Ferramenta estatística para análise de dados: comandos do software R

Daniele de Brito Trindade[1], Natália dos Santos Teixeira[1], Luzia Almeida Couto[2] & Jéssica Souza Coqueiro[3]

[1] Instituto Federal de Educação, Ciência e Tecnologia Baiano, Guanambi, Brasil

[2] Universidade Estadual de Santa Cruz, Ilhéus, Brasil

[3] Universidade Estadual do Sudoeste da Bahia, Itapetinga, Brasil

Correspondência: Luzia Almeida Couto, Universidade Estadual de Santa Cruz, Ilhéus, Brasil. E-mail: almeidacouto.luzia78@gmail.com

Recebido: Junho 17, 2022 Aceito: Julho 18, 2022 Publicado: Setembro 01, 2022

Resumo

Com o avanço tecnológico, as análises de dados se tornaram uma tarefa menos árdua. Porém, muitos pesquisadores, docentes e discentes apresentam dificuldades na análise e interpretação dos dados oriundos de suas atividades acadêmicas e pesquisas. Desta forma, o objetivo deste artigo é apresentar os principais comandos do software R para análise descritiva e exploratória, comparação de médias e validação dos pressupostos dos testes utilizados para analisar dados sensoriais oriundos do artigo proposto por Teixeira et al. (2020). Vale salientar que o objetivo deste artigo é apresentar os conceitos, de forma sucinta, dos testes utilizados no artigo de Teixeira et al. (2020) e os respectivos comandos do software R.

Palavras-chave: Medidas de posição; Medidas de dispersão; Teste para normalidade; Teste para homoscedasticidade; Teste de Kruskal-Wallis.

Abstract

With technological advancement, data analysis has become a less arduous task. However, many researchers, teachers and students have difficulties in the analysis and interpretation of data from their academic activities and research. Thus, the objective of this article is to present the main commands of the R software for descriptive and exploratory analysis, comparison of means and validation of the assumptions of the tests used to analyze sensory data from the article proposed by Teixeira et al. (2020). It is worth noting that the objective of this article is to present the concepts, in a succinct way, of the tests used in the article by Teixeira et al. (2020) and the respective R software commands.

Keywords: Position measures; Dispersion measures; Test for normality; Test for homoscedasticity; Kruskal-Wallis test.

Resumen

Con el avance tecnológico, el análisis de datos se ha convertido en una tarea menos ardua. Sin embargo, muchos investigadores, profesores y estudiantes tienen dificultades para analizar e interpretar los datos de sus actividades académicas y de investigación. Así, el objetivo de este artículo es presentar los principales comandos del software R para análisis descriptivo y exploratorio, comparación de medias y validación de los supuestos de las pruebas utilizadas para analizar datos sensoriales del artículo propuesto por Teixeira et al. (2020). Vale la pena mencionar que el objetivo de este artículo es presentar los conceptos, de forma sucinta, de las pruebas utilizadas en el artículo de Teixeira et al. (2020) y los respectivos comandos del software R.

Palabras clave: Mediciones de posición; medidas de dispersión; Prueba de normalidad; Prueba de homocedasticidad; Prueba de Kruskal-Wallis.

1. Introdução

Segundo a Associação Brasileira de Normas Técnicas (ABNT, 1993), a análise sensorial é definida como uma disciplina científica usada para evocar, medir, analisar e interpretar reações a características de alimentos e materiais percebidas pelos sentidos (visão, olfato, gosto, tato e audição).

Normalmente a análise sensorial é realizada por uma equipe montada para avaliar as características sensoriais de um produto para um determinado fim. Pode-se avaliar a seleção da matéria-prima que será utilizada em um novo produto, a textura, o sabor, a reação do consumidor, entre outros parâmetros que levarão o produto a ser aceito ou rejeitado pelo mercado, podendo este ser remodelado para agradar ao público alvo e obter êxito. Para alcançar o objetivo específico de cada análise, são utilizados métodos de avaliação diferenciados, visando a obtenção de respostas mais adequadas, ou seja, os métodos se moldam ao objetivo da análise. Finalizada a análise serão obtidos dados que devem ser transformados em resultados, segundo o teste sensorial aplicado, sendo estatisticamente avaliado concluindo assim a viabilidade do produto (Teixeira, 2009).

Existem diversas metodologias estatísticas que podem ser utilizadas para avaliar os resultados oriundos da análise sensorial. A análise de variância (ANOVA) é a técnica comumente utilizada. A análise de componentes principais (ACP) e a correlação entre as notas de cada julgador e as notas médias de todos os julgadores também são metodologias usadas para análise de dados sensoriais (Silva, Azevedo, 2009; Silva, Azevedo, 2006).

Vale salientar que, quando bem conduzidos, os experimentos sensoriais são extremamente eficientes e vão de encontro às necessidades do consumidor, subentendendo-se que o produto advindo da indústria alimentícia possui características sensoriais por ele desejadas. Além disso, a boa conduta pode contribuir no auxílio do suporte técnico em pesquisas, marketing e controle de qualidade do item (Camocardi, Ferreira, 2020; Rossini, Anzanello & Fogliatto, 2012).

Porém, os estudos com análise sensorial possuem entraves que merecem atenção, como o limitado escopo do vocabulário associado a difícil tarefa de traduzir percepções ou sensações, limitando a definição da qualidade sensorial do alimento, além das dificuldades em transformar os dados em resultados concretos.

Como dito anteriormente, a ANOVA é a ferramenta mais utilizada para interpretar os dados. Entretanto, nem sempre as notas atribuídas pelos julgadores atendem às exigências desta metodologia. Desta forma, há a necessidade de aplicação de um teste não paramétrico, denominado por teste de *Kruskal-Wallis*, para avaliar a igualdade das médias dos tratamentos estudados (Moraes, 1993; Alcântara, Freitas-Sá, 2018).

É importante que os dados estatísticos sejam interpretados de forma correta, visto o peso excessivo que estes possuem no estudo, ou seja, os cálculos estatísticos serão utilizados como provas irrefutáveis de conclusões discutíveis, sendo fato decisivo na pesquisa por se tratar da confiabilidade de todo aquele processo (Conceição, 2008).

Para facilitar esse processo de análise de dados, nos últimos anos, vários programas computacionais vêm sendo desenvolvidos, com o intuito de facilitar as análises das aplicações, apresentar uma eficiência computacional e fornecer resultados exatos e precisos (Sousa, 2000).

O uso de pacotes estatísticos para a análise de dados é de grande importância no que se refere à análise e a interpretação de resultados. Contudo, alguns programas apresentam um custo de aquisição relativamente elevado. Dentre os *softwares* de domínio público (sem custo financeiro) que podem ser utilizados para análise de dados em geral, encontra-se o Ambiente R (R Core Team, 2015).

O *software* R apresenta código fonte aberto, podendo ser modificado ou implementado com novos procedimentos desenvolvidos por qualquer usuário a qualquer momento. Além do que, o *software* R possui um grande número de colaboradores das mais diversas áreas do conhecimento (Venables, Smith, 2005)

O programa R é bastante difundido como um *software* estatístico, mas realiza várias outras tarefas, como por exemplo, análises de bioinformática, incluindo alinhamento de sequências e comparação com bases de dados, modelagem, produção de mapas e muitos outros.

De um modo geral, o *software* R é uma ferramenta excelente para armazenar e manipular dados, realizar cálculos, realizar testes estatísticos, análises exploratórias e produzir gráficos (Ritter, They & Konzen, 2019). Dessa forma, os dados sensoriais podem ser avaliados por meio deste programa facilitando a obtenção de resultados.

Pensando em simplificar o acesso aos códigos implementados no R, o presente artigo teve como objetivo apresentar os principais comandos para análise descritiva e exploratória, comparação de médias e validação dos pressupostos dos testes utilizados para analisar dados sensoriais oriundos do artigo proposto por Teixeira et al. (2020).

2. Referencial Teórico

2.1 Categorização das variáveis

Na Estatística é possível classificar as variáveis em qualitativas ou quantitativas. As variáveis qualitativas são oriundas de atributos ou qualidades e podem ser categorizadas como nominal ou ordinal. Já as variáveis quantitativas são observações relacionadas à mensuração ou contagem, sendo divididas em discretas ou contínuas (Morettin, Bussab, 2004).

Em diversos estudos, as variáveis quantitativas podem ser categorizadas, ou seja, é possível transformar uma variável quantitativa em qualitativa construindo duas ou mais categorias, com o objetivo de criar grupos mais homogêneos e facilitar a interpretação das respostas (Paulino, Da Motta Singer, 2006).

2.2 Testes paramétricos

2.2.1 Análise de Variância (ANOVA)

A análise de variância (ANOVA) é amplamente utilizada como ferramenta para expressar a variabilidade máxima que um conjunto de dados possui, a partir de uma soma de termos (Carpinetti, 2009). Sendo assim, por meio da utilização desse teste, é possível avaliar três ou mais grupos diferentes, presentes em um mesmo estudo, verificando a igualdade estatística entre as suas médias (Garcia-Marques, 1997).

Assim sendo, as hipóteses testadas são $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$, em que as médias são iguais nos *k* tratamentos e $H_1: \mu_i \neq \mu_{i'}$, com $i \neq i' = 1, \ldots, k$, ou seja, ao menos uma média difere das demais (Morettin, Bussab, 2004; Barbetta et al., 2010), em que H_0 e H_1 são, simultaneamente, as hipóteses nula e alternativa. A hipótese nula é rejeitada se o *p*-valor do teste for menor que o nível de significância (α) definido previamente no planejamento do estudo.

Antes de aplicar a ANOVA como método para o tratamento dos dados brutos é preciso atentar-se aos seus pressupostos, sendo os principais a normalidade e a homoscedasticidade, que são verificados sobre os resíduos dos dados, a fim de evitar a ocorrência de erros (Da Rocha et al., 2018).

2.2.2 Testes para normalidade

O pressuposto da normalidade indica que a distribuição dos dados deve ser normal, tendo como principal característica a curva em formato de sino quando demonstrada graficamente (Leotti, Coster & Riboldi, 2012). Esse pressuposto é comumente verificado pelo teste de *Shapiro-Wilk* (Shapiro & Wilk, 1965). Neste artigo, serão apresentados dois testes, também utilizados com frequência, denotados por Teste de *Anderson-Darling* (Anderson, Darling, 1954) e Teste de *Lilliefors* (Lilliefors, 1967; Lilliefors, 1969; Dallal, Wilkinson, 1986).

2.2.3 Teste de Anderson-Darling

O teste de *Anderson-Darling* é empregado para verificar se determinada amostra segue uma distribuição normal (H_0), ou não (H_1). Este teste utiliza observações reais sem agrupamento, além disso, é sensível ao surgimento de discrepâncias nas caudas da distribuição. Desse modo, suponha que uma amostra tenha sido retirada de uma definida população com função de distribuição acumulada (FDA) normal $F(x)$. A estatística de teste é dada por

$$A^2 = n \int_{-\infty}^{\infty} \frac{[F_n(x) - F(x)]}{F(x)(1 - F(x))} dF(x).$$

Outra forma mais simplificada de escrever A^2 é:

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \left[(2i - 1) \ln\left(F\left(x_{(i)}\right)\right) + (2(n - i) + 1) \ln\left(1 - F\left(x_{(i)}\right)\right) \right].$$

Assim, a transformação $F\left(x_{(i)}\right)$ leva $x_{(i)}$ em $U_{(i)}$, de uma amostra de tamanho *n* com distribuição Uniforme (0,1). Assim $A^2 = -n - \frac{1}{n} D$, em que $D = \sum_{i=1}^{n} \left[(2i - 1) \ln\left(U_{(i)}\right) + (2(n - i) + 1) \ln\left(1 - U_{(i)}\right) \right]$. Verifica-se a rejeição do teste por meio do *p*-valor, ou seja, se *p*-valor < α, rejeita H_0.

2.2.4 Teste de Lilliefors

O teste de *Lilliefors* é usual para averiguar a adesão dos dados a uma distribuição normal sem que haja a necessidade de especificar seus parâmetros (Barbetta et al., 2010). Este teste se assemelha ao teste de aderência de *Kolmogorov-Smirnov* (Kolmogorov, 1933; Smirnov, 1948), visto que são medidas as distribuições acumuladas

, $$S(x) = \frac{número\ de\ valores \leq x_i}{n}$$

em que *n* é o número de elementos da amostra e x_i é um valor qualquer da amostra e $F(x)$ que é a função de distribuição acumulada. Calcula-se a distância máxima *D* dentre elas e se compara com um valor tabelado, em função do nível de significância e do tamanho da amostra. A regra de decisão é baseada no nível descritivo do teste, ou seja, se *p*-valor < α, rejeita-se H_0, em que α é o nível de significância do teste.

2.3 Testes para homoscedasticidade

A homoscedasticidade avalia se os resíduos dos dados possuem variância mínima entre si, de forma que são, estatisticamente, semelhantes. Caso os resíduos apresentem variância entre si, a utilização da ANOVA torna-se inadequada, pois ocorrerá alteração no resultado do cálculo que fornecerá os mínimos quadrados (Royston, 1982). Um dos testes para verificação da homoscedasticidade é o teste de *Bartlett* (Bartlett, 1937). Porém, nesse presente artigo serão apresentados mais dois testes denotados por Teste de *Levene* (Levene, 1960) e Teste de *Fligner-Killeen* (ver Conover et al., 1981).

2.3.1 Teste de Bartlett

Um dos pressupostos para a validação dos resultados da ANOVA é a de homoscedasticidade, ou seja, que as variâncias sejam, estatisticamente, iguais em todos os *k* tratamentos (Morettin, Bussab, 2004). Um teste frequentemente aplicado na literatura para averiguar tal pressuposto é o teste de *Bartlett*. Aqui, as hipóteses são:

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 \text{ para todo par } i = 1, \dots, k;$$

$$H_1: \sigma_i^2 \neq \sigma_j^2 \text{ para algum } i, j = 1, \dots, k.$$

Para iniciar o teste, as seguintes informações são necessárias: as dimensões das amostras (n_i) e as variâncias amostrais (S_i^2), em que $i = 1, \dots, k$ e $n = n_1 + n_2 + \cdots + n_k$. Para a definição do teste é imprescindível adotar os passos que serão expostos a seguir. Primeiro, calcula-se a variância comum através da equação

$$S_c^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{n-k} = \frac{SQDen}{n-k} = QMDen.$$

Depois, são calculadas as equações $M = (n - l)\ln(S_c^2) - \sum_{i=1}^{l}(n_i - 1)\ln(S_i^2)$ e

$$C = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\left(\frac{1}{n_i - 1}\right) - \left(\frac{1}{n-k}\right)\right].$$

Por fim necessita-se definir a estatística que $\frac{M}{C}$ segue uma distribuição *qui*-quadrado com ($gl = k - 1$) graus de liberdade para amostras grandes (Morettin & Bussab, 2004).

Rejeita-se H_0 se $\frac{M}{C} \geq \chi_c^2$ em que χ_c^2 é o valor crítico encontrado na tabela *qui*-quadrado com $gl = k - 1$ e nível de significância $\alpha\%$ ou se *p*-valor < α.

2.3.2 Teste de Levene

O teste de *Levene* é mais robusto para desvios da normalidade do que o teste de *Bartlett*. Aqui, as hipóteses testadas são: $H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ para todo par $i = 1, \dots, k$ e $H_1: \sigma_i^2 \neq \sigma_j^2$ para algum $i, j = 1, \dots, k$. Este teste consiste em transformar os dados originais e aplicar o teste da ANOVA aos dados transformados. A transformação proposta por Levene (1960) é

$$z_{ij} = |x_{ij} - \bar{x}_{i.}|, i = 1, \dots, k \text{ e } j = 1, \dots, n_i$$

em que z_{ij} denota os dados transformados, x_{ij} representa os dados originais e $\bar{x}_{i.}$ define a média do *i*-ésimo tratamento, em relação aos dados originais. A estatística de teste é dada por

$$F^* = \frac{\frac{\sum_{i=1}^{k} \frac{n_i(\bar{z}_{i.} - \bar{z}_{..})^2}{k-1}}{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_{i.})^2}{\sum_{i=1}^{k} n_i - k}}}$$

em que $\bar{z}_{i.} = \sum_{i=1}^{k} \frac{z_{ij}}{n_i}$ e $\bar{z}_{..} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} z_{ij}}{\sum_{i=1}^{k} n_i}$

Após a transformação, aplica-se o teste da ANOVA. Rejeita-se H_0 se F for significativo ou se *p*-valor < α.

2.3.3 Teste de Fligner-Killeen

O teste de *Fligner-Killeen* é um teste não paramétrico muito robusto utilizado quando o pressuposto de normalidade não for satisfeito. Ele verifica se as variâncias são as mesmas em cada uma das amostras. Este teste, baseado na mediana, foi determinado em um estudo de simulação e para maiores detalhes ver Conover et al. (1981).

Ele é definido por uma regressão linear simples de *k* amostras que usa a classificação dos valores absolutos das amostras centradas e pondera em que *qnorm* é o quantil da normal.

$$a_{(i)} = qnorm\left(1 + \frac{i}{(n+1)}\right)$$

Rejeita-se a hipótese nula, se o *p*-valor for menor que o nível de significância do teste (α).

2.4 Testes não paramétricos

2.4.1 Teste de Kruskal-Wallis

O teste não paramétrico de *Kruskal-Wallis* é aplicado quando o objetivo é a comparação de três ou mais grupos independentes (Kruskal & Wallis, 1952; Neto, Stein, 2003). Ele testa a hipótese nula (H_0) de que todas as populações possuem funções de distribuição iguais contra a hipótese alternativa (H_1) de que ao menos uma das populações possuem funções de distribuição diferentes.

Considere *k* amostras de tamanho $N_1, N_2, \dots, N_k$. Além disso, suponha que os dados do conjunto de todas as amostras são atributos ou postos e as somas dos postos das *k* amostras são $R_1, R_2, \dots, R_k$, respectivamente. Se a estatística é definida como

$$H = \frac{12}{N(N+1)}\sum_{j=1}^{k}\frac{N_j^2}{Nj} - 3(N+1),$$

então, é possível mostrar que a distribuição amostral de H é muito próxima de uma distribuição *qui*-quadrado (χ^2) com $gl = k - 1$ graus de liberdade, desde que $N_1, N_2, \dots, N_k$, sejam todos pelo menos iguais a 5 (Speigel, Stephens, 2009).

O teste de *Kruskal-Wallis* é o análogo ao teste *F* utilizado na ANOVA com um fator. Enquanto o teste da ANOVA necessita das hipóteses de que todas as populações em confronto são independentes e normalmente distribuídas, o teste de *Kruskal-Wallis* não coloca nenhuma restrição sobre a comparação (Dunn, 1961; Dunn, 1964). Como critério de seleção, rejeita-se H_0 se a estatística de teste H for significativa ou se *p*-valor < α.

3. Material e Métodos

O conjunto de dados utilizado para a construção dos códigos apresentados neste trabalho é oriundo do artigo intitulado “Elaboração, análise descritiva e análise sensorial de biscoito nutritivo com polpa de maracujá

(*Passiflora edulis*)" (Teixeira et al., 2020). Os dados são referentes às notas de 40 provadores, não treinados e escolhidos ao acaso, sobre os atributos (cor, sabor, aroma e textura) e a notageral de três tipos de biscoitos (Biscoito 1 – biscoito de trigo e maracujá; Biscoito 2 – biscoito de polvilho e maracujá; Biscoito 3 – biscoito de polvilho e maracujá sem adição de manteiga ou outro tipo de gordura), de um experimento em Delineamento Inteiramente Casualizado (DIC). Para realização do presente estudo utilizou-se o *software* R, versão 3.5.0 (R Core Team, 2015), cujos resultados e discussão são detalhados em Teixeira et al. (2020).

4. Resultados e Discussão

Os comandos utilizados para analisar os dados considerando a variável **Cor** serão apresentados a seguir. Inicialmente, é necessário inserir os dados no *software* R e os comandos são apresentados abaixo. O objeto **Cor** guarda as notas dos 40 provadores para este atributo, considerando os três tipos de biscoitos estudados. As notas são organizadas por provador, ou seja, as três primeiras notas são referentes às notas do primeiro provador para os três tipos de biscoitos (Biscoito 1, Biscoito 2 e Biscoito 3) .

Cor = c(5,9,5,7,7,4,7,8,6,6,9,4,8,9,5,6,7,3,7,8,6,5,4,7,4,7,5,4,8,3,6,9,4,8,7,6,5,8,7,7,
7,7,5,7,8,8,9,2,6,9,3,7,8,5,9,8,5,5,9,5,3,7,6,7,7,4,8,8,3,6,9,4, 4,5,4,4,8,2,7,9,
5,5,8,5,6,9,3,5,9,4,7,8,3,6,7,6,8,8,7,5,9,5,3,8,8,4,6,3,3,7,4, 6,8,5, 5,9,7, 4,9,6)

Os objetos **Tipo** e **Provador** são criados para adicionar os tratamentos e a identificação dos provadores, respectivamente. Desta forma, os comandos são dados por:

Tipo = factor(rep(c("Biscoito 1", "Biscoito 2", "Biscoito 3"), 40))

Provador = factor(rep(1:40, rep(3, 40)))

O banco de dados é montado utilizando o comando **data.frame( )**. Dentro do **( )** são colocados os objetos que estruturam um DIC, ou seja, a variável em análise (**Cor**), o tratamento (**Tipo**) e as repetições (**Provador)**, como mostra o comando abaixo, indicado pela Figura 1. Além disso, o objeto que guarda o banco de dados é denotado por **dados_C**. Assim, o comando é:

dados_C = data.frame(Cor, Tipo, Provador); dados_C

```
>
> dados_C = data.frame(Cor, Tipo, Provador); dados_C
   Cor       Tipo Provador
1    5 Biscoito 1        1
2    9 Biscoito 2        1
3    5 Biscoito 3        1
4    7 Biscoito 1        2
5    7 Biscoito 2        2
6    4 Biscoito 3        2
7    7 Biscoito 1        3
8    8 Biscoito 2        3
9    6 Biscoito 3        3
10   6 Biscoito 1        4
```

Figura 1. Banco de Dados inserido no *software* R.

Fonte: Autoras, 2022.

O comando **attach( )** é usado para tornar o código mais limpo. Desta forma, para chamar as variáveis em estudo basta escrever os seus respectivos nomes. Caso o comando **attach(dados_C)** não seja usado, a variável **Cor** deve ser chamada no banco pelo comando **dados_C$Cor**. O comando para calcular as principais medidas descritivas (mínimo, 1º quartil, mediana, média, 3º quartil e máximo) é o **tapply(Cor, Tipo, summary)**, apresentado na Figura 2.

```
> attach(dados_C)
The following objects are masked _by_ .GlobalEnv:

    Cor, Provador, Tipo

>
> ###### medidas descritivas
> tapply(Cor, Tipo, summary)
$`Biscoito 1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   5.000   6.000   5.775   7.000   9.000

$`Biscoito 2`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.000   7.000   8.000   7.875   9.000   9.000

$`Biscoito 3`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.00    4.00    5.00    4.85    6.00    8.00
```

Figura 2. Saída do R com as Medidas Descritivas.

Fonte: Autoras, 2022.

Os comandos guardados nos objetos abaixo calculam, respectivamente, o desvio padrão (**sd_Cor**) e o coeficiente de variação (**CV_Cor**), que são medidas para avaliação a variabilidade dos dados, indicado pela Figura 3. Aqui, o objeto **mean_Cor** refere-se às médias por tratamento.

sd_Cor = tapply(Cor, Tipo, sd); sd_Cor

mean_Cor = tapply(dados_C$Cor, dados_C$Tipo, mean)

CV_Cor = 100*sd_Cor / mean_Cor; CV_Cor

```
>
> sd_Cor = tapply(Cor, Tipo, sd); sd_Cor
Biscoito 1 Biscoito 2 Biscoito 3
  1.560531   1.158857   1.577892
> mean_Cor = tapply(Cor, Tipo, mean)
>
> CV_Cor = 100*sd_Cor/mean_Cor; CV_Cor
Biscoito 1 Biscoito 2 Biscoito 3
  27.02218   14.71564   32.53386
```

Figura 3. Saída do R com os Desvios Padrão e os Coeficientes de Variação Considerando os Três Tratamentos.

Fonte: Autoras, 2022.

A Análise de Variância (ANOVA) é um teste utilizado para comparar a magnitude da variabilidade observada dentro das *k* amostras com uma medida da variabilidade entre as médias das *k* amostras. A ANOVA é calculada no *software* R utilizando os comandos **anova( )** ou **aov( )** e as saídas são apresentadas nas Figuras 4 e 5, respectivamente. Vale salientar que, no primeiro comando é ajustado o modelo de regressão linear através do comando **lm( )** antes de realizar o teste da ANOVA. Além disso, a variável **Cor** está sendo ajustada em relação às variáveis **Tipo** e **Provador**.

ajuste_C = lm(data = dados_C, Cor ~ .)

anova(ajuste_C)

```
> ajuste_C = lm(data = dados_C, Cor ~ .)
> anova(ajuste_C)
Analysis of Variance Table

Response: Cor
          Df  Sum Sq Mean Sq F value    Pr(>F)
Tipo       2 192.217  96.108 44.8576 1.08e-13 ***
Provador  39  77.333   1.983  0.9255    0.5971
Residuals 78 167.117   2.143
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Figura 4. Ajuste do Modelo de Regressão Linear e Análise de Variância.

Fonte: Autoras, 2022.

mod_C = aov(data = dados_C, Cor ~ Tipo)

summary(mod_C)

```
> mod_C = aov(data = dados_C, Cor ~ Tipo)
> summary(mod_C)
             Df Sum Sq Mean Sq F value   Pr(>F)
Tipo          2  192.2   96.11      46 1.82e-15 ***
Residuals   117  244.4    2.09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
`
```

Figura 5. Ajuste da Análise de Variância.

Fonte: Autoras, 2022.

Os pressupostos de normalidade e homoscedasticidade são necessários para validar o resultado da ANOVA. O primeiro passo é obter os resíduos do modelo da ANOVA utilizando o comando **residuo = mod_C$res**. Os testes de normalidade estão disponíveis na biblioteca **nordest**, desta forma é necessário o comando **library(nortest)**. Vale salientar que a biblioteca **nordest** precisa ser instalada antes de ser utilizada e o comando utilizado para a instalação de qualquer *library* é **install.packages( )**. Os testes para avaliar a normalidade dos dados de *Anderson-Darling* e *Lilliefors* são realizados pelos comandos **ad.test(residuo)** e **lillie.test(residuo)**, respectivamente.

```
> library(nortest)
>
> ad.test(mod_C$res)

        Anderson-Darling normality test

data:  mod_C$res
A = 1.8976, p-value = 7.182e-05

>
> lillie.test(mod_C$res)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  mod_C$res
D = 0.13475, p-value = 1.395e-05
```

Figura 6. Teste Para Verificação da Normalidade dos Dados.

Fonte: Autoras, 2022.

Os testes de homoscedasticidade de *Bartlett* é realizado pelo comando **bartlett.test(residuo ~ Tipo)**. Para realizar os testes de *Levene* e *Fligner-Killeen* é necessário o uso da biblioteca **car** utilizando o comando **library(car)**. Os comandos para os respectivos testes são **leveneTest(residuo ~ Tipo)** e **fligner.test(residuo ~ Tipo).**

```
> residuo = mod_C$res
>
> bartlett.test(residuo ~ Tipo)

        Bartlett test of homogeneity of variances

data:  residuo by Tipo
Bartlett's K-squared = 4.3732, df = 2, p-value = 0.1123

>
>
> library(car)
Carregando pacotes exigidos: carData
> leveneTest(residuo ~ Tipo)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   2  3.2099 0.04394 *
      117
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> fligner.test(residuo ~ Tipo)

        Fligner-Killeen test of homogeneity of variances

data:  residuo by Tipo
Fligner-Killeen:med chi-squared = 8.8475, df = 2, p-value = 0.01199
```

Figura 7. Teste para Verificação da Homoscedasticidade dos Dados.

Fonte: Autoras, 2022.

Os resultados apresentados em Teixeira et al. (2020), demostraram que o usual teste de *Tukey* não é indicado para comparar as médias, uma vez que o pressuposto de normalidade não foi satisfeito. Diante disso, uma alternativa é aplicar o teste não paramétrico de *Kruskal-Wallis*.

O *software* R disponibiliza o teste em três bibliotecas que precisam ser instaladas através dos comandos **install.packages("agricolae")**, **install.packages("pgirmess")** e **install.packages("dunn.test").** Em um conjunto de dados que não segue uma distribuição normal é necessário realizar uma análise não-paramétrica, neste caso, o teste de Kruskal-wallis, seguido do teste de Dunn. O pacote "agricolae" não tem o teste de Dunn como parte do repertório de suas funções. Logo, a função do teste de Dunn é adquirida por meio do pacote "dunn.test", que executa o teste de Dunn oferecendo como parte dos resultados o teste de Kruskal-Wallis.

Para chamar o pacote **agricolae**, utiliza-se o comando **library(agricolae)**. O teste de *Kruskal-Wallis*, disponível neste pacote, é realizado pelo comando **k_C = kruskal(Cor, Tipo); k_C**, e a saída é apresentada na Figura 8.

```
> library(agricolae)
Warning message:
package 'agricolae' was built under R version 3.5.1
>
> k_C = kruskal(Cor, Tipo);k_C
$`statistics`
     Chisq Df      p.chisq  t.value     MSD
  54.10123  2 1.786793e-12 1.980448 11.3419

$parameters
           test p.ajusted name.t ntr alpha
  Kruskal-Wallis      none   Tipo   3  0.05

$means
             Cor    rank      std  r Min Max Q25 Q50 Q75
Biscoito 1 5.775 52.5625 1.560531 40   3   9   5   6   7
Biscoito 2 7.875 91.9000 1.158857 40   4   9   7   8   9
Biscoito 3 4.850 37.0375 1.577892 40   2   8   4   5   6

$comparison
NULL

$groups
```

Figura 8. Teste não Paramétrico de *Kruskal-Wallis* Utilizando o Pacote **agricolae**.

Fonte: Autoras, 2022.

O gráfico com a comparação das médias utilizando o teste de *Kruskal-Wallis* (Gráfico 1) do pacote **agricolae** é construído através do comando **plot(k_C, main = " ")**.

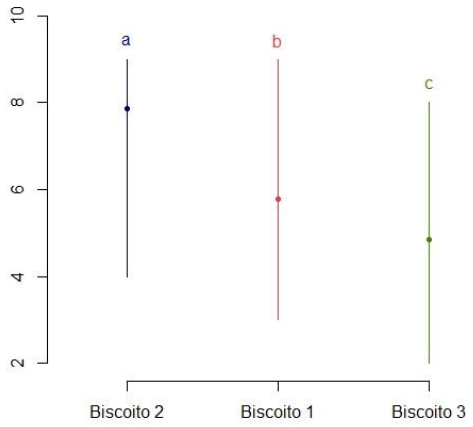

Gráfico 1. Comparação das Médias entre os Três Tipos de Biscoito, em Relação ao Atributo Cor.

Fonte: Autoras, 2022.

Para chamar o pacote **pgirmess**, utiliza-se o comando **library(pgirmess)**. O teste de *Kruskal-Wallis*, disponível neste pacote, é realizado pelo comando **k_C_2 = kruskalmc(Cor, Tipo); k_C_2** (Figura 9).

```
attr(,"class")
[1] "group"
> library(pgirmess)
>
> k_C_2 = kruskalmc(Cor, Tipo); k_C_2
Multiple comparison test after Kruskal-Wallis
p.value: 0.05
Comparisons
                      obs.dif critical.dif difference
Biscoito 1-Biscoito 2 39.3375     18.62079       TRUE
Biscoito 1-Biscoito 3 15.5250     18.62079      FALSE
Biscoito 2-Biscoito 3 54.8625     18.62079       TRUE
>
```

Figura 9. Teste de *Kruskal-Wallis* Utilizando o Pacote **pgirmess.**

Fonte: Autoras, 2022.

O pacote **dunn.test** é acionado pelo comando **library(dunn.test)** e o comando **k_C_3 = dunn.test(Cor, Tipo); summary(k_C_3)** realiza do teste de *Kruskal-Wallis* (Figura 10).

```
> library(dunn.test)
> k_C_3 = dunn.test(Cor, Tipo); summary(k_C_3)
  Kruskal-Wallis rank sum test

data: Cor and Tipo
Kruskal-Wallis chi-squared = 54.1012, df = 2, p-value = 0
```

Figura 10. Teste de *Kruskal-Wallis* Utilizando o Pacote **dunn.test.**

Fonte: Autoras, 2022.

No artigo Teixeira et al. (2020), as notas (X) dos atributos foram categorizadas em ruim ($0 \leq X < 3$), bom ($3 \leq X < 6$), ótimo ($6 \leq X < 9$) e excelente ($X \geq 9$). Inicialmente o banco de dados contendo as notas dos atributos (cor, sabor, textura e aroma) e da nota geral do Biscoito 1 é aberto no *software* R através do comando **read.table( )**. Vale salientar que no **( )** coloca-se o nome do banco e a extensão. Além disso, se as variáveis no banco de dados são nomeadas, deve-se colocar **header=TRUE**. Novamente o comando **attach( )** foi utilizado tornar o código de categorização mais otimizado.

dados1 = read.table("Amostra_1.txt", header = TRUE)

attach(dados1)

Os comandos utilizados para a categorização são apresentados abaixo e as respectivas saídas são visualizadas nas Figuras 11, 12 e 13.

cut(Cor, breaks= c(0,3,6,9,10),

labels = c("ruim", "bom", "otimo", "excelente"), right = FALSE)

```
> cut(dados1$Cor, breaks= c(0,3,6,9,10),
+ labels = c("ruim", "bom", "otimo", "excelente"), right = FALSE)
 [1] bom       otimo     otimo     otimo     otimo     otimo     otimo
 [8] bom       bom       bom       otimo     otimo     bom       otimo
[15] bom       otimo     otimo     otimo     excelente bom       bom
[22] otimo     otimo     otimo     bom       bom       otimo     bom
[29] otimo     bom       otimo     otimo     otimo     bom       bom
[36] bom       bom       otimo     bom       bom
Levels: ruim bom otimo excelente
```

Figura 11. Classificação das Variáveis que Serão Categorizadas.

Fonte: Autoras, 2022.

table(cut(Cor, breaks= c(0,3,6,9,10),

labels = c("0|- 3", "3|-6", "6|-9", ">= 9"), right = FALSE))

```
> table(cut(dados1$Cor, breaks= c(0,3,6,9,10),
+ labels = c("0|- 3", "3|-6", "6|-9", ">= 9"), right = FALSE))

0|- 3  3|-6  6|-9  >= 9
```

Figura 12. Classificação das Variáveis que Serão Categorizadas por Notas.

Fonte: Autoras, 2022.

cat_cor = cut(Cor, breaks=c(0,3,6, 9, 10), right = FALSE); cat_cor

levels(cat_cor) = c("Ruim","Bom","Otimo", "Excelente"); table(cat_cor)

```
>
> cat_cor = cut(dados1$Cor,breaks=c(0,3,6, 9, 10), right = FALSE); cat_cor
 [1] [3,6)  [6,9)  [6,9)  [6,9)  [6,9)  [6,9)  [6,9)  [3,6)  [3,6)  [3,6)
[11] [6,9)  [6,9)  [3,6)  [6,9)  [3,6)  [6,9)  [6,9)  [6,9)  [9,10) [3,6)
[21] [3,6)  [6,9)  [6,9)  [6,9)  [3,6)  [3,6)  [6,9)  [3,6)  [6,9)  [3,6)
[31] [6,9)  [6,9)  [6,9)  [3,6)  [3,6)  [3,6)  [3,6)  [6,9)  [3,6)  [3,6)
Levels: [0,3) [3,6) [6,9) [9,10)
>
> levels(cat_cor) = c("Ruim","Bom","Otimo", "Excelente"); table(cat_cor)
cat_cor
     Ruim       Bom     Otimo Excelente
        0        18        21         1
```

Figura 13. Classificação das Variáveis que Serão Categorizadas por Níveis e Notas.

Fonte: Autoras, 2022.

Para elaborar o gráfico de colunas das notas dos provadores em relação ao atributo **Cor** considerando o Biscoito 1, utiliza-se o comando abaixo. O resultado gráfico é apresentado no Gráfico 2.

barplot(table(Cor), col=c("red"), main="Cor", xlab = "Notas do Biscoito 1", ylim=c(0,25), sub = "a")

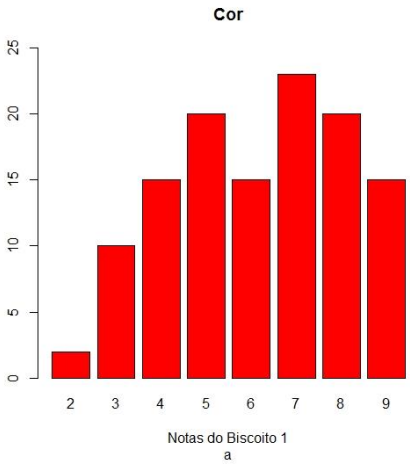

Gráfico 2. Gráfico de colunas Contendo as Notas dos Provadores em Relação ao atributo Cor considerando o Biscoito 1.

Fonte: Autoras, 2022.

Vale salientar que o objeto **col** é usado para definir a cor do gráfico. Desta forma, para colocar novas cores, os nomes as mesmas devem ser escritas em inglês.

Para construir o gráfico de setores (ou pizza) das notas dos provadores categorizadas (ruim, bom, ótimo e excelente) em relação ao atributo **Cor** considerando o Biscoito 1, utiliza-se o comando abaixo. O resultado gráfico é apresentado no Gráfico 3.

```
pie(table(cat_cor), main =" ")
text(locator(n=1), paste(round(prop.table(table(cat_cor))[2], digits = 2)*100, "%"))# bom
text(locator(n=1), paste(round(prop.table(table(cat_cor))[3], digits = 2)*100, "%")) # otimo
text(locator(n=1), paste(round(prop.table(table(cat_cor))[4], digits = 2)*100, "%"))# excelente
```

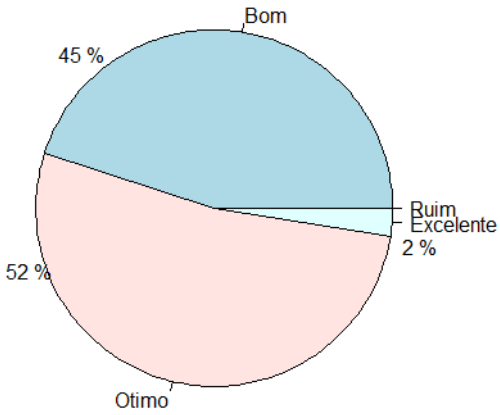

Gráfico 3. Gráfico e Setores Contendo as Notas dos Provadores Categorizadas em Relação ao atributo Cor considerando o Biscoito 1.

Fonte: Autoras, 2022.

5. Conclusões

O *software* R é uma ferramenta gratuita, de fácil manipulação e acessível para pesquisadores, docentes e discentes de todas as áreas do conhecimento. Há na literatura uma gama de materiais, tutoriais e notas de aula, que podem auxiliar na elaboração dos códigos para a implementação das análises dos dados oriundos de pesquisas. Entretanto, a proposta do artigo se torna relevante, pois com o auxílio das interpretações e análises, os pesquisadores se tornam mais confiantes em utilizar o R como programa estatístico.

6. Referências

Adriano, N. A. (2007). *O retorno acionário como fator determinante da estrutura de capital das empresas brasileiras de capital aberto*. Disponível em: https://repositorio.unb.br/handle/10482/3706. Acesso em: 17 jul. 2022.

Alcântara, M, Freitas-Sá, D. G. C. (2018). Metodologias sensoriais descritivas mais rápidas e versáteis – uma atualidade na ciência sensorial. *Brazilian Journal of Food Technology*, 21.

Anderson, T. W., Darling, D. A. (1954). A Test for Goodness of Fit. *J. Amer. Statist. Ass.*, 49, 765-769.

Associação Brasileira de Normas Técnicas – ABNT (1993). *Análise sensorial dos alimentos e bebidas: terminologia*. 8 pp.

Barbetta, P. A., Reis, M. M., Bornia, A. C. (2010). *Estatística Para Cursos de Engenharia e Informática***.** Editora. Atlas 3º Ed.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society, Series A*, 160, 268–282.

Carpinetti, L. C. R. (2009). *Planejamento e análise de experimentos*. Disponível em: http://repositorio.eesc.usp.br/handle/RIEESC/6043. Acesso em: 17 jul. 2022.

Conceicão, M. J. (2008). Leitura crítica dos dados estatísticos em trabalhos científicos. *Revista Brasileira Cirurgia Cardiovascular*, 23(3).

Conover, W. J., Johnson, E. M., Johnson, M. M. (1981). A Comparative Study of Tests for Homogeneity of Variances, With Applications to the Outer Continental Shelf Bidding Data. *Technometrics*, 23, 351-361.

Dallal, G. E., Wilkinson, L. (1986). "An Analytic Approximation to the Distribution of Lilliefors's Test Statistic for Normality". *The American Statistician*, 40(4), 294-296.

Da Rocha, K. R., Júnior, A. J. B. (2018). Anova medidas repetidas e seus pressupostos: análise passo a passo de um experimento. *Revista Eletrônica Perspectivas da Ciência e Tecnologia-ISSN: 1984-5693*, 10, 29.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56, 52–64.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6, 241–252.

Ferreira, P. L. (2005). Estatística descritiva e inferencial: breves notas. Disponível em: https://estudogeral.uc.pt/handle/10316/9961. Acesso em 17 jul. 2022.

Garcia-Marques, T. (1997). A hipótese de estudo determina a anális e estatística: Um exemplo com o modelo ANOVA. *Análise Psicológica*, 15(1), 19-28.

Kolmogorov, A. (1933). «Sulla determinazione empirica di una legge di distribuzione». *Giornale dell'istituto italiano degli attuari*. *4*, 83-91.

Kruskal, W. H., WALLIS, W. A. Use of ranks in one criterion variance analysis. *Journal of the American Statistical Association***,** v. 47, p. 583-621, 1952.

Leotti, V. B., Coster, R., Riboldi, J. Normalidade de variáveis: métodos de verificação e comparação de alguns testes não-paramétricos por simulação. *Revista HCPA*, 32(2), 227-234.

Levene, H. *(*1960). *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin et al. eds., Stanford University Press, pp. 278-292.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown". *Journal of the American Statistical Association*, 62(318), 399–402.

Lilliefors, H. W. (1969). On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown. *Journal of the American Statistical Association*, 64(325), 387–389.

Morettin, P. A., Bussab, W. de O. *Estatística Básica*. Saraiva, 5ª Ed, 2004.

Moraes, M. A. C. M. (1993). *Métodos para avaliação sensorial dos alimentos*. 8.ed. Campinas: UNICAMP. 93 p. (Série Manuais).

Neto, A. A. H., Stein, C. E. (2003). *Uma Abordagem dos Testes não Paramétricos com Utilização do Excel***.** Disponível em: http://www.mat.ufrgs.br/~viali/estatistica/mat2282/material/textos/artigo_11_09_ 2003.pdf. Acesso em 22 mai. 2020.

Paulino, C. D., Da Motta Singer, J. *Análise de dados categorizados*. Editora Blucher, 2006.

Royston, J. P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(2), 115-124.

Ritter, M. N., They, N. H., Konzen, E. (2019). *Introdução ao software estatístico R.* Universidade Federal do Rio Grande do Sul - UFRGS *campus* Litoral Norte, Imbé. Disponível em: http://professor.ufrgs.br/sites/default/files/matiasritter/files/apostila_introducao_ao_r_-_ritter_they_and_konzen.pdf. Acesso em 30 jun. 2020.

Rossini, K., Anzanello, M. J., Fogliatto, F. S. (2012). Seleção de atributos em avaliações sensoriais descritivas. *Produção*, 22, 380-390.

Silva, F. de A. S., Azevedo, C. A. V. de. (2009). Principal components analysis in the *software* Assistat-Statistical Attendance. In: World Congress on Computers in Agriculture, 7., 2009, Orlando. *Proceedings*... Reno, NV: American Society of Agricultural and Biological Engineers.

Sousa, M. H. de, Silva, N. N. da. (2000). Comparação de softwares para análise de dados de levantamentos complexos. *Revista Saúde Pública*, 34(6).

Shapiro, S. S., Wilk, M. B. (1965). *Testing The Normality of Several Samples.* (Unpublished Manuscript).

Smirnov, N. (1948). «Table for Estimating the Goodness of Fit of Empirical Distributions». *The Annals of Mathematical Statistics*, *19*(2), 279–281.

Silva, F. de A. S.; Azevedo, C. A. V. de. (2006). *A new version of the Assistat -Statistical Assistance Software*. In: World Congress on Computers in Agriculture, 4, 2006, Orlando. *Proceedings*... Reno, RV: American Society of Agricultural and Biological Engineers, 393-396.

Spiegel, M. R., Stephens, L., Nascimento, J. L. (2009). *Estatística*. Schaum. Bookman.

Teixeira, L. V. (2009). Análise Sensorial na Indústria de Alimentos. *Revista da Instituto de Laticínios "Cândido Tostes"*, 366, 12-21.

Teixeira, N. S., Trindade, D. B., Abrantes, M. F., Santos, H. C., Souza, M. T. D. (2020). Elaboração, análise descritiva e análise sensorial de biscoito nutritivo com poupa de maracujá (*Passiflora edulis*). *Global Science and Technology*, 13(1), 182-197.

Venables, W. N., Smith, D. M. (2005). R DEVELOPMENT CORE TEAM. An Introduction to R. Notes on R: Programming Environment for Data Analysis and Graphics. Version 2.2.0. Áustria.

Xavier, L. H., Dias, C. T. S. (2001). Acurácia do modelo univariado para análise de medidas repetidas por simulação multidimensional. *Scientia Agricola*, 58(2), 241-250.