INGEPRO – Inovação, Gestão e Produção
Março de 2011, vol. 03, nº. 03
ISSN 1984-6193
www.ingepro.com.br
equipe@ingepro.com.br

Comparação de rotas de coleta de leite usando métodos não-paramétricos

Enio Júnior Seidel <ejrseidel@hotmail.com>
Luis Felipe Dias Lopes <lflopes@smail.ufsm.br>
Angela Pellegrin Ansuj <angelaansuj@yahoo.com>

Resumo: O objetivo deste trabalho é desenvolver um estudo utilizando abordagens não-paramétricas univariada e multivariada para comparação entre grupos, que serão aplicadas em rotas de coleta de leite, com base nas variáveis físico-químicas do produto. Foram consideradas 81 observações coletadas no período de outubro a dezembro de 2007, em três rotas de coleta do leite denominadas de rota 1, rota 2 e rota 3, realizadas por uma usina de laticínios. As variáveis consideradas na análise foram: Água Excedente (%); Acidez (ºD); Gordura (%); Densidade (g/mL); Lactose (%) e Proteínas (%). Inicialmente, compararam-se as rotas utilizando o método não-paramétrico univariado. Por esse método, verificou-se diferença significativa entre as rotas apenas para a variável água excedente. Após, realizou-se a comparação pelo método multivariado, onde, verificou-se que não ocorreram diferenças significativas entre as rotas.
Palavras-chave: Comparação de rotas; Variáveis físico-químicas; Análise de variância univariada não-paramétrica; Análise de variância multivariada não-paramétrica.

Comparison of milk collection routes using nonparametric methods

Abstract: The objective of this work is to develop a study utilizing non-parametric univariate and multivariate approaches for comparison between milk collection routes, on the basis of physico-chemical variables of the product. 81 observations were collected in the period of October to December of 2007, in three milk collection routes named as route 1, route 2 and route 3, carried out by a dairy products factory. The variables considered in the analysis were: Excess Water (%); Acidity (ºD); Fat (%); Density (g/ml); Lactose (%) and Proteins (%). Initially, the routes were compared utilizing the non-parametric univariate approach. From the analysis, it is verified that there were significant differences between the routes only for the variable excess water. After, the routes were compared using multivariate approach, from which, it is verified that there were no significant differences occurred between the routes.
Keywords: Comparison of routes; Physico-chemical variables; Non-parametric univariate analysis of variance; Non-parametric multivariate analysis of variance.

1. Introdução

A comparação entre grupos, considerando uma única variável resposta, pode ser efetuada utilizando-se o procedimento não-paramétrico de análise de variância de Kruskal-

Wallis (GIBBONS; CHAKRABORTI, 1992) e o teste Wilcoxon-Mann-Whitney (SIEGEL; CASTELLAN JR, 2006) quando as pressuposições associadas ao procedimento paramétrico não são satisfeitas.

Contudo, quando múltiplas variáveis estão sendo medidas, utilizar uma abordagem univariada para comparar grupos exige a realização de vários testes univariados, o que dificulta a interpretação dos resultados, pois pode haver diferenças em relação a uma variável, mas não em relação à outra variável.

Desse modo, a incorporação de várias variáveis deve levar em conta o inter-relacionamento entre elas e melhorar a eficiência da análise dos dados. Segundo Pontes (2005), em geral, as diferenças entre grupos ou populações não dependem somente de uma variável, mas do conjunto delas.

Assim, a abordagem multivariada é a mais aconselhada quando se têm $p > 1$ variáveis respostas a serem consideradas para avaliar diferenças entre grupos. Neste caso, pode-se utilizar um procedimento multivariado não-paramétrico, se as pressuposições para a utilização de um procedimento paramétrico não forem satisfeitas.

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

Nesta pesquisa, o procedimento utilizado baseia-se no estudo realizado por Anderson (2001), onde se apresenta uma proposta de utilização de análise de variância multivariada permutacional.

O objetivo deste trabalho é desenvolver um estudo utilizando abordagens não-paramétricas univariada e multivariada para comparação de rotas de coleta de leite, com base nas variáveis físico-químicas do produto. Este trabalho se justifica pela busca em contribuir para uma maior difusão dos procedimentos multivariados não-paramétricos.

2. Metodologia da pesquisa

A presente pesquisa constitui-se de um estudo comparativo entre grupos de fornecedores de leite, caracterizados pelas rotas de coleta utilizadas por uma usina de laticínios, através das análises de variância não-paramétricas univariada e multivariada.

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

As variáveis consideradas foram: água excedente (%); acidez (ºD); gordura (%); densidade (g/mL); lactose (%) e proteínas (%).

Para testar a normalidade dos dados foram utilizados o teste de Shapiro Wilk (no caso univariado) e uma extensão do teste de Shapiro Wilk (no caso multivariado).

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

Após, foi utilizado o procedimento não-paramétrico multivariado, com a abordagem proposta por Anderson (2001), por meio da análise de variância multivariada permutacional.

Para a aplicação das técnicas e desenvolvimento do estudo utilizou-se o *software R* (R DEVELOPMENT CORE TEAM, 2007).

3. Análise de Variância Univariada Não-Paramétrica

O teste de Shapiro Wilk, ou teste W, é utilizado para verificar se os dados seguem uma distribuição normal. As hipóteses a serem testadas são:

H_0: os dados seguem distribuição normal;

H_1: os dados não seguem distribuição normal.

Rejeita-se a hipótese H_0 se o valor de W do teste for demasiadamente pequeno (SCHNEIDER; SCHNEIDER; SOUZA, 2009).

A técnica de Kruskal-Wallis testa a hipótese de que as k amostras provêm da mesma população ou de populações idênticas com a mesma mediana. 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.

onde: θ_j representa a mediana para o j - ésimo grupo.

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

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. Ao menor valor é atribuído o posto 1, ao seguinte menor valor é atribuído o posto 2 e ao maior valor é atribuído o posto n, onde o 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, somam-se estes valores para cada amostra. Com as 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.

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).

A estatística H é calculada pela expressão (GIBBONS; CHAKRABORTI, 1992):

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

onde: 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; $\bar{R}_j$ 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, a nova expressão para H é (GIBBONS; CHAKRABORTI, 1992):

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

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 de Wilcoxon-Mann-Whitney (SIEGEL; CASTELLAN JR, 2006).

4. Análise de Variância Multivariada Não-Paramétrica

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 aplica-se o teste de Shapiro-Wilk nesta nova variável.

Para comparar as rotas no caso multivariado toma-se uso da análise de variância multivariada permutacional. Este procedimento não-paramétrico leva em consideração medidas de distâncias entre pares de observações, que são comparadas dentro do mesmo grupo contra as 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).

Segundo Anderson (2001), a soma de quadrados total pode ser definida como:

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

A soma de quadrados dentro de grupo é dada por:

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

Em que ε_{ij} vale 1 (um) se as observações i e j são do mesmo grupo, e vale 0 (zero) se i e j não pertencem ao mesmo grupo.

Desse modo, a soma de quadrados entre grupos é:

$$SS_A = SS_T - SS_W$$

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

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

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 o valor F^{π} para todas estas reorganizações dos dados. Assim, o p-valor é definido por:

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

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

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

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

5. Resultados e Discussões

Nesta etapa, realizou-se a comparação entre as rotas de coleta por meio da análise de variância univariada não-paramétrica, pois ocorreu violação na pressuposição de normalidade dos dados. Os resultados do teste de normalidade se encontram na Tabela 1.

Tabela 1 – 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 1, verifica-se que a variável água excedente e acidez não seguem uma distribuição normal. Por isso, utilizou-se a análise de variância de Kruskal-Wallis para comparar as rotas. Os resultados estão na Tabela 2.

Tabela 2 – Análise de Variância de Kruskal-Wallis aplicado 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 2, é possível observar que ocorreu diferença significativa entre as rotas somente em relação a variável água excedente. Desse modo, passa-se à