Significâncias do teste de Shapiro Wilk aplicado aos dados 36

Tabela 5 – Significâncias do teste de Bartlett aplicado aos dados 37

Tabela 6 – Significâncias da análise de variância de Kruskal-Wallis aplicada aos dados 37

Tabela 7 – Significâncias do teste de Wilcoxon-Mann-Whitney aplicado aos dados 38

Tabela 8 – Correlações de Spearman entre as variáveis e suas respectivas significâncias 40

Tabela 9 – Significância do teste de Shapiro Wilk para normalidade multivariada dos dados 41

Tabela 10 – MANOVA com o traço de Pillai aplicada aos dados 42

Tabela 11 – Significância da análise de variância multivariada não-paramétrica, com base no estudo de Zwick (1985), aplicada aos dados 42

Tabela 12 – Significância da análise de variância multivariada não-paramétrica, com base no estudo de Anderson (2001), aplicada aos dados, considerando 1000 permutações 43

Tabela 13 – Significância da análise de variância multivariada não-paramétrica, com base no estudo de Anderson (2001), aplicada aos dados, considerando 5000 permutações 43

Tabela 14 – Média e desvio padrão (DP) das variáveis em cada uma das rotas de coleta 44

LISTA DE ILUSTRAÇÕES

Figura 1 – Tela inicial do *software* R 18

LISTA DE ABREVIATURAS E SIGLAS

ANOVA – Análise de Variância

MANOVA – Análise de Variância Multivariada

PERMANOVA – Análise de Variância Multivariada Permutacional

IN 51 – Instrução Normativa 51

RIISPOA – Regulamento de Inspeção Industrial e Sanitária de Produtos de Origem Animal

SUMÁRIO

1 INTRODUÇÃO 13
1.1 Objetivos 14
1.1.1 Objetivo geral 14
1.1.2 Objetivos específicos 14
1.2 Justificativa 15
1.3 Estrutura do trabalho 15

2 METODOLOGIA DA PESQUISA 16
2.1 Pesquisa desenvolvida 16
2.2 Coleta e tratamento dos dados 17
2.3 Procedimentos de análise dos dados 17
2.4 Limitação do estudo 18
2.5 Síntese do capítulo 18

3 COMPOSIÇÃO DO LEITE 19
3.1 Síntese do capítulo 21

4 MÉTODOS NÃO-PARAMÉTRICOS PARA COMPARAÇÃO DE GRUPOS – CASO UNIVARIADO 22
4.1 Teste de Shapiro Wilk 22
4.2 Teste de Bartlett 23
4.3 Análise de variância de Kruskal-Wallis 24
4.4 Teste de Wilcoxon-Mann-Whitney 26
4.5 Correlação Posto-ordem de Spearman 27
4.6 Síntese do capítulo 29

5 ANÁLISE DE VARIÂNCIA MULTIARIADA NÃO-PARAMÉTRICA 30
5.1 Teste de Normalidade Multivariada 30
5.2 Utilização da Análise de Variância Multivariada tradicional aplicada aos dados na forma de postos 30

5.3 Análise de Variância Multivariada Não-Paramétrica baseada em distâncias entre dados 32
5.4 Síntese do capítulo 34

6 RESULTADOS E DISCUSSÕES 35
6.1 Comparação das rotas de coleta utilizando a Análise de Variância de Kruskal-Wallis 35
6.2 Comparação das rotas de coleta utilizando a Análise de Variância Multivariada Não-Paramétrica 38
6.3 Síntese do capítulo 44

7 CONCLUSÕES 45
7.1 Sugestões para trabalhos futuros 46
7.2 Síntese do capítulo 46

REFERÊNCIAS BIBLIOGRÁFICAS 47

1 INTRODUÇÃO

A comparação entre grupos é efetuada sempre que se tem o objetivo de testar os efeitos de um determinado fator. Em experimentação podem-se alocar grupos aplicando em cada grupo um tratamento diferente. Porém, em outros casos os grupos são definidos a priori, como neste caso, as rotas de coleta já existentes.

Se for efetuada a comparação entre k grupos, considerando uma única variável dependente, pode ser utilizada a Análise de Variância (ANOVA) ou o procedimento não-paramétrico de Análise de Variância de Kruskal-Wallis.

Contudo, quando é utilizada uma abordagem univariada para comparar grupos, é necessário realizar vários testes univariados, fato que torna difícil a interpretação dos resultados e o julgamento sobre a diferença ou não entre os grupos, pois pode haver diferenças em relação a uma variável, mas pode não haver diferenças em relação a outra variável. Desse modo, a incorporação de várias variáveis leva em conta o inter-relacionamento entre elas e pode melhorar a eficiência da análise dos dados.

De acordo com Katz e Mcsweeney (1980), a vantagem de se utilizar um método multivariado é observada quando as variáveis dependentes são altamente correlacionadas.

Segundo Pontes (2005), de forma geral, as diferenças entre grupos ou populações não dependem somente de uma variável, mas sim de um conjunto delas.

Assim, a abordagem multivariada é a mais aconselhada quando se têm p variáveis a serem consideradas para avaliar diferenças entre grupos. Neste caso, pode-se utilizar a Análise de Variância Multivariada (MANOVA). Porém, quando as pressuposições para a utilização da MANOVA não são satisfeitas, um procedimento não-paramétrico pode ser utilizado.

Alguns trabalhos podem ser destacados no que tange a busca por um procedimento não-paramétrico para a análise de variância multivariada, como por exemplo, os trabalhos de Katz e Mcsweeney (1980), Zwick (1985) e Anderson (2001).

Nesta pesquisa, dois procedimentos são utilizados para a realização da análise de variância multivariada. O primeiro, envolve a transformação dos dados originais em postos e a posterior aplicação de uma estatística baseada no traço de Pillai (ZWICK, 1985). O segundo, se baseia no estudo realizado por Anderson (2001), onde apresenta uma proposta de utilização da Análise de Variância Multivariada Permutacional (PERMANOVA).

Estes procedimentos são utilizados para comparar rotas de coleta de leite de uma indústria de laticínios, considerando as variáveis físico-químicas: Água excedente (%); Acidez (ºD); Gordura (%); Densidade (g/mL); Lactose (%); e Proteínas (%).

1.1 Objetivos

1.1.1 Objetivo geral

Desenvolver um estudo sobre Análise de Variância Multivariada Não-Paramétrica para comparação de rotas de coleta de leite.

1.1.2 Objetivos específicos

- Estudar as abordagens não-paramétricas univariadas e multivariadas para comparação de grupos;
- Comparar as rotas de coleta quanto às variáveis físico-químicas do leite;
- Elaborar um material didático para a utilização dos métodos não-paramétricos de comparação de grupos no *software R*.

1.2 Justificativa

Os procedimentos paramétricos multivariados para comparação de grupos são bastante conhecidos e difundidos. Porém, o mesmo não acontece com os procedimentos não-paramétricos multivariados.

Poucos são os trabalhos desenvolvidos com esta abordagem e ainda são poucos os *softwares* que possuem comandos para realizar tais procedimentos para comparação de grupos no caso multivariado.

Assim, este trabalho se justifica pela busca em contribuir na maior difusão destes procedimentos não-paramétricos com uma aplicação na comparação de rotas de coleta de leite com base em variáveis físico-químicas.

1.3 Estrutura do trabalho

Esta pesquisa está estruturada em sete capítulos. Neste primeiro capítulo estão expostos os objetivos da pesquisa e sua justificativa.

No segundo capítulo, está definida a metodologia de pesquisa a ser utilizada.

No Capítulo 3, apresenta-se uma revisão de literatura sobre os aspectos físico-químicos do leite.

No Capítulo 4, apresenta-se a revisão de literatura sobre as técnicas estatísticas não-paramétricas univariadas para comparação de grupos.

No Capítulo 5, apresenta-se uma revisão de literatura sobre a análise de variância multivariada não-paramétrica.

No Capítulo 6, são desenvolvidas as técnicas propostas no estudo e desenvolvidas discussões sobre os resultados obtidos.

Por fim, no Capítulo 7, apresentam-se as conclusões com as devidas considerações.

2 METODOLOGIA DA PESQUISA

Segundo Marconi e Lakatos (2005), a pesquisa é um procedimento formal, com método de pensamento reflexivo, que requer um tratamento científico e se constitui no caminho para conhecer a realidade ou para descobrir verdades parciais.

De acordo com Miguel (2007), o objetivo estabelece a ação a ser conduzida e deve, portanto, ser fator determinante na escolha da abordagem metodológica.

O método é o conjunto das atividades sistemáticas e racionais, o qual, com maior segurança e economia, permite alcançar o objetivo traçando o caminho a ser seguido, detectando erros e auxiliando as decisões do pesquisador (MARCONI; LAKATOS, 2005). Para Inácio Filho (2004), metodologia consiste em um conjunto de procedimentos e técnicas utilizadas no processo de investigação, incluindo os aspectos relacionados a como fazer a pesquisa.

2.1 Pesquisa desenvolvida

O desenvolvimento do presente trabalho constitui-se, num primeiro momento, de pesquisa bibliográfica, na qual se busca embasamento teórico consistente sobre o tema, tanto para um maior aprofundamento no referencial teórico, quanto para balizar as discussões referentes aos resultados atingidos. A pesquisa bibliográfica abrange toda bibliografia já tornada pública em relação ao tema de estudo, e tem por objetivo colocar o pesquisador em contato direto com tudo que foi escrito, dito ou filmado sobre determinado assunto (MARCONI; LAKATOS, 2005).

A pesquisa bibliográfica disponibiliza referencial capaz de servir de base para a delimitação do tema e para a preparação do texto a ser elaborado. Conforme Miguel (2007, p. 222), “o referencial teórico serve para delimitar as fronteiras do que será investigado, proporcionar o suporte teórico para a pesquisa e também explicitar o grau de evolução sobre o tema estudado”.

Como segunda abordagem, o trabalho constitui-se de estudo comparativo entre grupos de fornecedores de leite, caracterizados pelas rotas de coleta. Para esta abordagem são utilizados métodos estatísticos para comparação de grupos, através das análises de variância univariada e multivariada não-paramétricas.

Assim, a presente pesquisa se desenvolve com a utilização conjunta de fundamentações teóricas existentes para a discussão do problema, encontradas através da pesquisa bibliográfica e de técnicas e procedimentos adequados para a abordagem de estudo de caso.

2.2 Coleta e tratamento dos dados

Foram consideradas 81 observações coletadas no período de outubro a dezembro de 2007, em três rotas de coleta de leite realizadas pela usina, denominadas de rota 1, rota 2 e rota 3. Foram considerados 13 fornecedores na rota 1; 34 fornecedores na rota 2 e; 34 na rota 3.

As variáveis consideradas na análise foram as seguintes: Água Excedente (%); Acidez (ºD); Gordura (%); Densidade (g/mL); Lactose (%) e Proteínas (%).

A seguir são definidos os procedimentos de análise utilizados.

2.3 Procedimentos de análise dos dados

Primeiramente, foram comparadas as rotas de coleta do leite utilizando métodos não-paramétricos univariados. Nessa fase foram utilizados os procedimentos de análise de variância de Kruskal-Wallis e o teste Wilcoxon-Mann-Whitney.

Após, foram utilizados procedimentos não-paramétricos multivariados para realizar a comparação das rotas de coleta do leite. Aqui, foram utilizadas a abordagem sugerida por Zwick (1985) onde se utiliza uma estatística baseada no traço de Pillai, e a abordagem proposta por Anderson (2001), utilizando a análise de variância multivariada permutacional (PERMANOVA).

Para a aplicação das técnicas e desenvolvimento do estudo utiliza-se o *software R*. A Figura 1 mostra o console do *R*.

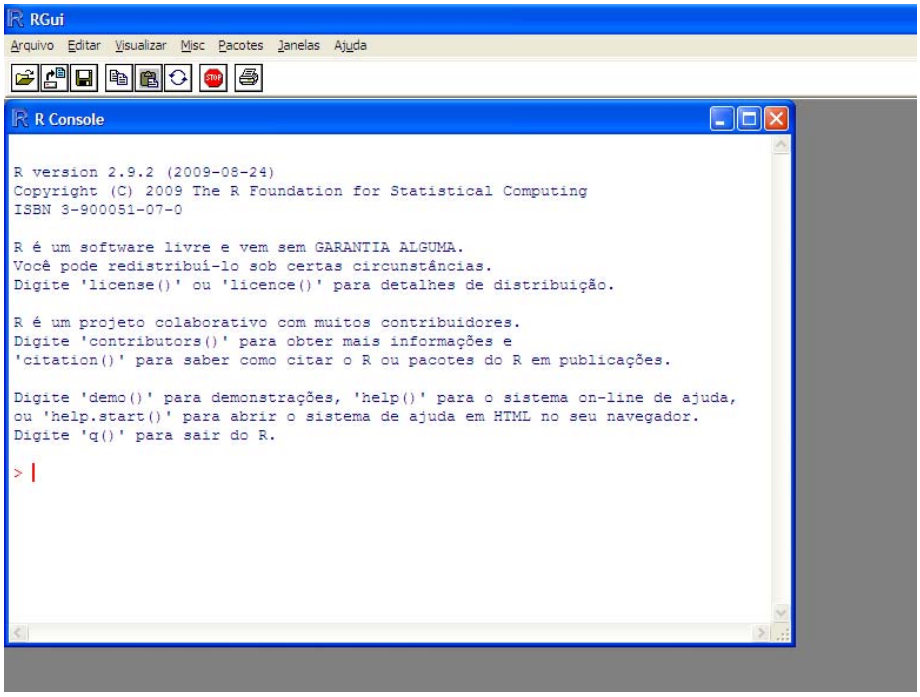

Figura 1 – Tela inicial do *software R*.

2.4 Limitação do estudo

A primeira limitação desta pesquisa foi a utilização de dados já coletados pela usina, ou seja, não se teve possibilidade de acompanhar o processo de amostragem e coleta dos dados para análise.

Uma segunda limitação foi o número reduzido de variáveis analisadas referentes aos aspectos qualitativos da matéria-prima. Uma gama maior de variáveis poderia contribuir e aumentar a qualidade da presente pesquisa.

2.5 Síntese do capítulo

Neste capítulo, foi realizada a explanação dos métodos de pesquisa a serem utilizados para desenvolvimento do presente estudo, dando ênfase à delimitação do tema, coleta e procedimentos de análise dos dados.

No capítulo seguinte, desenvolver-se-á a discussão sobre os aspectos físico-químicos do leite.

3 COMPOSIÇÃO DO LEITE

O leite é considerado o mais nobre dos alimentos, dada sua composição rica em proteínas, gordura, carboidratos, sais minerais e vitaminas (NOAL, 2006). Industrializado, o leite se apresenta em diversos tipos de produtos para consumo, devidamente controlados por normas de inspeção industrial e sanitária (TRONCO, 2008).

A Tabela 1 apresenta os padrões mínimos de composição química do leite de acordo com a Instrução Normativa 51 (IN 51). Salienta-se que esses padrões valem para todos os tipos de leite.

Tabela 1 – Padrões mínimos de composição química do leite de acordo com a Instrução Normativa 51.

Item	Requisito
Gordura	Mínimo igual a 3,0%
Proteína bruta	Mínimo igual a 2,9%
Sólidos não gordurosos	Mínimo igual a 8,4%

Em relação aos padrões físicos, a IN 51 estipula, independentemente do tipo de leite, o exposto na Tabela 2.

Tabela 2 – Padrões físicos normais do leite de acordo com a Instrução Normativa 51.

Item	Requisito
Acidez	Entre 14 e 18ºDornic
Densidade a 15ºC	Entre 1028 e 1034g/mL
Crioscopia	Máxima igual a -0,530ºH
Estabilidade Alizarol / Álcool 72%	Estável

A IN 51 foi criada pela necessidade de complementar e atualizar o Regulamento da Inspeção Industrial e Sanitária de Produtos de Origem Animal (RIISPOA) às condições atuais.

A boa qualidade do leite destinado ao consumo humano é fator de suma importância, “visto que o mesmo é considerado uma das principais fontes de nutrientes para uma grande parte da população” (KROLOW; RIBEIRO, 2006, p. 14).

O leite é um alimento de alto valor nutritivo em um meio aquoso, capaz de suprir as exigências tanto do homem quanto dos animais (SILVA, 1999). Sua composição nutricional pode ser observada na Tabela 3.

Tabela 3 – Composição nutricional do leite.

Componente	Percentual (%)
Água	87,5
Gordura	3,6
Caseína (Proteína)	3,0
Albumina (Proteína)	0,6
Lactose	4,6
Sais Minerais	0,7

Fonte: Adaptado de Silva (1999).

Segundo Ansuj (2000), no leite, os teores de alguns componentes variam expressivamente, como gordura e proteína, enquanto outros, como a lactose e os minerais, variam em menor proporção.

De acordo com Pereda et al. (2005), a gordura é o componente que mais varia, com concentração oscilando entre 3,2 e 6%.

Para Varnam e Sutherland (1995), muitos dos compostos que contribuem para o aroma e sabor do leite derivam da gordura. Porém, é preciso ter cuidado com a gordura do leite, pois segundo Belchior (2003), pouca gordura e muita proteína são as características procuradas pelos consumidores nos alimentos. Soares, Machado e Fonseca (2002) também tomam cuidado com o componente gordura, pois relatam que, nos últimos anos, o público em geral tem-se preocupado com o excesso de ingestão de calorias e de gorduras.

As proteínas têm papel nutritivo, mas também se destacam por possuir propriedades que permitem a aplicação de operações tecnológicas, como a esterilização, sem modificar de forma significativa o valor nutritivo e as propriedades sensoriais do leite (PEREDA et al., 2005).

Segundo Varnam e Sutherland (1995), as proteínas do leite são de dois tipos, proteínas do soro e caseínas, estas constituindo mais de 80% das proteínas totais do leite.

De acordo com Madalena (2000), a gordura e a proteína são os componentes do leite que possuem maior valor econômico.

Para Varnam e Sutherland (1995), a lactose é o principal constituinte sólido do leite. Sua concentração varia entre 4,2 e 5%. O conteúdo da lactose geralmente é mais baixo ao final da lactação e no leite de animais com mastite. Ainda, segundo os autores, a lactose influência na diminuição do ponto de congelamento e no aumento do ponto de ebulição.

Em se tratando de aspectos físicos do leite, segundo Vilela, Bressan e Cunha (2001), a acidez decorre de más condições de higiene e da conservação do leite à temperatura ambiente até a chegada à usina. Segundo os autores, calcula-se que haja perda diária de 2% do leite entregue na usina devido à acidez.

Conforme Figueiredo e Porto (2002, p.76), “a acidez superior à normal é proveniente da acidificação do leite pelo desdobramento da lactose, provocada por ação microbiológica”. Ela tende a aumentar consideravelmente, se o leite não for adequadamente manipulado e mantido refrigerado.

Para Machado et al. (1998 apud ANSUJ, 2000), diversos fatores podem alterar a composição do leite, como genética, ambiente, idade do animal, estágio de lactação, manejo de ordenha, sanidade, nutrição e doenças metabólicas. Porém, questões de manejo podem ser melhoradas para que se tenha um produto de melhor qualidade.

3.1 Síntese do capítulo

Neste capítulo, foram relatados os principais aspectos sobre o leite.

No próximo capítulo será apresentada a revisão da literatura sobre métodos estatísticos a serem utilizados para alcançar os objetivos propostos para a presente pesquisa.

4 MÉTODOS NÃO-PARAMÉTRICOS PARA COMPARAÇÃO DE GRUPOS – CASO UNIVARIADO

A seguir são abordadas as técnicas estatísticas utilizadas na análise dos dados da presente pesquisa.

4.1 Teste de Shapiro Wilk

O teste de Shapiro Wilk, ou teste W, é utilizado para verificar a hipótese de que os dados seguem distribuição normal.

As hipóteses a serem testadas são:

H_0: Os dados seguem distribuição normal;

H_1: Os dados não seguem distribuição normal.

A estatística de teste W é definida por (LOPES, 2007):

$$W = \frac{b^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4.1}$$

em que:

$$b = a_n(x_n - x_1) + a_{n-1}(x_{n-1} - x_2) + \ldots + a_{n-k-1}(x_{n-k-1} - x_k) = \sum_{i=1}^{k} a_{n-i+1}(x_{n-i+1} - x_i) \tag{4.2}$$

onde,

a_{n-i+1} são valores tabelados, e

$k = \frac{n}{2}$ se n for par, ou $k = \frac{(n+1)}{2}$ se n for ímpar.

e,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Para concluir sobre o teste utiliza-se uma tabela especifica para o teste e compara-se o valor calculado (W) com o valor tabelado (W_t).

Se $W < W_t$, rejeita-se a hipótese H_0.

4.2 Teste de Bartlett

A análise de variância paramétrica é relativamente robusta para o caso em que os tratamentos possuem variâncias homogênias. Assim, uma das pressuposições a serem verificadas é a homogeneidade das variâncias. Para isso pode-se utilizar o teste de Bartlett.

As hipóteses a serem testadas são:

H_0: $\sigma^2_1 = \sigma^2_2 = ... = \sigma^2_k$ As variâncias são homogêneas;

H_1: As variâncias não são homogêneas.

O procedimento do teste envolve uma estatística que se aproxima da distribuição Qui-Quadrado, com $k-1$ graus de liberdade, quando k amostras aleatórias vêm de populações normais independentes (MONTGOMERY, 1976). A estatística é:

$$\chi^2_0 = 2{,}3026\frac{q}{c} \tag{4.3}$$

onde,

$$q = (N-k)\log S^2_p - \sum_{i=1}^{k}(n_i - 1)\log S^2_i \tag{4.4}$$

e

$$c = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}(n_i - 1)^{-1} - (N-k)^{-1}\right) \tag{4.5}$$

e

$$S^2_p = \frac{\sum_{i=1}^{k}(n_i - 1)S^2_i}{N-k} \tag{4.6}$$

e S^2_i é a variância amostral da i–ésima população.

Se $\chi^2_0 > \chi^2_{\alpha;k-1}$, então rejeita-se H_0, onde $\chi^2_{\alpha;k-1}$ é o valor tabelado para uma significância α e $k-1$ graus de liberdade.

4.3 Análise de variância de Kruskal-Wallis

Quando mais de dois grupos de amostras independentes são objeto de comparação, Kruskal e Wallis sugeriram um método para este fim, o qual é geralmente conhecido como o método de análise de variância de Kruskal-Wallis (RODRIGUES, 1976).

A análise de variância de um fator de Kruskal-Wallis por postos é um teste extremamente útil para verificar se k amostras independentes provêm de populações diferentes. A questão é verificar se as diferenças entre as amostras significam genuínas diferenças entre as populações ou se elas representam apenas variações que seriam esperadas entre amostras aleatórias de uma mesma população (SIEGEL; CASTELLAN JR, 2006).

A técnica de Kruskal-Wallis testa a hipótese nula de que as k amostras provêm da mesma população ou de populações idênticas com a mesma mediana. Para especificar a hipótese nula e sua alternativa mais explicitamente, seja θ_j a mediana para o j-ésimo grupo. Então, as hipóteses a serem testadas são:

H_0: $\theta_1 = \theta_2 = ... = \theta_k$;

H_1: $\theta_i \neq \theta_j$ para alguns grupos i e j.

Ou seja, se a hipótese alternativa for verdadeira, pelo menos dois grupos têm medianas diferentes entre si.

A variável dependente a ser estudada deve ser medida, no mínimo, em escala ordinal (BISQUERRA; SARRIERA; MARTINEZ, 2004).

No cálculo do teste de Kruskal-Wallis, as n observações são substituídas por postos. Isto é, todos os escores de todas as k amostras são colocados juntos e organizados através de postos em uma única série. O menor valor é substituído pelo posto 1, o seguinte menor valor é substituído pelo posto 2 e o maior valor é substituído pelo posto n, em que n é o número total de observações independentes nas k amostras (SIEGEL; CASTELLAN JR, 2006). Caso haja empate entre escores, atribui-se o posto médio para esses escores (GONÇALVES, 2002).

Após a distribuição dos postos entre os valores, a soma originada por esses postos em cada amostra é encontrada. Dessas somas é possível encontrar o posto médio para cada amostra. De acordo com Siegel e Castellan Jr (2006), se as amostras são da mesma população ou de populações idênticas, os postos médios devem ser quase os mesmos.

O teste de Kruskal-Wallis trabalha com as diferenças entre os postos médios para determinar se elas são tão discrepantes que provavelmente não tenham vindo de amostras extraídas de uma mesma população.

A estatística do teste é denominada de H, tendo distribuição igual à do χ^2, com graus de liberdade iguais ao número de tratamentos menos 1 (RODRIGUES, 1976).

Calcula-se a estatística pela expressão:

$$H = \left[\frac{12}{n(n+1)} \sum_{j=1}^{k} n_j \overline{R}_j^2 \right] - 3(n+1) \tag{4.7}$$

em que k é o número de amostras; n_j é o número de casos na j-ésima amostra; n é o número de casos na amostra combinada (soma dos n_j's) e; $\overline{R}_j$ é a média dos postos na j-ésima amostra.

Quando ocorrem empates entre dois ou mais escores, deve-se ter cuidado, pois a variância da distribuição amostral de H é influenciada por empates. Para corrigir o efeito dos empates, H é calculado usando a seguinte correção (SIEGEL; CASTELLAN JR, 2006):

$$1 - \frac{\sum_{i=1}^{g} \left(t_i^3 - t_i \right)}{n^3 - n} \tag{4.8}$$

em que g é o número de agrupamentos de postos diferentes empatados; t_i é o número de postos empatados no i-ésimo agrupamento e; n é o número de observações através de todas as amostras.

Assim, a nova expressão para H corrigida para empates é:

$$H = \frac{\left[\frac{12}{n(n+1)} \sum_{j=1}^{k} n_j \overline{R}_j^2 \right] - 3(n+1)}{1 - \frac{\sum_{i=1}^{g} \left(t_i^3 - t_i \right)}{n^3 - n}} \tag{4.9}$$

O efeito da correção para empates é aumentar o valor de H e dessa forma tornar o resultado mais significante do que seria se nenhuma correção tivesse sido feita (SIEGEL; CASTELLAN JR, 2006).

Se a probabilidade associada com o valor observado para H é igual ou menor do que o nível de significância α preestabelecido, rejeita-se a hipótese H_0.

Desde que se verifiquem diferenças significativas entre k grupos através da análise de variância de Kruskal-Wallis, é interessante verificar quais desses k grupos diferem significativamente entre si. Para isso pode-se utilizar o teste U de Mann-Whitney.

4.4 Teste de Wilcoxon-Mann-Whitney

Desde que a variável dependente esteja medida em escala pelo menos ordinal, é possível aplicar o teste de Wilcoxon-Mann-Whitney para comparar se dois grupos independentes foram ou não extraídos da mesma população.

Este é um dos testes não-paramétricos mais poderosos, sendo uma alternativa para o teste paramétrico t quando se deseja evitar as suposições do teste t ou quando a mensuração é mais fraca do que uma dada na escala intervalar (SIEGEL; CASTELLAN JR, 2006).

Suponha-se que se tenham amostras de duas populações A e B. A hipótese nula é de que A e B possuem a mesma distribuição. A hipótese alternativa é que A é estocasticamente maior que B, ou pode-se estabelecer que B é estocasticamente maior que A (SIEGEL; CASTELLAN JR, 2006). Para uma prova bilateral, especifica-se que A e B são distintas.

Assim, as hipóteses a serem testadas são:

H_0: População A = População B;

H_1: População A $\neq$ População B.

Sejam n_1 o número de casos na amostra A e n_2 o número de casos na amostra B. Para aplicar o teste, primeiro combinamos as observações de ambos os grupos e os dispomos em postos em ordem crescente de tamanho.

Após, fazemos W_A igual a soma dos postos no grupo A e W_B a soma de postos do grupo B.

Se a hipótese nula é verdadeira, espera-se que a média dos postos em cada grupo seja quase a mesma. Porém, se a soma dos postos de um dos grupos for muito grande (ou muito pequena), então pode-se suspeitar que as amostras não foram extraídas da mesma população (SIEGEL; CASTELLAN, JR, 2006).

Quando n_1 e n_2 crescem em tamanho, a distribuição amostral de W_A aproxima-se da distribuição normal com média:

$$\mu_{W_A} = \frac{n_1(N+1)}{2} \tag{4.10}$$

e variância:

$$\sigma^2{}_{W_A} = \frac{n_1 n_2 (N+1)}{12} \tag{4.11}$$

Assim, temos:

$$z = \frac{W_A \pm (0{,}5 - \mu_{W_A})}{\sigma_{W_A}} \tag{4.12}$$

em que z segue distribuição normal com média zero e variância unitária.

Quando ocorrem empates, a aproximação pela distribuição normal deve ser corrigida fazendo a variância igual a (SIEGEL; CASTELLAN JR, 2006):

$$\sigma^2{}_{W_A} = \frac{n_1 n_2}{N(N-1)}\left(\frac{N^3 - N}{12} - \sum_{j=1}^{g}\frac{t_j{}^3 - t_j}{12}\right) \tag{4.13}$$

onde N é o total de observações, g é o número de agrupamentos de postos empatados e t_j é o número de postos empatados no j-ésimo agrupamento.

Se o valor z tem probabilidade associada não superior a α, rejeita-se a hipótese H_0.

4.5 Correlação Posto-ordem de Spearman

Freqüentemente se tem o objetivo de saber se dois conjuntos de escores estão relacionados e, caso estejam, qual o grau desta relação (SIEGEL; CASTELLAN, Jr., 2006).

O coeficiente de correlação posto-ordem de Spearman é uma medida de associação entre duas variáveis que requer que ambas as variáveis sejam medidas pelo menos em uma escala ordinal, de modo que os objetos ou indivíduos em estudo possam ser dispostos em postos em duas séries ordenadas (SIEGEL; CASTELLAN, Jr., 2006).

Para calcular o coeficiente de correlação deve-se primeiramente atribuir postos para a variável X e postos para a variavel Y, de modo que em cada variável, o menor valor receba posto 1 e o maior valor receba posto N. Após, é necessário calcular o valor d_i, o qual é a diferença entre os postos de X e Y para a i-ésima observação,

$d_i = X_i - Y_i$, com $i = 1,2,...,N$.

Com os valores d_i calculados e elevados ao quadrado, calcula-se o coeficiente de correlação pela expressão:

$$r_S = 1 - \frac{6\sum_{i=1}^{N} d_i^{\,2}}{N^3 - N} \tag{4.14}$$

Quando ocorrem empates nas observações para a mesma variável, a cada um deles é atribuído a média dos postos que eles teriam recebido se não tivessem ocorridos empates. Neste caso, é necessário proceder a uma correção levando em conta os empates (SIEGEL; CASTELLAN, Jr., 2006). O fator de correção é:

$T_X = \sum_{i=1}^{g} (t_i^{\,3} - t_i)$, para a variável X, e

$T_Y = \sum_{i=1}^{g} (t_i^{\,3} - t_i)$, para a variável Y,

onde g é o número de agrupamentos de diferentes postos empatados e; t_i é o número de postos empatados no i-ésimo agrupamento.

Assim, o cálculo do coeficiente de correlação considerando a ocorrência de empates é:

$$r_S = \frac{(N^3 - N) - (6\sum d^2) - \left[\frac{(T_X + T_Y)}{2}\right]}{\sqrt{(N^3 - N)^2 - (T_X + T_Y)(N^3 - N) + (T_X T_Y)}} \tag{4.15}$$

Para testar a significância da correlação entre a variáveis X e Y, formulamos as seguintes hipóteses:

H_0: Não há associação entre X e Y;

H_1: Existe associação entre X e Y.

Quando N é maior do que um valor em torno de 20 a 25, a significância do coeficiente de correlação pode ser testado pela estatística (SIEGEL; CASTELLAN, Jr., 2006):

$$z = r_S \sqrt{N-1} \tag{4.16}$$

em que z tem distribuição aproximadamente normal com média igual a zero e variância unitária.

Se o valor de z excede o valor crítico tabelado, rejeita-se H_0 em favor de H_1.

4.6 Síntese do capítulo

Neste capítulo, foram expostos os métodos estatísticos utilizados no presente estudo, dando ênfase aos seus procedimentos:

- Teste de Shapiro Wilk;
- Teste de Bartlett;
- Análise de variância de Kruskal-Wallis;
- Teste de Wilcoxon-Mann-Whitney;
- Correlação posto-ordem de Spearman.

No próximo capítulo são discutidos os métodos de análise de variância multivariada não-paramétrica.

5 ANÁLISE DE VARIÂNCIA MULTIARIADA NÃO-PARAMÉTRICA

A seguir são abordadas algumas formas de aplicação da análise de variância multivariada não-paramétrica.

5.1 Teste de Normalidade Multivariada

Considerando o caso univariado, se o interesse for testar a normalidade dos dados, um dos testes mais utilizados é o teste de Shapiro-Wilk.

No caso multivariado, uma possibilidade para testar a normalidade é a utilização da extensão multivariada do teste de Shapiro-Wilk. Esta extensão é baseada na generalização multivariada do teste proposto por Domanski em 1998 (CANTELMO; FERREIRA, 2007). Ainda, segundo os autores, esta generalização busca uma combinação linear das p variáveis originais e aplicar o teste de Shapiro-Wilk nesta nova variável.

Cantelmo e Ferreira (2007) concluíram que este procedimento de extensão do teste de Shapiro-Wilk não é tão eficiente para detecção de normalidade multivariada. Porém, como é a única alternativa implementada no *software* R, ela foi utilizada.

5.2 Utilização da Análise de Variância Multivariada tradicional aplicada aos dados na forma de postos

Para realizar a análise de variância multivariada com dados na forma de postos, deve-se atribuir postos às observações de cada variável, de forma separada. Isto é, deve-se atribuir postos para a primeira variável, após, deve-se atribuir os postos para a segunda variável, e assim por diante até a p-ésima variável a ser considerada na análise.

Atribuem-se postos ao conjunto de observações, do menor ao maior valor, com a menor observação tendo posto 1, a segunda maior observação tendo posto 2, e assim por diante, até que o maior valor do conjunto assuma o

maior posto, no caso o posto N. Postos médios são atribuídos nos casos de observações empatadas.

Este procedimento de atribuição de postos é o mesmo utilizado no procedimento para a análise de variância de Kruskal-Wallis.

Com os dados originais substituídos por seus respectivos postos, Zwick (1985) sugere que pode ser aplicada a análise de variância multivariada na sua forma tradicional, de modo a se obter o traço de Pillai para elaborar uma estatística que se aproxima da distribuição Qui-quadrado.

As hipóteses a serem testadas são:

H_0: Não existe diferença significativa entre os grupos;

H_1: existe diferença significativa entre os grupos.

A análise de variância multivariada (MANOVA) é a extensão multivariada das técnicas univariadas para avaliar as diferenças entre médias de grupos. Isto é, a MANOVA é utilizada para avaliar a significância estatística de diferenças entre grupos, onde a hipótese nula a ser testada é a de igualdade de vetores de médias sobre múltiplas variáveis dependentes ao longo de grupos (HAIR Jr. et al., 2005).

Se é desejável manter o controle sobre a taxa de erro experimental e existe pelo menos algum grau de intercorrelação entre as variáveis dependentes, então a MANOVA é apropriada. Desse modo, a MANOVA pode detectar diferenças combinadas não encontradas nos testes univariados (HAIR Jr. et al., 2005).

A análise de variância multivariada apresenta alguns critérios para avaliar as diferenças multivariadas ao longo de grupos. Os critérios mais conhecidos são a Maior Raiz Característica de Roy; o Lambda de Wilks; o Traço de Hotelling; e o Traço de Pillai.

De acordo com Hair Jr. et al. (2005), a medida a ser usada é a que for mais imune a violações das pressuposições inerentes a MANOVA e que ainda mantiver o maior poder. O traço de Pillai e o lambda de Wilks parecem ser os que melhor atendem tais necessidades, apesar de haver evidências de que o traço de Pillai é mais robusto se o tamanho da amostra diminui ou se a homogeneidade de covariâncias é violada.

O traço de Pillai é definido por:

$$V = tr\left[H\left(H+E\right)^{-1}\right] = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i} \quad (5.1)$$

E aproximado para F por:

$$F = \left(\frac{V}{s-V}\right)\left(\frac{2n+s+1}{2m+s+1}\right) \quad (5.2)$$

Segundo Zwick (1985), o traço de Pillai pode ser considerado para encontrar a seguinte estatística:

$$(N-1)V \quad (5.3)$$

em que N é o número de observações alocadas nos k grupos.

Esta estatística se aproxima de uma distribuição Qui-quadrado com $p(k-1)$ graus de liberdade. Em que p é o número de variáveis medidas.

Se $(N-1)V$ for maior que o valor $\chi^2_{\alpha;p(k-1)}$, rejeita-se a hipótese H_0 (ZWICK, 1985).

5.3 Análise de Variância Multivariada Não-Paramétrica baseada em distâncias entre dados

Este procedimento não-paramétrico leva em consideração medidas de distâncias entre pares de observações para construir uma estatística para comparar estas distâncias entre observações do mesmo grupo contra aquelas distâncias em diferentes grupos. Além disso, usam-se permutações de observações para obter a probabilidade associada com a hipótese nula de igualdade entre grupos (ANDERSON, 2001).

Pensando na ANOVA, a soma de quadrados total SS_T é calculada pela soma das diferenças quadradas entre as observações e sua média de grupo. SS_W é a soma de quadrados dentro de grupo e SS_A é a soma de quadrados entre grupos.

Uma extensão multivariada para SS_W pode ser definida por:

$$SS_W = \sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{k=1}^{p}\left(y_{ijk} - y_{i.k}\right)^2 \quad (5.4)$$

Esta extensão multivariada pode ser pensada em termos geométricos, onde SS_W é a soma das distâncias euclidianas quadradas entre cada observação e seu centróide de grupo. Neste caso, SS_W é dado por:

$$SS_W = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(y_{ij} - \bar{y}_{i.}\right)' \left(y_{ij} - \bar{y}_{i.}\right) \tag{5.5}$$

Segundo Anderson (2001), a soma de distâncias quadradas entre pontos e seu centróide é igual a soma de distâncias inter-pontos quadradas dividida pelo número de pontos. Assim, a soma de distâncias inter-pontos quadradas é:

$$SS_I = \frac{1}{n} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d^2{}_{ij} \varepsilon_{ij} \tag{5.6}$$

E a soma de quadrados total pode ser definida como:

$$SS_T = \frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d^2{}_{ij} \tag{5.7}$$

E a soma de quadrados dentro de grupos como:

$$SS_W = \frac{1}{n} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d^2{}_{ij} \varepsilon_{ij} \tag{5.8}$$

Desse modo, a soma de quadrados entre grupos é:

$$SS_A = SS_T - SS_W \tag{5.9}$$

E a pseudo estatística F para testar a hipótese multivariada é:

$$F = \frac{\left(\frac{SS_A}{k-1}\right)}{\left(\frac{SS_W}{N-k}\right)} \tag{5.10}$$

As somas de quadrados, quadrados médios e o pseudo F obtidas no caso multivariado podem ser interpretados da mesma maneira que na ANOVA (ANDERSON, 2001).

Fazendo as permutações nos dados originais podemos encontrar valor F^{π} para todas estas reorganizações dos dados. Assim, o p-valor é definido por:

$$p = \frac{(\text{Nº de } F^{\pi} \geq F)}{(\text{Total de } F^{\pi})} \tag{5.11}$$

Com k grupos e n repetições por grupo, o número de permutações (re-organizações) dos dados é dado por (CLARKE, 1993, apud, ANDERSON, 2001):

$$P = \frac{N!}{k!(n_1!n_2!...n_k!)} \tag{5.12}$$

Em geral, até 1000 permutações são suficientes para o teste considerando $\alpha = 0{,}05$ (MANLY, 1997, apud, ANDERSON, 2001).

5.4 Síntese do capítulo

Neste capítulo, foram expostos os métodos estatísticos utilizados para a aplicação da análise de variância não-paramétrica, a seguir descritos:

- Teste de normalidade multivariada;
- MANOVA aplicada aos dados em forma de postos conforme sugestão de Zwick (1985);
- Análise de variância multivariada não-paramétrica com base em distâncias entre os dados conforme sugestão de Anderson (2001).

No próximo capítulo serão apresentados os resultados da presente pesquisa juntamente com as discussões a respeito.

6 RESULTADOS E DISCUSSÕES

A seguir são apresentados os resultados e discussões referentes à presente pesquisa. Para o tratamento dos dados no decorrer deste capítulo, fez-se uso de ferramentas estatísticas descritas nos capítulos 4 e 5.

Este capítulo é dividido em duas partes. Na primeira parte, é realizada a comparação de grupos de forma univariada utilizando a análise de variância de Kruskal-Wallis.

Na segunda parte, é realizada a comparação de grupos utilizando a abordagem de análise de variância multivariada não-paramétrica.

Além disso, no decorrer do capítulo são apresentados todos os comandos utilizados para realizar as análises no *software* R.

6.1 Comparação das rotas de coleta utilizando a Análise de Variância de Kruskal-Wallis

Para carregar o banco de dados no console do R deve-se definir qual o diretório onde o arquivo com os dados está armazenado. Para mudar o diretório deve-se selecionar o *menu arquivo* e clicar em *Mudar dir...*, e assim alterar o diretório. Para visualizar os arquivos que estão salvos no diretório escolhido deve-se utilizar o comando *dir()*.

O arquivo contendo o banco de dados em formato *.txt* pode ser importado para o R através do comando *read.table()*. Neste caso, o primeiro banco de dados contém as variáveis a serem analisadas e será importado através do seguinte comando:

> variaveis <- read.table("VARIAVEIS_PRIM.txt", h=T).

O segundo banco de dados se refere aos grupos a serem comparados e é importado pelo comando:

> grupos <- read.table("GRUPOS_PRIM.txt", h=T).

O objetivo do estudo é realizar a comparação entre as rotas de coleta por meio da análise de variância. Porém, é necessário realizar a análise das

pressuposições de normalidade dos dados e homocedasticidade das variâncias.

Para verificar a existência de normalidade nos dados é utilizado o teste de Shapiro Wilk através dos comandos:

> shapiro.test(agua);

> shapiro.test(acidez);

> shapiro.test(gordura);

> shapiro.test(densidade);

> shapiro.test(lactose);

> shapiro.test(proteina).

Os resultados do teste de normalidade se encontram na Tabela 4.

Tabela 4 – Significâncias do teste de Shapiro Wilk aplicado aos dados.

Variável	Teste de Shapiro-Wilk (W)	p-valor
Água	0,8762	<0,0001
Acidez	0,9508	0,0036
Gordura	0,9707	0,0597
Densidade	0,9816	0,2984
Lactose	0,9823	0,3257
Proteína	0,9856	0,5027

Analisando a Tabela 4, verifica-se que as variáveis água excedente e acidez não seguem distribuição normal.

Além do teste de normalidade é necessário realizar a verificação da homocedasticidade de variâncias. Para isso utiliza-se o teste de Bartlett.

Para realizar o teste de Bartlett é necessário utilizar o seguinte comando:

> bartlett.test(agua ~ ROTA, data=grupos);

> bartlett.test(acidez ~ ROTA, data=grupos);

> bartlett.test(gordura ~ ROTA, data=grupos);

> bartlett.test(densidade ~ ROTA, data=grupos);

> bartlett.test(lactose ~ ROTA, data=grupos);

> bartlett.test(proteina ~ ROTA, data=grupos).

Os resultados do teste de homocedasticidade de variâncias estão expostos na Tabela 5.

Tabela 5 – Significâncias do teste de Bartlett aplicado aos dados.

Variável	Teste de Bartlett (χ^2)	p-valor
Água	2,4483	0,2940
Acidez	1,2307	0,5404
Gordura	4,1631	0,1247
Densidade	0,3635	0,8338
Lactose	4,5697	0,1018
Proteína	3,0399	0,2187

Observa-se na Tabela 5 que todas as variáveis apresentaram variâncias homocedásticas.

Porém, como algumas variáveis não atenderam a pressuposição de normalidade dos dados, utiliza-se a abordagem não-paramétrica para a comparação de grupos por meio da análise de variância de Kruskal-Wallis. Os comandos a serem utilizados para realizar a análise de variância são:

> kruskal.test(agua ~ ROTA, data=grupos);

> kruskal.test(acidez ~ ROTA, data=grupos);

> kruskal.test(gordura ~ ROTA, data=grupos);

> kruskal.test(densidade ~ ROTA, data=grupos);

> kruskal.test(lactose ~ ROTA, data=grupos);

> kruskal.test(proteina ~ ROTA, data=grupos).

Os resultados da realização da análise de variância de Kruskal-Wallis estão na Tabela 6.

Tabela 6 – Significâncias da análise de variância de Kruskal-Wallis aplicada aos dados.

Variável	Anova Kruskal-Wallis	p-valor
Água	6,9511	0,0309
Acidez	5,1001	0,0781
Gordura	2,1898	0,3346
Densidade	2,2987	0,3168
Lactose	0,7226	0,6968
Proteína	0,5784	0,7488

De acordo com a Tabela 6 é possível observar que ocorreu diferença significativa entre as rotas somente em relação a variável água excedente. Desse modo, passa-se à comparação das rotas duas a duas para verificar

quais rotas diferiram entre si. Para este procedimento utiliza-se o teste Wilcoxon-Mann-Whitney.

Para realizar o teste utiliza-se o comando wilcox.test(). Para comparar as rotas duas a duas em relação a variável água excedente utilizam-se os seguintes comandos:

> wilcox.test(agua_rota1, agua_rota2, alternative= "two.sided");

> wilcox.test(agua_rota1, agua_rota3, alternative= "two.sided");

> wilcox.test(agua_rota2, agua_rota3, alternative= "two.sided").

Em que *agua_rota1*, *agua_rota2* e *agua_rota3* são os dados relativos à variável água nas rotas 1, 2 e 3, respectivamente. Os resultados do teste Wilcoxon-Mann-Whitney estão na Tabela 7.

Tabela 7 – Significâncias do teste de Wilcoxon-Mann-Whitney aplicado aos dados.

Rotas	Teste Wilcoxon-Mann-Whitney (W)	p-valor
Rota1 x Rota2	303	0,0526
Rota1 x Rota3	240	0,6599
Rota2 x Rota3	387	0,0195

Observando a Tabela 7 verifica-se que as rotas 2 e 3 diferiram significativamente quanto a variável água excedente. Desse modo, verifica-se que de forma geral as rotas não apresentaram heterogeneidade. Pois ocorreram diferenças somente para uma das variáveis, enquanto que para as demais variáveis, não ocorreram diferenças significativas.

6.2 Comparação das rotas de coleta utilizando a Análise de Variância Multivariada Não-Paramétrica

Neste segundo momento passa-se a analisar as variáveis de forma conjunta. Primeiramente analisam-se as correlações entre as variáveis consideradas no estudo.

Como algumas variáveis não apresentaram normalidade, conforme o teste de Shapiro Wilk, opta-se por realizar a análise de correlação de

Spearman que é um procedimento não-paramétrico. Para realizar a análise de correlação de Spearman usa-se o seguinte comando:

> cor(variaveis, method= "spearman").

Este comando realiza a análise de correlação na matriz de dados originais, mas sem informar as significâncias das correlações. Para que se tenha a correlação entre as variáveis e sua significância deve-se utilizar os comandos:

> cor.test(agua, acidez, method= "spearman", alternative= "two.sided");
> cor.test(agua, gordura, method= "spearman", alternative= "two.sided");
> cor.test(agua, densidade, method= "spearman", alternative= "two.sided");
> cor.test(agua, lactose, method= "spearman", alternative= "two.sided");
> cor.test(agua, proteina, method= "spearman", alternative= "two.sided");
> cor.test(acidez, gordura, method= "spearman", alternative= "two.sided");
> cor.test(acidez, densidade, method= "spearman", alternative= "two.sided");
> cor.test(acidez, lactose, method= "spearman", alternative= "two.sided");
> cor.test(acidez, proteina, method= "spearman", alternative= "two.sided");
> cor.test(gordura, densidade, method= "spearman", alternative= "two.sided");
> cor.test(gordura, lactose, method= "spearman", alternative= "two.sided");
> cor.test(gordura, proteina, method= "spearman", alternative= "two.sided");
> cor.test(densidade, lactose, method= "spearman", alternative= "two.sided");
> cor.test(densidade, proteina, method= "spearman", alternative= "two.sided");
> cor.test(lactose, proteina, method= "spearman", alternative= "two.sided").

As correlações entre as variáveis estão na Tabela 8.

Tabela 8 – Correlações de Spearman entre as variáveis e suas respectivas significâncias.

Variável	Água	Acidez	Gordura	Densidade	Lactose	Proteína
Água	1,0000					
Acidez	-0,2325 p=0,0367	1,0000				
Gordura	-0,0384 p=0,7337	0,0489 p=0,6647	1,0000			
Densidade	-0,0706 p=0,5311	0,2256 p=0,0429	-0,2151 p=0,0538	1,0000		
Lactose	-0,1393 p=0,2148	0,2919 p=0,0082	-0,0470 p=0,6768	0,7262 p=0,0001	1,0000	
Proteína	-0,2215 p=0,0469	0,3003 p=0,0065	0,1676 p=0,1347	0,7930 p=0,0001	0,8520 p=0,0001	1,0000

Pela Tabela 8 verifica-se que ocorreram correlações significativas entre as variáveis água excedente e acidez, água excedente e proteína, acidez e densidade, acidez e lactose, acidez e proteína, densidade e lactose, densidade e proteína e entre lactose e proteína.

Assim, uma abordagem multivariada para a comparação das rotas pode ser utilizada.

Para realizar a análise de variância multivariada é necessário primeiramente realizar a verificação das pressuposições de normalidade multivariada e homocedasticidade de matrizes de variância e covariância.

Para realizar o teste de normalidade multivariada pode-se utilizar a extensão do teste de Shapiro Wilk para dados multivariados. Para realizar este teste utiliza-se o comando:

> mshapiro.test(dados).

Em que *dados* recebe a matriz de dados transposta através do comando:

> dados <- t(variaveis).

O comando *mshapiro.test()* está no pacote *mvnormtest* do R. Os resultados do teste de normalidade multivariada estão na Tabela 9.

Tabela 9 – Significância do teste de Shapiro Wilk para normalidade multivariada dos dados.

Variáveis	Teste MShapiro (W)	p-valor
Água Acidez Gordura Densidade Lactose Proteína	0,8002	p<0,0001

Pela Tabela 9 verifica-se que as variáveis não seguem distribuição normal multivariada. Desse modo, opta-se por uma abordagem não-paramétrica. A primeira abordagem sugere que se aplique a MANOVA na forma tradicional nos dados transformados em postos para encontrar a estatística traço de Pillai.

Assim, o primeiro passo é transformar os dados originais em postos. Para isso utiliza-se o comando *rank()* da seguinte maneira:

> ragua <- rank(agua);

> racidez <- rank(acidez);

> rgordura <- rank(gordura);

> rdensidade <- rank(densidade);

> rlactose <- rank(lactose);

> rproteina <- rank(proteina).

Este comando atribui postos da mesma forma que é atribuída na realização da análise de variância de Kruskal-Wallis.

Após realizar a transformação dos dados em postos deve-se juntar novamente as variáveis em forma de postos em um único banco de dados. Para isso utiliza-se o comando *cbind()* conforme abaixo:

> rvariaveis <- cbind(ragua, racidez, rgordura, rdensidade, rlactose, rproteina).

De posse dos dados em forma de postos é possível realizar a MANOVA através do comando *manova()*. Para realizar o teste com o traço de Pillai utilizam-se os seguintes comandos:

> resultado <- manova(rvariaveis ~ rotas);

> summary(resultado).

Os resultados para a MANOVA aplicada aos dados em forma de postos estão na Tabela 10.

Tabela 10 – MANOVA com o traço de Pillai aplicada aos dados.

Manova	Traço de Pillai	F aproximado	p-valor
Rotas	0,1134	1,5772	0,1658

Com o traço de Pillai calculado é possível encontrar a estatística definida por Zwick (1985), que aqui é denominada por *X*, conforme os comandos a seguir:

*> X < -(N-1)*V*,

em que *N* é o total de observações e *V* é o valor do traço de Pillai.

Para testar a significância desta estatística utiliza-se a distribuição qui-quadrado com p(k-1) gruas de liberdade, conforme o comando abaixo:

> pvalor <-pchisq(q=X,df=p(k-1), lower.tail =F).*

em que *p* é o número de variáveis e *k* é o número de grupos. O resultado do teste está exposto na Tabela 11.

Tabela 11 – Significância da análise de variância multivariada não-paramétrica, com base no estudo de Zwick (1985), aplicada aos dados.

Estatística sugerida por Zwick	p-valor
9,0704	0,6969

Pelo resultado da Tabela 11 verifica-se que não ocorreram diferenças significativas entre as rotas de coleta quanto ao conjunto de variáveis físico-químicas.

Considerando agora a segunda abordagem, onde realiza-se a análise de variância multivariada não-paramétrica com base em medidas de distância, deve-se, primeiramente, carregar no R o pacote *Vegan*.

No *Vegan*, utiliza-se o comando *adonis* para realizar a análise de variância multivariada não-paramétrica. Neste procedimento utiliza-se a distância euclidiana entre as observações. Primeiramente são consideradas

1000 permutações (rearranjos dos dados originais), conforme o comando a seguir:

> adonis(variaveis ~ ROTA, data=grupos, permutations=1000, method="euclidian").

Os resultados da análise estão na Tabela 12.

Tabela 12 – Significância da análise de variância multivariada não-paramétrica, com base no estudo de Anderson (2001), aplicada aos dados, considerando 1000 permutações.

CV	GL	SQ	QM	F	R2	p-valor
Rota	1	5,5105	5,5105	1,2475	0,0155	0,2597
Resíduos	79	348,9654	4,4173		0,9845	
Total	80	354,4759			1,0000	

Pela Tabela 12, percebesse que não ocorreram diferenças significativas entre as rotas de coleta considerando o conjunto de variáveis simultaneamente.

Considerando 5000 permutações têm-se os resultados da Tabela 13.

Tabela 13 – Significância da análise de variância multivariada não-paramétrica, com base no estudo de Anderson (2001), aplicada aos dados, considerando 5000 permutações.

CV	GL	SQ	QM	F	R2	p-valor
Rota	1	5,5105	5,5105	1,2475	0,0155	0,2843
Resíduos	79	348,9654	4,4173		0,9845	
Total	80	354,4759			1,0000	

De acordo com a Tabela 13 verifica-se que também não ocorreram diferenças significativas entre as rotas de coleta considerando 5000 permutações.

Mais uma vez se confirma o fato de que as rotas apresentaram-se semelhantes em relação as variáveis físico-químicas consideradas. Considerando a comparação das rotas de forma univariada verificou-se diferença significativa entre as rotas para a variável água excedente, porém quando se faz a comparação das rotas considerando todas as variáveis simultaneamente, verifica-se que não ocorreram diferenças significativas.

Para verificar o comportamento dos grupos de fornecedores utiliza-se a análise descritiva das variáveis através da média e do desvio padrão. Os comandos utilizados para esta análise são:

> mean(agua1);

> mean(agua2);

> mean(agua3);

> sd(agua1);

> sd(agua2);

> sd(agua3).

Em que *agua1*, *agua2*, e *agua3* são os dados referentes a variável água excedente observados nas rotas 1, 2 e 3, respectivamente. Para as demais variáveis, os comandos são semelhantes.

A Tabela 14 apresenta os valores das médias e desvios padrões das variáveis em cada rota de coleta.

Tabela 14 – Média e desvio padrão (DP) das variáveis em cada uma das rotas de coleta.

Grupo	Água excedente		Acidez		Gordura		Densidade		Lactose		Proteína	
	Média	DP	Média	DP	Média	DP	Média	DP	Média	DP	Média	DP
Rota 1	7,29	0,99	17,14	0,91	3,45	0,35	1027,79	1,14	4,41	0,09	3,28	0,06
Rota 2	6,56	1,48	17,56	1,11	3,39	0,47	1028,05	1,01	4,44	0,16	3,30	0,09
Rota 3	7,33	1,43	16,96	0,93	3,52	0,58	1028,28	0,99	4,44	0,12	3,30	0,07

Observando a Tabela 14 verifica-se que a rota 1 teve, em média, 7,29% de água excedente, acidez de 17,14 ºD, 3,45% de gordura, densidade de 1027,79g/mL, 4,41% de lactose e 3,28% de proteína nas amostras analisadas. Na rota 2 observou-se, em média, 6,56% de água excedente, acidez de 17,56ºD, 3,39% de gordura, densidade de 1028,05g/mL, 4,44% de lactose e 3,30% de proteína. Já na rota 3, observou-se que, em média, as amostras apresentaram 7,33% de água excedente, acidez de 16,96 ºD, 3,52% de gordura, densidade de 1028,28g/mL, 4,44% de lactose e 3,30% de proteína.

6.3 Síntese do capítulo

Neste capítulo, desenvolveu-se a aplicação das técnicas estatísticas, utilizando a metodologia proposta, juntamente com a explicação dos comandos utilizados.

No próximo capítulo serão apresentadas as conclusões observadas, com base nos resultados alcançados neste capítulo.

7 CONCLUSÕES

Este capítulo é dedicado às conclusões obtidas após o desenvolvimento do presente estudo.

O objetivo da pesquisa foi desenvolver um estudo sobre Análise de Variância Multivariada Não-Paramétrica para comparação de rotas de coleta de leite.

Para isso, usaram-se técnicas não-paramétricas univariadas e multivariadas para comparação de grupos.

Com a realização da análise de variância de Kruskal-Wallis, verificou-se que ocorreu diferença significativa entre as rotas somente em relação a variável água excedente.

Desse modo, verifica-se que de forma geral as rotas não apresentaram heterogeneidade, pois ocorreram diferenças somente para uma das variáveis, enquanto que para as demais variáveis, não ocorreram diferenças significativas.

Porém, uma abordagem multivariada pode ser usada para realizar as comparações entre as rotas de coleta. A primeira abordagem multivariada levou em consideração a aplicação da MANOVA aos dados em forma de postos. Com esta aplicação verificou-se que não ocorreram diferenças significativas entre as rotas quanto ao conjunto de variáveis físico-químicas.

Considerando a segunda abordagem, onde realizou-se a análise de variância multivariada não-paramétrica com base em medidas de distância, verificou-se que também não ocorreram diferenças significativas entre as rotas de coleta.

Considerando a comparação das rotas de forma univariada verificou-se diferença significativa entre as rotas para a variável água excedente, porém quando se fez a comparação das rotas considerando todas as variáveis simultaneamente, verificou-se que não ocorreram diferenças significativas.

Em relação a utilização do *software R*, foi possível mostrar os comandos a serem seguidos para realizar os procedimentos de análise realizados nesta pesquisa. Espera-se que o trabalho possa contribuir para pesquisadores que necessitem realizar comparações uni e multivariadas utilizando métodos não-paramétricos.

7.1 Sugestões para trabalhos futuros

Para trabalhos futuros, sugere-se a inclusão de mais variáveis e a utilização de um período maior de investigação.

Além disso, sugere-se um estudo mais aprofundado sobre as abordagens não-paramétricas de análise de variância multivariada e dos métodos de comparações múltiplas multivariadas.

7.2 Síntese do capítulo

Este capítulo apresentou as conclusões da presente pesquisa e as sugestões para trabalhos futuros.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON, M. J. A new method for non-parametric multivariate analysis of variance. **Austral Ecology**, 26, p. 32-46, 2001.

ANSUJ, A. P. **Melhoramento da qualidade de um processo de produção contínua utilizando técnicas estatísticas e os métodos de Taguchi**. 2000, 128f. Tese (Doutorado em Engenharia de Produção) – Universidade Federal de Santa Maria, Santa Maria, 2000.

BELCHIOR, F. Lácteos 100% saudáveis. **Leite & Derivados**, Ano XII, n. 69, p. 30-32, 2003.

BISQUERRA, R.; SARRIERA, J. C.; MARTINEZ, F. **Introdução à estatística: enfoque informático com o pacote estatístico SPSS**. Porto Alegre: Artmed, 2004.

BRASIL. **Instrução Normativa n. 51**. Brasília: MINISTÉRIO DA AGRICULTURA E ABASTECIMENTO, 2002. Disponível em: http://extranet.agricultura.gov.br/sislegis-consulta/consultarLegislacao.do?operacao=visualizar&id=8932. Acesso em: 04 ago. de 2008.

CANTELMO, N. F.; FERREIRA, D. F. Desempenho de testes de normalidade multivariados avaliado por simulação monte carlo. **Ciênc. Agrotec.**, v. 31, n. 6, p. 1630-1636, 2007.

FIGUEIREDO, M. G.; PORTO, E. Avaliação do impacto da qualidade da matéria-prima no processamento industrial do iogurte natural. **Indústria de Laticínios**, Ano 7, n. 41, p. 76-80, 2002.

GONÇALVES, C. F. F. **Estatística**. Londrina: Ed. UEL, 2002. 304 p.

HAIR Jr., J. F. et al. **Análise multivariada de dados**. 5 ed. Porto Alegre: Bookman, 2005.

INÁCIO FILHO, G. **A monografia na universidade.** 7 ed. Campinas: Papirus, 2004.

KATZ, B. M.; MCSWEENEY, M. A multivariate Kruskal-Wallis test with post hoc procedures. **Multivariate Behavioral Research**, 15, p. 281-297, 1980.

KROLOW, A. C. R.; RIBEIRO, M. E. R. **Obtenção de leite com qualidade e elaboração de derivados**. Pelotas: Embrapa Clima Temperado, 2006. 66 p.

LOPES, L. F. D. **Projetos de experimentos: parte I**. Apostila. PPGEP – UFSM, 2007.

MADALENA, F. E. Valores econômicos para a seleção de gordura e proteína do leite. **Rev. Bras. Zootec.**, v. 29, n. 3, p. 678-684, 2000.

MARCONI, M. A.; LAKATOS, E. A. **Fundamentos de metodologia científica**. 6ª ed. São Paulo: Atlas, 2005.

MIGUEL, P. A. C. Estudo de caso na engenharia de produção: estruturação e recomendações para sua condução. **Produção**, v. 17, n. 1, p. 216-229, 2007.

MONTGOMERY, D. C. **Design and analysis of experiments**. New York: John Wiley, 1976, 418 p.

NOAL, R. M. C. **Ações de melhoria contínua para incrementar a qualidade e produtividade na cadeia do leite**. 2006. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de Santa Maria, Santa Maria, 2006.

PEREDA, J. A. O. et al. **Tecnologia de alimentos: alimentos de origem animal**. v. 2, Porto Alegre: Artmed, 2005. 279 p.

PONTES, A. C. F. **Análise de variância multivariada com a utilização de testes não-paramétricos e componentes principais baseados em matrizes de postos**. 2005. Tese (Doutorado em Agronomia) – USP, Piracicaba, 2005.

RODRIGUES, A. **A pesquisa experimental em psicologia e educação**. 2 ed. Petrópolis: Editora Vozes, 1976.

SIEGEL, S.; CASTELLAN JR, N. J. **Estatística não-paramétrica para ciências do comportamento**. 2 ed. Porto Alegre: Artmed, 2006. 448 p.

SILVA, L. S. C. V. **Aplicação do controle estatístico de processos na indústria de laticínios lactoplasa: um estudo de caso**. 83 f. 1999. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de Santa Catarina, Florianópolis, 1999.

SOARES, F. M.; MACHADO, E. C.; FONSECA, L. M. Produção de queijos com teores reduzidos de gordura. **Indústria de Laticínios**, Ano 7, n. 41, p. 68-71, 2002.

TRONCO, V. M. **Manual para inspeção da qualidade do leite**. 3 ed. Santa Maria: Ed. da UFSM, 2008. 206 p.

VARNAM, A. H.; SUTHERLAND, J. P. **Leche y produtos lácteos: tecnologia, química y microbiologia**. Zaragoza: Editorial ACRIBIA, 1995. 476 p.

VILELA, D.; BRESSAN, M.; CUNHA, A. S. **Cadeia de lácteos no Brasil: restrições ao seu desenvolvimento**. Brasília: MCT/CNPq, Juiz de Fora: Embrapa Gado de Leite, 2001. 484 p.

ZWICK, R. Nonparametric one-way multivariate analysis of variance: a computational approach based on the Pillai-Bartlett trace. **Psychological Bulletin**, v. 97, n. 1, p. 148-152, 1985.