HOMERO RIBEIRO NETO

TESTES F E DE NORMALIDADE AVALIADOS SOB DIFERENTES CONDIÇÕES EXPERIMENTAIS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

Orientador: Nerilson Terra Santos

VIÇOSA - MINAS GERAIS
2023

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
R484t
2023

Ribeiro Neto, Homero, 1996-
Testes F e de normalidade avaliados sob diferentes condições experimentais / Homero Ribeiro Neto. – Viçosa, MG, 2023.
1 dissertação eletrônica (65 f.): il.

Orientador: Nerilson Terra Santos.
Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Estatística, 2023.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2023.214
Modo de acesso: World Wide Web.

1. Análise de variância. 2. Hipótese. 3. Testes. I. Santos, Nerilson Terra, 1966-. II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 519.538

Bibliotecário(a) responsável: Bruna Silva CRB-6/2552

HOMERO RIBEIRO NETO

TESTES F E DE NORMALIDADE AVALIADOS SOB DIFERENTES CONDIÇÕES EXPERIMENTAIS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 17 de fevereiro de 2023.

Assentimento:

Documento assinado digitalmente
gov.br
HOMERO RIBEIRO NETO
Data: 09/05/2023 10:54:46-0300
Verifique em https://validar.iti.gov.br

Homero Ribeiro Neto
Autor

Documento assinado digitalmente
gov.br
NERILSON TERRA SANTOS
Data: 09/05/2023 14:01:42-0300
Verifique em https://validar.iti.gov.br

Nerilson Terra Santos
Orientador

AGRADECIMENTOS

A Deus por todas as oportunidades e conquistas.

Aos meus pais (Homero José de Souza Ribeiro e Moema de Carvalho Ribeiro) por todo apoio emocional e financeiro, pela criação e amor.

Ao professor Nerilson Terra Santos, pela excelente orientação.

A todos os amigos e familiares que me apoiaram, como a minha irmã Lícia Carvalho Ribeiro.

Ao Departamento de Estatística da Universidade Federal de Viçosa e seus professores, pelo suporte e aprendizado.

À Universidade Federal de Viçosa, pela oportunidade de realizar a pós-graduação.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de estudos.

RESUMO

RIBEIRO NETO, Homero, M.Sc., Universidade Federal de Viçosa, fevereiro de 2023. **Testes F e de normalidade avaliados sob diferentes condições experimentais.** Orientador: Nerilson Terra Santos.

A pressuposição de normalidade dos erros experimentais é uma das exigências que se impõe para a aplicação de importantes procedimentos inferenciais, como o teste F da Análise de Variância (ANOVA), muito empregada em diversos campos científicos, como as Ciências Agrárias. Nesse sentido, resultados importantes e conhecidos da Estatística, como o Teorema Central do Limite, não impõem, teoricamente, muitas dificuldades para se obter, a partir de praticamente qualquer variável aleatória não normal, uma nova variável aleatória, que seja normal, com a finalidade de não violar essa pressuposição. No entanto, por questões de ordem prática, nem sempre é possível obter um número de repetições por tratamento suficientemente elevado para que o Teorema supracitado seja aplicado. Assim, algumas das alternativas mais empregadas são os testes de normalidade, para, com quantidades limitadas de observações amostrais, inferir a respeito da normalidade dos dados. Porém, as efetividades desses testes, assim como de outros testes de hipóteses, em termos de poder (probabilidade de rejeitar uma hipótese nula falsa) e nível de significância (probabilidade de rejeitar uma hipótese nula verdadeira cometendo o erro tipo I), são influenciadas pelas condições experimentais. Por isso, este trabalho foi realizado com o objetivo de comparar o desempenho dos testes de normalidade mais comuns em condições de igualdade (desigualdade) das médias dos tratamentos, homogeneidade (heterogeneidade) de suas variâncias residuais, número de repetições de cada um e simetria (assimetria) das distribuições de probabilidades dos erros experimentais. Foi possível também analisar o desempenho do próprio teste F, inclusive quando a pressuposição de normalidade foi violada. De maneira geral, foi possível concluir, ao realizar simulações, que o poder empírico dos testes de normalidade tende a cair quando a distribuição empírica dos erros experimentais é simétrica e o número total de observações é muito baixo, e que as taxas de erro tipo I, tanto dos testes de normalidade, quanto do teste F, tendem a aumentar quando as variâncias residuais dos tratamentos são heterogêneas.

Palavras-chave: Testes de Hipóteses. Nível de Significância. Erro Tipo I. Erro tipo II. Análise de Variância (ANOVA). Delineamento Inteiramente Casualizado (DIC). Distribuição Normal. Erros Experimentais.

ABSTRACT

RIBEIRO NETO, Homero, M.Sc., Universidade Federal de Viçosa, February, 2023. **F and normality tests analyzed under different experimental conditions** Adviser: Nerilson Terra Santos.

The assumption of normality of experimental errors is one of the requirements imposed for the application of important inferential procedures, such as the F test of Analysis of Variance (AOV), widely used in several scientific fields, such as Agricultural Sciences. In this sense, well-known statistics results, such as the Central Limit Theorem, do not theoretically impose many difficulties to obtain, from practically any non-normal random variable, a new random variable, which is normal, with the purpose to attend this assumption. However, due to practical reasons, it is not always possible to obtain a sufficiently high number of repetitions per treatment to apply this Theorem. Thus, some of the most used alternatives are the normality tests, to, with limited amounts of sample observations, infer about the normality of the data. But, the effectiveness of these tests, as well as other hypothesis tests, in terms of power (probability of rejecting a false null hypothesis) and significance level (probability of rejecting a true null hypothesis, type I error), are influenced by the experimental conditions. Therefore, this work was carried out with the objective of comparing the performance of the most common normality tests under conditions of equality (inequality) of the means of the treatments, homogeneity (heterogeneity) of their residual variances, number of repetitions of each one and symmetry (asymmetry) of experimental errors probability distributions. It was also possible to analyze the performance of the F test itself, when the assumption of normality was violated. In general, it was possible to conclude, when performing simulations, that the empirical power of normality tests tends to drop when the empirical distribution of experimental errors is symmetrical and the total number of observations is very low, and that Type I error rates, both for the normality tests and for the F test, tend to increase when the residual variances of the treatments are heterogeneous.

Keywords: Hypothesis Tests. Significance Level. Type I error. Type II error. Analysis of Variance (ANOVA). Completely Randomized Design (CRD). Normal Distribution. Experimental Errors.

SUMÁRIO

Capítulo 1: Introdução

Este trabalho teve como objetivo geral avaliar o nível de significância empírico (taxa de erro tipo I), assim como, o poder de testes de normalidade e do teste F quando as bases de dados são oriundas de delineamentos experimentais, sob diferentes condições. Com esse objetivo em foco, ele foi dividido em cinco capítulos. Neste Capítulo 1, é apresentada uma breve descrição do que será dissertado nos Capítulos 2, 3, 4 e 5.

No Capítulo 2, foram realizadas revisões a respeito da Distribuição Normal, do Teorema Central do Limite, dos testes de normalidade comumente utilizados, do Modelo Linear de Gauss Markov (MLGM) e da fundamentação do teste F.

Os Capítulos 3 e 4 foram elaborados no formato de artigo científico. No artigo do Capítulo 3, procedeu-se à comparação entre os testes de normalidade em face dos poderes empíricos apresentados por eles, sob diferentes condições experimentais, tais como a igualdade (desigualdade) das médias dos tratamentos, a homogeneidade (heterogeneidade) das suas variâncias residuais e o número de repetições de cada um. Já no artigo do Capítulo 4, as comparações entre os testes de normalidade foram feitas com base nas taxas empíricas de erro tipo I obtidas. Nos dois artigos, foram também avaliados os desempenhos do teste F, quanto às respectivas medidas de efetividade anteriormente mencionadas.

No Capítulo 5, são apresentadas as conclusões gerais visando à sugestão de um protocolo para que se possa utilizar os testes de normalidade quando a base de dados é oriunda de experimentos.

Em termos científicos, o objetivo principal, que norteou e motivou o desenvolvimento deste trabalho, foi o propósito de avaliar o teste F e os testes de normalidade de Kolmogorov-Smirnov (KS), Lilliefors (LI), Cramér-von Mises (CVM), Anderson-Darling (AD) e Shapiro-Wilk (SW) sob cenários experimentais com diferentes números de repetições por tratamento.

Os objetivos específicos deste trabalho foram de atestar se determinados padrões experimentais têm influência nas capacidades dos testes de normalidade de detectar a presença (ausência) de normalidade e do teste F de detectar a igualdade (diferença) entre as médias dos tratamentos. Os padrões experimentais avaliados foram:

- Heterogeneidade/ homogeneidade das variâncias residuais dos tratamentos;

- Igualdade/ desigualdade das médias dos tratamentos;

- Simetria/assimetria da distribuição de probabilidade dos erros experimentais alternativa à normal.

Capítulo 2: Fundamentação Teórica

1. Revisão Bibliográfica

1.1 Distribuição Normal e sua Importância na Estatística

A Distribuição Normal, também conhecida como Distribuição Gaussiana, é comum nos mais variados campos de pesquisa, pois diversas variáveis aleatórias apresentam uma distribuição que se aproxima de uma distribuição normal. Segundo Mood (1974), esse fato se explica pelo Teorema Central do Limite. Por outro lado, Casella e Berger (2002) justificam o frequente uso de distribuições normais ao formato de sino das suas curvas características, cuja simetria em torno da média faz com que essas distribuições sejam escolhas apropriadas para muitos modelos populacionais.

A Distribuição Normal é caracterizada pelos parâmetros média (μ) e variância (σ^2), os quais, segundo Casella e Berger (2002), tornam-na especial, pois eles fornecem as informações necessárias para caracterizar tanto o formato da sua curva de distribuição de probabilidades bem com a sua localização. Essa curva é descrita por meio da função densidade de probabilidade representada pela Equação 1:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty, \sigma > 0 \quad (1)$$

Dessa maneira, conforme atestam Casella e Berger (2002), apesar de muitas outras distribuições de probabilidades também apresentarem curvas em formato de sino, a maioria não confere a mesma tratabilidade analítica que as distribuições normais, o que pode ser observado na relativa facilidade para serem provadas propriedades dessas distribuições, para construir intervalos de confiança e testes de hipóteses.

Muito mais importante do que as propriedades supracitadas da normal, é o fato de que a pressuposição da distribuição normal é uma das exigências que se impõe para o uso de alguns procedimentos inferenciais. Exemplos desses procedimentos com essa exigência são os testes de hipóteses F e t. Portanto, o completo entendimento dessa distribuição, bem como o seu uso são de grande relevância para um profundo conhecimento de procedimentos inferenciais. Para se ter esse entendimento, é necessário que sejam conhecidos alguns conceitos como o de amostras aleatórias.

1.2 Amostras aleatórias

De acordo Mood (1974), um conjunto de n variáveis aleatórias $X_1, X_2, \dots, X_n$ é denominado de amostra aleatória se essas variáveis aleatórias forem independentes e identicamente distribuídas. A função densidade de probabilidade conjunta (ou função de verossimilhança) dessa amostra aleatória é definida pela Equação 2, em que $f(.)$ representa a função densidade comum a todas as n variáveis aleatórias.

$$f_{X_1,X_2,\dots,X_n}(X_1, X_2, \dots, X_n) = f_{X_1}(x_1).f_{X_2}(x_2) \dots f_{X_n}(x_n) = f(.)^n \quad (2)$$

Em alguns estudos, cada uma das X_i variáveis aleatórias pode não seguir uma distribuição normal. Contudo, de acordo com o Teorema Central do Limite, pode-se assegurar que, se determinadas condições forem satisfeitas, a distribuição da média amostral se aproxima de uma normal.

1.3 Teorema Central do Limite

Suponha a variável aleatória Z_n, definida na Equação 3, em que $\bar{X}_n$ representa a média amostral de uma amostra aleatória de tamanho n cuja função de verossimilhança é definida pela Equação 2:

$$Z_n = \frac{\bar{X}_n - E[\bar{X}_n]}{\sqrt{Var[\bar{X}_n]}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \quad (3)$$

De acordo com Mood (1974), o Teorema Central do Limite postula que Z_n converge para uma distribuição normal padrão quando $n \to \infty$.

A grande aplicabilidade do Teorema Central do Limite deve-se ao fato de que $f(.)$ pode corresponder a qualquer distribuição de probabilidades, desde que apresente variância finita. Portanto, partindo-se praticamente de qualquer distribuição não normal, a variável aleatória Z_n, obtida a partir da média amostral de uma amostra aleatória converge para a distribuição normal padrão desde que essa amostra aleatória tenha tamanho suficientemente grande, o qual depende de quão distante a distribuição $f(x_i)$ está de uma distribuição normal (MOOD, 1974).

1.4 Modelo Linear de Gauss-Markov

Nas ciências agrárias, são realizados com maior frequência experimentos com o objetivo de proceder às análises de variâncias (ANOVA), que se baseiam no teste F, realizado para avaliar a igualdade/desigualdade das médias dos tratamentos, os quais, por sua vez, representam os níveis dos fatores em estudo.

Contudo, antes de introduzir a estatística F e as bases teóricas do teste F, é crucial compreender os modelos estatísticos adotados para comparar os níveis do fator em estudo, para que não restem dúvidas a respeito da necessidade de se pressupor que os erros experimentais tenham uma distribuição normal, antes de realizar o teste F.

Dessa maneira, primeiramente, define-se o vetor $\boldsymbol{y}$ (Equação 4) como um vetor aleatório se cada elemento que o compõe for uma variável aleatória, ou seja, Y_i é uma variável aleatória $\forall\, i = 1,2, \dots, n$.

$$\boldsymbol{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} \tag{4}$$

Em seguida, define-se a média de um vetor aleatório como o vetor representado pela Equação 5 e a variância desse vetor como a diagonal principal da matriz da Equação 6.

$$E(\boldsymbol{y}) = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \dots \\ E(Y_n) \end{bmatrix} \tag{5}$$

$$V(\boldsymbol{y}) = \begin{bmatrix} V(Y_1) & \dots & Cov(Y_1, Y_n) \\ Cov(Y_2, Y_1) & \dots & Cov(Y_2, Y_n) \\ \dots & \dots & \dots \\ Cov(Y_n, Y_1) & \dots & V(Y_n) \end{bmatrix} \tag{6}$$

Em que, cada elemento i,j é calculado por meio da Equação 7 $\forall\, i = 1,2, \dots, n$ e $\forall\, j = 1,2, \dots, n$:

$$Cov(Y_i, Y_j) = E(Y_i Y_j) - E(Y_i)E(Y_j) \tag{7}$$

Considerando os vetores e matrizes apresentados anteriormente, segundo Searle e Gruber (2016), o Modelo Linear de Gauss-Markov (MLGM) é definido pela Equação 8.

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e} \tag{8}$$

Em que, $\boldsymbol{y}$ é um vetor aleatório $(n \times 1)$ (Equação 4 das variáveis aleatórias dependentes; $\boldsymbol{X}$ é uma matriz $(n \times p)$ de constantes com colunas correspondentes às variáveis explicativas (ou independentes), que também é conhecida como matriz de delineamento; $\boldsymbol{\beta}$ é um vetor $(p \times 1)$ de parâmetros desconhecidos e $\boldsymbol{e}$ é um vetor $(n \times 1)$ de erros aleatórios.

No caso, por exemplo, de um experimento instalado sob o Delineamento Inteiramente Casualizado (DIC) com um fator qualitativo com 5 níveis, isto é, 5 tratamentos e 2 repetições

por tratamento, o MLGM (Equação 8) é descrito conforme a Equação 9, em que μ é a média geral das variáveis dependentes e $\tau_i, i = 1;\ 2;\ 3; 4;\ 5$ é o efeito de cada tratamento.

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \\ Y_{41} \\ Y_{42} \\ Y_{51} \\ Y_{52} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \\ \tau_5 \end{bmatrix} + \begin{bmatrix} E_{11} \\ E_{12} \\ E_{21} \\ E_{22} \\ E_{31} \\ E_{32} \\ E_{41} \\ E_{42} \\ E_{51} \\ E_{52} \end{bmatrix} \tag{9}$$

De acordo com Searle e Gruber (2016), para esse modelo linear de Gauss-Markov (MLGM) (Equação 8), supõe-se que $E(\boldsymbol{e}) = \boldsymbol{0}$ e $V(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$, em que σ^2 é um parâmetro desconhecido. Portanto, ao se aplicar os operadores de esperança matemática e de variância diretamente à Equação 8, obtém-se, respectivamente, $E(\boldsymbol{y}) = \boldsymbol{X\beta}$ e $\mathrm{V}(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$. Dessa maneira, pode-se escrever que:

$$\boldsymbol{y} \sim (\boldsymbol{X\beta};\ \sigma^2\boldsymbol{I}) \tag{10}$$

No entanto, nota-se que, na Equação 10 não foi especificada nenhuma distribuição de probabilidade para $\boldsymbol{y}$, visto que, originalmente, o Modelo Linear de Gauss-Markov (MLGM) não é associado a nenhuma distribuição. Porém, quando esse modelo experimental é empregado como base para a realização de um estudo de análise de variância (ANOVA), baseada no teste F ou quando é empregado como base para a realização de outros testes de hipóteses, o modelo deve incorporar a pressuposição de normalidade, tornando-se Modelo Linear de Gauss-Markov Normal (MLGMN), visto que, como será demonstrado adiante, as estatísticas de teste são calculadas considerando amostragens sob distribuição normal. Dessa forma, a pressuposição de normalidade equivale à representação na Equação 11, em que a sigla $i.i.d.$ indica que os erros experimentais sejam independentes e identicamente distribuídos (amostra aleatória).

$$\boldsymbol{e} \sim N(\boldsymbol{0}; \sigma^2\boldsymbol{I}) \equiv E_1, \ldots, E_n \overset{i.i.d.}{\sim} N(0; \sigma^2) \tag{11}$$

Como consequência disso, tem-se que $Y_1, Y_2, \ldots, Y_n$ são também variáveis aleatórias independentes normais, conforme a Equação 12 (válida para o DIC), apesar de não necessariamente serem identicamente distribuídas, já que é possível que cada uma siga uma distribuição normal com uma média diferente.

$$Y_1, \ldots, Y_n \overset{i.}{\sim} N(\boldsymbol{x'}_{(\boldsymbol{i})}\boldsymbol{\beta}; \sigma^2) \tag{12}$$

Em que, $V(Y_i) = \sigma^2 \; \forall \; i = 1,2, \dots, n$ e $E(y_i) = \boldsymbol{x'}_{(i)}\boldsymbol{\beta} \; \forall \; i = 1,2, \dots, n$, sendo $\boldsymbol{x'}_{(i)}$a i-ésima linha de $\boldsymbol{X}$.

Sabendo disso, ainda é crucial compreender como o teste F é construído, para entender melhor o porquê de ter sido necessário que o Modelo Linear de Gauss-Markov (MLGM) incorporasse a pressuposição de normalidade dos erros experimentais, tornando-se o Modelo Linear de Gauss-Markov Normal (MLGMN), representado anteriormente.

1.5 O teste F e a pressuposição de normalidade

Suponha que cada uma das duas variáveis aleatórias, U e V, segue distribuição de qui-quadrado, com m e v graus de liberdade, respectivamente. Adicionalmente, suponha que as variáveis U e V sejam independentes. Então, a razão, W (Equação 13), entre essas variáveis segue uma distribuição de probabilidades F (MOOD, 1974).

$$W = \frac{U/m}{V/n} \tag{13}$$

Sabe-se também que, segundo Mood (1974) e Casella e Berger (2002), se $y_1, y_2, \dots, y_{m+1}$ é uma amostra aleatória de tamanho $m + 1$ advinda de uma distribuição normal com média μ_1 e variância σ^2, U é calculada de acordo com a Equação 14 e segue distribuição qui-quadrado com m graus de liberdade. Por outro lado, se $x_1, x_2, \dots, x_{n+1}$, com tamanho $n + 1$, também é uma amostra aleatória advinda de uma distribuição normal com média μ_2 e mesma variância σ^2, consecutivamente, V (Equação 15) segue distribuição qui-quadrado com n graus de liberdade:

$$U = \frac{\sum_{i=1}^{m+1}(Y_i - \bar{Y})^2}{\sigma^2} \tag{14}$$

$$V = \frac{\sum_{j=1}^{n+1}(X_j - \bar{X})^2}{\sigma^2} \tag{15}$$

Consequentemente, como a estatística do teste F (Equação 13) é definida pela razão de dois estimadores de variância (Equações 14 e 15), obtidos sob distribuição normal, a pressuposição de normalidade torna-se necessária para a realização desse teste. No caso da ANOVA, o estimador de variância do numerador do teste F, estima a variância entre as médias de tratamentos somada à variância que ocorre devido ao acaso e o estimador de variância do denominador estima apenas a variância que ocorre devido ao acaso.

A priori, conforme mencionado anteriormente, supõe-se ainda que as variâncias populacionais sejam ambas iguais a σ^2. Porém, quanto mais destoantes forem as estimativas obtidas pelos dois estimadores de variância, menos provável é a hipótese de que as variâncias populacionais sejam iguais. Considerando que a estatística do teste F, nas análises de variâncias (ANOVA), é uma razão entre o estimador da variância entre as médias de tratamentos somada à variância que ocorre devido ao acaso (numerador) e a variância apenas devida ao acaso (denominador), quanto maior for essa razão quando as estimativas forem obtidas, mais plausível é a hipótese de que a variância devida às diferenças entre as médias dos tratamentos é maior do que a devida às causas aleatórias. Por outro lado, se a razão F não for suficientemente grande, isto é, caso seja verificado que, estatisticamente, essas duas variâncias são iguais, assume-se que não há diferenças significativas entre as médias dos tratamentos, já que essas diferenças são equivalentes a diferenças aleatórias (acaso). Assim, as hipóteses do teste F da ANOVA podem ser escritas da seguinte maneira:

H_0: As médias dos tratamentos são idênticas;

H_a: A média de pelo menos um tratamento difere das demais.

Contudo, como os estimadores envolvidos no cálculo da estatística do teste F pressupõem a condição de normalidade, antes de realizar o teste F, são aplicados testes de normalidade para verificar se essa pressuposição não é violada.

1.6 Testes de normalidade

Os testes de normalidade podem ser classificados como sendo ou não testes de aderência. Os de aderência são aqueles, cujo objetivo é verificar se os dados amostrais se adequam a uma distribuição teórica, que, no caso de um teste de normalidade de aderência, é uma distribuição teórica Normal. Exemplos desses testes são os testes de Kolmogorov-Smirnov (KS), Lilliefors (LI), Cramér-von Mises (CVM) e Anderson-Darling (AD). Já os testes de normalidade, que não são de aderência, não comparam funções de distribuição empírica e teórica. Geralmente, comparam dois estimadores, como é o caso do teste de Shapiro-Wilk (SW), que compara dois estimadores de variância, de maneira que a discrepância entre eles indique se os dados podem ser considerados como originados de uma distribuição normal.

1.6.1 Teste de Kolmogorov-Smirnov

Em 1933, Kolmogorov e Smirnov introduziram seu teste homônimo objetivando, de acordo com Barbetta et al. (2004) e Campos (1976), verificar se uma distribuição desconhecida dos dados, $F_0(x)$, é suficientemente próxima da teórica conhecida, $F(x)$. Para aplicá-lo, deve-

se definir as distribuições acumuladas $F(x)$ (teórica) e $S(x)$ (empírica), calculada conforme a Equação 16, em que x_i é um valor qualquer advindo da amostra aleatória $x_1, x_2, \dots, x_n$.

$$S(x_i) = \frac{número\ de\ valores \leq x_i}{n} \tag{16}$$

Desta forma, as hipóteses a serem testadas são as seguintes:

H_0: Os dados advêm de $F(x)$, ocorre aderência e $F(x) = F_0(x)$;

H_a: Os dados não advêm de $F(x)$, não ocorre aderência e $F(x) \neq F_0(x)$.

Assim, para optar por uma dessas hipóteses, obtém-se os valores teóricos de $F(x_i)$ para cada valor x_i $(i = 1,2, \dots, n)$, sendo possível, pois $F(x)$ é a distribuição normal definida a priori. O próximo passo, então, é o cálculo de todas as diferenças absolutas entre $S(x_i)$ e $F(x_i)$ e entre $S(x_{i-1})$ e $F(x_i)$, definindo-se, como estatística do teste (d), a máxima diferença absoluta obtida (Equação 17), sendo que, quanto menor for o valor da estatística, maior será a aproximação da distribuição empírica com a teórica:

$$d = \max\{|F(x_i) - S(x_i)|, |F(x_i) - S(x_{i-1})|\} \tag{17}$$

Em seguida, d é comparado ao valor tabelado d_c, o qual é obtido em função do nível de significância α e do tamanho n da amostra. A regra de decisão desse teste é:

Se $d < d_c$, H_0 não é rejeitada (ocorre aderência à distribuição teórica). Caso contrário, se $d \geq d_c$, H_0 é rejeitada (não ocorre aderência à distribuição teórica).

Ressalta-se ainda, conforme pontua Campos (1976), que se a variável aleatória X estudada for contínua, o teste de Kolmogorov-Smirnov é exato. Porém, caso ela seja discreta, o teste é apenas aproximado.

1.6.2 Teste de Lilliefors

É uma adaptação do teste de Kolmogorov-Smirnov, apresentando os mesmos objetivos, hipóteses e estatística (Equação 17). No entanto, a diferença ocorre, conforme afirmam Barbetta et al. (2004), pelo fato de que, nessa situação, como os parâmetros média e variância da distribuição teórica não são conhecidos, eles precisam ser estimados respectivamente, pelas Equações 18 e 19:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{18}$$

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \tag{19}$$

Além disso, outra diferença é o fato de os cálculos das estatísticas se basearem na variável reduzida Z_i (Equação 20), ao invés da variável original X_i. Por todas essas diferenças, os valores críticos tabelados do teste de Lilliefors são diferentes daqueles do teste de Kolmogorov-Smirnov, sendo:

$$Z_i = \frac{X_i - \bar{X}}{S} \tag{20}$$

Contudo, uma das relativas limitações deste teste é o fato de que, segundo Campos (1976), a não rejeição da hipótese de nulidade apenas indica que a adoção da distribuição normal é uma aproximação aceitável. Portanto, não se trata de um teste exato.

1.6.3 Teste de Cramér-von Mises

O teste de Cramér-von Mises é uma alternativa ao teste de Kolmogorov-Smirnov, pois apesar de ambos avaliarem as diferenças entre as distribuições empírica e teórica, a estatística de teste de Cramér-von Mises é calculada de outra maneira. No entanto, apresentam as mesmas hipóteses. Assim, sendo $x_{(1)} \leq \cdots \leq x_{(n)}$ as observações ordenadas de uma amostra aleatória advinda de uma população com função de distribuição acumulada $F_0(x)$ e sendo as distribuições teórica e empírica, respectivamente $F(x)$ e $S(x)$, as hipóteses desse teste são:

H_0: $F(x) = F_0(x)$;

H_a: $F(x) \neq F_0(x)$.

Dessa maneira, segundo Thadewald e Büning (2007), a estatística CM é definida pela Equação 21, mas equivale à representada na Equação 22.

$$CM = n\int_{-\infty}^{\infty}[S(x) - F(x)]^2 f(x)dx \tag{21}$$

$$CM = \frac{1}{12n} + \sum_{i=1}^{n}\left[F\left(x_{(i)}\right) - \frac{2i-1}{2n}\right]^2 \tag{22}$$

Nesse caso, H_0 é rejeitada se $CM \geq c_{1-\alpha}$, sendo $c_{1-\alpha}$ o valor crítico tabelado, cujas tabelas são disponibilizadas por Anderson e Darling (1952, p.203).

1.6.4 Teste de Anderson-Darling

Conforme Yap e Sim (2011), o teste desenvolvido por Anderson e Darling (1952) é uma adaptação do teste de Cramér-von Mises (CVM), pois se diferencia do teste CVM por dar mais

peso aos valores mais extremos da distribuição, aumentando a estatística calculada e a sensibilidade do teste. Porém, o que torna essa adaptação mais complexa e específica é o fato de que os valores críticos devem ser calculados, empiricamente, para cada distribuição teórica definida a priori, de acordo com a sua função densidade de probabilidade. Nesse sentido, a estatística AD (Equação 23) é uma média ponderada do quadrado da diferença entre a função acumulada empírica e a hipotética e caso o peso (Equação 24) fosse igual a 1, essa estatística seria a CM (Equação 21). Porém, o peso fica maior quanto mais extremo for o intervalo considerado.

$$AD = n\int_{-\infty}^{\infty}[S(x) - F(x)]^2\Psi\big(F(x)\big)f(x)dx \tag{23}$$

$$\Psi\big(F(x)\big) = \{F(x)[1 - F(x)]\}^{-1} \tag{24}$$

A estatística AD pode também ser calculada conforme a Equação 25, sendo $x_{(1)} \leq \cdots \leq x_{(n)}$ as observações ordenadas, P_i a função acumulada da distribuição de probabilidade teórica e log o logaritmo neperiano. Para ambas as formas de cálculo, as hipóteses testadas são: H_0: $F(x) = F_0(x)$; H_a: $F(x) \neq F_0(x)$. Assim, se a estatística AD for maior ou igual ao valor crítico específico, a hipótese de nulidade é rejeitada.

$$AD = -\sum_{i=1}^{n}\left\{\frac{[2i-1][logP_i + log(1 - P_{n+1-i})]}{n}\right\} - n \tag{25}$$

1.6.5 Teste de Shapiro-Wilk

Ao invés de realizar as comparações diretas anteriores entre funções de distribuição empírica e teórica, o teste de Shapiro e Wilk (1965) compara dois estimadores de variância, de maneira que a discrepância entre eles indique se pode ou não inferir que a amostra aleatória seja advinda de uma população normal. Logo, apesar de não ser um teste de aderência, o teste de Shapiro-Wilk também é um teste de normalidade.

Inicialmente, ordena-se a amostra aleatória de uma normal padrão de forma crescente $X_1 \leq X_2 \leq \cdots \leq X_n$. Então, define-se $\boldsymbol{m}' = (m_1, m_2, \ldots, m_n)$ como o vetor de valores esperados dessas variáveis aleatórias e $\boldsymbol{V} = (v_{ij})$ como a sua matriz $n \times n$ de covariâncias. Assim, sendo $Y_1, \ldots, Y_n$ uma outra amostra aleatória ordenada, o teste objetiva verificar se essa última amostra é advinda de uma N($\mu; \sigma^2$). Logo, se essa hipótese for verdadeira, de acordo com Shapiro e Wilk (1965), cada Y_i pode ser escrito conforme a Equação 26 e, a partir do teorema dos mínimos

quadrados generalizados, o melhor estimador não viesado da variância σ^2 de distribuições simétricas, como a normal, é representado pela Equação 27.

$$Y_i = \mu + \sigma x_i \tag{26}$$

$$\hat{\sigma}^2 = \left(\frac{\boldsymbol{m'V^{-1}y}}{\boldsymbol{m'V^{-1}m}}\right)^2 \tag{27}$$

Porém, $(n-1)\sigma^2$ também tem estimador representado pela Equação 28.

$$S^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{28}$$

Shapiro e Wilk (1965), então, definiram a estatística do teste como uma razão envolvendo esses estimadores, apresentada na Equação 29.

$$W = \frac{R^4\hat{\sigma}^2}{C^2S^2} = \frac{b^2}{S^2} = \frac{(\boldsymbol{a'y})^2}{S^2} = \frac{(\sum_{i=1}^{n} a_i y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{29}$$

Em que:

$$R^2 = \boldsymbol{m'V^{-1}m} \tag{30}$$

$$C^2 = \boldsymbol{m'V^{-1}V^{-1}m} \tag{31}$$

$$\boldsymbol{a'} = (a_1, \ldots, a_n) = \frac{1}{(\boldsymbol{m'V^{-1}V^{-1}m})^{1/2}}\boldsymbol{m'V^{-1}} \tag{32}$$

$$b = \frac{R^2\hat{\sigma}}{C} \tag{33}$$

Segundo Shapiro e Wilk (1965), b (Equação 33) é o melhor estimador linear não viesado da inclinação da reta de regressão dos valores observados e ordenados y_i em função dos valores esperados m_i da normal padrão, sendo C definida de tal modo que os coeficientes lineares sejam normalizados. Assim, quanto menor a inclinação, menor a estatística W, enfraquecendo a suposição de que a relação da Equação 26 é válida, indicando não normalidade dos dados.

Caso contrário, se a amostra aleatória $Y_1, \ldots, Y_n$ for originada de uma distribuição normal, o numerador b^2 e o denominador S^2 da estatística W do teste estarão, ambos, conforme Shapiro e Wilk (1965), estimando a mesma variância σ^2 tornando a razão aproximadamente uma constante próxima do valor máximo unitário. Portanto, valores baixos de W, menores do que os valores críticos, levam à rejeição da hipótese nula de que $Y_1, \ldots, Y_n$ advém de uma normal. Shapiro e Wilk (1965, p.605) fornecem uma tabela dos valores W correspondentes às

probabilidades bilaterais 1%, 2%, 5% ,10% e 50% sob H_0 (normalidade) para que, ao comparar o valor W calculado com o crítico, opte-se por rejeitar ou não a hipótese de nulidade.

1.7 Outras distribuições de probabilidades e seus efeitos nos testes de normalidade

Presume-se que os testes de normalidade de aderência apresentem maior poder, isto é, maior capacidade de rejeitar a hipótese nula falsa, quando a variável aleatória em análise segue distribuição de probabilidades com função densidade, que resulte em uma curva com formato substancialmente diferente do formato em sino da distribuição normal teórica. O motivo disso deve-se ao fato de que se as curvas diferem muito, a diferença entre as funções acumuladas empírica e teórica tende a ser maior, do que se as curvas forem pouco diferentes. Assim, como as estatísticas dos testes de normalidade de aderência são diretamente proporcionais a essas diferenças, as estatísticas também tenderão a aumentar, resultando na rejeição da hipótese de normalidade.

Porém, como será explicado a seguir, é possível que os dados sigam uma distribuição de probabilidades, que não é normal, mas que se assemelha muito a uma curva de distribuição normal, o que pode ser o caso de distribuições Gama e Gama Inversa. Dessa maneira, é possível que os testes de normalidade sejam afetados, perdendo poder.

1.7.1 Distribuições Gama e Gama Inversa

As distribuições Gama, diferentemente das distribuições normais, apenas são válidas para variáveis aleatórias positivas, ou seja, para $X \in [0, \infty)$.Contudo, a depender dos valores de seus parâmetros $\alpha > 0$ e $\beta > 0$, as curvas das distribuições Gama podem ser muito parecidas com as das distribuições normais, também apresentando formato de sino e sendo simétricas em torno da média. Ressalta-se, entretanto, que, caso os valores dos parâmetros sejam alterados, essas curvas podem se tornar extremamente assimétricas e, consequentemente, muito diferentes das curvas normais:

$$f(x) = \frac{\beta^{\alpha}}{\Gamma[\alpha]} x^{\alpha-1} e^{-\beta x}, 0 < x < \infty \tag{34}$$

De acordo com Mood (1974) e Gelman et al. (2013), se a variável aleatória X segue uma distribuição Gama (Equação 34), então, pode-se provar que a variável Y obtida pelo inverso de X, ou seja, $Y = 1/X$ segue uma distribuição Gama Inversa com função densidade de probabilidade definida pela Equação 35. As relações entre os parâmetros α e β da Gama Inversa, para obter a média e a variância de Y, são, respectivamente, representadas pelas Equações 36 e 37.

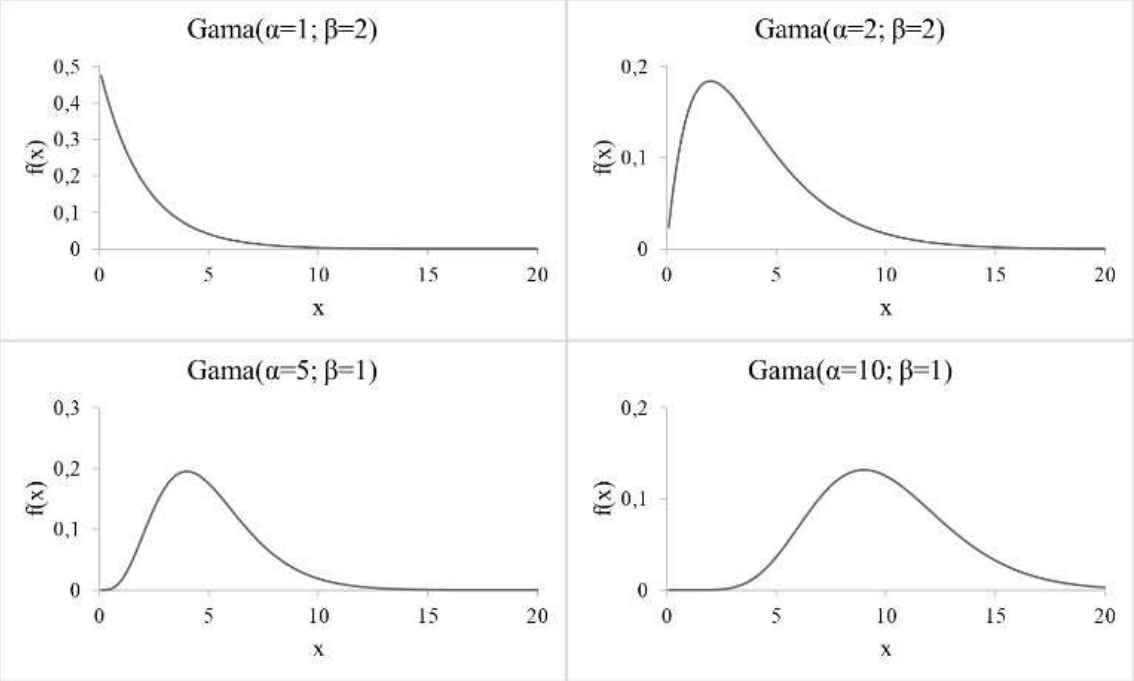
$$f(y) = \frac{\beta^{\alpha}}{\Gamma[\alpha]} y^{-(\alpha+1)} e^{-\beta/y}, 0 < y < \infty \quad (35)$$

$$E(Y) = \frac{\beta}{\alpha - 1}, \alpha > 1 \quad (36)$$

$$V(Y) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \alpha > 2 \quad (37)$$

Dependendo dos valores assumidos pelos parâmetros α e β, as distribuições Gama Inversa, assim como as distribuições Gama, podem apresentar curvas com formatos similares ou não aos de uma distribuição normal tal como exemplificado na Figura 1.

Figura 1- Curvas de densidade de probabilidade associadas às distribuições Gama com parâmetros α = 1, β= 2; Gama com parâmetros α = 2, β= 2; Gama com parâmetros α = 5, β= 1; Gama com parâmetros α = 10, β= 1;



Observa-se, nos gráficos da Figura 1, que, para cada par de parâmetros α e β, a curva da distribuição Gama apresenta um formato específico. Enquanto as curvas das distribuições Gama (α=1; β=2), Gama (α=2; β=2) e Gama (α=5; β=1) são claramente assimétricas e se destoam consideravelmente das distribuições normais, a curva da Gama (α=10; β=1) é aproximadamente simétrica e semelhante a uma curva normal.

Consequentemente, espera-se que, caso os dados submetidos aos testes de normalidade sejam originados de distribuições Gama ou Gama Inversa simétricas, os testes de normalidade falhem mais, não rejeitando a hipótese nula de normalidade, visto que, as suas estatísticas calculadas poderão não ser capazes de permitir a diferenciação da distribuição normal teórica da distribuição Gama empírica, já que ambas são visivelmente parecidas.

1.8 O teste qui-quadrado de independência

O teste qui-quadrado de independência é construído com o propósito de estudar a relação de dependência (associação) entre duas variáveis categóricas. Pode ser empregado, por exemplo, para avaliar a relação entre determinada condição experimental, como a simetria (assimetria) da distribuição empírica e o nível de poder empírico (elevado ou baixo) apresentado por um teste de normalidade. Dessa maneira, para a realização desse teste, são elaboradas tabelas de contingência, preenchidas com as frequências observadas para cada combinação de categorias, sendo que na horizontal são definidas as categorias de uma variável aleatória e na vertical as categorias da outra. As hipóteses testadas são as seguintes:

H_0: As variáveis são independentes;

H_a: As variáveis não são independentes.

Então, para investigar a concordância entre as frequências observadas e as frequências esperadas, é utilizada a estatística da Equação 38, a qual, segundo Siegel e Castellan Jr. (2006) segue uma distribuição de qui-quadrado com $k = (h-1)(n-1)$ graus de liberdade, se as frequências esperadas puderem ser calculadas sem necessidade de se estimar parâmetros ou com $k = (h-1)(n-1) - r$ graus de liberdade, se r parâmetros populacionais forem estimados para se calcular as frequências esperadas.

$$\chi^2_{cal} = \sum_{i=1}^{h} \sum_{j=1}^{n} \frac{(F_{oij} - F_{eij})^2}{F_{eij}} \tag{38}$$

Em que, h é o total de níveis de uma variável aleatória; n é o total de níveis da outra variável aleatória; F_{oij} equivale à frequência observada para a combinação de níveis i e j das duas variáveis aleatórias; F_{eij} é a frequência esperada para a combinação de níveis i e j das duas variáveis aleatórias supondo independência entre elas.

Ressalta-se ainda que a frequência esperada (F_{eij}) para o teste qui-quadrado de independência é calculada conforme a Equação 39.

$$F_{eij} = \frac{(T_i)(T_j)}{hn} \quad (39)$$

Em que, T_i é a soma de frequências observadas do nível i de uma das variáveis aleatórias; T_j é a soma de frequências observadas do nível j da outra variável aleatória.

1.9 Estudos de comparações entre os testes de normalidade

Como já mencionado, a pressuposição de normalidade é importante para diversos métodos inferenciais, tais como aqueles que se baseiam no teste F, cuja estatística é calculada considerando amostragens sob distribuição normal. No entanto, como nem sempre é possível obter um número suficientemente grande de repetições por tratamento, considerando quão destoante é a distribuição empírica da distribuição teórica normal, não é sempre viável aplicar o Teorema Central do Limite com a finalidade de garantir a condição de normalidade. Por isso, os testes de normalidade se estabelecem como alternativas relativamente simples. Assim, antes de empregar o teste F, por exemplo, algum teste de normalidade é realizado com o objetivo de verificar se pode-se considerar que os erros experimentais são originados de uma distribuição normal. Contudo, há diversas opções de testes de normalidade e para escolher pelo menos um deles, o pesquisador precisa se amparar em comparações, que indiquem qual(quais) é(são) mais adequado(s) de acordo com as condições experimentais adotadas.

Na literatura, são encontrados diversos estudos comparativos entre os testes de normalidade, realizados sob diferentes condições, sendo os valores empíricos do nível de significância (probabilidade de rejeitar a hipótese nula de normalidade, sendo ela verdadeira) e do poder do teste (probabilidade de rejeitar a hipótese nula de normalidade quando ela é falsa) utilizados como medidas de efetividade. Contudo, esses estudos, em geral, apresentam uma limitação em comum. Na maioria deles, os experimentos não são simulados considerando um modelo estatístico adequado, como o Modelo Linear de Gauss-Markov Normal (MLGMN). Consequentemente, não são instalados sob delineamentos experimentais, como o Delineamento Inteiramente Casualizado (DIC). Isso impossibilita a análise da relação entre a igualdade (desigualdade) das médias de cada nível do fator em estudo, quanto à efetividade de cada teste. Impossibilita também a avaliação da relação entre a homogeneidade (heterogeneidade) de variâncias residuais dos níveis do fator em estudo quanto à efetividade dos testes, já que, em quase todos esses estudos, apenas um nível do fator é considerado.

Em seu estudo, Torman et al. (2012) obtiveram amostras de seis tamanhos distintos (10, 30, 50, 100, 500 e 1000) a partir de distribuições de probabilidade distintas (Normal padrão; Qui-quadrado com três graus de liberdade; t de Student com dez graus de liberdade; Gama com α=10 e β=1/3; Exponencial com β=1) e para cada tamanho amostral simularam 5000 amostras. Então, para cada uma das amostras nas diferentes condições, procederam-se aos testes de normalidade de Kolmogorov-Smirnov, Lilliefors, Cramér-von Mises, Shapiro-Wilk, dentre outros. Para analisar a eficiência desses testes, contabilizaram os percentuais de acertos, isto é, a rejeição correta de H_0 falsa, correspondente ao poder empírico do teste e a não rejeição de H_0 quando ela é verdadeira, que equivale à subtração do nível de significância empírico de 1.

Torman et al. (2012) concluíram que, para todas as amostras originadas de distribuições não normais, independentemente do tamanho amostral, o percentual de acertos do teste de Kolmogorov-Smirnov foi o menor. No entanto, quando as amostras foram obtidas a partir de populações normais, o teste de Kolmogorov-Smirnov foi o único que não rejeitou, para nenhuma amostra, a hipótese de normalidade, que é verdadeira, ou seja, apresentou uma eficiência de 100% nesse caso. Torman et al. (2012) pontuaram ainda que, de maneira geral, contabilizando o percentual de acertos em todas as condições de estudo, os testes de Shapiro-Francia e de Shapiro-Wilk foram os que apresentaram a maior efetividade com 72,41% e 72,15%, respectivamente. Já o pior resultado geral foi obtido pelo teste de Kolmogorov-Smirnov com 44,78% de acertos, percentual fortemente influenciado pelo desempenho muito baixo desse teste quando as amostras foram originadas da distribuição t de Student. No entanto, esse trabalho tem a limitação de não ter estudado a efetividade dos testes quando as distribuições apresentam outros parâmetros (apenas um tipo de distribuição normal, a padrão, foi testada, por exemplo) e quando as amostras originadas delas têm outros tamanhos mais compatíveis com os experimentos realizados na prática.

Em contraposição, no estudo realizado por Razali e Yap (2011), as amostras obtidas apresentaram 15 tamanhos distintos (10, 15, 20, 25, 30, 40, 50, 100, 200, 300, 400, 500, 1000, 1500 e 2000). As distribuições simétricas a partir das quais as amostragens foram realizadas foram: Uniforme no intervalo (0,1); Beta com α=2 e β=2; t de Student com trezentos, dez, sete e cinco graus de liberdade; Laplace com μ=0 e σ=1. Já as distribuições assimétricas alternativas foram: Beta com α=6 e β=2; Beta com α=2 e β=1; Qui-quadrado com 20 e com 4 graus de liberdade; Gamas com α=4 e β=5 e com α=1 e β=5.

Nesse estudo de Razali e Yap (2011), objetivando comparar o poder dos testes de Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling e Lilliefors, os níveis de significância adotados foram tanto α=0,05, quanto α=0,10. Para cada tamanho amostral foram, então, realizadas 50000 simulações, empregando compiladores FORTRAN. Assim como no estudo de Torman et al. (2012), Razali e Yap (2011) concluíram que, em geral, dentre os quatro testes considerados, o teste de Shapiro-Wilk apresentou os melhores resultados, ou seja, maiores poderes e o teste de Kolmogorov-Smirnov os menores poderes. Os autores demonstraram ainda que o teste de Lilliefors foi mais potente do que o de Kolmogorov-Smirnov em todas as situações e que o teste de Anderson-Darling teve resultados próximos dos de Shapiro-Wilk. Porém, todos os testes tiveram baixos poderes para amostras de tamanhos pequenos.

Ressalta-se, então, que Razali e Yap (2011) observaram que ao aumentar o tamanho amostral, o poder de todos os testes também aumentou progressivamente de maneira que, no caso da amostragem feita a partir da distribuição Gama assimétrica (α=4, β=5) para as amostras de tamanho 2000, todos os testes apresentaram um poder de 100%, ou seja, em todas as 50000 simulações, a hipótese falsa de normalidade foi rejeitada nessa situação. Já para as amostras advindas da distribuição qui-quadrado com quatro graus de liberdade, o poder de 100% em todos os testes ocorreu para amostras de tamanho 300.

Em comparação aos estudos anteriores, Farrell e Stewart (2006) adotaram alguns métodos mais categóricos para proceder às comparações entre os testes, apesar de terem considerado apenas tamanhos amostrais pequenos (10, 15, 20, 25 e 30). Um desses métodos categóricos foi a classificação objetiva das distribuições não-normais, baseando-se nos valores padronizados de assimetria e curtose delas. Essa categorização é justificada pelo fato de que Farrell e Stewart (2006) compararam os resultados dos testes de normalidade para uma variedade bem maior de distribuições não-normais (foram 48), a partir das quais as amostragens foram feitas. Assim, os autores classificaram as distribuições não-normais em simétricas de cauda curta, simétricas de cauda longa e assimétricas.

Além disso, mesmo que os valores críticos de alguns dos testes comparados estejam disponíveis na literatura, antes de realizar os testes de normalidade, Farrell e Stewart (2006) obtiveram esses valores empiricamente. Então, os valores críticos foram baseados em 1000000 de amostras simuladas de uma distribuição normal padrão. Desse modo, para obtê-los, a estatística do respectivo teste foi computada para cada uma das 1000000 amostras advindas da normal padrão. Essas estatísticas foram, a seguir, ordenadas para criar uma distribuição

empírica. Como os testes de Anderson-Darling e de Lilliefors são unilaterais à direita, para os níveis de significância estudados (α=0,05 e α=0,10), os seus valores críticos equivalem aos quantis de 0,90 e de 0,95 da distribuição empírica. Para os outros testes, unilaterais à esquerda, como o de Shapiro-Wilk, os valores críticos foram os correspondentes aos quantis de 0,05 e de 0,10.

Feito isso, Farrell e Stewart (2006) prosseguiram à simulação de um total de 10000 amostras para cada um dos cinco tamanhos amostrais estudados de cada uma das 48 distribuições não normais. Para o grupo de distribuições simétricas de calda curta, concluíram que o teste SHU de Spiegelhalter, especializado nesse tipo de distribuição, foi o que apresentou maior poder. Notou-se também que os testes de Shapiro-Wilk e de Lilliefors, em geral, não apresentaram elevado poder para essas distribuições, nas quais não são especializados, o que já era esperado, considerando os tamanhos amostrais pequenos. Já para as distribuições amostrais simétricas de cauda longa, como a Cauchy, em geral, os resultados dos testes de Shapiro-Wilk e de Lilliefors melhoraram, mesmo considerando os tamanhos amostrais limitados. Por fim, para as distribuições simétricas de cauda longa, o teste de Shapiro-Wilk foi um dos que apresentou resultados razoáveis.

Arnastauskaitė et al. (2021) realizaram um estudo ainda mais extenso a respeito do poder empírico de diferentes testes de normalidade do que o de Farrell e Stewart (2006), visto que além de compararem os resultados obtidos para bases de dados de diversas distribuições de probabilidade, consideraram 40 diferentes testes de normalidade.

Inclusive, diferentemente de Farrell e Stewart (2006), Arnastauskaitė et al. (2021) não compararam os resultados dos testes de normalidade apenas das amostras pequenas, pois obtiveram, por simulação, 1000000 amostras para cada um dos seis diversos tamanhos amostrais estudados (32, 64, 128, 256, 512 e 1024) provenientes de onze distribuições de probabilidade (Beta; Cauchy; Laplace; Logística; Student; Qui-quadrado; Gama; Gumbel; Lognormal; Weibull e Normal padrão modificada). O nível de significância adotado foi α=0,05. Concluíram que, para amostras grandes (n=1024), o poder empírico médio para os testes de Kolmogorov-Smirnov, Lilliefors, Cramér-von Mises e Shapiro-Wilk foi de, respectivamente, 0,939; 0,947; 0,949 e 0,962. Ou seja, mesmo com o teste de Shapiro-Wilk apresentando o melhor resultado, esses resultados não são muito discrepantes. Porém, para amostras pequenas (n=32), a discrepância é um pouco maior e os resultados anteriores, na mesma ordem, são 0,585; 0,669; 0,591 e 0,718. Assim, de acordo com os autores, percebe-se que alguns testes perderam

mais poder que os outros nesse último caso. Entretanto, o teste de Shapiro-Wilk continuou sendo o melhor.

Dessa maneira, Arnastauskaitė et al. (2021) perceberam ainda que, o poder empírico médio obtido para esses testes, em geral, foi maior quando as distribuições alternativas eram assimétricas. Nesses casos, para amostras de tamanho n=32, os testes de Kolmogorov-Smirnov, Lilliefors, Cramér-von Mises e Shapiro-Wilk resultaram, respectivamente, nos poderes empíricos médios 0,582; 0,671; 0,594 e 0,753 e, para as de tamanho grande (n=1024) os poderes aumentaram muito, respectivamente, para 0,945; 0,976; 0,957 e 0,991.

Diferentemente da metodologia anterior, Ogunleye et al. (2018) avaliaram também o erro tipo I empírico médio (nível de significância empírico médio), o qual, quanto mais próximo for do nível de significância teórico, melhor. Para computar essa medida, os autores simularam 5000 amostras aleatórias para cada tamanho amostral (10; 20; 30; 40; 50; 100; 200; 300; 400; 500; 1000) e aplicaram os testes de Anderson-Darling, Qui-quadrado, Kolmogorov-Smirnov e Shapiro-Wilk. Por fim, ranquearam, para cada tamanho amostral, os erros médios obtidos. Dessa maneira, atribuíram o ranque 1 ao teste com erro médio resultante mais próximo de α=0,05 e ranque 4 ao que mais se afastou. Para empates, a média das posições foi feita. Por fim, para cada teste e todos os tamanhos amostrais, os ranques foram somados e como o teste de Shapiro-Wilk resultou na menor soma (igual a 24), em geral, foi o mais efetivo. O segundo mais efetivo, nesse sentido, foi o teste de Kolmogorov-Smirnov (soma de 25,5). Os menos efetivos foram os de Qui-quadrado e Anderson-Darling (somas de 31,5 e 29).

Para comparar os poderes empíricos, Ogunleye et al. (2018) simularam 5.000 amostras das distribuições Uniforme (0;1), Beta (2;2), Gama (4;5), Binomial (5.000; 0,5) e Poisson (4). Novamente, concluíram que o teste de Shapiro-Wilk foi o melhor, já que apresentou maior poder tanto para amostras pequenas, quanto grandes. Inclusive, quando o tamanho amostral foi aumentado para 100, o poder empírico médio desse teste se aproximou de 100% e quando n=200, atingiu 100%. O segundo teste com maiores poderes, em geral, foi o de Anderson-Darling. Observaram também que os resultados dos quatro testes melhoraram quando a distribuição alternativa foi a Gama (4;5). Para comparar de maneira generalizada todas as distribuições contínuas estudadas, os autores recorreram ao ranqueamento dos testes, com o teste com maior poder recebendo ranque 1, para cada tamanho amostral. Assim, de maneira geral, o teste de Shapiro-Wilk foi o mais poderoso (menor somatório de ranques), quando a hipótese alternativa foi uma distribuição contínua e não normal.

Porém, por fim, para as distribuições alternativas discretas, Binomial e Poisson, as conclusões de Ogunleye et al. (2018) não foram idênticas às das contínuas, pois o teste Qui-quadrado superou os outros três testes em todos os tamanhos amostrais. O teste de Anderson-Darling foi o segundo mais poderoso, seguido pelo teste de Shapiro-Wilk, superando o teste de Kolmogorov-Smirnov.

Outro estudo semelhante é o de Doulah (2019), em que, a partir de 10.000 amostras simuladas no R versão 3.6.1, para cada tamanho amostral (10, 30, 50, 100, 200, 500 e 1000), foram comparados os poderes empíricos médios de 27 testes de normalidade. Essas amostras foram obtidas de populações com distribuições simétricas, Uniforme (0, 1), t (10), t (5), Beta (4; 4), Laplace (0; 1), Logística (2; 1), e assimétricas, Weibull (2; 3), Gompertz (10; 0,001), Gama (1; 5), Lognormal (0; 1), Exponencial (1) e Qui-quadrado (5). Concluiu-se que, para as distribuições simétricas Uniforme (0; 1), t (10) e Beta (4; 4), os melhores testes foram os de Agostino-Pearson, Jarque Bera e os de curtose. Para as demais distribuições simétricas, o teste robusto de Jarque Bera e os testes de Geary e Jarque Bera apresentaram os melhores resultados. Já, para as distribuições assimétricas Weibull (2; 3), Gompertz (10; 0,001) e Gama (1; 5), o teste de Shapiro-Wilk foi um dos melhores. Para as demais distribuições assimétricas, o teste de Shapiro-Wilk foi o mais poderoso.

Doulah (2019) pontua ainda que o teste de Shapiro-Wilk teve boa performance em todas as situações, excetuando-se as de amostras pequenas, caso em que nenhum teste teve resultados considerados bons. Além disso, os poderes médios dos testes de Lilliefors, Crámer-von Mises e Anderson-Darling também foram satisfatórios em grande parte das condições. Para a distribuição alternativa Uniforme (0, 1) com amostras de tamanho igual a 1.000, por exemplo, o poder empírico médio do teste de Lilliefors foi de 98,54% e os dos testes de Crámer-von Mises e Anderson-Darling foram ambos de 100%. Apesar de resultados próximos terem sido obtidos para as outras distribuições simétricas, esses três testes tiveram resultados superiores para as distribuições assimétricas, apresentando poder médio de 100% para o maior tamanho amostral (1.000).

Por fim, diante de tantos testes e condições, como se apresentou anteriormente, ficou evidente que uma das dúvidas que motivam tantos estudos de comparação entre testes de normalidade é de saber qual é a melhor opção. Essa escolha, porém, não deve se restringir aos poderes empíricos dos testes. Nesse sentido, segundo Miot (2017) há um consenso de que os testes de normalidade sofrem influência do tamanho amostral quanto à sua eficiência,

considerando que, em amostras pequenas (4 a 30 observações), as taxas de erro tipo I são inflacionadas, sendo os testes de Shapiro-Wilk e de Shapiro-Francia os preferidos, por serem mais capazes de controlá-las. Porém, quando se aumenta o tamanho das amostras, principalmente acima de 500 observações, o desempenho de todos os testes de normalidade quanto às taxas de erro tipo I tende a melhorar, com a ressalva de que, em compensação, as taxas de erro tipo II aumentam.

2. Referências Bibliográficas

ANDERSON, T. W., DARLING, D. A. **Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes**. **The Annals of Mathematical Statistics**. [S.l: s.n.], 1952.

ARNASTAUSKAITĖ, J., RUZGAS, T., BRAŽĖNAS, M. "An exhaustive power comparison of normality tests", **Mathematics**, v. 9, n. 7, 1 abr. 2021. DOI: 10.3390/math9070788.

BARBETTA, P. A., REIS, M. M., BORNIA, A. C. **Estatística: para cursos de engenharia e informática**. São Paulo, Atlas, 2004.

CASELLA, G., BERGER, R. L. **Statistical Inference**. 2nd. ed. [S.l.], Duxbury/Thomson Learning, 2002.

CAMPOS, Humberto. **Estatística Experimental Não-Paramétrica**. 2ª ed. Piracicaba, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, 1976.

DOULAH, S.U., Md. "A Comparison among Twenty-Seven Normality Tests", **STM Journals**, v. 8, n. 3, p. 41–59, 2019.

FARRELL, P., STEWART, K.R. "Comprehensive study of tests for normality and symmetry: Extending the Spiegelhalter test", **Journal of Statistical Computation and Simulation**, v. 76, n. 9, p. 803–816, 1 set. 2006. DOI: 10.1080/10629360500109023.

GELMAN, A., CARLIN, J. B., STERN, H. S., *et al.* **Bayesian Data Analysis**. Third ed. [S.l.], CRC Press, 2013.

MIOT, H. A. **Avaliação da normalidade dos dados em estudos clínicos e experimentais**. **Jornal Vascular Brasileiro**. [S.l.], Sociedade Brasileira de Angiologia e Cirurgia Vascular., 2017.

MOOD, A. M. **INTRODUCTION TO THE THEORY OF STATISTICS**. 3. ed. [S.l.], McGraw-Hill, Inc., 1974.

OGUNLEYE, L. I., OYEJOLA, B. A., OBISESAN, K. O. "Comparison of Some Common Tests for Normality", **International Journal of Probability and Statistics**, v. 7, n. 5, p. 130–137, 2018.

RAZALI, N. M., YAP, B.W. **Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests**. [S.l: s.n.], 2011. v. 2.

SEARLE, S. R., GRUBER, M. H. J. **Linear Models**. 2nd. ed. [S.l.], Wiley Series in Probability and Statistics, 2016.

SHAPIRO, S. S., WILK, M B. **An Analysis of Variance Test for Normality (Complete Samples)**. **Biometrika**. [S.l: s.n.], 1965.

SIEGEL, S., CASTELLAN JR., N. J. **ESTATÍSTICA NÃO-PARAMÉTRICA PARA CIÊNCIAS DO COMPORTAMENTO**. 2.ed. ed. Porto Alegre, RS., ARTMED EDITORA S.A., 2006.

THADEWALD, T., BÜNING, H. "Jarque-Bera Test and its Competitors for Testing Normality - A Power Comparison", **Journal of Applied Statistics**, v. 34, n. 1, p. 87–105, 2007.

TORMAN, V. B. L., COSTER, R., RIBOLDI, J. "Normalidade de variáveis: métodos de verificação e comparação de alguns testes não-paramétricos por simulação", **Revista Clinical & Biomedical Research**, v. 32, n. 2, 2012.

YAP, B. W., SIM, C. H. "Comparisons of various types of normality tests", **Journal of Statistical Computation and Simulation**, v. 81, n. 12, p. 2141–2155, dez. 2011. DOI: 10.1080/00949655.2010.520163.

Capítulo 3: Avaliação do poder empírico dos testes F e de normalidade sob diferentes condições experimentais

1. Resumo

A pressuposição de normalidade é importante para métodos inferenciais como o teste F da análise de variância (ANOVA), muito empregada nas Ciências Agrárias. Por isso, os testes de normalidade são aplicados para atestar a normalidade dos erros experimentais. Porém, os estudos já existentes, que objetivam avaliar qual(quais) o(s) teste(s), que apresentaram maiores poderes, em geral, não aplicam nenhum delineamento experimental. Assim, não são capazes de analisar o efeito de diferentes condições experimentais nos poderes. Com esse objetivo, neste trabalho, aplicando o Delineamento Inteiramente Casualizado (DIC), avaliou-se o efeito da simetria(assimetria) das distribuições empíricas da variável resposta para cada tratamento, da igualdade(desigualdade) das suas médias e da homogeneidade(heterogeneidade) de suas variâncias no poder empírico dos testes de normalidade e do teste F. Para isso, aplicaram-se os testes de normalidade aos conjuntos de resíduos experimentais de cada uma das 10 mil iterações simuladas e o teste F aos conjuntos de valores simulados da variável resposta. Concluiu-se que, na maioria dos cenários, o poder aumentou com o aumento do número de repetições por tratamento, que a condição de simetria diminui o poder dos testes de normalidade, que testes como os de Shapiro-Wilk e de Anderson-Darling são, em geral, mais poderosos do que o teste

de Kolmogorov-Smirnov e que o teste F é extremamente robusto à normalidade, sendo, entretanto, levemente afetado pela violação da condição de homogeneidade de variâncias, perdendo poder.

Palavras-chave: Análise de Variância (ANOVA); Delineamento Inteiramente Casualizado (DIC); Distribuição Normal; Erros Experimentais; Anderson-Darling; Cramér-von Mises; Kolmogorov-Smirnov; Lilliefors; Shapiro-Wilk.

2. Introdução

É de conhecimento geral que os estudos estatísticos buscam encontrar respostas, padrões e conclusões a partir de análise de dados. Nesse sentido, quando esses dados são numéricos, para interpretá-los, uma das alternativas é avaliar qual a distribuição de probabilidades, que seguem. As distribuições de probabilidades normais são algumas das mais incidentes, o que pode ser explicado pelo Teorema Central do Limite, visto que partindo-se praticamente de qualquer distribuição não normal, é possível que a distribuição de probabilidade da média amostral se aproxime de uma distribuição normal. As únicas exigências, de acordo com Mood (1974), são de que a distribuição não normal tenha variância finita e que a média amostral seja calculada a partir de um tamanho amostral suficientemente grande.

Porém, a grande incidência das distribuições normais não é o único fator, que as destacam. Muitos procedimentos inferenciais, como a análise de variância (ANOVA), pressupõem a normalidade dos erros experimentais, principal motivo pelo qual as distribuições normais são tão importantes. Nesse caso, a razão pela qual a ANOVA pressupõe a normalidade diz respeito ao teste F, cuja distribuição de probabilidades, a distribuição F, de Fisher-Snedecor, é construída tendo como base a estatística F, a razão entre dois estimadores de variância (entre tratamentos no numerador e, dentro dos tratamentos, no denominador), cada um seguindo uma distribuição conhecida de qui-quadrado.

Assim, como esses estimadores são calculados aplicando somas de quadrados de variáveis aleatórias e, de acordo com Mood (1974), pode-se provar facilmente que os somatórios de quadrados de variáveis, que seguem distribuição normal, seguem distribuição qui-quadrado, é possível compreender o quão relevante é a pressuposição de normalidade, visto que não é possível garantir que os somatórios dos quadrados de variáveis não normais seguirão distribuição qui-quadrado e, consequentemente, não é possível afirmar que a razão entre eles seguirá a distribuição F, a partir da qual, as decisões a respeito das hipóteses do teste F são tomadas.

Portanto, por serem pressuposições de vários procedimentos inferenciais, como o teste F, segundo Casella e Berger (2002), as distribuições normais se destacam principalmente, pela sua tratabilidade analítica. Além disso, algumas das outras vantagens das distribuições normais são decorrentes do fato de que suas curvas são completamente descritas com apenas dois parâmetros, a média (μ) e a variância (σ^2), e independentemente desses parâmetros, sempre são curvas simétricas em torno da média. Ressalta-se ainda que, variáveis obtidas por meio da subtração ou da soma de outras variáveis aleatórias normais e independentes, também seguem distribuição normal, caso da média amostral.

Em contraposição, outras distribuições de probabilidades não usufruem nem das mesmas vantagens das distribuições normais, nem da mesma tratabilidade analítica. Conforme Gelman et al. (2013), para que seja descrita a curva de densidade de probabilidade de uma variável, que segue distribuição Gama Inversa, por exemplo, são necessários quatro parâmetros. Além da média ou esperança matemática e da variância, são necessários os parâmetros α e β, maiores do que zero. E, apesar de ser possível que a curva de uma distribuição Gama Inversa também seja simétrica e muito parecida com a de uma distribuição normal, nem sempre isso acontece, visto que, a depender dos parâmetros α e β, o formato da curva pode mudar totalmente tornando-se extremamente assimétrica. Além disso, não é comum encontrar procedimentos inferenciais, que precisem pressupor que os dados seguem distribuição Gama Inversa.

Pelo contrário, como a normalidade é um dos pressupostos de diversos procedimentos inferenciais, para verificar se ele é atendido, podem ser aplicados testes de aderência de normalidade, como os de Kolmogorov-Smirnov (KS), Lilliefors (LI), Cramér-von Mises (CVM) e Anderson-Darling (AD).

Nesses testes, são comparadas uma distribuição teórica normal com a distribuição empírica dos dados. Quanto maior a diferença observada entre essas distribuições acumuladas, menor a probabilidade de os dados seguirem uma distribuição normal, como atesta Anderson e Darling (1952).

Porém, podem também ser aplicados testes de normalidade, que se baseiam em outras metodologias de comparação para se tomar a decisão de rejeitar ou não a hipótese de normalidade, como o teste de Shapiro e Wilk (1965) (SW). Esse teste se baseia na comparação de dois estimadores de variância, de maneira que a discrepância entre eles permita inferir se os dados seguem uma distribuição normal.

Posto isso, com tantas opções de testes de normalidade, questiona-se qual ou quais deles são mais poderosos. Para responder a essa pergunta, o pesquisador deve definir os métodos inferenciais a partir dos quais tomará decisões e estar ciente das condições experimentais, sob as quais os resultados são obtidos.

Nesse contexto, considerando as Ciências Agrárias, é extremamente comum a realização de Análises de Variância (ANOVA), baseadas no teste F, o qual pressupõe que os erros são independentes, seguem uma distribuição normal e apresentam variâncias homogêneas. De acordo com estudos como o de Nguyen et al. (2019), o teste F é consideravelmente sensível à violação da pressuposição de homogeneidade das variâncias. No entanto, não está claro como a violação da pressuposição de normalidade afeta o poder desse teste, quando está atrelado a um delineamento experimental.

Estudos anteriores desenvolvidos por Uyanto (2022), Arnastauskaitė et al. (2021), Islam (2021) e Doulah (2019) envolvendo a comparação de alguns testes de normalidade apresentaram conclusões compatíveis entre si. Contudo, foram realizados com tamanhos amostrais simulados, em geral, muito maiores do que as adotados em delineamentos utilizados em experimentos aplicados às Ciências Agrárias.

No estudo de Uyanto (2022), por exemplo, praticamente todos os testes de normalidade comparados apresentaram uma tendência clara de aumento do poder com o aumento do tamanho amostral, apresentando os maiores poderes apenas para o maior tamanho amostral (100). Concluiu-se ainda que o teste KS foi um dos mais conservadores da hipótese de normalidade, tanto quando a distribuição empírica dos dados era simétrica, quanto quando era assimétrica e que, de maneira geral, os testes SW, AD, CVM e LI foram, respectivamente, em ordem decrescente, mais poderosos.

Arnastauskaitė et al. (2021) obtiveram resultados semelhantes aos de Uyanto (2022) com os maiores poderes sendo obtidos para o maior tamanho amostral (1.024) e com o teste SW sendo o mais poderoso. Conclusões essas compatíveis às de Doulah (2019), cujos resultados apontam para os maiores poderes obtidos para um tamanho amostral igual a 1.000, para o teste SW sendo um dos mais poderosos e os testes AD, CVM e LI apresentando resultados satisfatórios para os maiores tamanhos amostrais.

Portanto, o presente trabalho tem por objetivo geral avaliar se esses testes de normalidade apresentam resultados satisfatórios quando aplicados a tamanhos amostrais compatíveis com a realidade de experimentos agrícolas. Os objetivos específicos são de avaliar

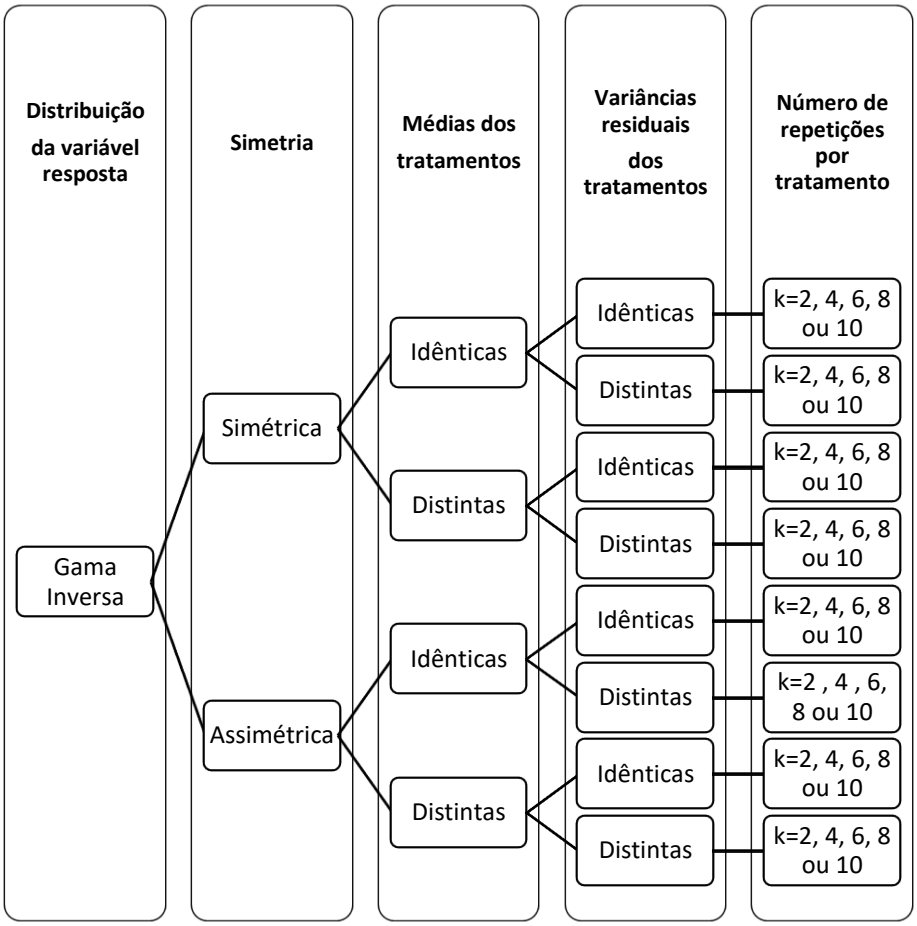
o efeito que a simetria (assimetria) das distribuições empíricas não normais, a igualdade (desigualdade) das suas médias e a homogeneidade (heterogeneidade) de suas variâncias têm no poder empírico dos testes de normalidade e do teste F.

3. Metodologia

Foram simulados diferentes subcenários para a obtenção dos valores de uma variável resposta qualquer, considerando diferentes distribuições de probabilidades gama inversa, oriundos de um experimento instalado segundo um Delineamento Inteiramente Casualizado (DIC) com cinco tratamentos e k repetições.

Os cenários se diferenciaram quanto à ocorrência ou não de simetria da gama inversa (distribuições Gama Inversa aproximadamente simétricas foram consideradas simétricas para efeitos de estudo), igualdade ou não entre as médias dos tratamentos, homogeneidade ou não dentro dos tratamentos. Por outro lado, os subcenários de cada cenário se diferenciaram quanto ao número k de repetições por tratamento conforme organograma da Figura 1, que representa um Fatorial $2 \times 2 \times 2 \times 5$.

Figura 1- Organograma de formação de cada padrão de cenário e subcenário, sob o DIC, definido por uma distribuição de probabilidade gama considerada simétrica ou assimétrica da variável resposta, condição de igualdade ou não das médias dos tratamentos, de homogeneidade ou não de variâncias dentro dos tratamentos e pelo número k de repetições por tratamento (Fatorial $2 \times 2 \times 2 \times 5$).



A partir da combinação das médias dos tratamentos e das variâncias dentro de tratamentos de acordo com a Tabela 1, foi possível criar oito cenários (C1, C2, C3, C4, C5, C6, C7 e C8). Foram geradas 10.000 iterações para cada um dos cinco números k de repetições por tratamento, definidos como subcenários de cada um desses cenários. Portanto, foram avaliados 40 subcenários. Os valores escolhidos dos parâmetros média e desvio padrão dos tratamentos foram os que permitiram a obtenção de parâmetros α e β, que tornaram as curvas das gamas inversas simétricas ou assimétricas. Além disso, nos cenários com desigualdade de parâmetros, optou-se pela mesma progressão, tornando o segundo valor, o dobro do primeiro, o terceiro o triplo do primeiro, assim sucessivamente, com o objetivo de padronizar as comparações.

Tabela 1-Valores das médias e dos desvios padrão estabelecidos para a simulação dos cenários considerando diferentes distribuições Gama Inversa.

Gama Inversa	Média	Desvio padrão	Cenário
Assimétrica	$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = 0{,}25$	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5 = 0{,}144.$	C1
		$\sigma_1 = 0{,}144; \sigma_2 = 0{,}288; \sigma_3 = 0{,}432; \sigma_4 = 0{,}576; \sigma_5 = 0{,}720$	C2
	$\mu_1 = 0{,}25; \mu_2 = 0{,}50; \mu_3 = 0{,}75; \mu_4 = 1{,}00; \mu_5 = 1{,}25$	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5 = 0{,}144.$	C3
		$\sigma_1 = 0{,}144; \sigma_2 = 0{,}288; \sigma_3 = 0{,}432; \sigma_4 = 0{,}576; \sigma_5 = 0{,}720$	C4
Simétrica	$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = 100$	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5 = 1.$	C5
		$\sigma_1 = 1; \sigma_2 = 2; \sigma_3 = 3; \sigma_4 = 4; \sigma_5 = 5$	C6
	$\mu_1 = 100; \mu_2 = 200; \mu_3 = 300; \mu_4 = 400; \mu_5 = 500$	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5 = 1.$	C7
		$\sigma_1 = 1; \sigma_2 = 2; \sigma_3 = 3; \sigma_4 = 4; \sigma_5 = 5$	C8

Considerando o modelo estatístico do DIC, para cada um dos 8 cenários foram simulados cinco subcenários, sendo cada subcenário caracterizado por um número k=2, 4, 6, 8 e 10 de repetições por tratamento. Para cada um desses subcenários, 10.000 iterações foram simuladas. Para cada iteração w, tal que $w = 1, 2, \cdots, 10.000$, foi obtido um conjunto w de 5k resíduos, conforme a Equação 1, em que $\hat{\varepsilon}_{ijw}$ é o resíduo obtido para o valor observado y_{ijw} da variável resposta na iteração w para a repetição j do tratamento i:

$$\hat{\varepsilon}_{ijw} = y_{ijw} - \hat{\mu}_{iw} \quad (1)$$

Os testes de normalidade (KS, LI, CVM, AD e SW) foram, então, aplicados separadamente aos conjuntos w de resíduos, compostos por $5k$ resíduos cada um.

Considerando que a distribuição de probabilidades dos erros experimentais não é normal e um nível de 5% de significância, o poder de cada um dos testes de normalidade em rejeitar a hipótese de normalidade foi estimado a partir de uma medida doravante denominada de poder empírico do teste ($\hat{P}$) calculada conforme estabelecido na Equação 2 a partir do p-valor de cada um dos testes de normalidade aplicados a cada um dos 10.000 conjuntos de resíduos sendo:

$$\hat{P} = \frac{número\ de\ p - valores \leq 0{,}05}{10.000} \quad (2)$$

Diferentemente dos testes de normalidade, os testes F foram aplicados aos valores da variável resposta dos tratamentos, não aos resíduos, e apenas nos cenários com desigualdade de médias. Entretanto, seus poderes empíricos ($\hat{P}$) também foram calculados por meio da Equação 2, considerando como hipótese de nulidade, a igualdade das médias dos 5 tratamentos e um nível de 5% de significância.

Com o objetivo de qualificar os testes de normalidade e o teste F, quanto ao poder, cada um foi classificado como:

- Não poderoso, se $\hat{P} < 0{,}75$;
- Poderoso, se $\hat{P} \geq 0{,}75$.

O desejável seria que o limiar utilizado nessa classificação fosse maior que 0,75 ou mesmo que se aproximasse do valor ideal igual a 1. Mas, para isso, seria necessário utilizar um número bem superior às 10.000 iterações utilizadas nesse estudo. Contudo, não foi possível utilizar um número maior de iterações devido às limitações do hardware utilizado.

Para avaliar a independência entre os níveis de poder de um teste e os fatores experimentais estudados (igualdade ou não de médias, homogeneidade (heterogeneidade) de variâncias e simetria (assimetria) da distribuição gama inversa), isto é, para avaliar como as condições experimentais categóricas afetam os níveis de poder, foram aplicados testes de qui-quadrado para independência, com estatísticas calculadas, de acordo com Siegel e Castellan Jr. (2006).

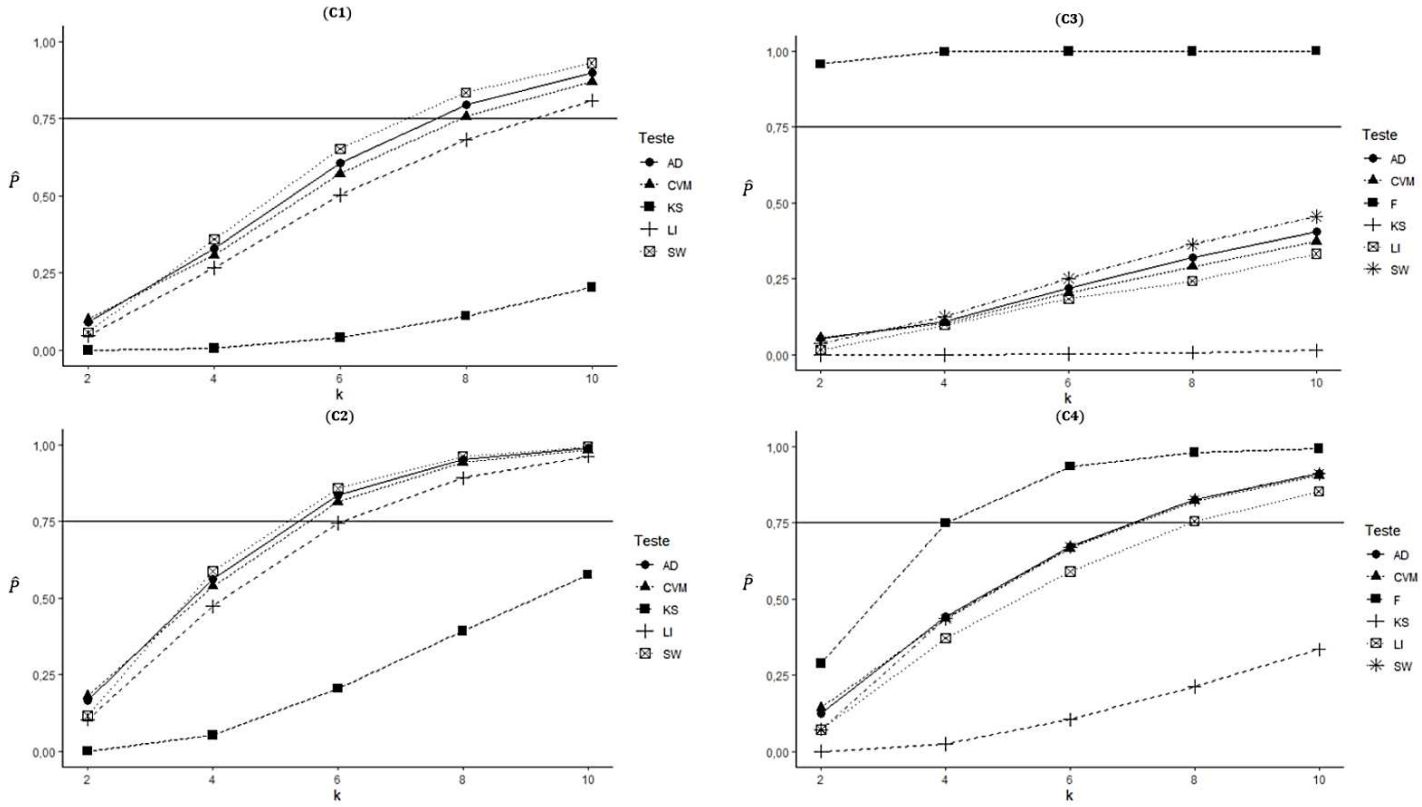
Para a simulação das iterações bem como para as aplicações dos testes de normalidade e F foi utilizado o RStudio versão 4.0.2 (R Core Team, 2020).

4. Resultados e Discussão

4.1 Resultados

O poder empírico de cada um dos testes de normalidade e do teste F (nos cenários com desigualdade de médias de tratamentos), em função do número de repetições por tratamento, obtido a partir das iterações simuladas para os cenários C1, C2, C3 e C4 (Tabela 1) é apresentado na Figura 2. Nessa figura, a reta horizontal, interceptando o poder empírico ($\hat{P}$) em 0,75 indica o limiar, acima do qual, considerou-se um teste como poderoso.

Figura 2- Poder empírico ($\hat{P}$) dos testes de Anderson-Darling (AD), Kolmogorov-Smirnov (KS), Crámer-von Mises (CVM), Lilliefors (LI) e Shapiro-Wilk (SW) e do teste F, em função do número de repetições (k) por tratamento, aplicados, respectivamente, aos resíduos desses tratamentos e aos valores observados da variável resposta, os quais advêm das iterações simuladas dos subcenários dos cenários C1, C2, C3 e C4, todos relativos a distribuições empíricas gamas inversas assimétricas. Nos gráficos com os resultados de cada cenário avaliado, a reta horizontal $\hat{P}$=0,75 indica o limiar, acima do qual, o teste é considerado poderoso.



Percebe-se, na Figura 2, que, sob distribuições assimétricas com médias idênticas e desvios padrão também idênticos (C1), o poder empírico de todos os testes aumentou linearmente com o aumento do número de repetições por tratamento. Essa tendência de aumento do poder empírico dos testes de normalidade com o aumento do número de observações também foi constatada em estudos, como os de Arnastauskaitė et al. (2021), Razali e Yap (2011), Ogunleye et al. (2018) e de Doulah (2019).

Ressalta-se ainda que o teste que mais se distanciou dos demais e que apresentou os menores poderes empíricos em cada uma das k repetições, foi o teste KS, comportando-se de

maneira extremamente conservadora em relação à hipótese de normalidade. Essa constatação a respeito do teste KS é compatível com as conclusões obtidas por Torman et al. (2012), que em seu estudo constataram que o teste KS foi o menos poderoso de todos os testes, enquanto o teste SW foi o mais poderoso.

Em contraposição aos poderes do teste KS, o presente trabalho permite também constatar, a partir dos resultados do cenário C1 na Figura 2, que os maiores poderes empíricos foram observados para o maior número de repetições (k=10). Nessa condição, o teste SW se destaca como o mais poderoso, seguido, respectivamente, pelos testes AD, CVM, LI e, por último, KS. Além disso, apenas foram atingidos valores satisfatórios para o poder empírico ($\hat{P}$>0,75) para um total de observações igual ou superior a 40 (k=8). Para k=2 (10 observações), k=4(20 observações) e k=6 (30 observações) nenhum teste foi classificado como poderoso segundo o limiar de poder empírico adotado.

A comparação entre os poderes empíricos apresentados nos gráficos dos cenários C1 e C3 (Figura 2), que se diferem apenas na igualdade ou não das médias dos tratamentos, permite concluir que, sob distribuição assimétrica e tratamentos com médias diferentes e desvios padrão idênticos (C3), todos os testes de normalidade apresentaram quedas expressivas em seus poderes empíricos, para praticamente todos os números de repetições (k). Nesses subcenários, em que as médias são diferentes, nenhum teste de normalidade pôde ser considerado poderoso ($\hat{P}$≥0,75), nem mesmo quando foram utilizadas 10 repetições (50 valores observados) no experimento. Contudo, apesar da aparente diferença de padrões observadas entre os resultados dos cenários C1 e C3, de acordo com o teste de qui-quadrado (Tabela 2), os níveis de poder dos testes de normalidade independem da igualdade das médias dos tratamentos para distribuições assimétricas e variâncias homogêneas.

Tabela 2- Resultados dos testes qui-quadrado de independência relacionando os níveis de poder empírico (alto para $\hat{P}$≥0,75 e baixo para $\hat{P}$<0,75) apresentados pelos testes de normalidade de Anderson-Darling (AD), Kolmogorov-Smirnov (KS), Crámer-von Mises (CVM), Lilliefors (LI) e Shapiro-Wilk (SW) com as condições de igualdade (ou não) de médias dos tratamentos, de acordo com o banco de dados dos cenários C1 e C3.

Teste	χ^2
AD	$2{,}50^{ns}$
KS	-
CVM	$2{,}50^{ns}$
LI	$1{,}11^{ns}$
SW	$2{,}50^{ns}$

^{ns}: p-valor>0,05.

Diferentemente dos subcenários do cenário C1, como os subcenários C3 foram caracterizados pela desigualdade das médias dos tratamentos, também foi possível calcular os poderes empíricos do teste F. Assim, na Figura 2, pode-se também observar que, considerando o cenário C3, o teste F foi poderoso ($\widehat{P}$≥0,75) mesmo para um número pequeno de repetições por tratamento (k=2). Esse resultado indica que, sendo as variâncias homogêneas, o teste F é capaz de detectar diferenças entre as médias mesmo quando a pressuposição de normalidade não é satisfeita.

Quanto aos resultados dos subcenários do cenário C2, percebe-se, na Figura 2, que, sendo as médias dos tratamentos idênticas e os desvios padrão distintos, quando k≥6, praticamente todos os testes de normalidade, com exceção do KS, foram poderosos ($\hat{P}$≥0,75). Assim como se observou para os subcenários C1 e C3, foi também observada, para C2, uma tendência de aumento dos poderes empíricos dos testes de normalidade em decorrência do aumento do número de repetições por tratamento.

Porém, observa-se, para os subcenários C2, um relativo aumento dos poderes empíricos dos testes de normalidade em comparação com C1. Contudo, esse aumento não foi tão relevante ao ponto de ser possível afirmar que os níveis de poder aumentam quando os desvios padrão dos tratamentos não são idênticos, o que pode ser observado por meio dos resultados dos testes qui-quadrado para os testes de normalidade (Tabela 3). Não há, portanto, uma relação significativa de dependência entre os níveis de poder e a homogeneidade (heterogeneidade) das

variâncias quando se trata das distribuições assimétricas com médias idênticas dos tratamentos (cenários C1 e C2).

Tabela 3- Resultados dos testes qui-quadrado de independência relacionando os níveis de poder empírico (alto para $\hat{P}\geq 0{,}75$ e baixo para $\hat{P}<0{,}75$) apresentados pelos testes de normalidade de Anderson-Darling (AD), Kolmogorov-Smirnov (KS), Crámer-von Mises (CVM), Lilliefors (LI) e Shapiro-Wilk (SW) com as condições de homogeneidade (ou heterogeneidade) das variâncias dos tratamentos, de acordo com o banco de dados dos cenários C1 e C2.

Teste	χ^2
AD	$0{,}40^{ns}$
KS	-
CVM	$0{,}40^{ns}$
LI	$0{,}48^{ns}$
SW	$0{,}40^{ns}$

^{ns}: p-valor>0,05.

Por fim, para distribuições gamas inversas assimétricas, também foram apresentados, na Figura 2, os poderes empíricos dos subcenários do cenário C4, isto é, com médias de tratamentos distintas e desvios padrão também distintos. Os resultados plotados mostram que, nesses subcenários, também há uma tendência de aumento do poder empírico de todos os testes de normalidade (AD, CVM, LI, SW e KS) com o aumento do número de repetições (k) e de observações. Porém, novamente, o teste KS se destoa dos demais, por apresentar, para todos os valores de k, poderes empíricos mais baixos. Para k≥8, todos os testes de normalidade foram poderosos, exceto o KS, sendo perceptível o empate entre os mais poderosos (SW, AD e CVM).

Além disso, sob distribuições assimétricas com médias diferentes, o poder empírico do Teste F parece tender a ser menor no cenário C4 (com heterogeneidade de variâncias) do que no cenário C3 (com homogeneidade de variâncias), o que faz sentido considerando que, para calcular a estatística F, o estimador da variância do resíduo, no denominador, considera uma média ponderada das estimativas de variância dentro de cada tratamento. Assim, se essas estimativas forem destoantes, o estimador comum da variância do resíduo tende a não ser representativo. No entanto, o teste qui-quadrado de independência para esses cenários indica que não há uma relação de dependência significativa entre a homogeneidade (heterogeneidade) de variâncias dos tratamentos e os níveis de poder do teste F ($\chi^2=2{,}50^{ns}$[1]). Isso pode ser explicado pelo fato de que apenas quando k=2, nos subcenários C4, o poder empírico ficou

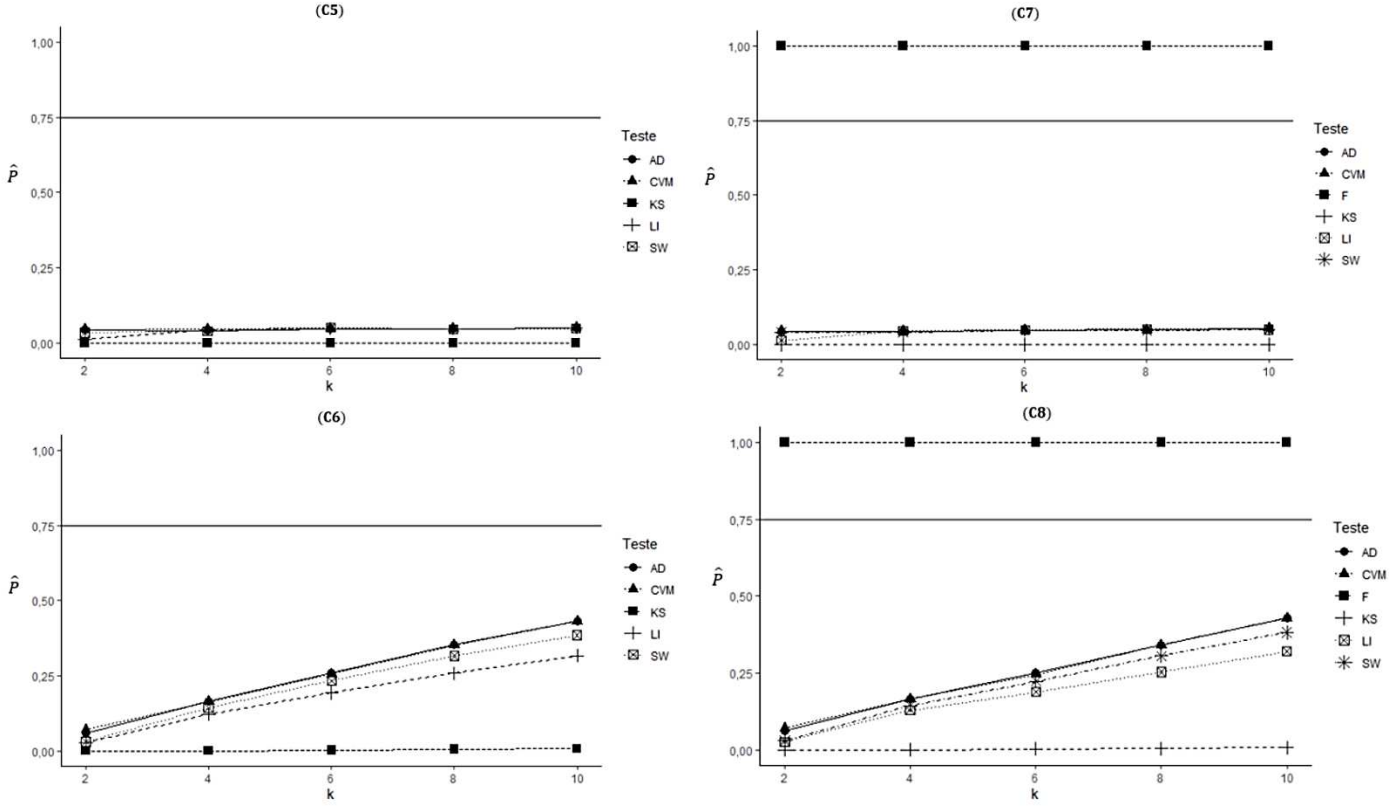
[1] ns: p-valor>0,05

visivelmente abaixo do limiar de 0,75 (Figura 2). Portanto, é inegável a capacidade do teste F de detectar diferenças entre as médias dos tratamentos, mesmo quando a pressuposição de normalidade não é satisfeita. Posto isso, conclui-se que são raras as situações em que o teste F não é poderoso e, caso k>4, independentemente da homogeneidade das variâncias, para os cenários de distribuições assimétricas, é muito improvável que não se atinja $\hat{P} \geq 0{,}75$.

Por conseguinte, as análises de todos os cenários anteriores, com valores observados e resíduos advindos de distribuições Gama Inversa assimétricas, permitem concluir que, para todos os testes de normalidade e até para o teste F no cenário C4, há uma tendência observável de aumento do poder empírico com o aumento do número k de repetições e do total de observações, como já havia sido confirmado por Arnastauskaitė et al. (2021) e Ogunleye et al. (2018). No entanto, ainda é necessário avaliar individualmente e visualmente o comportamento dos poderes empíricos dos testes de normalidade e do teste F em subcenários simulados com distribuições gamas inversas simétricas. Assim, um dos questionamentos ainda não respondidos e que será respondido é se a condição de simetria ou de assimetria da distribuição de probabilidade dos dados influencia no nível de poder dos testes de normalidade.

Dessa maneira, na Figura 3, é apresentado o poder empírico de cada um dos testes de normalidade e do teste F (nos cenários com desigualdade de médias de tratamentos), em função do número de repetições por tratamento, obtido a partir das iterações simuladas para os subcenários dos cenários C5, C6, C7 e C8 (Tabela 1).

Figura 3- Poder empírico ($\hat{P}$) dos testes de Anderson-Darling (AD), Kolmogorov-Smirnov (KS), Crámer-von Mises (CVM), Lilliefors (LI) e Shapiro-Wilk (SW) e do teste F, em função do número de repetições (k) por tratamento, aplicados, respectivamente, aos resíduos desses tratamentos e aos valores observados da variável resposta, os quais advêm das iterações simuladas dos subcenários dos cenários C5, C6, C7 e C8, todos relativos a distribuições empíricas gamas inversas simétricas. Nos gráficos com os resultados de cada cenário avaliado, a reta horizontal $\hat{P}$=0,75 indica o limiar, acima do qual, o teste é considerado poderoso.



Os resultados (Figura 3) indicam que, sob distribuições simétricas com médias idênticas e desvios padrão também idênticos (C5), independentemente do número de repetições, todos os testes de normalidade não foram poderosos ($\hat{P}$<0,75), apresentando poderes empíricos extremamente baixos, estáveis e próximos de zero ou iguais a zero, resultados totalmente diferentes dos observados para distribuições assimétricas, o que já era esperado considerando que as distribuições gamas inversas simétricas se aproximam de distribuições normais. Nesse sentido, o teste KS apresentou poderes empíricos ainda mais baixos do que todos os outros,

sendo exatamente iguais a zero para todos os números de repetição (k) avaliados. Em comparação aos cenários com distribuições assimétricas (Figura 2), fica evidente a queda generalizada dos poderes empíricos para distribuições gamas inversas simétricas (Figura 3). Portanto, conclui-se que, com base nos resultados anteriores, os testes de normalidade são muito sensíveis à ocorrência de simetria na distribuição de probabilidades dos erros experimentais, fato esse corroborado pelos testes qui-quadrado de independência entre o nível de poder empírico dos testes de normalidade e a simetria das distribuições simuladas (Tabela 4), que geraram estatísticas significativas.

Tabela 4- Resultados dos testes qui-quadrado de independência relacionando os níveis de poder empírico (alto para $\hat{P}\geq 0{,}75$ e baixo para $\hat{P}<0{,}75$) apresentados pelos testes de Anderson-Darling (AD), Kolmogorov-Smirnov (KS), Crámer-von Mises (CVM), Lilliefors (LI), Shapiro-Wilk (SW) e pelo teste F (médias diferentes), com as condições de simetria (ou não) da distribuição de probabilidade dos valores observados considerando todos os cenários

Teste	χ^2
AD	$6{,}23^{*}$
KS	-
CVM	$6{,}23^{*}$
LI	$3{,}66^{*}$
SW	$6{,}23^{*}$
F	$2{,}22^{ns}$

^{ns}: p-valor>0,05; *: p-valor≤0,05.

Foram também avaliados (Figura 3) os poderes empíricos dos testes de normalidade e do teste F a partir de subcenários simulados com distribuições gamas inversas simétricas com médias distintas e desvios padrão idênticos, ou seja, os subcenários do cenário C7 (Tabela 1).

Sob distribuições simétricas e variâncias homogêneas, ao comparar os cenários com médias idênticas (C5) com os de médias diferentes (C7) observa-se que quatro dos cinco testes de normalidade continuaram apresentando baixos poderes empíricos, pouca variação com a mudança do número k de repetições por tratamento e sobreposição da maior parte dos resultados, enquanto o teste KS continuou apresentando poderes empíricos ainda mais baixos do que os demais, corroborando a conclusão de que os testes de normalidade são muito sensíveis e prejudicados pela condição de simetria das distribuições dos dados, conforme os resultados dos testes qui-quadrado de independência (Tabela 4).

Em contraposição, especificamente para os subcenários de C7, o teste F apresentou poderes empíricos iguais a 1 para todos os valores de k, indicando que, em todas as respectivas simulações, pelo teste F, a hipótese de nulidade de igualdade das médias dos tratamentos foi corretamente rejeitada. A tendência observada para o poder empírico do teste F (Figura 3) indica que o poder do teste F independe da condição de simetria mesmo que os testes de normalidade não apresentem poder suficiente para detectar a falta de normalidade, o que também é confirmado pelo teste qui-quadrado (Tabela 4). Essa confirmação corrobora, inclusive, o que já evidenciava a comparação dos resultados do teste F para os cenários simulados de gamas inversas assimétricas com médias diferentes e desvios padrão iguais (C3 da Figura 2) com os simulados de gamas inversas simétricas com médias diferentes e desvios padrão iguais (C7 da Figura 3), pois, em ambos os casos, os poderes empíricos do teste F foram muito próximos de 1.

Ademais, com o objetivo de avaliar também a influência da heterogeneidade de variâncias dos tratamentos nos poderes empíricos dos testes de normalidade, observam-se os resultados das simulações dos subcenários de C6 (Figura 3), com valores da variável resposta e resíduos decorrentes de distribuições simétricas com médias idênticas e desvios padrão distintos.

No gráfico da Figura 3, observa-se que, considerando C6, para os testes AD, CVM, LI e SW, há uma clara tendência linear de aumento dos poderes empíricos com o aumento do número de repetições k por tratamento. O único teste de normalidade que não apresentou essa tendência foi o KS, cujos valores se mantiveram bem estáveis, variando muito pouco próximos de zero. Apesar de os demais testes de normalidade terem atingido seus maiores poderes empíricos para k=10, nenhum teste foi poderoso (todos apresentaram $\widehat{P}<0,75$), o que, como já mencionado anteriormente, é corroborado pelo fato de a estatística do teste qui-quadrado de independência ser significativa para a relação entre os níveis de poder e a condição de simetria das distribuições dos dados (Tabela 4), demonstrando que, quando os dados advêm de distribuições simétricas, os poderes empíricos dos testes de normalidade caem significativamente. Porém, em comparação com os cenários simulados das distribuições simétricas com desvios padrão idênticos (C5 e C7), para os cenários com desvios padrão distintos (C6 e C8), os poderes empíricos da maior parte dos testes de normalidade aumentaram suscitando o questionamento sobre se há uma relação entre os níveis de poder e a homogeneidade (ou heterogeneidade) das variâncias. Contudo, apesar dos aumentos observados, já que os testes continuaram sendo não poderosos ($\widehat{P}<0,75$), não é possível avaliar

se há uma dependência significativa entre a homogeneidade(heterogeneidade) das variâncias e o nível de poder, pois uma das colunas da tabela de contingência é nula, o que resulta em frequências esperadas nulas, e na impossibilidade do cálculo da estatística χ^2.

Porém, também é fácil notar, sem precisar do apoio do teste qui-quadrado, que os poderes empíricos dos testes de normalidade para os cenários oriundos de distribuições simétricas com médias idênticas e desvios padrão distintos (C6) são muito semelhantes aos obtidos para os cenários oriundos de distribuições simétricas com médias distintas e desvios padrão distintos (C8), indicando que a igualdade (ou desigualdade) de médias não teve influência observável nos níveis de poder empírico dos testes de normalidade nesses cenários.

Constata-se, além disso, que sob distribuições Gama Inversa simétricas com médias distintas e desvios padrão distintos (C8 na Figura 3), nenhum teste de normalidade foi poderoso ($\widehat{P}$<0,75), mais uma vez demonstrando como esses testes são sensíveis à condição de simetria.

Já o teste F, para C8, continuou sendo poderoso para todos os valores de k, demonstrando mais uma vez que não há dependência entre os níveis de poder empírico do teste F e a condição de simetria(assimetria) da distribuição de probabilidade dos dados, o que foi confirmado pelo teste qui-quadrado (Tabela 4). É perceptível também, sem a necessidade da realização de testes que, para as distribuições simétricas avaliadas, não há relação de dependência entre o nível de poder e a homogeneidade (heterogeneidade) de variâncias em relação ao teste F, visto que tanto para o cenário em que as médias são diferentes e os desvios padrão são idênticos (C7), quanto para o cenário em que as médias são diferentes e os desvios padrão são distintos (C8), os poderes empíricos do teste F foram iguais a 1, para todos os números (k) de repetições por tratamento.

4.2 Discussão

Os resultados apresentados anteriormente foram obtidos adotando metodologia que se difere da aplicada em muitos dos estudos de comparação entre testes de normalidade já realizados. Em geral, nesses outros estudos, os valores simulados da variável resposta não estão atrelados a nenhum delineamento experimental. Além disso, diferentemente do presente trabalho, avaliam a normalidade dos próprios valores simulados da variável resposta e não dos erros experimentais associados. E por não adotarem delineamentos experimentais, a maioria desses estudos não avalia também o efeito da igualdade ou da desigualdade das médias dos tratamentos, nem da homogeneidade ou heterogeneidade de suas variâncias como foi feito neste trabalho. Apesar disso, estudos como os de Arnastauskaitė et al. (2021), Razali e Yap (2011),

Ogunleye et al. (2018), Doulah (2019), Torman et al. (2012) e Uyanto (2022) apresentam resultados compatíveis com os deste trabalho, visto que também demonstram uma tendência clara de aumento do poder empírico dos testes de normalidade AD, CVM, KS, LI e SW com o aumento do número de observações, principalmente quando os dados são originados de distribuições de probabilidade assimétricas.

No entanto, neste trabalho, essa tendência de aumento do poder com o aumento do número de observações foi totalmente contrariada quando os resíduos foram obtidos a partir de cenários simulados com distribuições gamas inversas consideradas simétricas, isto é, que se aproximam do formato da curva de uma distribuição normal, com igualdade de médias e de desvios padrão dos tratamentos (C5 da Figura 3) e, também, quando foram originados de gamas inversas consideradas simétricas com desigualdade de médias e homogeneidade dos desvios padrão dos tratamentos (C7 da Figura 3). Nesses cenários, todos os testes de normalidade apresentaram poderes empíricos estáveis e extremamente baixos. Entretanto, nos cenários com gamas inversas simétricas e heterogeneidade de variâncias (C6 e C8 da Figura 3), os poderes empíricos foram um pouco superiores, apesar de continuarem mais baixos do que aqueles dos cenários das gamas inversas assimétricas. Esses resultados coincidem com as constatações de Farrell e Stewart (2006) de que, quando as distribuições empíricas são simétricas, os testes de normalidade apresentam poderes mais baixos quando as curvas apresentam caudas curtas, isto é, baixos valores de desvio padrão (C5 e C7) e poderes mais elevados apenas quando as curvas apresentam caudas mais longas (C6 e C8).

Ressalta-se, inclusive, que Farrell e Stewart (2006) concluíram que a modificação proposta no teste SW por Rahman e Govindarajulu (1997) foi capaz de aumentar o seu poder quando a distribuição empírica dos dados é simétrica de cauda curta. No entanto, esses resultados não estão atrelados a nenhum delineamento experimental, como acontece no presente trabalho, em que se constatou que, para distribuições simétricas, nem mesmo o teste SW, que apresentou alguns dos maiores poderes empíricos nos cenários de assimetria, foi poderoso. Para estudos futuros, então, uma das possíveis recomendações é a de avaliar se a modificação proposta por Rahman e Govindarajulu (1997) é também eficaz para testar se os erros experimentais seguem distribuição normal e se provocam melhorias consideráveis nos resultados dos cenários de simetria. Deve-se procurar também possíveis modificações, que possam provocar aumentos de poder dos demais testes de normalidade (AD, CVM, LI e KS) nesses casos, visto que também foram ineficazes.

Outra constatação neste estudo é a de que o teste KS foi o que apresentou os piores resultados, já que, em nenhum cenário simulado, apresentou poder empírico superior a 0,75 e, em todos os cenários, foi o teste com os menores poderes empíricos e o mais destoante dos demais. Esses resultados são compatíveis com os de Ogunleye et al. (2018), Torman et al. (2012) e Uyanto (2022), que observaram que o teste KS se mostrou bastante conservador da hipótese de nulidade.

Com este trabalho, foi possível concluir também que, ao ser respeitada a pressuposição de homogeneidade de variâncias dos tratamentos, o teste F se mostrou extremamente poderoso, para todos os tamanhos amostrais, e muito capaz de rejeitar corretamente a hipótese de igualdade de médias dos tratamentos. Essa conclusão é condizente com as constatações de Nguyen et al. (2019) de que, sob homogeneidade de variâncias dos tratamentos, o teste F tem performance melhor do que métodos não paramétricos de verificação da condição de igualdade das médias, tanto quanto ao nível elevado de poder, quanto ao nível de significância empírico.

Assim, com o presente trabalho, assim como perceberam Nguyen et al. (2019), observou-se uma certa sensibilidade do teste F à não homogeneidade das variâncias, visto que, o único cenário em que o poder do teste F foi inferior ocorreu quando essa pressuposição foi desrespeitada e o número de repetições por tratamento foi o menor. Porém, com o presente trabalho, foi possível constatar adicionalmente que a violação da pressuposição de normalidade sozinha não foi capaz de diminuir os poderes empíricos do teste F, o que atesta a robustez desse método inferencial a essa violação.

5. Conclusões

De maneira geral, pode-se concluir que os resultados deste estudo indicam que:

- O aumento do número total de observações no experimento, ou seja, o aumento do número de repetições k por tratamento, resulta no aumento do poder empírico dos testes de normalidade, principalmente nos cenários simulados com distribuições assimétricas. Por outro lado, em cenários simulados com distribuições simétricas, o aumento do número de observações do experimento tende a aumentar o poder dos testes de normalidade apenas quando os desvios padrão dos tratamentos são diferentes e apresentam valores maiores. Porém, mesmo nesses casos, os testes de normalidade não foram poderosos.

- Os poderes empíricos dos testes de normalidade tendem a ser menores quando a distribuição de probabilidades dos dados não é normal, porém, simétrica.

-O teste F foi poderoso tanto quando a distribuição dos valores da variável resposta foi simétrica, quanto quando foi assimétrica. Nem a condição de simetria ou assimetria da distribuição não normal, nem a de homogeneidade ou heterogeneidade das variâncias, quando o número de repetições por tratamento foi superior a quatro, foram capazes de diminuir significativamente, ao nível de 5% de significância, os poderes empíricos do teste F.

-Em praticamente todos os cenários, o poder empírico do teste F se manteve elevado e estável. Apenas quando os resíduos foram originados de distribuições gamas inversas assimétricas com médias de tratamentos diferentes e variâncias heterogêneas (C4), é que se observou, para quatro ou menos repetições por tratamento, poder empírico do teste F inferior ou igual a 0,75. Já para k>4, até mesmo no cenário C4, todos os poderes empíricos do teste F foram próximos de 1.

6. Referências Bibliográficas

ANDERSON, T. W., DARLING, D. A. **Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes**. **The Annals of Mathematical Statistics**. [S.l: s.n.], 1952.

ARNASTAUSKAITĖ, J., RUZGAS, T., BRAŽĖNAS, M. "An exhaustive power comparison of normality tests", **Mathematics**, v. 9, n. 7, 1 abr. 2021. DOI: 10.3390/math9070788.

CASELLA, G., BERGER, R. L. **Statistical Inference**. 2nd. ed. [S.l.], Duxbury/Thomson Learning, 2002.

DOULAH, M. S. U. "A Comparison among Twenty-Seven Normality Tests", p. 41–59, 2019a. Disponível em: www.stmjournals.com.

DOULAH, Md. S. U. "A Comparison among Twenty-Seven Normality Tests", **STM Journals**, v. 8, n. 3, p. 41–59, 2019b.

FARRELL, P., STEWART, K.R. "Comprehensive study of tests for normality and symmetry: Extending the Spiegelhalter test", **Journal of Statistical Computation and Simulation**, v. 76, n. 9, p. 803–816, 1 set. 2006. DOI: 10.1080/10629360500109023.

GELMAN, A., CARLIN, J. B., STERN, H. S., *et al.* **Bayesian Data Analysis**. Third ed. [S.l.], CRC Press, 2013.

ISLAM, T. U. "Min-max approach for comparison of univariate normality tests", **PLoS ONE**, v. 16, n. 8, 1 ago. 2021. DOI: 10.1371/journal.pone.0255024.

MOOD, A. M. **INTRODUCTION TO THE THEORY OF STATISTICS**. 3. ed. [S.l.], McGraw-Hill, Inc., 1974.

NGUYEN, D., PHAM, T. V., KIM, E., *et al.* "Empirical Comparison of Tests for One-Factor ANOVA Under Heterogeneity and Non-Normality: A Monte Carlo Study",

Journal of Modern Applied Statistical Methods, v. 18, n. 2, 2019. DOI: 10.22237/jmasm/1604190000.

OGUNLEYE, L. I., OYEJOLA, B. A., OBISESAN, K. O. "Comparison of Some Common Tests for Normality", **International Journal of Probability and Statistics**, v. 7, n. 5, p. 130–137, 2018.

RAHMAN, M. M., GOVINDARAJULU, Z. "A modification of the test of Shapiro and Wilk for normality", **Journal of Applied Statistics**, v. 24, p. 219–235, 1997.

RAZALI, N.M., YAP, B.W. **Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests**. [S.l: s.n.], 2011. v. 2.

SHAPIRO, S. S., WILK, ; M. B. **An Analysis of Variance Test for Normality (Complete Samples)**. **Biometrika**. [S.l: s.n.], 1965.

SIEGEL, S., CASTELLAN JR., N. J. **ESTATÍSTICA NÃO-PARAMÉTRICA PARA CIENCIAS DO COMPORTAMENTO**. 2.ed. ed. Porto Alegre, RS., ARTMED EDITORA S.A., 2006.

TORMAN, V. B. L., COSTER, R., RIBOLDI, J. "Normalidade de variáveis: métodos de verificação e comparação de alguns testes não-paramétricos por simulação", **Revista Clinical & Biomedical Research**, v. 32, n. 2, 2012.

UYANTO, S. S. "An Extensive Comparisons of 50 Univariate Goodness-of-fit Tests for Normality", **Austrian Journal of Statistics**, v. 51, p. 45–97, 22 ago. 2022. DOI: 10.17713/ajs.v51i3.1279.

Capítulo 4: Avaliação das taxas de erro tipo I empíricas dos testes F e de normalidade sob diferentes condições experimentais

1. Resumo

Nas Ciências Agrárias, alguns dos testes mais comuns e aplicados são os testes de normalidade, tendo em vista que permitem a verificação da pressuposição de normalidade do teste F da Análise de Variância (ANOVA), quando não é possível obter amostras muito grandes. Porém, ao realizar esses ou quaisquer outros testes de hipóteses, o pesquisador está sujeito a cometer o erro tipo I ou o erro tipo II, cujas probabilidades são inversamente proporcionais e afetadas pelas diferentes condições experimentais em que os métodos inferenciais são aplicados. Diante disso, o objetivo deste trabalho foi avaliar a taxa de erro tipo I empírica sob igualdade (desigualdade) das médias dos tratamentos, homogeneidade (heterogeneidade) de variâncias e ao variar o número de repetições por tratamento. Foram, então, simulados os subcenários de cada cenário experimental definido. Para cada uma das 10 mil iterações de cada subcenário, foi gerado um conjunto de valores da variável resposta e de resíduos experimentais, aplicados aos testes convenientes. Observou-se que, tanto o teste F, quanto os testes de

normalidade comparados, com exceção do teste de Kolmogorov-Smirnov, apresentaram taxas de erro tipo I empíricas maiores e, que, para os testes de normalidade, cresceram ainda mais com o aumento do número de repetições, quando a pressuposição de homogeneidade de variâncias foi violada. Por outro lado, quando não houve essa violação, as taxas foram estáveis e próximas do nível de significância teórico para todos os testes de hipóteses analisados.

Palavras-chave: Testes de Hipóteses; Nível de Significância; Erro Tipo II; Análise de Variância (ANOVA); Delineamento Inteiramente Casualizado (DIC); Distribuição Normal; Erros Experimentais; Resíduos Experimentais; Anderson-Darling; Cramér-von Mises; Kolmogorov-Smirnov; Lilliefors; Shapiro-Wilk;

2. Introdução

Na Estatística, alguns dos principais métodos decisórios são os testes de hipóteses, que consistem na escolha por uma das duas hipóteses, conhecidas como hipótese de nulidade (H_0) e hipótese alternativa (H_a). Então, para optar por uma dessas hipóteses, o pesquisador se baseia nas distribuições do estimador e da estatística do teste, sob a pressuposição de que H_0 é verdadeira.

Assim, uma das principais razões pelas quais esses testes são tão práticos e aplicáveis deve-se ao fato de que, ao mesmo tempo que suas hipóteses versam sobre os parâmetros populacionais, as decisões sobre qual delas escolher baseiam-se nas amostras, o que representa economia de recursos, visto que mapear toda uma população pode ser uma tarefa extremamente trabalhosa, custosa e, em muitos casos, impossível.

No entanto, de acordo com Bussab e Morettin (2010), ao ser definida a região crítica de rejeição da hipótese de nulidade (H_0), cuja probabilidade é conhecida como nível de significância α do teste de hipóteses, o pesquisador, inevitavelmente, fica sujeito à possibilidade de cometer dois tipos de erros.

A probabilidade do erro tipo I é definida pelo próprio nível de significância α, associada a valores da estatística do teste presentes na região crítica. Isto é, esse erro acontece quando se rejeita H_0, sendo H_0 verdadeira e pode acontecer, pois, por mais que os valores presentes na região crítica sejam pouco prováveis, é possível que a estatística do teste assuma algum desses valores, sob H_0.

Por outro lado, a probabilidade do erro tipo II, definida por β, acontece quando não se rejeita H_0 sendo H_0 falsa e pode acontecer, pois, por mais que não façam parte da região crítica, os

valores presentes na região de não rejeição de H_0 podem fazer parte de outra distribuição, que parcialmente se sobrepõe à distribuição de H_0.

Dessa maneira, segundo Bussab e Morettin (2010), as probabilidades α e β de cometer esses erros são inversamente proporcionais, ou seja, quando se aumenta uma, a outra é diminuída e vice-versa. Porém, sempre existirão, seja qual for o teste de hipóteses realizado.

Nesse sentido, quando se diz respeito às Ciências Agrárias, alguns dos testes de hipóteses mais comuns e, obviamente, suscetíveis a esses erros, são os testes de normalidade. Esses testes, em geral, são aplicados para verificar se os erros experimentais associados aos valores da variável resposta dos tratamentos seguem uma distribuição normal, satisfazendo a pressuposição de normalidade do teste F da Análise de Variância (ANOVA).

Portanto, faz-se necessário compreender melhor essa pressuposição e a natureza dos testes supracitados. Para isso, deve-se lembrar de que, quando o pesquisador realiza uma ANOVA, seu objetivo é o de comparar, em termos das médias populacionais, diferentes tratamentos, visto que a hipótese de nulidade (H_0) do teste F é a de que essas médias são iguais. Porém, para que as comparações sejam fidedignas, é necessário considerar as condições experimentais sob as quais os experimentos são realizados, o que é feito de acordo com um modelo estatístico.

Consequentemente, ressalta-se que a estatística do teste F da ANOVA é definida como a razão entre dois estimadores de variância, em que cada um deles segue uma distribuição qui-quadrado. No numerador dessa razão, encontra-se o estimador da variância entre as médias de tratamentos e dentro dos tratamentos e no denominador apenas o estimador da variância dentro de tratamentos, devida a variações aleatórias, conhecidas como erros experimentais.

Dessa maneira, segundo Mood (1974) e Casella e Berger (2002), os estimadores, que compõem a estatística do teste F seguem distribuições qui-quadrado, pois são calculados a partir de somatórios dos quadrados de variáveis aleatórias, que seguem distribuições normais. Por isso, caso a pressuposição de normalidade dos erros experimentais seja violada, espera-se que esses estimadores não sigam distribuições qui-quadrado. Assim, a razão entre eles não seguiria a distribuição F de Fisher-Snedecor, o que geraria consequências nas taxas dos erros tipo I e tipo II do teste F.

Por conseguinte, o modelo que, de acordo com Searle e Gruber (2016), incorpora essa pressuposição de normalidade e que permite comparar os parâmetros populacionais é o Modelo Linear de Gauss-Markov Normal (MLGMN). Por causa disso, os modelos experimentais das

análises de variância são desdobramentos do MLGMN, que pressupõe que os erros experimentais seguem uma distribuição normal multivariada com média nula e variâncias homogêneas.

Desta forma, quando, por exemplo, se adota o Delineamento Inteiramente Casualizado (DIC) para a realização de experimentos, o modelo estatístico adotado é um caso do MLGMN (Equação 1).

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad (1)$$

Em que, y_{ij} é o valor observado da variável resposta do tratamento i na repetição j, μ é a média populacional geral, τ_i equivale ao efeito do tratamento i e ε_{ij}, ao erro experimental, que segue uma distribuição normal e é associado ao valor observado y_{ij}.

Portanto, para avaliar se a pressuposição de normalidade dos erros experimentais é satisfeita, comumente podem ser empregados os testes de normalidade de Anderson-Darling (AD), Cramér-von Mises (CVM), Kolmogorov-Smirnov (KS), Lilliefors (LI) e Shapiro-Wilk (SW). De acordo com Anderson e Darling (1952), Thadewald e Büning (2007) , Barbetta et al. (2004) e Campos (1976), os testes AD, CVM, KS e LI são considerados testes de aderência, pois comparam a distribuição de probabilidade acumulada da distribuição teórica normal com a distribuição de probabilidade empírica da variável em análise. Nesses testes, a hipótese de nulidade (H_0) estabelece que ocorre aderência da distribuição empírica a uma distribuição normal teórica. A decisão sobre H_0 é tomada com base num nível de significância (α) e no tamanho da amostra utilizada para a obtenção da distribuição empírica.

Em contraposição, apesar de o teste SW também ter por objetivo decidir se a hipótese de distribuição normal deve ser rejeitada, esse teste apresenta estatística calculada de maneira diferente dos demais testes de normalidade supracitados. Segundo Shapiro e Wilk (1965), seu teste homônimo se baseia na razão entre dois estimadores de variância, um geral e outro aplicado para distribuições simétricas, como as normais. Assim, quanto menor a discrepância entre os dois, maiores as chances de a hipótese de normalidade ser verdadeira.

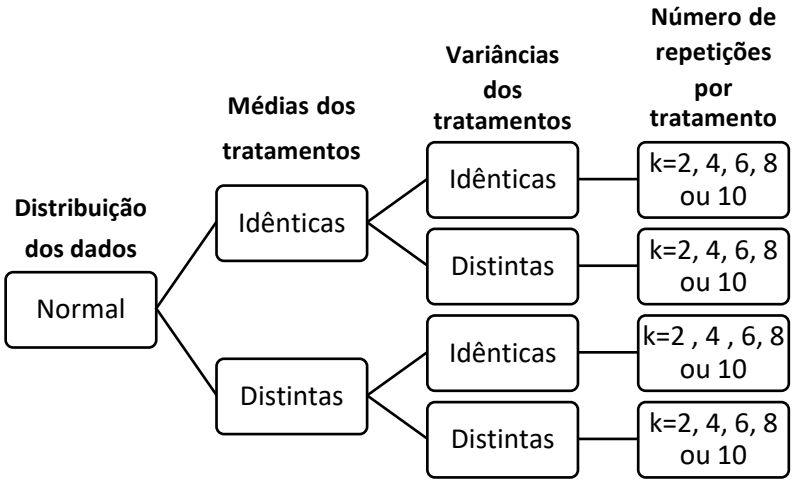
Reconhecendo a importância dos testes de normalidade e do teste F e, também, os erros relacionados às decisões equivocadas, este trabalho teve por objetivo geral avaliar a taxa empírica do erro tipo I sob diversas condições experimentais. Especificamente, as condições experimentais avaliadas foram a igualdade (desigualdade) das médias de tratamentos, homogeneidade (heterogeneidade) de variâncias e o número de repetições por tratamento.

Dessa maneira, avaliaram-se os testes de normalidade em situações mais próximas das que acontecem nas Ciências Agrárias, inclusive, adotando números de repetição, que gerassem um total de observações mais próximo ao que ocorre em experimentos agrícolas.

3. Metodologia

Foram simulados quatro cenários para possíveis valores de uma variável resposta (y_{ij}) em um experimento instalado segundo o delineamento inteiramente casualizado (DIC) com cinco tratamentos e k repetições por tratamento conforme apresentado na Figura 1.

Figura 1- Cenários e subcenários simulados definindo uma distribuição normal e considerando igualdade (desigualdade) de médias dos tratamentos, homogeneidade (heterogeneidade) de variâncias dos tratamentos e um número de repetições por tratamento (k=2, 4, 6, 8, 10).



Nas simulações, foi estabelecido que cada um dos tratamentos deveria seguir uma distribuição normal específica. Com essa finalidade, foi definido que os quatro cenários se diferenciassem quanto à igualdade (desigualdade) de médias de tratamentos e à homogeneidade (heterogeneidade) de variâncias e desvios padrão dentro de tratamentos. Os parâmetros dessas diferentes distribuições normais utilizadas na simulação são apresentados na Tabela 1.

Tabela 1 - Valores das médias e dos desvios padrão estabelecidos para a simulação dos cenários considerando diferentes distribuições normais

Distribuição	Média	Desvio Padrão	Cenário
Normal	$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = 100$	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5 = 1$	C1
		$\sigma_1 = 1; \sigma_2 = 2; \sigma_3 = 3; \sigma_4 = 4; \sigma_5 = 5$	C2
	$\mu_1 = 100; \mu_2 = 200; \mu_3 = 300; \mu_4 = 400; \mu_5 = 500$	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5 = 1$	C3
		$\sigma_1 = 1; \sigma_2 = 2; \sigma_3 = 3; \sigma_4 = 4; \sigma_5 = 5$	C4

Para cada um dos quatro cenários (Tabela 1) foram simulados subcenários que se diferenciavam quanto ao número k de repetições por tratamento, tal que, k = 2, 4, 6, 8 e 10.

Para cada um desses subcenários, foram simuladas 10.000 iterações. Para cada iteração w tal que $w = 1, 2, \cdots, 10.000$, foi obtido um conjunto w de 5k valores de resíduos experimentais de acordo com a Equação 2.

$$\hat{\varepsilon}_{ijw} = y_{ijw} - \hat{\mu}_{iw} \quad (2)$$

Em que, $\hat{\varepsilon}_{ijw}$ é o resíduo obtido para o valor observado y_{ijw} da variável resposta na iteração w para a repetição j do tratamento i; $\hat{\mu}_{iw}$ é a média dos valores da variável resposta do tratamento i na iteração w tal que $w = 1, 2, \cdots, 10.000$; $i = 1, \cdots, 5$; $j = 1, \cdots, k$ e $k = 2, 4, 6, 8, 10$.

Em seguida, cada um dos w conjuntos de resíduos foi submetido a cada um dos testes de normalidade (AD, CVM, KS, LI e SW). O p-valor de cada um destes testes, em cada iteração, foi devidamente registrado para o cálculo da taxa empírica do erro tipo I ($\hat{\alpha}$) de cada teste, em cada subcenário, por meio da Equação 3.

$$\hat{\alpha} = \frac{número\ de\ p - valores \leq 0{,}05}{10.000} \quad (3)$$

Adicionalmente, cada conjunto w de valores da variável resposta (não dos resíduos) dos subcenários dos cenários C1 e C2 foi submetido ao teste F. O p-valor do teste F em cada iteração dos subcenários foi devidamente registrado para o cálculo da taxa empírica do erro tipo I ($\hat{\alpha}$) por meio da Equação 3.

Com base nos resultados obtidos, foram plotados gráficos relacionando o número de repetições (k) por tratamento e as taxas empíricas do erro tipo I ($\hat{\alpha}$) de cada cenário, separadamente.

Com o objetivo de qualificar as taxas empíricas do erro tipo I dos testes de normalidade e do teste F, o nível de significância empírico de cada um destes testes foi classificado como:

- Se $\hat{\alpha} > 0{,}05$, o teste cometeu uma probabilidade alta de erro tipo I;
- Se $\hat{\alpha} \leq 0{,}05$, o teste cometeu uma probabilidade baixa de erro tipo I.

Para analisar as possíveis relações de dependência entre a homogeneidade (heterogeneidade) de variâncias dos tratamentos e as taxas empíricas do erro tipo I ($\hat{\alpha}$), foi utilizado o teste de qui-quadrado para independência. Nos casos em que o teste de qui-quadrado foi significativo e o tamanho amostral foi inferior a 40 e/ou pelo menos uma classe apresentou frequência esperada inferior a 5, empregou-se a correção de Yates de acordo com Fukunaga et al. (2018).

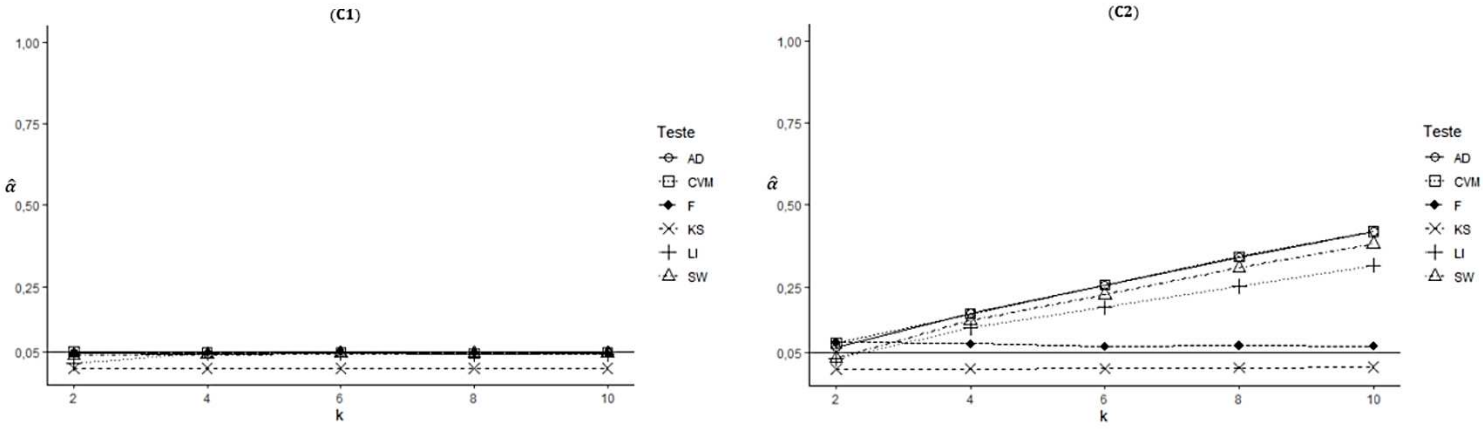
Para a simulação das iterações bem como para as aplicações dos testes de normalidade e F foi utilizado o RStudio versão 4.0.2 (R Core Team, 2020).

4. Resultados e discussão

4.1 Resultados

Os resultados, isto é, as taxas empíricas de erro tipo I, obtidas para os subcenários de C1 e C2 são apresentados na Figura 2.

Figura 2- Taxa empírica de erro tipo I ($\hat{\alpha}$) dos testes AD, CVM, KS, LI, SW e F em função do número de repetições (k) por tratamento dos subcenários dos cenários C1 e C2, que se diferem apenas quanto às variâncias residuais dos tratamentos (homogeneidade em C1 e heterogeneidade em C2).A reta horizontal para 0,05 indica o nível de significância teórico (α) adotado em todos os testes.



Todos os testes de normalidade apresentaram valores satisfatórios de taxa empírica de erro tipo I no cenário C1 (Figura 2), quando as médias e desvios padrão dos tratamentos foram idênticos, visto que, para os testes AD, CVM, LI e SW, para praticamente todos os números de repetições (k) por tratamento, os valores de $\hat{\alpha}$ foram muito próximos do nível teórico de significância (α=0,05). Para o teste KS, os valores de $\hat{\alpha}$ foram todos visivelmente inferiores ao limiar de 0,05 e muito próximos de zero. Esses resultados confirmam, inclusive, o caráter conservador do teste KS, em consonância com os resultados obtidos por Torman et al. (2012). Dessa maneira, apesar de, em estudos como os de Torman et al. (2012), Arnastauskaitė et al. (2021), Razali e Yap (2011) e Ogunleye et al. (2018), o teste KS apresentar os menores poderes empíricos, em contraposição, esse teste é o que menos erra rejeitando a hipótese de normalidade, quando ela não deve ser rejeitada. O teste F também apresentou resultados

satisfatórios para a taxa empírica de erro tipo I, uma vez que o seu valor ficou abaixo ou muito próximo de 0,05 para todos os números de repetições avaliados.

Porém, esses resultados satisfatórios não se mantiveram para as simulações realizadas para o cenário C2 (Figura 2). Ao contrário dos resultados de $\hat{\alpha}$ quando tanto as médias quanto os desvios padrão dos tratamentos são idênticos (C1), para os subcenários com desvios padrão distintos (C2), as taxas empírica de erro tipo I ($\hat{\alpha}$) claramente aumentaram e superaram o limiar teórico de α=0,05 na maioria dos subcenários. Além disso, para todos os testes, exceto o F e o KS, o nível de significância aumentou com o aumento do número de repetições por tratamento (k), ao invés de permanecer estável. Esse resultado sugere uma possível relação de dependência entre homogeneidade (heterogeneidade) de variâncias dentro de tratamentos e $\hat{\alpha}$ dos testes de hipóteses comparados. De acordo com os testes de qui-quadrado, cujos resultados são apresentados na Tabela 2, essa dependência entre a homogeneidade de variâncias e o nível de significância foi confirmada para todos os testes, exceto para o teste KS.

Tabela 2- Testes de qui-quadrado de independência entre as taxas empíricas de erro tipo I (elevadas se $\hat{\alpha}$>0,05 e baixas se $\hat{\alpha}\leq$0,05) apresentadas pelos testes F, Anderson-Darling (AD), Kolmogorov-Smirnov (KS), Crámer-von Mises (CVM), Lilliefors (LI) e Shapiro-Wilk (SW) com as condições de homogeneidade (ou heterogeneidade) de variâncias dos tratamentos, de acordo com o banco de dados dos cenários C1 e C2.

Teste	χ^2
AD	6,40*
KS	-
CVM	3,75*
LI	3,75*
SW	3,75*
F	3,75*

*: p-valor≤0,05.

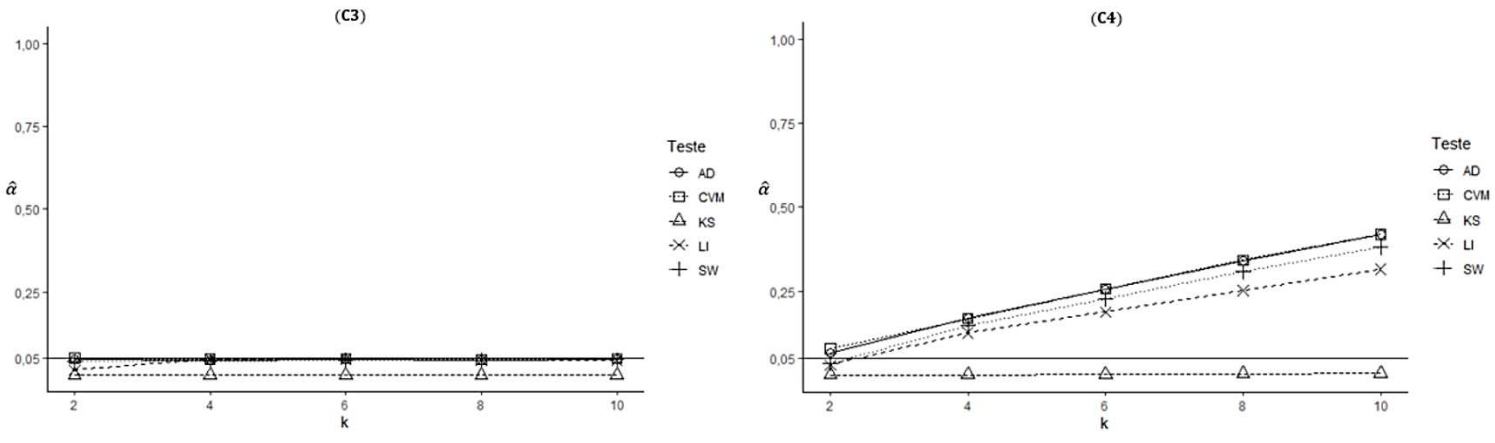
Portanto, as taxas empíricas de incidência do erro do tipo I dos testes de normalidade analisados e também do teste F tendem a ser maiores quando a pressuposição de homogeneidade das variâncias não é satisfeita. Ressalta-se ainda que o teste KS destoou dos demais testes de normalidade e não apresentou, em nenhum dos subcenários, nível de significância empírico superior ao limiar teórico de 0,05 e, por isso, não foi possível realizar o teste qui-quadrado para esse teste, já que uma das linhas da tabela de contingência foi nula.

Porém, nem seria necessário, pois o padrão de comportamento do teste KS foi essencialmente o mesmo em C1 e C2 (Figura 2), apresentando valores de $\hat{\alpha}$ estáveis e próximos de zero. Esse resultado, portanto, indica que a heterogeneidade de variâncias não teve influência na taxa de erro empírica do teste de normalidade KS, o que pode ser explicado pelo fato de que os parâmetros da distribuição teórica (μ, σ^2) são conhecidos e especificados quando se realiza o teste KS, que se comporta como um teste exato.

Já para o teste F, o resultado do teste qui-quadrado de independência (Tabela 2) demonstra que quando os desvios padrão e, consequentemente, as variâncias dos tratamentos são distintas, a incidência do erro tipo I aumenta, isto é, aumenta a probabilidade de rejeitar a hipótese nula de igualdade entre as médias dos tratamentos, quando ela não deveria ser rejeitada, o que reforça a importância da pressuposição de homogeneidade de variâncias dos tratamentos das análises de variância (ANOVA). Porém, diferentemente do que foi observado para os testes de normalidade, os aumentos dos valores de $\hat{\alpha}$ para o teste F, observados para C2 no gráfico da Figura 2 foram menores.

Outro ponto avaliado foi o efeito da desigualdade de médias dos tratamentos no valor de $\hat{\alpha}$ apresentado pelos testes de normalidade AD, KS, CVM, LI e SW. Assim, foram plotados, no gráfico da Figura 3, os resultados relativos à aplicação desses testes, quando os resíduos advêm das iterações das simulações dos subcenários dos cenários C3 e C4 (Tabela 1).

Figura 3- Taxa empírica de erro tipo I ($\hat{\alpha}$) dos testes de normalidade de AD, CVM, KS, LI e SW em função do número de repetições (k) por tratamento dos subcenários dos cenários C3 e C4, que se diferem apenas quanto às variâncias residuais dos tratamentos (homogeneidade em C3 e heterogeneidade em C4), pois todos foram simulados sob desigualdade de médias. A reta horizontal para 0,05 indica o nível de significância teórico (α) adotado em todos os testes.



Percebe-se que a desigualdade de médias dos tratamentos não foi capaz de aumentar as taxas empíricas ($\hat{\alpha}$) do erro tipo I quando as variâncias residuais dos tratamentos foram homogêneas, pois, para todos os testes de normalidade, nos subcenários de C3 na Figura 3, quase todos os valores de $\hat{\alpha}$ ou foram muito próximos de 0,05 ou foram inferiores. Em comparação com os resultados dos cenários com igualdade de médias e homogeneidade de variâncias (C1 na Figura 2), os resultados dos subcenários de C3 foram muito semelhantes.

Entretanto, ao comparar os resultados das simulações dos subcenários de C3 (médias distintas e variâncias homogêneas) com os obtidos para C4 (médias distintas e variâncias heterogêneas) na Figura 3, percebe-se claramente que os valores de $\hat{\alpha}$ aumentam e esse aumento é proporcional ao aumento do número de repetições por tratamento. Para investigar a relação entre a homogeneidade (heterogeneidade) de variâncias dentro de tratamentos e as taxas empíricas de erro tipo I, quando as médias dos tratamentos são diferentes, foi aplicado o teste qui-quadrado de independência para cada um dos testes de normalidade, considerando o banco de dados de C3 e C4, que se diferem apenas quanto à homogeneidade ou não das variâncias dos tratamentos. Os resultados desses testes de qui-quadrado são apresentados na Tabela 3.

Tabela 3- Testes de qui-quadrado de independência entre as taxas empíricas de erro tipo I (elevadas se $\hat{\alpha}$>0,05 e baixas se $\hat{\alpha}\leq$0,05) apresentadas pelos testes de Anderson-Darling (AD), Kolmogorov-Smirnov (KS), Crámer-von Mises (CVM), Lilliefors (LI) e Shapiro-Wilk (SW) com as condições de homogeneidade (ou heterogeneidade) de variâncias dos tratamentos, de acordo com o banco de dados dos cenários C3 e C4.

Teste	χ^2
AD	6,40*
KS	-
CVM	3,75*
LI	3,75*
SW	3,75*

*: p-valor≤0,05.

Por conseguinte, os resultados da Tabela 3 confirmam que, para todos os testes de normalidade estudados, há uma relação significativa de dependência entre as taxas empíricas de erro tipo I e a condição de homogeneidade (heterogeneidade) das variâncias dentro de tratamentos nos cenários, demonstrando que, quando as variâncias são heterogêneas, aumenta significativamente a incidência do erro tipo I.

4.2 Discussão

Em geral, muitos dos estudos de comparação entre os testes de normalidade, quanto às taxas empíricas de erro tipo I ($\hat{\alpha}$), apresentam a mesma limitação de não realizar as simulações, baseando-se em algum delineamento experimental. Por isso, as taxas calculadas são relacionadas a uma única distribuição normal, não sendo possível avaliar o efeito da homogeneidade (heterogeneidade) das variâncias dos tratamentos, nem da igualdade (desigualdade) de suas médias, como foi feito no presente trabalho. Apesar disso, os resultados de estudos, tais quais os de Ogunleye et al. (2018), Öztuna et al. (2006), Keskin (2006) e Torman et al. (2012) são compatíveis com os resultados obtidos nos cenários com homogeneidade de variâncias dos tratamentos (C1 e C3).

Ogunleye et al. (2018) constataram que, em geral, os testes que apresentaram taxas empíricas de erro tipo I mais próximas do nível teórico de 5% de significância foram, respectivamente, o teste SW, seguido do teste KS e do teste AD. Mas, as diferenças entre os valores observados não foram expressivas. Além disso, foi observada uma certa estabilidade dos níveis de significância empíricos, considerando tamanhos amostrais variando de 10 a 100 e um total de 5.000 iterações para cada um. No presente trabalho, conclusões semelhantes a essas foram obtidas por meio dos resultados dos cenários C1 e C3.

Öztuna et al. (2006) também concluíram que não houve uma diferença muito grande entre as taxas de erro do tipo I dos testes LI e SW. Variando o tamanho amostral de 5 a 200, as taxas foram relativamente estáveis, padrão de comportamento muito parecido com o que se observou no presente estudo para os cenários em que as variâncias dos tratamentos foram idênticas (C1 e C3). No entanto, como o experimento de Öztuna et al. (2006) não foi instalado atrelado a nenhum delineamento experimental, não foi capaz de avaliar, por exemplo, a influência da homogeneidade (heterogeneidade) das variâncias de tratamentos, já que as simulações foram realizadas para apenas um nível de um fator (apenas um tratamento).

As conclusões de Keskin (2006) foram semelhantes às dos estudos supracitados, visto que a taxa de erro tipo I do teste SW ficou em torno de 0,05 para todos os tamanhos amostrais, que variaram de 10 a 150. O total de iterações para cada um desses tamanhos amostrais foi igual a 100.000. Contudo, mesmo com essa grande quantidade de iterações, também não foi possível concluir qual é a influência de diferentes condições experimentais nos resultados.

Em contraposição, neste trabalho, foi possível concluir que a heterogeneidade de variâncias dos tratamentos tem efeito significativo no nível de significância de quase todos os testes de

normalidade avaliados (AD, CVM, LI e SW), tanto quando as médias dos tratamentos são iguais, quanto quando as médias são diferentes, de acordo com os resultados dos testes de qui-quadrado para a independência, respectivamente nas Tabelas 2 e 3. Esses resultados confirmam que, quando as variâncias dos tratamentos são heterogêneas, as taxas de erro tipo I aumentam significativamente, superando o limiar de 0,05 quando as médias dos tratamentos são idênticas (C2) e também quando são diferentes (C4). Nessas condições, os valores de $\hat{\alpha}$ ficam cada vez maiores, quando o número de repetições por tratamento aumenta.

Em relação ao teste F, os resultados do presente estudo são concordantes com os que foram obtidos por Nguyen et al. (2019), Kulkarni e Patil (2021) e Kelter (2021). Em geral, constatou-se que, quando as variâncias dos tratamentos são homogêneas (C1), o teste F é muito efetivo, pois apresenta taxas empíricas de erro tipo I ($\hat{\alpha}$) muito próximas de 0,05, o nível de significância teórico adotado. Em contraposição, quando as variâncias dos tratamentos são heterogêneas (C2), para todos os números de repetições, os valores de $\hat{\alpha}$ foram superiores a 0,05 e o teste F foi menos efetivo. A influência da heterogeneidade das variâncias no aumento das taxas de erro tipo I foi confirmada, inclusive, pelos testes de qui-quadrado para a independência, que foram significativos (Tabela 2).

Em consonância com esses resultados, Nguyen et al. (2019) concluíram que, sob condições de homogeneidade de variâncias, o teste F teve resultados satisfatórios quanto às taxas de erro tipo I, com a maior parte dos valores abaixo ou iguais a 0,05. Sob todas as condições de heterogeneidade de variâncias dos tratamentos, no entanto, a performance do teste F não foi satisfatória, visto que a maior parte das taxas de erro tipo I calculadas foram superiores a 0,05. Essas taxas de erro tipo I aumentaram ainda mais quando a disparidade entre as variâncias dos tratamentos aumentou. Nessas situações, métodos não paramétricos ou semi-paramétricos, como o teste de Wilcox proposto por Wilcox (1988) e Wilcox (1989), assim como o teste de Welch proposto por Welch (1951), apresentaram resultados melhores que o teste F, sendo mais capazes de controlar as taxas de erro tipo I.

Kulkarni e Patil (2021) concluíram de maneira até mais geral que muitos dos testes de hipóteses convencionais, como o teste F, exibem altas taxas de erro tipo I, sob condições paramétricas específicas, principalmente quando há um aumento da quantidade de grupos sendo comparados, cada um com um pequeno tamanho amostral e quando há heterogeneidade entre as variâncias de cada grupo. E, como o aumento das taxas de erro tipo I tem relação direta com a diminuição da confiança da análise de poder desses testes, nesse artigo, Kulkarni e Patil

(2021) propõem a construção de testes alternativos baseados em razões entre funções de verossimilhança integradas em relação aos parâmetros problemáticos, objetivando diminuir as taxas de erro tipo I nas condições mais críticas.

Já Kelter (2021), diferentemente do presente trabalho, estudou os padrões de comportamento de testes de hipóteses para duas amostras. Concluiu que, tanto os testes de hipóteses frequentistas para duas amostras, quanto as suas versões bayesianas, têm seus erros do tipo II reduzidos e, consequentemente, o erro tipo I aumentado, quando se aumenta o tamanho amostral. Concluiu também que os testes frequentistas apresentam taxa de erro tipo I superior à das suas versões bayesianas. Kelter (2021) pontuou ainda que até poderia ser sugerida a adoção um nível de significância teórico inferior a 0,05 com o objetivo de obter taxas inferiores de erro tipo I, mas isso aumentaria inevitavelmente as taxas de erro tipo II.

Na verdade, o que os resultados desse estudo de Kelter (2021) e do presente trabalho evidenciam é o fato de que, quando os tamanhos amostrais ou números de repetições por tratamento aumentam, as amostras de cada tratamento ficam mais representativas e a variabilidade dentro de tratamentos aumenta. Porém, o estimador da variância do erro experimental não acompanha esse aumento, pois é uma média ponderada, que atribuiu os mesmos pesos para a variância dentro de cada tratamento, justamente devido ao pressuposto de homogeneidade de variâncias dentro dos tratamentos. Assim, como esse estimador fica subestimado e como ele se encontra no denominador da estatística do teste F, os valores dela ficam maiores do que deveriam, o que gera, como consequência, a rejeição equivocada da hipótese de nulidade de igualdade de médias, aumentando a taxa de erro tipo I.

5. Conclusões

Este estudo permitiu concluir que:

- Em geral, o teste KS, com parâmetros μ e σ^2 conhecidos, foi o que apresentou menores taxas empíricas $\hat{\alpha}$ de erro tipo I, visto que, em todos os cenários, com igualdade ou não de médias dos tratamentos e homogeneidade ou heterogeneidade de suas variâncias, para nenhum número de repetições (k) por tratamento, seus valores de $\hat{\alpha}$ ultrapassaram o limiar de significância de 0,05, diferentemente dos outros testes de hipóteses comparados (AD, CVM, LI, SW e F).

-Quando houve homogeneidade de variâncias residuais dos tratamentos, todos os testes de normalidade estudados e também o teste F apresentaram taxas empíricas de erro tipo I ($\hat{\alpha}$)

próximas ou inferiores a 0,05 e, portanto, satisfatórias, diferentemente de quando essas variâncias foram heterogêneas.

-Foi confirmado pelos testes de qui-quadrado para a independência, que tanto quando as médias dos tratamentos são idênticas, quanto quando são distintas, há uma relação de dependência significativa entre os níveis de significância empíricos ($\hat{\alpha}$) e a condição de homogeneidade (heterogeneidade) das variâncias. Assim, quando os desvios padrão dos tratamentos são distintos, todos os testes de hipóteses, exceto o KS, têm a incidência de erro tipo I maior.

6. Referências Bibliográficas

ANDERSON, T. W., DARLING, D. A. **Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes**. **The Annals of Mathematical Statistics**. [S.l: s.n.], 1952.

ARNASTAUSKAITĖ, J., RUZGAS, T., BRAŽĖNAS, M. "An exhaustive power comparison of normality tests", **Mathematics**, v. 9, n. 7, 1 abr. 2021. DOI: 10.3390/math9070788.

BARBETTA, P. A., REIS, M. M., BORNIA, A. C. **Estatística: para cursos de engenharia e informática**. São Paulo, Atlas, 2004.

BUSSAB, W. O., MORETTIN, P. A. **Estatística Básica**. 6ª ed. ed. São Paulo, Editora Saraiva, 2010.

CASELLA, G., BERGER, R. L. **Statistical Inference.** 2nd. ed. [S.l.], Duxbury/Thomson Learning, 2002.

CAMPOS, Humberto. **Estatística Experimental Não-Paramétrica**. 2ª ed. Piracicaba, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo., 1976.

FUKUNAGA, E. T., GUIBU, I. A., MORAES, J. C., *et al.* **Bases de Estatística para Profissionais de Saúde**. São Paulo, Memnon: CEALAG- Centro de Estudos Augusto Leopoldo Ayrosa Galvão., 2018.

KELTER, R. "Analysis of type I and II error rates of Bayesian and frequentist parametric and nonparametric two-sample hypothesis tests under preliminary assessment of normality", **Computational Statistics**, v. 36, n. 2, p. 1263–1288, 1 jun. 2021. DOI: 10.1007/s00180-020-01034-7.

KESKIN, S. "Comparison of Several Univariate Normality Tests Regarding Type I Error Rate and Power of the Test in Simulation based Small Samples", **Journal of Applied Science Research**, v. 2, n. 5, p. 296–300, 2006.

KULKARNI, H. v., PATIL, S. M. "Uniformly implementable small sample integrated likelihood ratio test for one-way and two-way ANOVA under heteroscedasticity and normality", **AStA Advances in Statistical Analysis**, v. 105, n. 2, p. 273–305, 1 jun. 2021. DOI: 10.1007/s10182-021-00404-w.

MOOD, A. M. **INTRODUCTION TO THE THEORY OF STATISTICS**. 3. ed. [S.l.], McGraw-Hill, Inc., 1974.

NGUYEN, D., PHAM, T. V., KIM, E., *et al.* "Empirical Comparison of Tests for One-Factor ANOVA Under Heterogeneity and Non-Normality: A Monte Carlo Study", **Journal of Modern Applied Statistical Methods**, v. 18, n. 2, 2019. DOI: 10.22237/jmasm/1604190000.

OGUNLEYE, L. I., OYEJOLA, B. A., OBISESAN, K. O. "Comparison of Some Common Tests for Normality", **International Journal of Probability and Statistics**, v. 7, n. 5, p. 130–137, 2018.

ÖZTUNA, D., ELHAN, A. H., TÜCCAR, E. "Investigation of Four Different Normality Tests in Terms of Type 1 Error Rate and Power under Different Distributions", **Turkish Journal of Medical Sciences**, v. 36, n. 3, p. 171–176, 2006. Disponível em: https://journals.tubitak.gov.tr/medical/vol36/iss3/7. Acesso em: 11 dez. 2022.

RAZALI, N. M., BEE WAH, Y. **Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests**. [S.l: s.n.], 2011. v. 2.

SEARLE, S. R., GRUBER, M. H. J. **Linear Models**. 2nd. ed. [S.l.], Wiley Series in Probability and Statistics, 2016.

SHAPIRO, S. S., WILK, M B. **An Analysis of Variance Test for Normality (Complete Samples)**. **Biometrika**. [S.l: s.n.], 1965.

THADEWALD, T., BÜNING, H. "Jarque-Bera Test and its Competitors for Testing Normality - A Power Comparison", **Journal of Applied Statistics**, v. 34, n. 1, p. 87–105, 2007.

TORMAN, V. B. L., COSTER, R., RIBOLDI, J. "Normalidade de variáveis: métodos de verificação e comparação de alguns testes não-paramétricos por simulação", **Revista Clinical & Biomedical Research**, v. 32, n. 2, 2012.

WELCH, B. L. "On the comparison of several mean values: An alternative approach", **Biometrika**, v. 38, n. 3/4, p. 330–336, 1951. DOI: 10.2307/2332579. Disponível em: http://biomet.oxfordjournals.org/.

WILCOX, R. R. "A new alternative to the ANOVA F and new results on James's second-order method", **British Journal of Mathematical and Statistical Psychology**, v. 41, n. 1, p. 109–117, 1988. DOI: 10.1111/j.2044-8317.1988.tb00890.x.

WILCOX, R. R. "Adjusting for Unequal Variances When Comparing Means in One-Way and Two-Way Fixed Effects ANOVA Models", **Source: Journal of Educational and Behavioral Statistics**, v. 14, n. 3, p. 269–278, 1989. DOI: 10.3102/10769986014003269.

Capítulo 5: Conclusões Gerais

Com este trabalho, foi possível avaliar o comportamento de alguns dos testes de normalidade mais comuns e do teste F, sob igualdade (desigualdade) entre as médias dos

tratamentos, homogeneidade (heterogeneidade) das variâncias residuais e simetria (assimetria) das distribuições empíricas dos valores simulados das variáveis resposta e dos resíduos. Dessa maneira, os resultados apresentados nos artigos dos Capítulos 3 e 4 permitem que sejam feitas algumas recomendações para o emprego ou não desses testes.

Para aumentar o poder empírico do teste de normalidade escolhido, os resultados obtidos no Capítulo 3 demonstram que se deve empregar o maior número possível de repetições por tratamento, dar preferência pelos testes de Shapiro-Wilk, Anderson-Darling e Crámer-von Mises, em detrimento do teste de Lilliefors e, principalmente, em detrimento do teste de Kolmogorov-Smirnov, o mais conservador dos testes de normalidade comparados. Porém, caso seja verificada, por meio de métodos descritivos, uma considerável simetria dos resíduos experimentais e, mesmo assim, deseje-se verificar de maneira inferencial a pressuposição de normalidade, nenhum desses testes de normalidade comparados para um total de até 50 observações, será uma boa alternativa, pois apresentam baixos poderes empíricos nesses casos de simetria.

Quanto à taxa empírica de erro tipo I, os resultados do Capítulo 4 permitem concluir que a principal pressuposição que deve ser atendida, antes de realizar tanto os testes de normalidade, quanto o teste F da ANOVA, é a de homogeneidade de variâncias dos tratamentos, pois quando as variâncias dos tratamentos são heterogêneas, tanto as taxas de erro tipo I dos testes de normalidade, quanto as do teste F superam o nível de significância teórico, em praticamente todos os cenários. Inclusive, em relação ao teste F, caso a pressuposição de homogeneidade de variâncias não possa ser atendida, recomenda-se a sua não aplicação, sendo melhor, em conformidade com Nguyen et al. (2019), optar por métodos não paramétricos ou semi-paramétricos, como o teste de Wilcox proposto por Wilcox (1988), assim como o teste de Welch proposto por Welch (1951), pois tendem a ser mais capazes de controlar as taxas de erro tipo I.

Portanto, ao realizar uma ANOVA qualquer, a primeira pressuposição, que deve ser verificada, é a de homogeneidade de variâncias dos tratamentos, por meio, por exemplo de algum teste de hipóteses para múltiplas variâncias. Caso essa pressuposição não seja atendida, não se recomenda realizar nenhum dos testes de normalidade comparados, nem empregar o teste F da ANOVA. Nesse caso, recomenda-se o emprego de algum método não paramétrico para comparação das médias dos tratamentos. Caso contrário, caso haja homogeneidade de variâncias dos tratamentos, a ANOVA é recomendada, pois o teste F apresenta taxas de erro

tipo I controladas e elevados níveis de poder, independentemente do número de repetições por tratamento, até mesmo quando a pressuposição de normalidade não é atendida.