Univesidade Federal Rural do Rio de Janeiro

Instituto de Ciências Exatas

Departamento de Matemática

A Utilização de Alguns Testes Estatísticos para Análise da Variabilidade do Preço do Mel nos Municípios de Angra dos Reis e Mangaratiba, Estado do Rio de Janeiro

Patrícia Araújo Scudino

Orientador: Prof. Mestre Wagner de Souza Tassinari

Seropédica - RJ

2008

Patrícia Araújo Scudino

A Utilização de Alguns Testes Estatísticos para Análise da Variabilidade do Preço do Mel nos Municípios de Angra dos Reis e Mangaratiba, Estado do Rio de Janeiro

Sob a orientação do Prof. Mestre Wagner de Souza Tassinari

Monografia submetida como requisito parcial para obtenção do grau de Licenciado e Bacharel em Matemática.

Seropédica

Junho - 2008

Patrícia Araújo Scudino

A Utilização de Alguns Testes Estatísticos para Análise da Variabilidade do Preço do Mel nos Municípios de Angra dos Reis e Mangaratiba, Estado do Rio de Janeiro

Monografia submetida como requisito parcial para obtenção do grau de Licenciado e Bacharel em Matemática, submetida à aprovação da banca examinadora composta pelos seguintes membros:

Prof. Mestre Wagner de Souza Tassinari

Profª. Dr. Maria Cristina Lorenzon

Prof. Dr. Celso Guimarães Barbosa

Seropédica, 2008.

Agradeço primeiramente a Deus, por todas as oportunidades que tem me dado.

Aos meus pais pelo total apoio, amor e incentivo.

A minha irmã pelo carinho e paciência.

A meus familiares em geral.

Aos meus amigos e professores da universidade.

Em especial ao meu orientador Wagner Tassinari pela total dedicação.

E a professora Maria Cristina Lorenzon pelo apoio.

A todos, muito obrigada!

Amo vocês!!

Resumo

O estado do Rio de Janeiro é um dos maiores centros consumidores de mel no país. Em dez anos a classe produtora dobrou, mas a produção de mel, em torno de 400 toneladas, continua estagnada, favorecendo a importação de muitas marcas de méis de outros estados. Para o Sebrae, o estado do Rio apresentou uma alta devastação ambiental e índices muito pobres de suporte à agricultura familiar, fatores estes que contribuem para a improdutividade. Dentro do estado, a região da Costa Verde é uma das menos expressivas na produção apícola. Este estudo tem por objetivo analisar a variabilidade do preço do mel entre diferentes tipos de estabelecimentos, localizados nos Municipios de Angra dos Reis e Mangaratiba. Para explicar tal fenômeno foram aplicados alguns testes estatísticos não-paramétricos. Nas análises, foi observado que há uma grande variabilidade do preço do mel entre os diferentes tipos de estabelecimentos e embalagens, entre os municípios estudados, fontes de origem e inspeção.

Sumário

Lista de Tabelas

Introdução

A cadeia produtiva da criação de abelhas propicia a geração de inúmeros postos de trabalho, empregos e fluxo de renda, principalmente no que diz respeito à agricultura familiar, que desamparada, encontrou nesta atividade uma diversificação de sua produção. Além disso, a oscilação do comércio externo de mel pressiona o agronegócio apícola a se reestruturar, promovendo um desevolvimento do comércio interno. No Brasil o preço médio do mel é de $R\$2,83$ e no estado do Rio de Janeiro é de $R\$2,27$.

O perfil do consumidor de produtos apícolas foi delineado no mercado da região da Costa Verde - RJ, Brasil mas especificamente em Angra dos Reis e Mangaratiba. Entre janeiro e julho de 2007, 354 estabelecimentos foram pesquisados. Os aspectos deste perfil avaliados no mercado foram: origem das marcas, tipos de produtos, peso, preço, embalagens e florada. A região da Costa Verde apresentou uma vasta gama de produtos e um mercado consumidor promissor. O objetivo dessa monografia é a utilização de alguns testes estatísticos para analisar a variabilidade do preço do mel entre os estabelecimentos nestas duas cidades.

Essa monografia está dividida em quatro capítulos. No primeiro, serão apresentadas alguns tipos de variáveis e alguns métodos utilizados na análise explo-

ratória de dados (EDA). No segundo, serão apresentados alguns testes estatíticos paramétricos e não-paramétricos. Alguns desses testes serão ultilizados para verificação de Normalidade nos dados (Kolmogorov-Smirnov, Shapiro Wilk e Anderson-Darling). Após serão apresentados os testes Paramétricos como o teste T-Student e Análise de variância (ANOVA). E por último alguns testes não-paramétricos como teste do sinal, Wilcoxon-Mann-Whitney e Kruskal-Wallis. No terceiro capítulo será feita uma contextualização sobre o mercado do mel. Já o quarto capítulo será composto pela análise dos dados obtidos através dos métodos estatíticos já apresentados no segundo capítulo.

Finalmente será possível concluir quais são os fatores que explicar a variabilidade do preço do mel nos estabelecimentos em Angra dos Reis e Mangaratiba.

Para avaliar as variáveis que influenciam na variabilidade do preço do mel nestes municípios, foram avaliados: a embalagem, o tipo de estabelecimento, a composição, o município, a origem e a inspeção. Foram visitados e entrevistados 354 estabelicimentos em Angra dos Reis e Mangaratiba, dois importantes municípios de comércio.

Capítulo 1

Estatística Descritiva

Em sua essência, a Estatística é a ciência que apresenta processos próprios para coletar, apresentar e interpretar adequadamente conjuntos de dados, sejam eles numéricos ou não. Pode-se dizer que seu objetivo é o de apresentar informações sobre dados em análise para que se tenha maior compreensão dos fatos que os mesmos representam (BUSSAB e MORRETIN, 2002).

A estatística descritiva é a etapa inicial da análise utilizada para descrever e resumir os dados. Neste capítulo será comentado um pouco dos tipos de medidas de tendência central e dispersão como: média, mediana, quartis e coeficiente de variação.

1.1 Variáveis contínuas e discretas

Uma característica importante nas variáveis é de quão precisamente elas podem ser avaliadas. Isto é, de acordo com sua mensuração elas podem se classificar em contínuas, como, idade, altura, etc., que podem assumir qualquer valor dentro de

um intervalo contínuo. E discretas, que assumem valores inteiros provindos de uma contagem, como, por exemplo, número de filhos por família. Neste caso, não sendo possível utilizar a idéia de contínuo, isto é, obter frações desse evento.

O cumprimento dos requisitos de normalidade condiciona a escolha do pesquisador, a utilizar as estatísticas paramétricas, cujos testes são em geral mais eficientes do que os da estatística não-paramétrica e, conseqüentemente, devem ter a preferência do pesquisador, quando o seu emprego for permitido.

Para avaliar a normalidade da distribuição dos dados podemos utilizar os seguintes testes: Kolmogorov-Smirnov, Shapiro Wilk e Anderson Darling.

1.2 Média Aritmética $(\overline{X})$

A medida de tendência central, mais comumente usada para descrever resumidamente um conjunto de dados, tabelados ou não, é a média aritmética simples. Ela é um valor típico, ou representativo, de um conjunto de dados (SPIEGEL, 1993). Ou podemos dizer que é a razão entre a soma de todos os valores e o número de termos da série.

A média aritmética, em alguns casos, não é uma boa medida de tendência central, pois, se os dados apresentarem algum valor discrepante isso influenciará na posição da média. Quando isto ocorre, a mediana é a medida mais adequada. Dada a variavel X, com os seus n valores distintos, isto é, x_1, ..., x_n ,

temos que média aritmética de X, pode ser escrita:

$$\overline{X} = \frac{x_1 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.1}$$

$$\mu = \frac{x_1 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.2}$$

1.3 Mediana (Md)

A mediana é uma quantidade que, como a média, também procura caracterizar o centro da distribuição de freqüências quando os valores são dispostos em ordem crescente ou decrescente em magnitude. É o valor que divide o conjunto ordenado de valores em duas partes com igual número de elementos, ou seja, 50% das observações ficam acima da mediana e 50% ficam abaixo. Será denotada por Md. Para calcularmos a mediana deve-se, em primeiro lugar, ordenar os dados para que se possa localizar a posição da mediana e assim encontrar seu valor. O número que indica a ordem ou posição em que se encontra o valor correspondente à mediana é denominado elemento mediano (EMd).

Para determinar a mediana é preciso ordenar os dados; em seguida aplique um dos processos:

a) A variável em estudo é discreta e n é ímpar. Neste caso a mediana será o valor da variável que ocupa a posição:

$$EMd = \frac{n\ +\ 1}{2} \tag{1.3}$$

b) A variável em estudo é discreta e n é par. Neste caso a mediana, por convenção, será a média aritmética dos valores que ocupam as posições:

$$EMd = \frac{n}{2} \quad e \quad \frac{n\ +\ 2}{2} \tag{1.4}$$

1.4 Quartil

Um Quartil é qualquer um dos três valores que divide o conjunto ordenado de dados em quatro partes iguais, e assim cada parte representa $\frac{1}{4}$ da amostra. O Primeiro Quartil chamado de quartil inferior, é o valor aos 25% da amostra. O Segundo Quartil, é igual a mediana com o valor até 50% da amostra. O Terceiro Quartil, chamado quartil superior é o valor a partir do qual se encontram 25% dos valores ordenados, ou seja, valor aos 75% da amostra.

1.5 Coeficiente de Variação (CV)

É uma medida relativa de dispersão utilizada para comparar o grau de concentração em torno da média em percentual. Então;

$$CV_{amostra} = \frac{S}{\overline{X}} \times 100 \tag{1.5}$$

$$CV_{populacao} = \frac{\sigma}{\mu} \times 100 \tag{1.6}$$

Se

$CV \leq 15\%$, ocorre uma baixa dispersão, sendo considerada homogênea ou estável.

$15\% \leq CV \leq 30\%$, apresenta uma dispersão média.

$CV \geq 30\%$, apresenta uma dispersão alta, sendo considerada heterogênea.

Capítulo 2

Alguns Testes Estatísticos

A inferência estatística preocupa-se em estimar o verdadeiro valor desconhecido dos parâmetros de uma população e testar hipóteses com respeito aos parâmetros estimados, ou a natureza da distribuição da população. Existem duas classifcações dos testes de hipóteses: os paramétricos (conhece a distribuição dos dados) e os não paramétricos (não se conhece a distribuição dos dados). O pesquisador em sua terefa de analisar os dados necessita identificar quais testes serão utilizados e, antes de tudo, identificar sua hipótese na pesquisa e escolher a técnica de coleta de dados (CARVALHO, 2007).

2.1 Testes de Hipóteses

Nos estudos em pesquisas quantitativas, são formuladas hipóteses acerca de uma dada amostra, que serão submetidas a testes específicos. De acordo com Devore (2006), uma hipótese estatística é uma alegação ou afirmação sobre o valor de um único parâmetro, ou sobre os valores de vários parâmetros, ou sobre a forma de

uma distribuição de probabilidade inteira.

Nos testes de hipóteses, existem duas suposições contraditórias em consideração. O objetivo é decidir, com base nas informações da amostra, qual das duas hipóteses está correta. Então, no teste de hipóteses estatísticas, o problema será formulado de modo que uma das alegações seja inicialmente favorecida. Tal alegação não será rejeitada em favor da alegação alternativa, a menos que a evidência da amostra contradiga e forneça forte apoio à afirmação alternativa (LEVIN, 1987).

2.1.1 Hipótese nula H_0 e Hipótese alternativa H_1

A hipótese nula H_0 é a alegação inicialmente assumida como verdadeira. A hipótese alternativa H_1 é a afirmação contraditória a H_0.

A hipótese nula será rejeitada em favor da hipótese alternativa somente se a evidência da amostra sugerir que H_0 seja falsa. Se a amostra não contradiz fortemente H_0, continua-se a acreditar na verdade da hipótese nula. As duas conclusões possíveis de uma análise do teste de hipóteses são, então, rejeitar H_0 ou não rejeitar H_0 (DEVORE, 2006).

2.1.2 Erros do tipo I e II

Se uma hipótese for rejeitada quando deveria ser aceita, diz-se que foi cometido o erro do tipo I. Se, por outro lado, for aceita uma hipótese que deveria ser rejeitada, diz-se que foi cometido um erro do tipo II. Em ambos os casos ocorreu uma decisão errada ou um erro de julgamento.

2.1.3 Nível de significância e *p*-valor

Para testar uma hipótese estabelecida, a probabilidade máxima com o qual se pode correr o Erro do tipo I é denominada nível de significância do teste (SPIEGEL, 1993). Normalmente, o nível de significância é representado por α e, geralmente, é especificado antes da extração das amostras e das hipóteses, de modo que os resultados obtidos não influenciem a escolha. Usualmente são escolhidos os seguintes níveis $\alpha = 0,01$ ou $0,05$, isto é, se escolhido o índice de $0,01$, então existe 1 chance em 100, da hipótese ser rejeitada. Da mesma maneira podemos dizer que existe uma confiança de 99% de que se tome a decisão certa. Supondo que a hipótese nula seja verdadeira e que a probabilidade de se obter um efeito devido ao erro amostral seja menor do que 1%, o achado é dito significativo. Se a probabilidade for maior que 1%, o achado é dito não-significativo (DANCEY & REIDEY, 2006). Na resposta dos testes de hipóteses, um valor é comparado com o nível de significância previamente escolhido, sendo chamado de p-valor ou valor p, isto é, valor do poder do teste. O p-valor (nível de significância observado) é o menor nível de significância em que H_0 seria rejeitada, quando um procedimento de teste específico é usado em um determinado conjunto de dados. Assim, quando $p - valor \leq \alpha$ implica na rejeição de H_0 no nível α. Ou se $p - valor > \alpha$ implica na não rejeição de H_0 no nível α. Então, em vários estudos as respostas poderão vir referenciando o nível de significância ou $p - valor$.

2.2 Testes de Normalidade

Os testes paramétricos necessitam de alguns pressupostos, a população da qual as amostras são retiradas devem ser normalmente distribuída. Então, se deve sempre verificar antes da análise se os dados da amostra são aproximadamente normais para se decidir pelo uso de um teste paramétrico.

Para isso, se utilizam alguns testes de normalidade, dentre eles destacamos Kolmogorov-Smirnov, Shapiro-wilk e Anderson-Darling.

2.2.1 Kolmogorov-Smirnov

Um dos pressupostos de testes estatísticos paramétricos diz respeito à distribuição normal dos dados nas variáveis das populações. Quando se retira uma amostra para esses modelos de testes, deve-se supor que as unidades do universo em questão apresentem distribuição normal. Será apresentado o teste de normalidade Kolmogorov-Smirnov para uma amostra, (SIEGEL & CASTELLAN JR, 2006). Este teste é um teste de aderência. Verifica o grau de concordância entre distribuição de um conjunto de valores (escores observados) e alguma distribuição teórica, ou seja, verificar se os dados seguem a distribuição normal. O teste Kolmogorov-Smirnov admite que a distribuição da variável que está sendo testada seja contínua. O teste utiliza a distribuição de freqüência acumulada, que ocorreria dada a distribuição teórica, e a compara com a distribuição de freqüência acumulada observada. A distribuição teórica representa o que seria esperado sob H_0. Então, verifica-se se as distribuições teórica e observada mostram divergência.

Seja $F_0(X)$ uma função especificada de distribuição de freqüências relativas

acumuladas, a distribuição teórica sob H_0. Para qualquer valor de X, o valor de $F_0(X)$ é a proporção de casos esperados com escores menores ou iguas a X.

Seja S_N a distribuição de frequências relativas acumuladas observada de uma amostra aleatória de N observações. Se X_i é um escore qualquer possível, então $S_N(X_i) = \frac{F_i}{N}$, onde F_i é o número de observações menores ou iguais a X_i. $F_0(X_i)$ é a proporção esperada de observações menores ou iguais a X_i. As hipóteses do teste são descritas como:

H_0: A amostra provém de uma distribuição teórica específica (neste caso: distribuição normal);

H_1: A amostra não provém de uma distribuição teórica específica (neste caso: distribuição não normal).

A estatística do teste espera que quando H_0 é verdadeira, as diferenças entre $S_N(X_i)$ e $F_0(X_i)$ sejam pequenas e estejam dentro do limite dos erros aleatórios. O teste focaliza o maior dos desvios chamado de desvio máximo:

$$D = max|F_0(X_i) - S_N(X_i)|, \qquad i = 1, 2, ..., N$$

Mas, deve-se verificar a hipótese através do poder do teste $p - valor$. Então verifica-se a normalidade da amostra:

Se $D = max|F_0(X_i) - S_N(X_i)| < D_{(N,\alpha)}$ é não rejeitada H_0; isto é, a amostra provém da distribuição normal.

Se $D = max|F_0(X_i) - S_N(X_i)| > D_{(N,\alpha)}$ é rejeitada H_0; isto é, a amostra não provém da distribuição normal.

Com $D \geq \frac{1.36}{\sqrt{N}}$, para $\alpha = 0,05$; $\quad D \geq \frac{1.63}{\sqrt{N}}$, para $\alpha = 0,01$;

2.2.2 Teste de Shapiro-Wilk

O teste Shapiro-Wilk, calcula uma variável estatística (W) que investiga se uma amostra aleatória provém de uma distribuição normal.

A variável W é calculada da seguinte forma:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2.1}$$

sendo,

- x_i os valores ordenados de amostras (x_1 é o menor).

- a_i constantes geradas a partir de meio, variâncias e covariâncias da ordem estatística de uma amostra de tamanho n e uma distribuição normal.

Sendo X uma característica em estudo, então formula-se as hipóteses:

H_0: X tem distribuição Normal;

H_1: X não tem distribuição Normal.

2.2.3 Teste de Anderson-Darling

O teste Anderson-Darling (STEPHENS, 1974) é usado para testar se uma amostra de dados provém de uma determinada distribuição. Trata-se de uma modificação do teste Kolmogorov-Smirnov (KS). O Teste KS é de distribuição gratuita, no sentido de que os valores críticos não dependem da distribuição específica para calcular valores críticos. Isto tem a vantagem de permitir um exame mais sensível e a desvantagem de que os valores críticos devem ser calculados para cada distribuição.

O teste Anderson-Darling é definido como:

$$A^2 = -N - S \tag{2.2}$$

sendo

$$S = \sum_{i=1}^{N} \frac{(2_i - 1)}{N} [logF(Y_i) + log(1 - F(Y_{N+1-i})] \tag{2.3}$$

onde F é a distribuição cumulativa dos dados.

As hipóteses do teste são descritas como:

H_0: Os dados seguem uma distribuição especificada;

H_1: Os dados não seguem uma distribuição especificada.

Os valores críticos para o teste Anderson-Darling, são dependentes da distribuição específica, sendo testada. Valores tabulados e fórmulas foram publicados por Stephens para algumas distribuições específicas (normal, lognormal, exponencial, Weibull, logística, extremo valor tipo 1, dupla exponencial, uniforme, generalizada pareto).

Testar a hipótese de que a distribuição é feita de uma forma específica é rejeitada se a estatística de ensaio, A^2 for superior ao valor crítico.

2.3 Testes Paramétricos

Testes estatísticos paramétricos especificam certas condições sobre a distribuição das respostas na população, da qual a amostra da pesquisa foi retirada. Essas

condições devem ser testadas para que os resultados de um teste paramétrico sejam significativos. Os dados devem seguir a distribuição normal para que se tenha uma interpretação apropriada de testes e, também, que as variáveis, ou escores a serem analisados, resultem de medidas em pelo menos uma escala intervalar. Então, como mencionado no item anterior, é de suma importância verificar a normalidade dos dados.

2.3.1 Teste t de Student em duas amostras independentes

O teste t para duas amostras é usado quando temos duas condições e se precisa saber se as diferenças entre as médias das amostras são grandes o suficiente para que se possa concluir que as diferenças ocorrem somente devido à influência da variável independente. Ele avalia as diferenças significativas entre as médias $\mu_1 - \mu_2$ das duas condições (DANCEY & REIDY, 2006).

Ambas as populações são normais de modo que as amostras aleatórias de uma distribuição amostral $X_1, X_2, ..., X_m$ e $Y_1, Y_2, ..., Y_n$, com X'_s e Y'_s independentes entre si.

A estatística do teste com distribuição da população normal e variável padronizada:

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}. \tag{2.4}$$

As hipóteses seguem a seguinte estrutura:

$H_0 : \mu_1 = \mu_2$, não existe diferença entre as médias das populações;

$H_1 : \mu_1 \neq \mu_2$, existe diferença entre as médias das populações;

Hipótese alternativa	Região de rejeição ao nivel α
$H_1 : \mu_1 - \mu_2 > 0$	$t \geq t_{\alpha,v}$
$H_1 : \mu_1 - \mu_2 < 0$	$t \leq t_{\alpha,v}$
$H_1 : \mu_1 \neq \mu_2$	$out \geq t_{\alpha/2,v}$ ou $t \leq t_{\alpha/2,v}$

Existem muitos problemas, em que o tamanho da amostra é pequeno e as variâncias da população possuem valores desconhecidos. Nesses casos não se poderá aplicar o teste Z para duas amostras, justificando a grande aplicação do teste t de Student (DEVORE, 2006).

2.3.2 Análise da variância (ANOVA)

Em muitas pesquisas comparação entre várias médias se torna necessário. Um procedimento será de utilizar o teste t de Student em duas variáveis, que torna a estatística demorada e trabalhosa. O mais recomendado é utilizar a análise da variância (ANOVA). Ela deve seguir algumas condições, como apresentar os dados com distribuição normal e haver homogeneidade das variâncias.

A ANOVA procura verificar se existem diferenças entre as médias dos grupos. Faz isso determinando a média geral e verificando o quão diferente cada média individual é da média geral (DANCEY & REIDY, 2006). A ANOVA de fator único concentra-se na comparação de mais de duas médias populacionais ou tratamentos. Seja I = número de populações ou tratamentos que serão comparados; e $\mu_1, \mu_2, \mu_3, ..., \mu_i$ as médias populacionais ou médias dos tratamentos;

Então, as hipóteses de interesse são

$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_i$

H_1: Pelo menos duas médias diferentes;

Para obter a estatística do teste é necessário conhecer

A soma dos quadrados totais $(SQT) = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_{ij} - \overline{x})^2$

A soma dos quadrados dos tratamentos $(SQTr) = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_i - \overline{x})^2$

A soma dos quadrados dos resíduos $(SQR) = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_{ij} - \overline{x})^2 - (x_i - \overline{x})^2$

$a =$ número de tratamentos; $b =$ número de repetições

Quadrados médios dos tratamentos $QMT = \frac{SQTr}{a-1}$

Quadrados médios dentro dos tratamentos (resíduos) $QMR = \frac{SQR}{a(b-1)}$

De uma maneira prática o teste F é apresentado pela tabela 2.1 da análise da variância (ANOVA):

Tabela 2.1: Análise da variância - ANOVA

Causa de Variação	g.l.	SQ	QM	F
Tratamento	$a-1$	$(SQT) = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_{ij} - \overline{x})^2$	$QMTr = \frac{SQTr}{a-1}$	$\frac{QMTr}{QMR}$
Resíduos	$a(b-1)$	$(SQR) = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_{ij} - \overline{x})^2 - (x_i - \overline{x})^2$	$QMR = \frac{SQR}{a(b-1)}$	com $a-1$ e $a(b-1)$ g.l.
Total	$ab-1$	$(SQT) = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_{ij} - \overline{x})^2$		

2.4 Testes não-paramétricos

Um teste estatístico não-paramétrico é baseado em um modelo que especifica somente condições muito gerais e nenhuma a respeito da forma específica da distribuição, da qual a amostra foi extraída. E, diferentemente dos testes

paramétricos, os testes não-paramétricos podem ser usados em dados medidos em uma escala nominal. (SIEGEL & CASTELLAN JR, 2006).

2.4.1 Teste do Sinal

O teste do Sinal é utilizado na análise de dados emparelhados. Situações em que o pesquisador desejar determinar se duas condições são diferentes. O nome do teste dos sinais se deve ao fato de utilizar sinais negativos e positivos em lugar dos dados numéricos. A lógica do teste é que as condições podem ser consideradas iguais quando as quantidades de sinais positivos e negativos forem aproximadamente iguais. Isto, é a proporção de sinais positivos equivale a 50%, ou seja, $p = 0,5$.

Então, temos como hipóteses:

H_0: Não há diferença entre os grupos, ou seja, $p = 0,5$

H_1: Há diferença , ou seja, uma das alternativas:

a) $p \neq 0,5$

b) $p < 0,5$

c) $p > 0,5$

O teste do sinal não faz suposição sobre distribuição das diferenças, mas leva em conta apenas o sinal da diferença ignorando a grandeza dessas diferenças. Esse teste não é frequentemente usado na prática.

2.4.2 Teste Wilcoxon-Mann-Whitney

O teste de Wilcoxon-Mann-Whitney é usado para testar se dois grupos independentes foram extraídos da mesma população (SIEGEL & CASTELLAN JR, 2006). É um dos testes não-paramétricos mais poderosos, sendo uma alternativa para o teste t de Student, que necessita que os dados apresentem uma distribuição normal. A variável em estudo pode ser mensurada pelo menos em um nível ordinal.

A hipótese nula H_0 é que dados amostrais de duas populações, X e Y, tenham a mesma distribuição. A hipótese alternativa H_1 é de que se a probabilidade de um escore de X seja diferente de Y, isto é, ela será diferente de meio. Seja m o número de casos na amostra do grupo X e n o número de casos na amostra do grupo Y. Assumidos que as duas amostras são independentes. Para aplicar o teste de Wilcoxon, combinam-se as observações ou escores de ambos os grupos e organizam-se os postos em ordem crescente de tamanho. A estatística desse teste é a soma dos postos no primeiro e segundo grupo, dada por:

W_X = soma dos postos do primeiro grupo;

W_Y = soma dos postos do segundo grupo;

$N = m + n$.

$$W_X + W_Y = \frac{N(N+1)}{2}. \tag{2.5}$$

Se H_0 é verdadeira, a média dos postos em cada um dos dois grupos é quase a mesma. Se a soma dos postos para um grupo é muito grande (ou muito pequena), pode-se suspeitar que as amostras não foram extraídas da mesma população.

Assim, temos que

H_0: não existe diferença entre os dois grupos em relação às probabilidades das respostas;

H_1: existe diferença entre os dois grupos em relação às probabilidades das respostas.

Assim,

Hipótese nula

$H_o : P[X > Y] = \frac{1}{2}$ $\quad P_{cal} > P_{tab}$ não rejeita-se H_0

Hipótese alternativa

$H_1 : P[X > Y] < \frac{1}{2}$

$H_1 : P[X > Y] > \frac{1}{2}$ rejeita-se H_0

$H_1 : P[X > Y] \neq \frac{1}{2}$

Uma observação a ser feita é de quando $m > 10$ ou $n > 10$, a distribuição amostral de W_X aproxima-se rapidamente da distribuição normal, com média 1 e variância unitária.

2.4.3 Teste de Kruskal-Wallis

A análise da variância de um fator de Kruskall-Wallis por postos é usado para decidir se K amostras independentes provêm de populações diferentes. O teste de Kruskall-Wallis verifica a hipótese nula H_0 de que as K amostras provêm da mesma população ou de populações idênticas com a mesma mediana. Então, dada θ_j a mediana para o j-ésimo grupo ou amostra.

O teste de Kruskall-Wallis trabalha com as diferenças entre os postos médios para determinar se elas são tão discrepantes que, provavelmente, não tenham vindo

de amostras que sairam da mesma população. A estatística é definida por

$$KW = \frac{12}{N(N-1)} \sum_{j=1}^{N} n_j(\overline{R}_j - \overline{R})^2, \tag{2.6}$$

sendo

K = número de amostras dos grupos;

n_j = número de casos na j-ésima amostra;

N = número de casos na amostra combinada (a soma dos nj's);

R_j = soma dos postos na j-ésima amostra ou grupo;

$\overline{R_j}$ = média dos postos na j-ésima amostra ou grupo;

$\overline{R} = \frac{(N+1)}{2}$ = média dos postos na amostra combinada (a grande média).

Logo, as hipóteses são definidas por

$H_0 : \theta_1 = \theta_2, ..., \theta_j$ (todos os grupos têm medianas iguais).

$H_1 : \theta_i \neq \theta_J$ (pelo menos um par de grupos tem medianas diferentes).

Portanto, Se $KW_{cal} < KW_{tab}$, não rejeita-se H_0,

Se $KW_{cal} \geq KW_{tab}$ rejeita-se H_0.

Capítulo 3

Mercado do Mel

A atividade apícola teve início no Brasil, com a chegada dos imigrantes italianos ainda no período colonial. Mas foi em 1956 que a apicultura começou a progredir, com a introdução das abelhas africanas - Apis mellifera L. pelo geneticista Dr. Warwick Estevam Kerr. Através dos cruzamentos entre as abelhas africanas e as italianas, temos um híbrido conhecido popularmente como abelha africanizada. Os inúmeros trabalhos na área de produção e melhoramento genético dessa espécie, aliado ao clima favorável ao seu desenvolvimento, fizeram com que em cinqüenta anos a apicultura desse um salto fabuloso de 4.000 ton/ano para 40.000 ton/ano (SEBRAE, 2006).

A apicultura é uma das atividades capazes de causar impactos positivos, tanto sociais quanto econômicos, além de contribuir para a manutenção e preservação dos ecossistemas existentes. A cadeia produtiva da apicultura propicia a geração de inúmeros postos de trabalho, empregos e fluxo de renda, principalmente no ambiente da agricultura familiar, sendo dessa forma, determinante na melhoria da

qualidade de vida e fixão do homem no meio rural.

A produção mundial de mel teve uma tendência crescente nos últimos 20 anos, atribuídas a um aumento no número de colméias e da produção por colônia apesar das flutuações, em regiões e países (industrializados e não industrializados). O consumo também aumentou durante os últimos anos, sendo atribuido ao aumento geral nos padrões de vida e também a um interesse maior pelos produtos naturais e saudáveis.

O estado do Rio de Janeiro é um dos maiores centros consumidores de mel do país. Em dez anos a classe produtora dobrou, mas a produção de mel, em torno de 400 toneladas, continua estagnada, favorecendo a importação de muitas marcas de méis de outros estados. Para o Sebrae, o estado do Rio apresentou uma alta devastação ambiental e índices muito pobres de suporte à agricultura familiar, fatores estes que contribuem para a improdutividade. Dentro do estado, a região da Costa Verde é uma das menos expressivas na produção apícola (RIO BRANCO, 2008).

O termo Costa Verde refere-se a faixa de vegetação costeira, localizada ao sul do litoral fluminense, composta por mais de duas mil praias e quase 400 ilhas. A vegetação é formada pela floresta tropical (Mata Atlântica), apresentando fragmentos com diferentes graus de preservação. O clima é tropical úmido, com temperatura média entre 22°C e 25°C, área total de $2.118,5 km^2$ e uma população de 188.305 habitantes (FIBGE,2000). Os setores econômicos que mais se destacam nesta região são: indústria naval, maricultura, náutico, portuário e turismo.

Capítulo 4

Resultados

Ao se observar as medidas de tendência central e de dispersão, apresentadas na tabela 4.1, é possível verificar que grande parte da variabilidade do preço do mel está relacionada ao tipo de embalagem, a composição do mel, ao município de venda, a origem do mel e ao tipo de inspeção. Em cada variável analisada, em algumas categorias, existem ocorrências extremas (outliers), ou seja, o preço do mel em alguns estabelecimentos ficam muito distantes do padrão da distribuição dos outros preços.

De acordo com as figuras boxplots, em anexo, e a tabela 4.1, verifica-se que no município de Angra dos Reis, a média dos produtos está acima dos de Mangaratiba. Dentre os produtos avaliados, o composto de mel é o que apresenta-se mais caro por unidade de peso ($p - valor < 0,001$), este produto representa uma mistura de mel com extratos comumente de conotação terapêutica e isto pode favorecer a alta no seu preço. A média do preço do mel vendido em embalagens de vidro, é superior

à media geral do preço do mel, pelo vidro ser um produto mais vulnerável e de maior custo. A famácia é a indústria do medicamento, o que torna o consumidor mais propenso a gastar pela necessidade presente, e portanto, o preço do mel é mais elevado do que nos outros estabelecimentos. E o mel de origem em SP, também tem um preço superior em relação às demais origens. A inspeção (SIM) teve maior média, por ser a inspeção feita pelo município, os conumidores tem mais confiança na qualidade do produto. Para verificar a suposição de normalidade na variável preço do mel, foi utilizado o teste de Shapiro-Wilk e verificado que não segue uma distribuição normal ($p - valor < 0.001$). E portanto para verificar os possíveis fatores que possam influenciar na variabilidade do preço do mel foram utilizados os testes não-paramétricos de Wilcoxon-Mann-Whitney e de Kruskal Wallis.

Para avaliar se existe diferença do preço do mel entre os diferentes tipos de embalagens foi aplicado o teste de Wilcoxon-Mann-Whitney e de fato foi verificado que existe uma diferença significativa no preço do mel ($p - valor < 0.001$) entre o produto com embalagrm de vidro e de plástico (Tabela 4.3). Fazendo o mesmo para as categorias composição do mel e município de venda observa-se também uma diferença significativa entre suas categotias ($p - valor < 0.001$).

Ao aplicar o teste de Kruskal Wallis, para a variável preçoentre os tipos de estabelecimentos (famácia, supermercado, hortifruti, feira e produtos naturais), diferentes locais de origem (MG, SP, RJ, ES, SC, RN, PE, CE) e diferentes tipos de inspeção sanitária (Serviço de Inspeção Federal (SIF), Serviço de Inspeção Estadual (SIE), Relacionamento no Serviço Inspeção Estadual (SIE/ER), Serviço de Inspeção Estadual do Rio de Janeiro (SIE/RJ), Relacionamento no Serviço

de Inspeção Federal (SIF/ER), Serviço de Inspeção Municipal (SIM), não é inspecionado), conclui-se que existe pelo menos uma diferença significativa ($p - valor < 0.001$) entre os preços nestas categorias.

Tabela 4.1: Dados descritiva do preço do mel em reais (R$) por um grama

		Média	Mediana	Q_1	Q_3	CV(%)	Máximo	Mínimo	n
Embalagem	Vidro	0.02803	0.0252	0.0199	0.0297	50.49946	0.0899	0.0084	126
	Plástico	0.02419	0.0204	0.01298	0.0317	70.85572	0.01818	0.0019	228
Tipo de Estabelecimento	Farmácia	0.02823	0.0254	0.019	0.03448	52.53276	0.0899	0.0019	202
	Supermercado	0.02037	0.0182	0.0111	0.0266	52.28276	0.0648	0.0078	121
	Feira	0.0125	0.0111	0.0111	0.01205	23.616	0.019	0.0111	7
	Hortifruti	0.02537	0.024	0.0225	0.02755	20.41781	0.0311	0.021	3
	Produtos Naturais	0.03404	0.0246	0.0185	0.0317	109.10693	0.1818	0.01	21
Composição	Mel	0.02049	0.019	0.013	0.0248	51.53733	0.0736	0.0019	226
	Mel Composto	0.03448	0.02965	0.02293	0.0375	58.96171	0.1818	0.0073	126
Município	Angra	0.02676	0.024	0.01715	0.03167	55.26905	0.0899	0.0045	25
	Mangaratiba	0.02265	0.0195	0.0143	0.02785	83.92935	0.1818	0.0019	104
Origem	MG	0.02089	0.02	0.01422	0.0263	47.22222	0.0019	0.08 222	
	RJ	0.03196	0.03	0.0211	0.0375	46.58802	0.0068	0.0789	97
	SP	0.05022	0.0363	0.033	0.04942	77.69812	0.0078	0.1818	18
	ES	0.02185	0.0153	0.01395	0.0232	77.29977	0.0099	0.0232	4
	SC	0.034	0.034	0.034	0.034	—	0.034	0.034	1
	RN	0.02555	0.02555	0.0148	0.0148	2.49078	0.0251	0.026	2
	PE	0.0148	0.0148	0.0148	0.0148	—	0.0148	0.0148	1
	CE	0.0125	0.0125	0.0125	0.0125	—	0.0125	0.0125	1
Inspeção	SIF	0.02281	0.02075	0.0153	0.02912	51.42481	0.0019	0.08	242
	SIE	0.0469	0.0469	0.0469	0.0469	—	0.0469	0.0469	1
	SIE/ER	0.02548	0.02123	0.02123	0.0317	23.8226	0.0169	0.033	8
	SIE/RJ	0.03189	0.0311	0.0238	0.03975	47.50705	0.0068	0.0736	67
	SIE/ER	0.01783	0.01775	0.0166	0.0199	15.98429	0.0126	0.0211	8
	SIM	0.06	0.06	0.06	0.06	—	0.06	0.06	1
	Não é Inspecionado	0.0285	0.0285	0.0285	0.0285	—	0.0285	0.0285	1
Total		0.02555	0.022	0.016	0.0312	63.51109	0.0019	0.1818	354

Fonte: UFRRJ, 2007

Tabela 4.3: Testes não-paramétricos Wilcoxon Mann-Whitney e Kruskal Wallis com $\alpha = 5\%$

Teste não-paramétrico	Variáveis	p-valor
Wilcoxon Mann-Whitney	Embalagem	< 0.001
	Composição	< 0.001
	Municipio	< 0.001
Kruskal Wallis	Tipo de estabelecimento	< 0.001
	Inspeção	< 0.001
	Origem	< 0.001

Capítulo 5

Conclusão

Foram analisadas as variações de preço de mel e produtos, que estão relacionados com o consumo. A embalagem de vidro é mais cara do que a de plástico. Embora sejam ambos materiais recicláveis, o preço da grama de cada material se diferencia, sendo o vidro um material mais valorizado e portanto mais caro.

Nas lojas de produtos naturais e farmácias, o mel é vendido mais caro ($p - valor < 0.001$) do que nos demais tipos de estabelecimento, por serem estabelecimentos que vendem os produtos em menor quantidade e portanto compram do fornecedor uma quantidade pequena e consequentemente mais cara do que os outros estabelecimentos que compram em grande quantidade.

Em Angra dos Reis, o mel é vendido mais caro do que em Mangaratiba ($p - valor < 0.001$), devido ao padrão vida das pessoas que vivem neste municipio e também por ser uma cidade turística, na qual pessoas de poder aquisitivo elevado visitam com maior frequência.

O mel composto, se torna mais caro que o mel puro, pois geralmente são misturados à ele extratos de conotação terapeutica. Dessa forma o consumidor é mais atraído por ele, pois o utiliza como medicamento.

As principais vantagens da utilização dos testes não paramétricos são: simples; não dependem da distribuição da população da qual a amostra foi selecionada; não exigem que as populações originais sejam normalmente distribuídas; os dados não precisam ser quantitativos basta que tenham uma escala ordinal e o uso de postos faz as técnicas não-paramétricas menos sensíveis aos erros de medidas do que os testes tradicionais (testes paramétricos).

Referências Bibliográficas

[1] R Development Core Team (2007). **R: A language and environment for statistical computing. R Foundation for Statistical Computing.** Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[2] BUSSAB,W. O.; MORRETTIN,P. A. **Estatística Básica.** 5ª ed. São Paulo: Saraiva, 2002.

[3] SPIEGEL, M. R. **Estatística**. [Tradução: CONSENTINO, P.] (Coleção Schaum), São Paulo: Makron Books, 1993.

[4] CARVALHO, R.L. **Apresentação e Descrição dos Testes Paramétricos e Não-paramétricos Aplicados as Ciências Humanas e Sociais.** Monografia de Licenciatura em Matemática, UFRRJ, 2007.

[5] LEVIN, J.**Estatística Aplicada a Ciências Humanas** 2ª ed. São Paulo: Harbra, 1987.

[6] DANCEY, C. P.; REIDY, J. **Estatística sem Matemática para Psicologia: usando SPSS para Windows. [Tradução VIALI, L.].** 3ª ed. Porto Alegre: Artmed, 2006.

[7] SIEGEL, S.; CASTELLAN JR, N. J. **Estatística não-paramétrica para ciências do comportamento**; [Tradução: CARMONA, S. I. C.], 2ª ed. Porto Alegre: Artmed, 2006.

[8] STEPHENS, M. A.**FED para a Bondade de Estatística Fit e algumas comparações, Jornal da Associação Americana Estatística** Vol. 69, 730-737, 1974.

[9] DEVORE, J. L. **Probabilidade e Estatística: para Engenharia e Ciências. [Trad. SILVA, J. P. N.].** São Paulo: Pioneira Thomson Learning, 2006.

[10] SEBRAE. Desafios da Apicultura brasileira. **Revista SEBRAE Agronegócio** n.3 24-25, 2006.

[11] RIO BRANCO C. **Comercialização e Marketing de méis de abelhas na Região da Costa Verde, Rio de Janeiro, Sudeste do Brasil.** Artigo em andamento, 2008.

[12] FIBGE. **FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA.** Geografia do Brasil, Rio de Janeiro: FIBGE, 2000.

ANEXOS

ANEXO 1

```
banco = read.table('dado17042008.csv', sep="\t", header=T)

save.image('dadomel240408.RData')
load("dadomel240408.RData")

edit(banco) # Planilha do banco
attach(banco)
summary(preco1grama) # Sumário estatístico da variável
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00190 0.01600 0.02200 0.02555 0.03120 0.18180

sd(preco1grama)/mean(preco1grama)*100
[1] 0.6351109

# Testes de normalidade p/ nossa variável de interesse -
preço por 1 g de mel

# Primeira forma) Simulação da distribuição normal
# X ~ N(mi, S)
mean(preco1grama)
[1] 0.02555452

sd(preco1grama)
[1] 0.01622995


# Simulando a distribuição normal
x = rnorm(354, 0.02555452, 0.01622995)
mean(x)
[1] 0.02555452

sd(x)
[1] 0.01579409

boxplot(preco1grama, x, main="Variabilidade do Preço do Mel")

# Coeficiente de variação

% CV = sd(banco$preco1grama)/mean(banco$preco1grama) * 100

# Segunda forma) Utilizando o artifício gráfico - QQplot

qqnorm(preco1grama, main="Variabilidade do preço do Mel por uma grama")
```

```
## drawing the QQplot
qqnorm(x, main= "Variabilidade do Preço do Mel")

# Terceira forma através do teste shapiro.wilk

shapiro.test(preco1grama)
Shapiro-Wilk normality test

data:  preco1grama
W = 0.773, p-value < 2.2e-16

# Ho: A distribuição da variável pertence a uma dist. normal
# H1: A distribuição da variável não pertence a uma dist. normal

# p-valor < 0.05, rejeita-se a Ho


## Assumindo que iremos utilizar os testes não paramétricos
p/ as variáveis:

# Tipo de estabeleciemento - Tipoestab

banco$Tipoestab = factor(banco$Tipoestab, labels=c("Farmacia","Feira",
"Supermercado", "HortiFruti", "ProdutosNaturais"))

tapply(banco$preco1grama, banco$Tipoestab, summary)
# Média do preco em cada estabelecimento

 $Farmacia
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00190 0.01900 0.02540 0.02823 0.03448 0.08990

$Feira
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01110 0.01110 0.01110 0.01250 0.01205 0.01900

$Supermercado
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00780 0.01110 0.01820 0.02037 0.02660 0.06480

$HortiFruti
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.02100 0.02250 0.02400 0.02537 0.02755 0.03110

$ProdutosNaturais
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
0.01000 0.01850 0.02460 0.03404 0.03170 0.18180

tapply(banco$preco1grama, banco$Tipoestab, mean)

 Farmacia            Feira     Supermercado
      0.02823119       0.01250000       0.02037355

HortiFruti ProdutosNaturais
0.02536667       0.03403810

table(banco$Tipoestab)

        Farmacia            Feira     Supermercado
             202                7              121
HortiFruti ProdutosNaturais
    3               21

boxplot(banco$preco1grama ~ banco$Tipoestab,
main="Variabilidade do Preço do Mel entre os Estabeleciemntos",
ylab="Preço do Mel por Uma Grama em R$" )
abline(h=mean(banco$preco1grama), col="red") # Vericando o preco medio
geral do mel em relação aos estabelecimentos

tapply(banco$preco1grama, banco$Tipoestab, sd)
          Farmacia            Feira     Supermercado
     0.014839224      0.002952400      0.010657554
HortiFruti ProdutosNaturais
0.005186842      0.037140133

CV = sd(banco$Tipoestab)/mean(banco$Tipoestab) * 100

# Utilizando o kruskal wallis

# Ho: As médias entre as amostras são iguais
# H1: Pelo menos umas das médias é diferente

kruskal.test(banco$preco1grama ~ banco$Tipoestab)
Kruskal-Wallis rank sum test

data:  banco$preco1grama by banco$Tipoestab
Kruskal-Wallis chi-squared = 37.9691, df = 4, p-value = 1.137e-07


# Embalagem - embala
banco$embala = factor(banco$embala, labels=c("Vidro","Plastico"))
```

```
tapply(banco$preco1grama, banco$embala, mean) # Média do preco em cada
 Vidro   Plastico
0.02803254 0.02418509

table(banco$embala)  # Quantidade de produtos (n)

   Vidro Plastico
     126      228

boxplot(banco$preco1grama ~ banco$embala,
main="Variabilidade do Preço do Mel entre as Embalagens",
ylab="Preço do Mel por Uma Grama em R$" )

# Utilizando WILCOXON

# Ho: As médias entre as amostras são iguais
# H1: Existe diferença entre as médias das amostras
# p-valor < 0.05, rejeita-se Ho

wilcox.test(banco$preco1grama ~ banco$embala)
 Wilcoxon rank sum test with continuity correction

data:  banco$preco1grama by banco$embala
W = 17528, p-value = 0.0005998
alternative hypothesis: true location shift is not equal to 0

# Composição - composicao
banco$composicao = factor(banco$composicao, labels=c("Mel","Mel Composto"))

tapply(banco$preco1grama, banco$composicao, mean) # Média do preco em cada

 Mel Mel Composto
  0.02048761   0.03448413

tapply(banco$preco1grama, banco$composicao, sd)
         Mel Mel Composto
  0.01056828   0.02033484

tapply(banco$preco1grama, banco$composicao, summary)
$Mel
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00190 0.01300 0.01900 0.02049 0.02480 0.07360

$`Mel Composto`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00730 0.02293 0.02965 0.03448 0.03750 0.18180
```

```
table(banco$composicao)

         Mel Mel Composto
         226          126

boxplot(banco$preco1grama ~ banco$composicao,
main="Variabilidade do Preço do Mel em relação a Composição",
ylab="Preço do Mel por Uma Grama em R$" )

# Utilizando WILCOXON

# Ho: As médias entre as amostras são iguais
# H1: Existe diferença entre as médias das amostras
# p-valor < 0.05, rejeita-se Ho

> wilcox.test(banco$preco1grama ~ banco$composicao)

 Wilcoxon rank sum test with continuity correction

data:  banco$preco1grama by banco$composicao
W = 6041, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0


# Municipio - municip

banco$municip = factor(banco$municip, labels=c("Angra","Mangaratiba"))
tapply(banco$preco1grama, banco$municip, mean) # Média do preco em cada

Angra Mangaratiba
  0.0267628   0.0226500

 tapply(banco$preco1grama, banco$municip, sd)
      Angra Mangaratiba
 0.01479337  0.01901706

tapply(banco$preco1grama, banco$municip, summary)
$Angra
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00450 0.01715 0.02400 0.02676 0.03167 0.08990

$Mangaratiba
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00190 0.01430 0.01950 0.02265 0.02785 0.18180
```

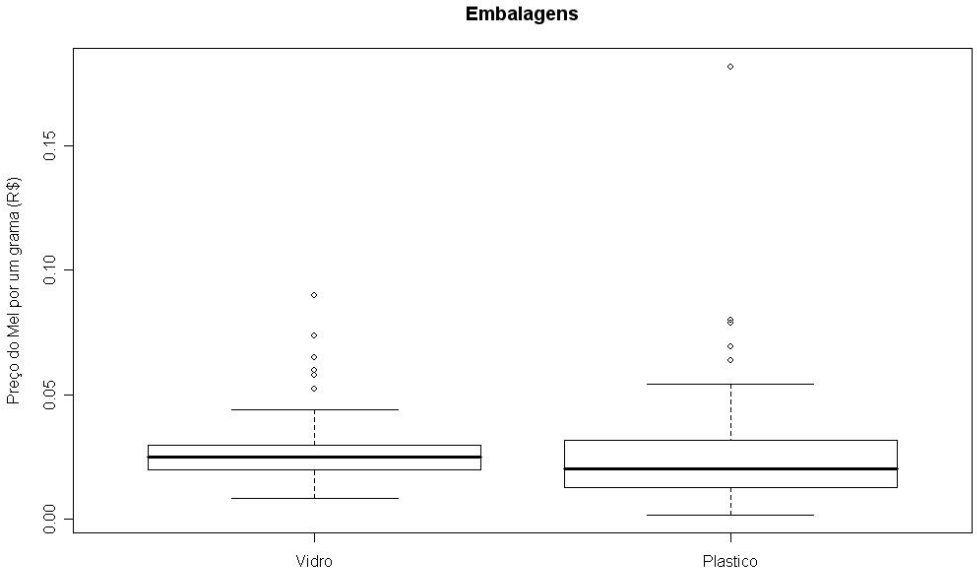
```
table(banco$municip)

      Angra Mangaratiba
        250         104

boxplot(banco$preco1grama ~ banco$municip,
main="Variabilidade do Preço do Mel nos Municipios",
ylab="Preço do Mel por Uma Grama em R$" )

# Utilizando WILCOXON

# Ho: As médias entre as amostras são iguais
# H1: Existe diferença entre as médias das amostras
# p-valor < 0.05, rejeita-se Ho

Wilcoxon rank sum test with continuity correction

data:  banco$preco1grama by banco$municip
W = 15969, p-value = 0.000712
alternative hypothesis: true location shift is not equal to 0


# Inspeção - inspecao

banco$inspecao = factor(banco$inspecao,
labels=c("Sif","Sie","sie/er","sie/rj","sif/er","visa sim",
"não é inspecionado"))
tapply(banco$preco1grama, banco$inspecao, mean) # Média do preco em cada

Sif                 Sie              sie/er              sie/rj
        0.02280826          0.04690000          0.02547500      0.03188657
sif/er            visa sim       não é inspecionado
         0.01782500          0.06000000          0.02850000

tapply(banco$preco1grama, banco$inspecao, summary)
$Sif
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00190 0.01530 0.02075 0.02281 0.02912 0.08000

$Sie
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0469  0.0469  0.0469  0.0469  0.0469  0.0469

$`sie/er`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01690 0.02123 0.02460 0.02548 0.03170 0.03300
```

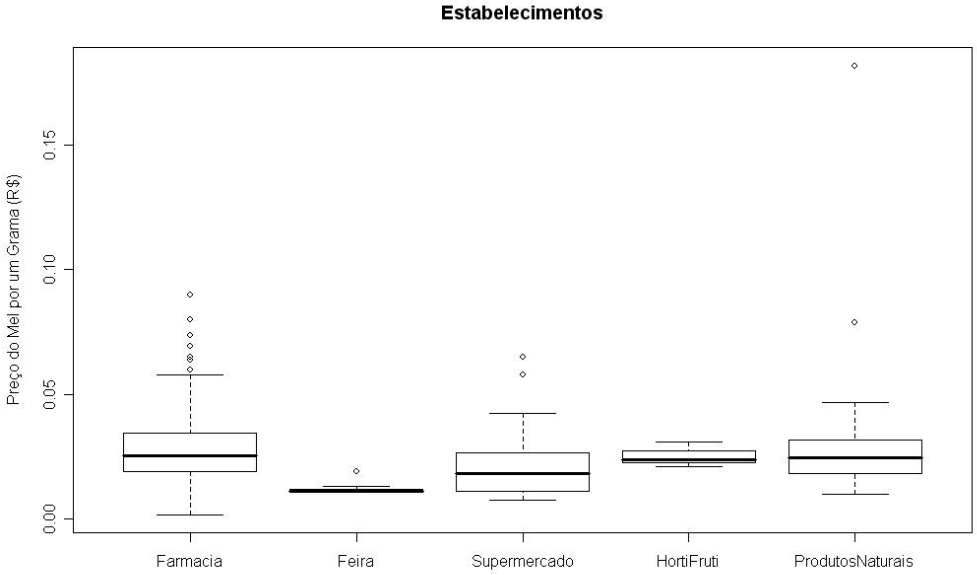
```
$‘sie/rj‘
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00680 0.02380 0.03110 0.03189 0.03975 0.07360

$‘sif/er‘
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01260 0.01660 0.01775 0.01783 0.01990 0.02110

$‘visa sim‘
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.06    0.06    0.06    0.06    0.06    0.06

$‘não é inspecionado‘
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0285  0.0285  0.0285  0.0285  0.0285  0.0285

table(banco$inspecao)

              Sif                Sie             sie/er          sie/rj
              242                  1                  8              67
 sif/er           visa sim        não é inspecionado
   8                  1                  1

boxplot(banco$preco1grama ~ banco$inspecao,
main="Variabilidade do Preço do Mel em relação a Inspeção",
ylab="Preço do Mel por Uma Grama em R$" )

tapply(banco$preco1grama, banco$inspecao, sd)

# Utilizando o kruskal wallis

# Ho: As médias entre as amostras são iguais
# H1: Pelo menos umas das médias é diferente

kruskal.test(banco$preco1grama ~ banco$inspecao)
data:  banco$preco1grama by banco$inspecao
Kruskal-Wallis chi-squared = 32.8212, df = 6, p-value = 1.135e-05

#Origem - origem

banco$origem = factor(banco$origem,
labels=c("MG","RJ","SP","ES","SC","RN","PE","CE"))
tapply(banco$preco1grama, banco$origem, summary)
$MG
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

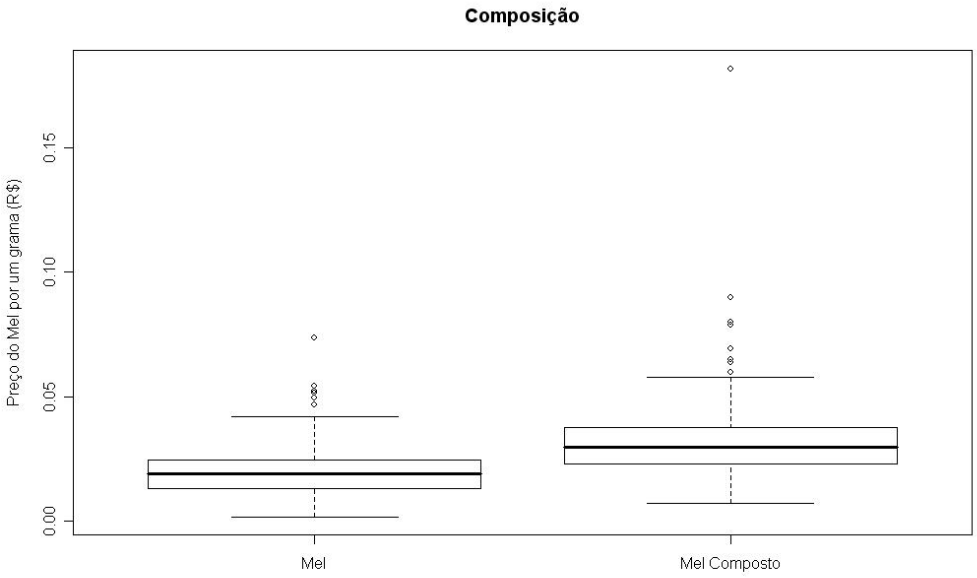
```
0.00190 0.01422 0.02000 0.02089 0.02630 0.08000

$RJ
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00680 0.02110 0.03000 0.03196 0.03750 0.07890

$SP
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00780 0.03300 0.03630 0.05022 0.04942 0.18180

$ES
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00990 0.01395 0.01530 0.02185 0.02320 0.04690

$SC
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.034   0.034   0.034   0.034   0.034   0.034

$RN
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.02510 0.02533 0.02555 0.02555 0.02577 0.02600

$PE
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0148  0.0148  0.0148  0.0148  0.0148  0.0148

$CE
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0125  0.0125  0.0125  0.0125  0.0125  0.0125

boxplot(banco$preco1grama ~ banco$origem,
main="Variabilidade do Preço do Mel nas Origens",
ylab="Preço do Mel por Uma Grama em R$" )

 table(banco$origem)

 MG  RJ  SP  ES  SC  RN  PE  CE
222  97  18   4   1   2   1   1

tapply(banco$preco1grama, banco$origem, mean)
        MG         RJ         SP         ES         SC         RN
0.02088559 0.03196186 0.05022222 0.02185000 0.03400000 0.02555000

PE         CE
0.01480000 0.01250000
```

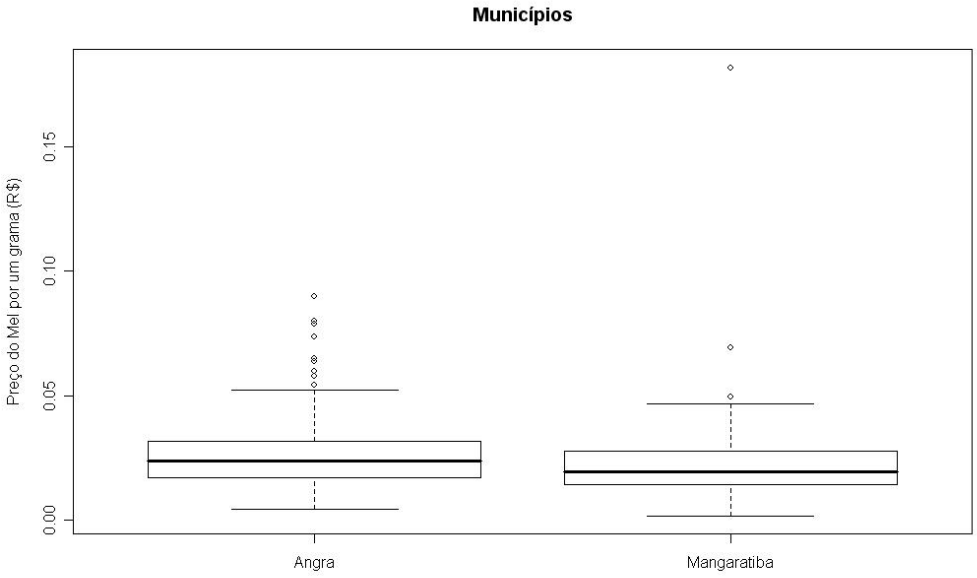
```
tapply(banco$preco1grama, banco$origem, sd)
          MG           RJ           SP           ES           CE
0.0098652262 0.0148934371 0.0390204706 0.0168928979           NA

SC                   RN           PE
0.0006363961          NA           NA

# Utilizando o kruskal wallis

# Ho: As médias entre as amostras são iguais
# H1: Pelo menos umas das médias é diferente

kruskal.test(banco$preco1grama ~ banco$origem)

 Kruskal-Wallis rank sum test

data:  banco$preco1grama by banco$origem
Kruskal-Wallis chi-squared = 71.1443, df = 7, p-value = 8.672e-13
```

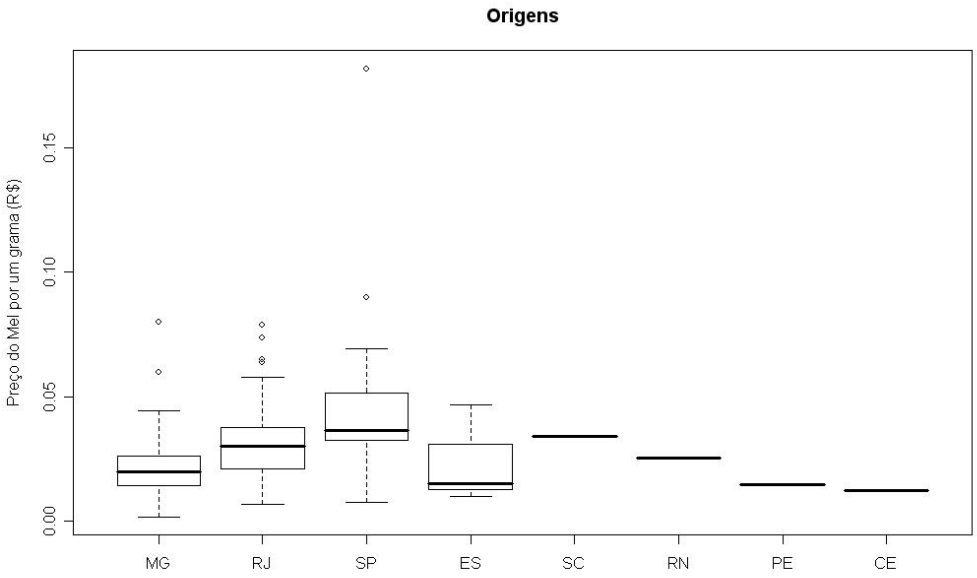
ANEXO 2: Variabilidade do Preço do Mel

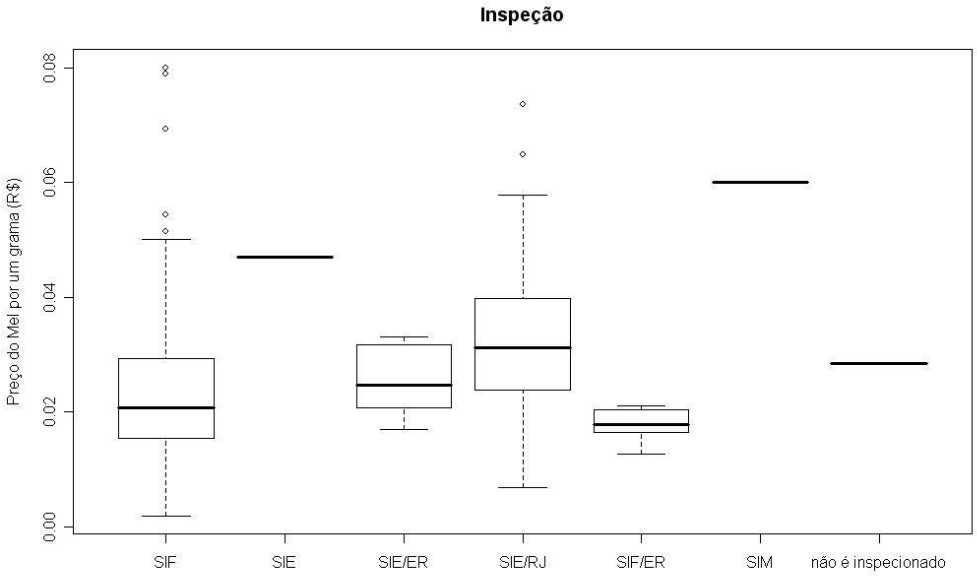





Composição
Preço do Mel por um grama (R$)
0.00
0.05
0.10
0.15
Mel
Mel Composto


Municípios
Preço do Mel por um grama (R$)
0.00
0.05
0.10
0.15
Angra
Mangaratiba


Origens
Preço do Mel por um grama (R$)
0.00
0.05
0.10
0.15
MG
RJ
SP
ES
SC
RN
PE
CE


Inspeção
Preço do Mel por um grama (R$)
0.00
0.02
0.04
0.06
0.08
SIF
SIE
SIE/ER
SIE/RJ
SIF/ER
SIM
não é inspecionado