

**UNIVERSIDADE FEDERAL DE VIÇOSA**

**Transformações de dados e análise de variância: teoria, técnicas e impactos**

Jhennifer dos Santos Nascimento  
*Doctor Scientiae*

**VIÇOSA - MINAS GERAIS  
2024**

**JHENNIFER DOS SANTOS NASCIMENTO**

**Transformações de dados e análise de variância: teoria, técnicas e impactos**

Tese 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 *Doctor Scientiae*.

Orientador: Paulo Cesar Emiliano

**VIÇOSA - MINAS GERAIS  
2024**

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

T

N244t Nascimento, Jhennifer dos Santos, 1996-  
2024 Transformações de dados e análise de variância: teoria,  
técnicas e impactos / Jhennifer dos Santos Nascimento. – Viçosa,  
MG, 2024.  
1 tese eletrônica (86 f.): il. (algumas color.).

Orientador: Paulo César Emiliano.  
Tese (doutorado) - Universidade Federal de Viçosa,  
Departamento de Estatística, 2024.

Inclui bibliografia.

DOI: <https://doi.org/10.47328/ufvbbt.2025.001>

Modo de acesso: World Wide Web.

1. Análise de variância - Métodos de simulação.  
2. Unidades de medida. 3. Agronomia - Estatísticas. I. Emiliano,  
Paulo César, 1982-. II. Universidade Federal de Viçosa.  
Departamento de Estatística. Programa de Pós-Graduação em  
Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 519.538

# JHENNIFER DOS SANTOS NASCIMENTO

## Transformações de dados e análise de variância: teoria, técnicas e impactos

Tese 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 *Doctor Scientiae*.

APROVADA: 16 de dezembro de 2024.

Assentimento:

---

Jhennifer dos Santos Nascimento  
Autora

---

Paulo Cesar Emiliano  
Orientador

Essa tese foi assinada digitalmente pela autora em 02/01/2025 às 11:42:13 e pelo orientador em 02/01/2025 às 16:50:20. As assinaturas têm validade legal, conforme o disposto na Medida Provisória 2.200-2/2001 e na Resolução nº 37/2012 do CONARQ. Para conferir a autenticidade, acesse <https://siadoc.ufv.br/validar-documento>. No campo 'Código de registro', informe o código **KZ1T.W4AR.5YUR** e clique no botão 'Validar documento'.

**À minha grande e amada família, dedico este trabalho.**

## AGRADECIMENTOS

Meu período de doutorado se resume às palavras "saudade" e "gratidão". Saudade de casa quando estava longe da minha família, e saudade dos amigos incríveis que fiz durante essa etapa da minha vida. Sempre orei para que Deus guiasse todos os meus passos, me levasse a um lugar onde pudesse realizar meus sonhos e me aproximasse ainda mais Dele. E assim Ele fez. Hoje, ao olhar para minha trajetória, sei e sinto que vivi a ação de Deus em cada dia, em cada decisão e em cada passo.

Acredito que cada pessoa colocada em meu caminho foi escolhida a dedo por Deus. Thaynara, Jaqueline, Alice, Wagner, Luciano, Samantha, Marco e Laucinéia: obrigada pela companhia diária e pelo companheirismo. Obrigada pelas conversas divertidas, pelas trocas construtivas e por todo o apoio e suporte emocional que me deram com tanto carinho. A amizade de vocês é preciosíssima, saibam que vocês estarão para sempre no meu coração.

Agradeço à minha família, especialmente ao meu esposo, Lucas, que sempre apoiou meus sonhos de forma incondicional, torcendo e acreditando em mim até mesmo quando eu mesma não conseguia. Família, foi por vocês que cheguei aqui, por vocês que tive a ousadia de sonhar alto e a força para continuar. Obrigada pelo incentivo, pelas orações e pelos sacrifícios que fizeram para que eu me tornasse a mulher que sou hoje. Mãe, pai, vó, tias e meus irmãos, vocês moldaram o meu caráter e fizeram de mim quem sou. Amo vocês incondicionalmente e sempre levarei comigo o amor e os valores que me ensinaram.

Meu agradecimento também ao Júnior, secretário do programa de pós-graduação do qual faço parte. Foi o seu chamado que me motivou a ingressar no doutorado em plena pandemia. Você é um grande amigo, e sou imensamente grata pelo impulso e apoio que me deu.

Expresso minha imensa gratidão ao meu orientador, Paulo Emiliano. Sua orientação foi transformadora em minha vida. Obrigada por ser um orientador tão prestativo e dedicado. Saiba que me inspiro não só no profissional que você é, mas também na sua humildade. Aprender com sua maneira de ser tem sido um privilégio.

Sou grata ao programa PPESTBIO, à UFV e, em especial, ao GESTBIO. Obrigada pelo acolhimento. Tenho certeza de que cresci muito como pessoa e como profissional ao fazer parte de vocês.

Agradeço à agência de fomento CAPES pela bolsa de estudos. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001. O suporte financeiro foi fundamental para a conclusão dessa jornada.

A Ele, toda a minha gratidão, pois nada seria possível sem Sua graça. Nos momentos de fraqueza, encontrei força em Suas promessas, e nos momentos de dúvida, encontrei paz em Sua presença. Tudo posso porque Ele me fortalece e, com Deus ao meu lado, nenhuma adversidade é grande demais. Sua presença me sustentou em cada passo desta jornada, e hoje, olho para trás com o coração cheio de gratidão por Sua fidelidade e amor incondicional.

**“All my life, I've been carried by grace. You held me steady so I wouldn't give up. You opened doors that nobody could shut. It's not coincidence and it's not luck. It's nothing short of a miracle I'm here”. (Joe L Barnes)**

## RESUMO

NASCIMENTO, Jhennifer dos Santos, D.Sc., Universidade Federal de Viçosa, dezembro de 2024. **Transformações de dados e análise de variância: teoria, técnicas e impactos.** Orientador: Paulo Cesar Emiliano.

A transformação de um conjunto de dados para outra escala matemática é uma técnica comum utilizada em análises estatísticas do tipo ANOVA, com o propósito de satisfazer os pressupostos de normalidade e homogeneidade de variâncias para tornar a análise de variância válida. Esta técnica começou a ser utilizada em 1898 por Edgeworth e permanece nas publicações científicas até os dias atuais. As transformações de dados em análises estatísticas, especialmente para modelos lineares, têm o objetivo de tornar os dados e os termos de erro do modelo mais próximos de uma distribuição normal, sendo possível utilizá-las para corrigir heterogeneidade de variâncias ou para tornar um modelo aditivo. Surgem questionamentos a respeito desta técnica, tais como: se a interpretação dos dados deve ser feita na nova escala, ou se deve ser aplicada a inversa da transformação para a interpretação ser realizada na escala original; se é mais adequado utilizar uma análise não paramétrica, que não necessita de pressuposições de normalidade e homogeneidade de variâncias. O principal objetivo deste estudo é avaliar a capacidade de correção da não normalidade e heterogeneidade de variâncias das principais transformações de dados utilizadas pela comunidade científica de ciências agrárias. Para isso, realizamos duas revisões sistemáticas de literatura de trabalhos que realizaram análise de variância e utilizaram transformações de dados para validar seus pressupostos, a primeira fornece uma noção de como a técnica vem sendo aplicada na área de ciências agrárias como um todo, e a segunda para investigar o uso da técnica em quatro de seus subcampos (entomologia, fitotecnia, florestal e solos). Ambas as revisões identificaram as transformações logarítmica, raiz quadrada, arco seno e Box-Cox como as mais utilizadas na área para a validação dos pressupostos da análise de variância, além de erros de execução e reporte da técnica. Com base nisso, realizamos um estudo de simulação para avaliar a eficácia dessas quatro transformações na correção da não normalidade e heterogeneidade, destacando a superioridade da transformação Box-Cox na correção simultânea dos pressupostos; elaboramos diretrizes práticas para orientar pesquisadores na execução da ANOVA; e por fim, desenvolvemos um aplicativo web para permitir que pesquisadores sem afinidade com programação possam ter acesso ao método de triagem que utilizamos nas revisões sistemáticas executadas.

Palavras-chave: revisão sistemática ; ciências agrárias ; pressuposições da anova ;  
simulação ; diretrizes

## ABSTRACT

NASCIMENTO, Jhennifer dos Santos, D.Sc., Universidade Federal de Viçosa, December, 2024. **Data transformations and variance analysis: theorie, techniques and impacts.** Adviser: Paulo Cesar Emiliano.

The transformation of a dataset to another mathematical scale is a common technique used in ANOVA-type statistical analyses, aimed at satisfying the assumptions of normality and homogeneity of variances to validate the analysis of variance. This technique began to be used in 1898 by Edgeworth and remains prevalent in scientific publications to this day. Data transformations in statistical analyses, particularly for linear models, aim to make the data and error terms of the model closer to a normal distribution. They can be employed to correct heterogeneity of variances or to make a model additive. Questions arise regarding this technique, such as whether data interpretation should be performed on the new scale or if the inverse transformation should be applied to interpret results on the original scale; or whether it is more appropriate to use a non-parametric analysis that does not require assumptions of normality and homogeneity of variances. The primary objective of this study is to assess the ability of the most commonly used data transformations in the agricultural sciences research community to correct non-normality and heterogeneity of variances. To this end, we conducted two systematic literature reviews of studies that performed analysis of variance and used data transformations to validate their assumptions. The first review provides an overview of how the technique has been applied in agricultural sciences as a whole, while the second focuses on its use in four subfields (entomology, crop science, forestry, and soil science). Both reviews identified logarithmic, square root, arcsine, and Box-Cox transformations as the most frequently used techniques for validating variance analysis assumptions, alongside errors in the execution and reporting of the technique. Based on these findings, we conducted a simulation study to evaluate the effectiveness of these four transformations in correcting non-normality and heterogeneity, highlighting the superiority of the Box-Cox transformation in simultaneously addressing both assumptions. We developed practical guidelines to assist researchers in conducting ANOVA and created a web application to provide access to our systematic review screening method for researchers with no programming expertise.

Keywords: literature review; agrarian sciences; anova assumptions; simulation; guidelines

# SUMÁRIO

INTRODUÇÃO GERAL . . . . .	12
1 REVISÃO DE LITERATURA . . . . .	17
1.1 Análise de variância . . . . .	17
1.2 Testes para verificação das pressuposições da análise de variância . . . . .	20
1.2.1 Normalidade . . . . .	20
1.2.2 Homogeneidade das variâncias . . . . .	26
1.3 Transformações . . . . .	29
1.3.1 Transformação logarítmica . . . . .	31
1.3.2 Transformação raiz quadrada . . . . .	33
1.3.3 Transformação arco seno . . . . .	35
1.3.4 Transformação Box-Cox . . . . .	36
1.3.5 Transformação potência . . . . .	36
1.3.6 Transformação recíproca . . . . .	37
1.3.7 Transformação de Johnson . . . . .	37
1.4 Revisão sistemática . . . . .	38
2 TRANSFORMAÇÃO DE DADOS ESTATÍSTICOS NAS CIÊNCIAS AGRÁRIAS PARA ANÁLISE DE VARIÂNCIA: UMA REVISÃO SISTEMÁTICA . . . . .	46
2.1 Introdução . . . . .	47
2.2 Métodos . . . . .	49
2.2.1 Coleta de dados . . . . .	49
2.2.2 Análise descritiva . . . . .	50
2.3 Resultados . . . . .	51
2.3.1 Coleta de dados . . . . .	51
2.3.2 Análise descritiva . . . . .	51

2.4	Discussão . . . . .	54
2.4.1	Coleta de dados . . . . .	54
2.4.2	Análise descritiva . . . . .	54
2.5	Conclusões . . . . .	56
3	TRANSFORMAÇÕES DE DADOS PARA VIOLAÇÕES DAS SUPOSIÇÕES DE ANOVA EM SUBCAMPOS DAS CIÊNCIAS AGRÁRIAS: UMA REVISÃO SISTEMÁTICA, ESTUDO DE SIMULAÇÃO E DIRETRIZES PRÁTICAS . . . . .	60
3.1	Introdução . . . . .	61
3.2	Revisão sistemática . . . . .	62
3.2.1	Metodologia da revisão sistemática . . . . .	62
3.2.2	Resultados e discussão da revisão sistemática . . . . .	65
3.3	Estudo de simulação . . . . .	73
3.3.1	Metodologia do estudo de simulação . . . . .	73
3.3.2	Resultados e discussão do estudo de simulação . . . . .	74
3.4	Diretrizes para o uso de transformações de dados na validação da análise de variâncias . . . . .	75
3.5	Conclusão . . . . .	79
	CONCLUSÃO GERAL . . . . .	82

# INTRODUÇÃO GERAL

Nas ciências agrárias, a análise de variância (ANOVA) e as transformações de dados têm sido frequentemente utilizadas em conjunto desde pelo menos 1956 até 2024 (Nascimento *et al.*, 2024b). A ANOVA é empregada para detectar diferenças entre múltiplos grupos, enquanto as transformações de dados são aplicadas para ajudar na validação dos pressupostos da ANOVA, como a normalidade dos resíduos e a homogeneidade das variâncias. Embora seja uma metodologia consolidada, há registros que seu uso continua a ser explorado em diversos subcampos, refletindo a necessidade de avaliar como a comunidade acadêmica tem aplicado essa metodologia. Este trabalho visa explorar essa interseção, revisando sistematicamente a aplicação das transformações de dados na literatura acadêmica na área de ciências agrárias e em seus subcampos entomologia, fitotecnia, florestal e solos.

A ANOVA é utilizada para testar a hipótese de igualdade estatística entre um conjunto de médias populacionais, sendo preferível ao teste t devido ao aumento da probabilidade de ocorrência do erro de tipo I ao comparar múltiplas médias duas a duas utilizando o teste t, conforme indicado por Mohr *et al.* (2021). Para validar a análise de variância, é necessário satisfazer pressuposições tais como aditividade dos efeitos do tratamento e ambientais, independência, homocedasticidade das variâncias e normalidade dos erros experimentais. Quando essas pressuposições não são atendidas, podem-se aplicar transformações de dados para validar a análise conforme destacado por Cochran (1947), Doncaster e Davey (2007), Christensen (2017), Ramachandran e Tsokos (2021), Assis *et al.* (2023).

As transformações de dados, inicialmente propostas por Edgeworth (1898) e referidas como Método de Translação, surgiram da necessidade de relacionar distribuições observadas à distribuição normal. Segundo Bartlett (1947), o propósito usual da transformação é alterar a escala das medições para tornar a análise de variância válida. No entanto, uma transformação é bem-sucedida apenas se satisfizer os pressupostos de homogeneidade das variâncias, estrutura aditiva e normalidade dos erros através da mesma, conforme afirmado por Atkinson (1985).

Após a aplicação de qualquer transformação de dados, é fundamental verificar se ela garante a validade dos pressupostos subjacentes à análise, como enfatizado por Quinn e Keough (2002) e Pierre *et al.* (2018). Afinal, uma transformação que não satisfaça os pressupostos exigidos é inutilizável, pois não apenas compromete a validade da análise original, mas também adiciona um segundo nível de erro, decorrente da interpretação inadequada dos dados transformados.

Embora as transformações sejam amplamente utilizadas, Oliveira *et al.* (2018) destacam que

elas recebem de críticas teóricas e práticas. Box e Cox (1964) apontam que elas podem alterar a interpretação dos resultados, enquanto Wilcox (2017) ressalta que, apesar de corrigirem problemas como a não normalidade ou a heterocedasticidade, nem sempre são eficazes, podendo introduzir novos problemas, como falsos positivos ou negativos. Portanto, como destaca Rupert (2001), é crucial que os pesquisadores considerem cuidadosamente as implicações antes de aplicar transformações em seus dados, avaliando se os benefícios superam os riscos.

As transformações de dados são amplamente reconhecidas na literatura por razões históricas e metodológicas. Originalmente, quando os métodos estatísticos paramétricos, como a análise de variância, ganharam popularidade no século XX, os testes não paramétricos ainda eram limitados em disponibilidade e poder estatístico para amostras pequenas ou médias (Conover, 1999). Diante disso, as transformações de dados surgiram como uma solução prática para lidar com violações dos pressupostos de normalidade e homogeneidade de variâncias. Autores como Box e Cox, nos anos 1960, consolidaram essa abordagem, contribuindo para sua disseminação como uma prática da época.

Mesmo após o desenvolvimento e popularização dos testes não paramétricos, as transformações de dados continuaram a ser amplamente utilizadas devido a vantagens como: (1) permitem a aplicação de métodos paramétricos, que tendem a ser mais poderosos em detectar efeitos quando os pressupostos são atendidos (Zar, 2010); (2) preservam a comparabilidade entre estudos, uma vez que boa parte da literatura consolidada baseia-se em análises paramétricas (Sokal e Rohlf, 2012); e (3) são amplamente documentadas, ensinadas e aplicadas, formando uma tradição metodológica que permanece válida em muitas áreas, como as ciências agrárias e biológicas (Gomez e Gomez, 1984). Fazendo com que, a prática de transformação de dados reflita tanto sua origem histórica quanto sua aplicabilidade em diferentes contextos científicos.

Testes paramétricos, cuja estatística é baseada em uma função de distribuição de probabilidade, dependem do cumprimento de pressupostos que assegurem as condições necessárias para os parâmetros dessa distribuição. Quando esses pressupostos são violados, a validade da análise é comprometida. Embora, a rigor, a análise possa ser considerada inválida, há casos em que ela apresenta robustez suficiente para resistir a determinadas violações, dependendo do tipo de teste e do pressuposto em questão, como destacado por Wilcox (2017). Com base em Zar (2010), testes não paramétricos não dependem de pressupostos rígidos sobre a distribuição dos dados, sendo baseados em ordens, ranks ou outras características não relacionadas diretamente aos parâmetros da distribuição. Os testes não paramétricos se popularizaram entre 1930 e 1970 com as obras de Wilcoxon (1945), Mann e Whitney (1947), Siegel (1956) em conjunto com o avanço computacional nas décadas de 1960 e 1970.

Nas ciências agrárias, como evidenciado no estudo de Nascimento *et al.* (2024a), que realizou uma revisão sistemática da literatura para mapear o uso de transformações de dados na validação dos pressupostos da ANOVA, documenta que essa técnica continua sendo amplamente empregada até os dias atuais, reafirmando a importância da investigação sobre essa técnica na literatura. Dando continuidade a essa linha de investigação, o presente trabalho explora a análise de variância,

as técnicas de transformação de dados e a metodologia de revisão sistemática, seguindo as diretrizes dos Itens de relatório preferidos para revisões sistemáticas e meta-análises (PRISMA), conforme estabelecido por Moher *et al.* (2009).

O principal objetivo deste trabalho é avaliar as transformações de dados utilizadas para validar os pressupostos da análise de variância nas ciências agrárias, com foco em quatro subcampos específicos: entomologia, fitotecnia, florestal e solos. Para isso, foram realizadas duas revisões sistemáticas da literatura: a primeira aborda a aplicação de transformações nas ciências agrárias como um todo (Capítulo 2), enquanto a segunda investiga seu uso em subcampos específicos (Capítulo 3). Com base nas informações obtidas, conduzimos um estudo de simulação para avaliar a eficácia das transformações mais frequentes na correção de pressupostos estatísticos, como normalidade e homogeneidade das variâncias. Além disso, também foram identificados equívocos metodológicos recorrentes e propostas diretrizes práticas que visam minimizar falhas na aplicação e no reporte de análises de variâncias com e sem transformações de dados.

Como parte desta pesquisa, foi desenvolvido um aplicativo projetado especificamente para realizar a triagem em revisões sistemáticas, baseado na busca de termos no texto completo dos trabalhos. Embora esse processo possa ser realizado diretamente em Python, o aplicativo foi criado com o objetivo de tornar essa funcionalidade acessível também a pesquisadores sem conhecimentos em programação. Diferentemente de outros aplicativos web, que geralmente abrangem todo o processo de revisão sistemática em uma única interface, este se concentra exclusivamente na etapa de triagem. Essa abordagem oferece uma solução específica para situações em que a informação de interesse está localizada no corpo do texto, em vez de estar presente no resumo, título ou palavras-chave. O aplicativo está disponível de forma gratuita em <https://systematicreviewscreeningapp.streamlit.app/>.

Este trabalho apresenta a presente introdução geral ao estudo, contextualizando sua relevância e objetivos. O Capítulo 1 traz uma revisão de literatura que aborda a análise de variância, os testes para a verificação dos pressupostos de normalidade e homogeneidade de variâncias, as transformações de dados e a metodologia de revisão sistemática. No Capítulo 2, é apresentada a revisão sistemática realizada na área de ciências agrárias. O Capítulo 3 explora a revisão sistemática aplicada aos subcampos de entomologia, fitotecnia, florestal e solos, além de descrever o estudo de simulação para avaliar a eficácia das transformações de dados e propor diretrizes para a execução e o reporte adequado da análise de variância. Por fim, o trabalho apresenta as conclusões gerais, destacando as contribuições científicas e metodológicas desta pesquisa, bem como o Anexo 3.5, que reúne definições de diversos termos estatísticos mencionados no texto, para orientar leitores que não estão familiarizados com a linguagem da área.

# Referências Bibliográficas

ASSIS, J. P.; DE LIMA, I. R. P.; FRANÇA, J. A.; DE SOUSA, R. P.; LINHARES, P. C. F.; CUSTÓDIO, T. N.; DE SOUSA, R. P.; RODRIGUES, W. M. *Transformação de dados aplicada à estatística*. 1st ed. Brasil: Pantanal Editora, 2023. 942 p.

ATKINSON, A. C. *Plots, transformations and regression*. 1st ed. USA: Oxford Science Publications, 1985. 282 p.

BARTLETT, M. S. The use of transformations. *Biometrics*, v. 3, p. 39–52, 1947.

BOX, G. E. P.; COX, D. R. An analysis of transformations. *Journal of the Royal Statiscian Society*, v. 26, p. 211–252, 1964.

CHRISTENSEN, R. *Analysis of variance, design, and regression*. 2nd ed. Albuquerque: CRC Press, 2017. 606 p.

COCHRAN, W. G. When the assumptions for the analysis of variance are not satisfied. *Biometrics*, v. 3, p. 22–38, 1947.

CONOVER, W. J. *Practical nonparametric statistics*. 3rd ed. New York: Wiley, 1999. 584 p.

DONCASTER, C. P.; DAVEY, A. J. H. *Analysis of variance and covariance: how to choose and construct models for the life sciences*. 1st ed. New York: Cambridge University Press, 2007. 302 p.

EDGEWORTH, F. Y. A suggestion for using powerful and informative tests of normality. *Journal of the Royal Statiscian Society*, v. 61, p. 534–544, 1898.

GOMEZ, K. A.; GOMEZ, A. A. *Statistical procedures for agricultural research*. 2nd ed. New York: Wiley, 1984. 704 p.

MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, v. 18, n. 1, p. 50–60, 1947.

MOHER, D.; LIBERATI, A.; TETZLAFF, J.; ALTMAN, D. G. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *British Medical Journal*, v. 6, p. 332–336, 2009.

- MOHR, D. L.; WILSON, W. J.; FREUND, R. J. *Statistical methods*. 4th ed. Chennai: Academic Press, 2021. 784 p.
- NASCIMENTO, J. S. N.; SILVA, J.; BERNARDES, R. C.; COSTA, G.; EMILIANO, P. C. Statistical data transformation in agrarian sciences for variance analysis: a systematic review. *F1000Research*, p. 1–12, 2024a.
- NASCIMENTO, J.; GONÇALVES, L. B.; BARBOSA, W. F.; BELO, L. P.; SILVA, J. F.; COSTA, G. S.; EMILIANO, P. C. Data transformations for ANOVA assumption violations in agrarian sciences subfields: a systematic review, simulation study, and practical guidelines. *DELLOS*, p. 1–27, 2024b.
- OLIVEIRA, J. P. R.; SANTANA, D. G.; PEREIRA, V. J.; SANTOS, C. M. Data transformation: an underestimated tool by inappropriate use. *Acta Scientiarum*, v. 40, p. 1–11, 2018.
- PIERRE, A. P.; SHIKON, V.; SCHNEIDER, D. C. Count data in biology: data transformation or model reformation? *Ecological Evolution*, v. 8, p. 3077–3085, 2018.
- QUINN, G. P.; KEOUGH, M. J. *Experimental design and data analysis for biologists*. 1st ed. New York: Cambridge University Press, 2002. 537 p.
- RAMACHANDRAN, K. M.; TSOKOS, C. P. *Mathematical statistics with applications in R*. 3rd ed. London: Academic Press, 2021. 704 p.
- RUPERT, D. Statistical analysis, special problems of: transformations of data. *Society Behavioral Science*, v. 20, p. 15007–15014, 2001.
- SIEGEL, S. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill, 1956. 312 p.
- SOKAL, R. R.; ROHLF, F. J. *Biometry: the principles and practice of statistics in biological research*. 4th ed. New York: W.H. Freeman, 2012. 937 p.
- WILCOXON, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, v. 1, n. 6, p. 80–83, 1945.
- WILCOX, R. R. *Introduction to robust estimation and hypothesis testing*. 4th ed. Academic Press, 2017. 810 p.
- ZAR, J. H. *Biostatistical analysis*. 5th ed. Upper Saddle River, NJ: Pearson, 2010. 944 p.

# Capítulo 1

## REVISÃO DE LITERATURA

### 1.1 Análise de variância

A análise de variância (ANOVA) surgiu com a necessidade de testar a hipótese de igualdade estatística de um conjunto de médias populacionais. Segundo Mohr *et al.* (2021), é possível usar o teste  $t$  repetidamente para amostras independentes até que todos os pares possíveis de médias populacionais tenham sido comparadas, entretanto esta abordagem gera problemas, tais como, a quantidade de testes a serem feitos e o aumento do verdadeiro nível de significância da análise, que seria mais alto do que o nível de significância usado nos testes individuais.

Conforme Herzog *et al.* (2019), a probabilidade de ocorrência do erro de tipo I é dada por  $1 - (1 - \alpha)^m$ , para  $m$  testes independentes. Por tanto, se a título de exemplo, utilizar o teste  $t$  para verificar a igualdade de 6 médias com nível de significância  $\alpha = 0,05$ , seria necessário realizar 15 testes, resultantes da combinação de seis, dois a dois ( $C_{6,2}$ ), cada um com nível de confiança  $(1 - \alpha) = 0,95$ , para comparar todos os pares. Com isso, assumindo independência dos testes, a probabilidade de ocorrência do erro de tipo I seria  $1 - (1 - 0,05)^{15} = 0,54$ , em vez de 0,05, como preespecificado, e por este motivo é mais adequado utilizar os procedimentos da análise de variância para a comparação múltipla de médias.

Segundo Quinn e Keough (2002), o termo análise de variância é comumente usado para se referir ao caso em que as variáveis explicativas são qualitativas, geralmente no contexto de experimentos planejados, em que as categorias de cada fator são os grupos ou tratamentos experimentais, e análise de regressão para o caso em que as variáveis explicativas são quantitativas. Os autores ainda ressaltam que tanto a “regressão clássica” quanto a “ANOVA clássica” mensuram e particionam a variação da variável resposta, que é explicada ou predita por meio de um modelo linear ajustado<sup>1</sup>..

Historicamente, no início do século XX, com o objetivo de analisar estatisticamente as causas da variabilidade humana, cuja herdabilidade genética segue o esquema de herança de Mendel (1866), Fisher (1919) utilizou o quadrado do desvio padrão, nomeado na época como quantidade de variação

---

<sup>1</sup>A variação total da variável resposta é decomposta em diferentes componentes, dependendo do modelo ajustado. Isso é feito para identificar quanto da variação observada pode ser atribuída aos efeitos explicados pelo modelo e quanto é devido ao erro residual (ou variação não explicada).

da população normal, como medida de variabilidade dos desvios causados por duas fontes de variações genéticas independentes, e decompôs a variabilidade total produzida atribuindo à essas duas fontes suas respectivas porcentagens da variação total. Neste artigo, Fisher mencionou pela primeira vez o termo “análise de variância”, atualmente conhecido como ANOVA, que posteriormente em 1925, na primeira edição do livro *Statistical methods for research workers*, foi estabelecido que a análise de variância consiste em separar a variação atribuível a um grupo de causas mensurável, da variação atribuível a outros grupos.

Speed (1987) realiza uma análise crítica do conceito de análise de variância, questionando seu significado e levantando dúvidas sobre a interpretação dessa técnica. Ele examina como a ANOVA se relaciona a somas de quadrados, modelos lineares e testes  $F$ , destacando as complexidades conceituais envolvidas. Além disso, Speed cita a discussão histórica entre Fisher e outros estatísticos, como Wishart (1934), na qual Fisher argumenta que a ANOVA não deve ser vista como um teorema matemático, mas como um método prático para organizar os cálculos aritméticos necessários na análise estatística.

Eisenhart (1947) os problemas abordados pela análise de variância em duas categorias: na Classe I, os parâmetros de interesse são as médias, e o foco está nas inter-relações entre elas; já na Classe II, os parâmetros analisados são variâncias, com modelos matemáticos distintos para atender às diferentes questões dos dados. A análise de variância tradicional se enquadra principalmente na Classe I, pois seu objetivo principal é testar hipóteses sobre médias e identificar diferenças estatisticamente significativas entre elas. Eisenhart também discute os pressupostos necessários para a validade da ANOVA, como independência, normalidade e homogeneidade das variâncias. As demonstrações das somas de quadrados que compõem a tabela da ANOVA podem ser encontradas na obra clássica de Snedecor e Cochran (1989).

Cochran (1947) aborda de forma sucinta as pressuposições necessárias para a validação da análise de variância e os efeitos causados por sua violação. Segundo ele, os efeitos dos tratamentos e os efeitos ambientais devem ser aditivos, os erros experimentais devem ser independentes, apresentar variâncias homogêneas (homocedasticidade) e seguir uma distribuição normal. Quando essas condições não são atendidas, surgem consequências específicas: a não normalidade e a heterogeneidade dos erros resultam em perda de eficiência na estimativa dos efeitos dos tratamentos, reduzindo o poder dos testes  $F$  e  $t$ ; a não independência dos erros compromete a acurácia das estimativas das médias dos tratamentos; e a não aditividade dificulta a separação dos efeitos, prejudicando a interpretação do modelo.

Eisenhart (1947) complementa Cochran (1947) assegurando que a normalidade junto às demais pressuposições é suficiente para a realização de testes de significância e, portanto, quando as quatro pressuposições são satisfeitas, todos os procedimentos usuais da análise de variância para estimar, testar e inferir sobre a existência de relações lineares entre médias populacionais são estritamente válidos. Quando descumpridos, Bartlett (1947) indica as transformações raiz quadrada, logarítmica, arco seno entre outras, como ferramentas para violações dos pressupostos de homogeneidade, normalidade e efeitos aditivos.

Três das quatro pressuposições da análise de variância são avaliadas no erro experimental, que devem ser normais, independentes e com variâncias homogêneas. Segundo Taylor (1997), nenhuma grandeza (comprimento, tempo, temperatura e etc.) pode ser medida com total certeza, no entanto, as incertezas podem ser reduzidas até que sejam extremamente pequenas, mas eliminá-las completamente é impossível. Cochran (1947) salienta que este tipo de erro inclui todos os elementos de variação que não são atribuídos aos efeitos de tratamentos ou características do ambiente que podem ser identificadas e controladas. Portanto, ao descrever um fenômeno de natureza aleatória com um modelo linear haverá um erro experimental associado, já que conforme Olive (2017), um modelo estatístico é usado para fornecer uma aproximação para as características da população que gerou os dados.

O erro estatístico está diretamente relacionado com a precisão de uma estimativa. Conforme Kish (1965), precisão é definida como o inverso do erro total, incluindo viés e variância, ou seja, quanto maior o erro, menor a precisão de uma estimativa. Na experimentação, independente do delineamento utilizado, o erro estará presente. E com isso, para obter estimativas precisas é necessário utilizar os melhores estimadores possíveis. Conforme Casella e Berger (2002), sob normalidade, os estimadores de mínimos quadrados ordinários do modelo da ANOVA são diferenças entre médias, que por sua vez é um estimador não viesado e com variância mínima. Por isso a análise de variância é tão utilizada na comunidade científica, pois possui ótimos estimadores, quando não há quebra de pressupostos.

Lix *et al.* (1996) e Acutis *et al.* (2012) destacam que o uso da ANOVA para testar a igualdade de médias entre grupos é frequente em pesquisas educacionais e psicológicas. Nas ciências agrárias, essa técnica também se consolidou como um método padrão para a análise de resultados experimentais, auxiliando os pesquisadores a transferir informações confiáveis, oriundas de pesquisas científicas, e a desenvolver serviços de assistência eficazes para agricultores e fazendeiros.

O teste  $F$  de Snedecor (1934), foi estabelecido como uma razão de variâncias por Fisher e Yates (1938). Conforme Miller (1984), pode-se mostrar que a razão de duas variâncias amostrais resulta na distribuição conhecida como  $F$ , podendo esta ser utilizada para determinar a probabilidade de obter uma diferença tão grande ou maior do que a diferença observada entre as variâncias de duas amostras aleatórias. Cochran e Cox (1992) ressaltam que o teste  $F$  é relativamente robusto a pequenos desvios da suposição de normalidade, no entanto, na presença de heterocedasticidade e quando o número de repetições não é igual, o teste tende a ser mais sensível tendendo a rejeitar mais a hipótese de nulidade.

Desse modo ao realizar uma ANOVA, o teste  $F$  é utilizado para verificar se múltiplas variâncias diferem estatisticamente, que por sua vez permite concluir a respeito da igualdade entre as médias <sup>2</sup>. Mohr *et al.* (2021) evidenciam que ao realizar este tipo de análise para fazer inferências sobre dife-

---

<sup>2</sup>O teste  $F$ , utilizado na ANOVA, compara a variância entre grupos (variância explicada pelo modelo) com a variância dentro dos grupos (variância residual). Essa relação é expressa pela razão  $F = \frac{\text{variância entre grupos}}{\text{variância dentro dos grupos}}$ . Um valor elevado de  $F$  indica que as diferenças entre os grupos são maiores do que o esperado pelo acaso, permitindo rejeitar a hipótese nula de igualdade entre as médias dos grupos e concluir pela existência de diferenças estatisticamente significativas. Essa relação permite verificar se a variância entre os grupos é significativamente maior do que seria esperado pelo acaso, possibilitando concluir sobre a igualdade ou diferença estatística entre as médias dos grupos.

renças entre médias, é tomado como base a comparação da variância entre as médias das populações com a variância entre observações amostrais dentro das populações individuais. Os autores fazem as demonstrações necessárias e alguns exemplos, mais detalhes sobre a análise de variância podem ser encontrados em Murphy e Myers (2004); Greene e D'Oliveira (2005); Herzog *et al.* (2019).

Assis *et al.* (2023) destacam que fatores como a presença de outliers, assimetria extrema, variabilidade maior em certos tratamentos e variâncias dependentes das médias são os principais responsáveis por distúrbios na análise de variância. Esses fatores podem causar não normalidade e heterogeneidade nos resíduos dos grupos comparados. Para mitigar esses problemas, vários autores sugerem a transformação dos dados, incluindo Cochran (1947), Doncaster e Davey (2007), Christensen (2017), Ramachandran e Tsokos (2021) e Assis *et al.* (2023).

## 1.2 Testes para verificação das pressuposições da análise de variância

Antes de qualquer análise estatística ser realizada é essencial fazer verificações preliminares de seus dados para se assegurar de que as premissas da análise escolhida são atendidas (Quinn e Keough, 2002). Nesta seção, serão vistos os testes mais comuns para a verificação das pressuposições, normalidade e homogeneidade de variâncias de resíduos, e as condições para a aplicação de cada teste.

Diferentemente das pressuposições de normalidade e homogeneidade de variâncias residual, que serão vistas nas Subseções 1.2.1 e 1.2.2, conforme Christensen e Bedrick (1997), a pressuposição de independência não é uma propriedade da população em questão, mas uma propriedade obtida pela maneira que a população é amostrada. Deseja-se amostras que possuem independência entre grupos e independência das observações dentro de cada grupo, portanto para evitar os dois tipos de dependência é necessário garanti-las durante a coleta da amostra. As técnicas de amostragem tratadas em Bolfarine e Bussab (2005), garantem que as observações serão obtidas independentemente, com igual probabilidade, e retiradas de uma mesma população, pois deseja-se amostras que produzam resultados confiáveis e livres de viés.

### 1.2.1 Normalidade

A normalidade dos resíduos é uma pressuposição fundamental para muitos procedimentos estatísticos, como regressão linear, testes  $t$  e análise de variância, pois garante a validade das inferências feitas com base na amostra (Ghasemi e Zahediasl, 2012). Além disso, a normalidade dos resíduos também é importante para alguns testes de homogeneidade de variâncias, que serão abordados na Subseção 1.2.2. De acordo com Montgomery (2013), essa pressuposição pode ser verificada visualmente por meio de gráficos, como o histograma dos resíduos, que indica uma distribuição normal quando apresenta um formato de sino, e o gráfico de probabilidade normal, que deve formar uma linha reta em casos de normalidade. No entanto, embora esses métodos gráficos ofereçam uma visão preliminar da forma da distribuição, eles não garantem a normalidade nem testam formalmente se

a diferença entre a distribuição observada dos resíduos e a distribuição normal é estatisticamente significativa (Öztuna *et al.*, 2006). Assim, desvios consideráveis da normalidade podem ter consequências graves, exigindo análises adicionais e complementares, como a aplicação de testes de hipóteses formais (Montgomery, 2013).

### 1.2.1.1 Teste Anderson-Darling

O teste Anderson-Darling, desenvolvido por Anderson e Darling (1954), é usado para verificar a adequação de uma amostra a uma distribuição específica, comumente a distribuição normal. As hipóteses para o teste de Anderson-Darling são as seguintes:

$H_0$  : Os dados seguem distribuição especificada.

$H_1$  : Os dados não seguem distribuição especificada.

Para a amostra aleatória  $X_1, X_2, \dots, X_n$ , com  $n$  observações ordenadas e  $F(x_i)$ , a função de distribuição acumulada da distribuição normal, que é a distribuição particular de interesse neste tópico, a estatística do teste é dada por

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

Segundo os autores, se  $W_n^2$  for muito grande, a hipótese de que os dados seguem uma distribuição normal deve ser rejeitada com base em valores críticos obtidos de uma distribuição qui-quadrado ajustada, a qual depende do tamanho da amostra  $n$  e do nível de significância escolhido. Esse procedimento é particularmente eficaz quando a distribuição verdadeira difere da distribuição normal nas caudas. Além disso, Hintze (2011) conduziu um estudo de simulação comparando vários testes de normalidade, no qual o teste de Anderson-Darling se mostrou tão poderoso quanto o teste de Shapiro-Wilk em diversas situações.

### 1.2.1.2 Teste Shapiro-Wilk

O teste de Shapiro-Wilk foi originalmente desenvolvido por Shapiro e Wilk (1965) para testar a normalidade de amostras pequenas, originalmente limitado a amostras com até 20 observações. O teste é baseado na correlação entre os dados da amostra e os escores normais, em que valores baixos dessa correlação indicam que a amostra pode não ser proveniente de uma distribuição normal, como afirmado por Ghasemi e Zahediasl (2012).

Royston (1992, 1995) propôs extensões e aproximações que permitiram a aplicação do teste de Shapiro-Wilk para tamanhos de amostras maiores. Essas adaptações fizeram com que o teste pudesse ser aplicado a amostras de tamanho ilimitado, enquanto o teste original era restrito a amostras pequenas.

Conforme Öztuna *et al.* (2006), as hipóteses e estatística do teste são dadas por

$H_0$  : A amostra aleatória  $X_1, X_2, \dots, X_n$  é originária de uma distribuição normal.

$H_1$  : A amostra aleatória  $X_1, X_2, \dots, X_n$  não é originária de uma distribuição normal.

$$W = \frac{\left( \sum_{i=1}^n a_i X_{(i)} \right)^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

em que,  $X_{(i)}$  é a  $i$ -ésima estatística de maior ordem, ou seja,  $i$ -ésima observação ordenada;  $\bar{X}$  é a média amostral;  $a_{(i)}$  são os coeficientes calculados de acordo com as estatísticas de ordem da distribuição normal, e  $n$  é o número de observações.

Conforme Hintze (2011), esse teste de normalidade é a razão de duas estimativas da variância de uma distribuição normal baseada em uma amostra aleatória de  $n$  observações, sendo o numerador proporcional ao quadrado do melhor estimador linear do desvio padrão e o denominador a soma dos quadrados das observações em torno da média amostral. A estatística  $W$  pode ser interpretada como o quadrado do coeficiente de correlação de Pearson entre as observações ordenadas e um conjunto de pesos que são usados para calcular o numerador, que são proporcionais às esperanças das estatísticas de ordem da distribuição normal. Portanto, quanto mais próximo  $W$  estiver de 1, mais próximo de uma distribuição normal será a distribuição amostral.

A estatística  $W$  calculada pelo teste de Shapiro-Wilk, é comparada com valores críticos que são obtidos empiricamente a partir de simulações numéricas, conforme descrito por Shapiro e Wilk (1965) e expandido por Royston (1992). Esses valores críticos dependem do tamanho da amostra  $n$  e do nível de significância  $\alpha$ . Se o valor de  $W$  for menor do que o valor crítico correspondente, a hipótese nula de normalidade é rejeitada (Ghasemi e Zahediasl, 2012).

### 1.2.1.3 Teste Martinez-Iglewicz

Desenvolvido por Martinez e Iglewicz (1981), o teste proposto utiliza a mediana e uma estatística baseada no estimador *biweight* de escala. Esse estimador, introduzido por Tukey (1977), reduz o impacto de valores muito distantes da mediana, sendo adequado para dados com *outliers* e distribuições simétricas com caudas pesadas. Dada uma amostra aleatória  $X_1, X_2, \dots, X_k$ , de tamanho  $k$ , as hipóteses e estatística de teste são expressos por:

$H_0$  : Os dados seguem uma distribuição normal.

$H_1$  : Os dados não seguem uma distribuição normal.

$$I = \frac{\sum_{i=1}^k (X_i - \tilde{X})^2}{(n-1)S_b^2}$$

em que, a estatística  $I$  é a razão entre a soma dos quadrados das distâncias entre os valores da amostra  $X_i$  e a mediana amostral  $\tilde{X}$ , dividida pelo estimador de escala *Tukey's biweight*  $S_b^2$  ajustado por  $(n-1)$ , que é um estimador robusto da variância (Martinez e Iglewicz, 1981) e,

$$S_b^2 = \frac{n \sum_{|z_i| < 1} (X_i - \tilde{X})^2 (1 - z_i^2)^4}{\left[ \sum_{|z_i| < 1} (1 - z_i^2)(1 - 5z_i^2) \right]^2}, \text{ em que } \begin{cases} z_i = \frac{(X_i - \tilde{X})}{k \cdot \text{med}|X_i - \tilde{X}|}, \text{ para } |z_i| < 1; \\ z_i = 0, \text{ caso contrário;} \end{cases}$$

em que,  $z_i$  é a normalização da diferença entre o valor da amostra  $X_i$  e a mediana  $\tilde{X}$ , dividida pela mediana das distâncias absolutas entre os valores da amostra e a mediana, ou seja, a mediana dos desvios absolutos. Para valores de  $|z_i| \geq 1$ ,  $z_i$  é truncado para zero, o que reflete a exclusão de pontos distantes na computação do estimador.

A estatística  $I$  calculada pelo teste de Martinez-Iglewicz é comparada com valores críticos obtidos a partir de simulações numéricas, conforme descrito por Martinez e Iglewicz (1981). Esses valores críticos são dependentes do tamanho da amostra  $n$  e dos níveis de significância, e foram determinados para diversos tamanhos de amostra. Se o valor da estatística  $I$  se desviar significativamente do valor crítico, a hipótese nula de normalidade é rejeitada.

De acordo com Hintze (2011), valores da estatística de teste  $I$  próximos de 1 indicam que a distribuição dos dados é normal. Martinez e Iglewicz (1981) destacam que esse teste é particularmente poderoso para distribuições simétricas com caudas pesadas; contudo, para distribuições assimétricas, seu poder é reduzido. Esse teste também é abordado em Lee *et al.* (2016).

#### 1.2.1.4 Teste Kolmogorov-Smirnov

Conforme Hollander *et al.* (2013), o teste de Kolmogorov-Smirnov (K-S) é baseado na diferença máxima entre a função de distribuição acumulada observada e a função de distribuição acumulada esperada de uma distribuição teórica, que pode ser normal ou outra. Quanto menor essa diferença, maior a probabilidade de que a distribuição em análise se ajuste à distribuição teórica escolhida, mas o teste não se limita apenas à normalidade. Conforme os autores, as hipóteses e estatística  $D$  do teste K-S para testar normalidade são dadas por

$H_0$  : A amostra segue uma distribuição normal.

$H_1$  : A amostra não segue uma distribuição normal.

$$D_{KS} = \sup_x |F_n(x) - F(x, \mu, \sigma)|$$

em que,  $F(x, \mu, \sigma)$  é a função de distribuição acumulada teórica da distribuição normal com média  $\mu$  e variância  $\sigma^2$  e  $F_n(x)$  é a função de distribuição observada ou empírica dos dados observados.

Os autores destacam que, para aplicar o teste K-S original, os parâmetros populacionais  $(\mu, \sigma)$  precisam ser conhecidos. Caso contrário, eles podem ser substituídos por estimativas amostrais. No entanto, essa substituição resulta no teste de Kolmogorov-Smirnov corrigido por Lilliefors (1967), com estatística de teste dada por

$$D_L = \sup_x |F_n(x) - F^*(x)|$$

em que,  $F^*(x)$  é a função de distribuição acumulada da normal com média amostral  $\bar{x}$  e variância amostral  $s^2$ .

A estatística  $D$  calculada pelo teste de Kolmogorov-Smirnov é comparada com valores críticos obtidos a partir de simulações numéricas, conforme descrito por Lilliefors (1967). Esses valores críticos são dependentes do tamanho da amostra  $n$  e dos níveis de significância, e foram determinados para diferentes cenários. Quando os parâmetros da distribuição normal não são conhecidos, o teste de Kolmogorov-Smirnov modificado por Lilliefors é utilizado, e a hipótese nula é rejeitada se a estatística  $D$  exceder esses valores críticos.

Ghasemi e Zahediasl (2012) e Barton e Peat (2014) apontam que o teste de Kolmogorov-Smirnov é altamente sensível a valores extremos, e que a correção de Lilliefors torna o teste menos conservador.

### 1.2.1.5 Teste D'Agostino

D'Agostino e Belanger (1990) descrevem três testes conhecidos como os testes de normalidade de D'Agostino. O primeiro possui a estatística  $Z_s$  (Equação 1.1), baseado na estimativa  $\sqrt{b_1}$  do coeficiente de assimetria  $\sqrt{\beta_1}$ . O segundo, com estatística  $Z_k$  (Equação 1.2), é baseado na estimativa  $b_2$  do coeficiente  $\beta_2$  de curtose. E o terceiro, que possui estatística de teste  $K^2$  do teste *omnibus* D'Agostino (Equação 1.3), é baseado nas estimativas de ambos os coeficientes de assimetria e curtose da distribuição normal.

A definição das estimativas de  $\sqrt{\beta_1}$  e  $\beta_2$ , fundamenta-se nos trabalhos de D'Agostino e Pearson (1973) e foi posteriormente refinada por D'Agostino e Belanger (1990). D'Agostino e Pearson (1973) descrevem essas estimativas para capturar desvios de normalidade via assimetria e curtose em testes empíricos, enquanto o trabalho de D'Agostino e Belanger (1990) amplia esses conceitos para o teste *omnibus*, que combina ambas as estatísticas para uma verificação mais robusta da normalidade. Com base em ambas as referências, para uma amostra aleatória  $X_1, X_2, \dots, X_n$ , de tamanho  $n$ , as estimativas de  $\sqrt{\beta_1}$  e  $\beta_2$  são, respectivamente

$$\sqrt{b_1} = \frac{m_3}{m_2^{3/2}}, \quad b_2 = \frac{m_4}{m_2^2}, \quad \text{com } m_k = \frac{\sum_{i=1}^n (X_i - \bar{X})^k}{n} \quad \text{e } \bar{X} = \frac{\sum_{i=1}^n X_i}{n}.$$

Hintze (2011), assim como Ferreira (2018), ressaltam que os valores teóricos dos coeficientes de assimetria e curtose para uma distribuição normal são, respectivamente, zero e três. Portanto, se

$\sqrt{b_1}$  e  $b_2$  diferirem de zero e três, respectivamente, há indícios de que os dados não seguem uma distribuição normal.

Segundo D'Agostino e Belanger (1990) e Hintze (2011), as hipóteses e a estatística do teste de assimetria D'Agostino, são dados por

$H_0$  : Os dados originam-se de uma distribuição normal.

$H_1$  : Não normalidade devido a assimetria não normal.

$$Z_s = \delta \ln \left[ \frac{T}{\alpha} + \sqrt{\left(\frac{T}{\alpha}\right)^2 + 1} \right], \text{ em que } \begin{cases} \delta = \frac{1}{\sqrt{\ln W}}; \\ W^2 = -1 + \sqrt{2(C-1)}; \\ C = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}; \\ T = \sqrt{b_1 \left(\frac{(n+1)(n+3)}{6(n-2)}\right)}; \\ b_1 = \frac{m_3^2}{m_2^3}; \\ \alpha = \sqrt{\frac{2}{W^2-1}}. \end{cases} \quad (1.1)$$

Do teste de curtose D'Agostino, dados por

$H_0$  : Os dados originam-se de uma distribuição normal.

$H_1$  : Não normalidade devido a curtose não normal.

$$Z_k = \frac{\left(1 - \frac{2}{9A}\right) - \left(\frac{1 - \frac{2}{A}}{1 + G\sqrt{\frac{2}{A-4}}}\right)^{\frac{1}{3}}}{\sqrt{\frac{2}{9A}}}, \text{ em que } \begin{cases} A = 6 + \frac{8}{E} \left(\frac{2}{E} + \sqrt{1 + \frac{4}{E^2}}\right); \\ E = \frac{6(n^2-5n+2)}{(n+7)(n+9)} \sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}}; \\ G = \frac{b_2 - \left(\frac{3n-3}{n+1}\right)}{\sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}}; \\ b_2 = \frac{m_4}{m_2^2}. \end{cases} \quad (1.2)$$

Por fim, o teste de *omnibus* D'Agostino, que é capaz de detectar desvios da normalidade por meio da assimetria ou curtose, são dados por

$H_0$  : Os dados originam-se de uma distribuição normal.

$H_1$  : Não normalidade devido a assimetria ou curtose não normal.

$$K^2 = Z_s^2 + Z_k^2. \quad (1.3)$$

D'Agostino e Belanger (1990) destacam que, sob a hipótese nula de normalidade populacional, as estatísticas  $Z_s$  e  $Z_k$  têm distribuição aproximadamente normal e  $K^2$  tem distribuição aproximadamente qui-quadrado com 2 graus de liberdade,  $\chi_2^2$ . Essas estatísticas calculadas são então comparadas com essas distribuições teóricas para determinar se a hipótese de normalidade deve ser rejeitada. Se os valores de  $Z_s$  ou  $Z_k$  forem extremos em relação à distribuição normal, ou se  $K^2$  exceder o valor crítico da distribuição qui-quadrado, isso indica um desvio significativo da normalidade.

Além disso, os autores recomendam o uso da estatística  $K^2$  para amostras com  $n > 50$ .

## 1.2.2 Homogeneidade das variâncias

Nesta seção serão abordados os principais testes de hipóteses para a verificação do pressuposto de homogeneidade de variâncias de  $k$  populações.

### 1.2.2.1 Teste de Bartlett

O teste de Bartlett (1937) é utilizado para verificar a homogeneidade das variâncias entre diferentes populações. Esse teste é baseado em  $k$  amostras aleatórias e é abordado em literaturas tais como Rajagopalan (2006), Kanji (2006) e Mohr *et al.* (2021). Segundo esses autores, as condições para a aplicação do teste são que as populações sejam independentes e normalmente distribuídas.

As hipóteses estatísticas do teste de Bartlett, em que  $H_0$  é a hipótese nula, assumindo que as variâncias das  $k$  populações são iguais, enquanto  $H_1$  representa a hipótese alternativa, que postula que pelo menos uma das  $k$  variâncias testadas difere das demais, são dadas por

$$\begin{aligned} H_0 : \sigma_1^2 &= \sigma_2^2 = \dots = \sigma_k^2, \\ H_1 : &\text{pelo menos um } \sigma_i^2 \text{ difere dos demais, para } i = 1, \dots, k. \end{aligned}$$

Conforme Montgomery (2013), o teste de Bartlett é pouco robusto à falta de normalidade, e, caso haja dúvidas quanto à validade dessa pressuposição, o teste não deve ser utilizado. A estatística do teste é aproximada pela distribuição qui-quadrado com  $k - 1$  graus de liberdade e é dada por:

$$\chi_0^2 = 2,3026 \frac{q}{c} \quad (1.4)$$

em que,

$$\begin{aligned} q &= (N - k) \log_{10} S_p^2 - \sum_{i=1}^k (n_i - 1) \log_{10} S_i^2, \\ c &= 1 + \frac{1}{3(k-1)} \left( \sum_{i=1}^k (n_i - 1)^{-1} - (N - k)^{-1} \right), \\ S_p^2 &= \frac{\sum_{i=1}^k (n_i - 1) S_i^2}{N - k}, \\ N &= \sum_i n_i, \end{aligned}$$

e  $S_i^2$  é a variância amostral da  $i$ -ésima população e  $N$  o total de observações nas  $k$  amostras.

A constante 2,3026 ajusta o cálculo de  $q$  quando se usa o logaritmo de base 10 ( $\log_{10}$ ) em vez do logaritmo natural ( $\ln$ ). Se a estatística for utilizada diretamente com  $\ln$ , essa constante pode ser omitida.

Ainda conforme Montgomery (2013), a quantidade  $q$  é grande quando a variância amostral  $S_i^2$  é alta, e é igual a zero quando  $S_i^2$  são iguais, e portanto, a hipótese  $H_0$  é rejeitada, se e somente se, a estatística  $\chi_0^2$  é maior que o  $\alpha$ -ésimo quantil da distribuição qui-quadrado com  $k - 1$  graus de liberdade,  $\chi_{\alpha, k-1}^2$ .

### 1.2.2.2 Teste de Cochran

Segundo Kanji (2006),  $k$  amostras normalmente distribuídas e de igual tamanho são as condições para a aplicação do teste de Cochran (1941). Conforme, Mayr e Eye (2005), as hipóteses estatísticas do teste de Cochran são dadas por

$$\begin{aligned} H_0 : \sigma_1^2 &= \sigma_2^2 = \dots = \sigma_k^2, \\ H_1 : \sigma_i^2 &\neq \sigma_j^2, \text{ para pelo menos um par } (i, j). \end{aligned}$$

e sua estatística de teste, denotada por  $C$ , é dada por

$$C = \frac{s_{max}^2}{\sum_{i=1}^k s_i^2}.$$

em que,  $s_{max}^2$  é a maior variâncias das  $k$  amostras e  $\sum_{i=1}^k s_i^2$  é a soma das variâncias de todas amostras.

Conforme Mayr e Eye (2005), devido a definição matemática simples da estatística  $C$  do teste de Cochran, este teste é chamado de teste fácil de usar, *easy-to-use test*. Cochran (1941) elaborou uma tabela com os valores críticos da distribuição teórica da razão  $C$ , que se encontra em seu trabalho, para que fosse possível testar a hipótese de nulidade de homocedasticidade de variâncias das  $k$  amostras consideradas. Se o valor calculado por  $C$  exceder o valor crítico tabelado, utilizando os graus de liberdade  $k$  e  $n - 1$ , em que  $n$  é o número de observações em cada grupo, a hipótese nula de homogeneidade das variâncias é rejeitada (WANG *et al.*, 2017).

### 1.2.2.3 Teste de Hartley

Criado por Hartley (1950), este teste foi desenvolvido com o intuito de ser um teste alternativo para o teste de Bartlett. A estatística  $F$  deste teste consiste em calcular a razão da máxima variância pela mínima variância das  $k$  amostras consideradas e, assim como o teste de Cochran, devido a forma simplificada de sua estatística, o teste de Hartley também é conhecido como *easy-to-use test*.

$$F_{max} = \frac{S_{max}^2}{S_{min}^2},$$

em que,  $S_{max}^2$  e  $S_{min}^2$  são, respectivamente, a máxima e a mínima das variâncias amostradas.

Segundo Kanji (2006), as populações das  $k$  amostras devem ser normalmente distribuídas e com tamanhos aproximadamente iguais. Se a razão  $F_{max}$  observada exceder ao valor crítico, conforme apresentado por Kanji (2006), a hipótese nula de variâncias homogêneas deve ser rejeitada.

Mayr e Eye (2005), ressaltam que os testes de Hartley e Cochran levam aos mesmos resultados, no entanto o teste de Cochran tende a ser mais poderoso devido ao uso de mais informações dos dados. Além disso, o teste de Cochran pode ser preferível ao teste de Hartley pois segundo os autores, funciona melhor quando uma variação de tratamento é substancialmente maior do que todas as outras variações de tratamento.

#### 1.2.2.4 Teste de Levene

O teste para homogeneidade de variâncias de Levene (1960), assim como o teste de Bartlett, é usado para verificar se  $k$  amostras possuem variâncias estatisticamente iguais. Este teste requer observações independentes e que a variável aleatória estudada seja quantitativa. Segundo Filliben e Heckert (2003), o teste de Levene é mais robusto a desvios da normalidade quando comparado ao teste de Bartlett, e se houver indícios de normalidade dos dados, o teste de Bartlett terá maior poder, conforme Lim e Loh (1996). As hipóteses de nulidade e alternativa do teste de Levene são dadas por

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2,$$

$$H_1 : \sigma_i^2 = \sigma_j^2, \text{ para pelo menos um par } (i, j).$$

Dada uma variável aleatória  $X$ , com amostra de tamanho  $N$ , dividida em  $k$  subgrupos, em que  $N_i$  é o tamanho da amostra do  $i$ -ésimo subgrupo, a estatística do teste de Levene conforme Filliben e Heckert (2003) e Wang *et al.* (2017), é dada por

$$W = \frac{(N - k) \sum_{i=1}^k N_i (\bar{Z}_i - \bar{Z}_{..})^2}{(k - 1) \sum_{i=1}^k \sum_{j=1}^{N_i} (Z_{ij} - \bar{Z}_i)^2}$$

em que,  $\bar{Z}_i$  é a média do  $i$ -ésimo grupo de  $Z_{ij}$ ;  $\bar{Z}_{..}$  é a média geral de todos os  $Z_{ij}$ ; e  $Z_{ij}$  foi originalmente definido por Levene (1960) como  $Z_{ij} = |X_{ij} - \bar{X}_i|$ , em que  $\bar{X}_i$  é a média do  $i$ -ésimo subgrupo da variável  $Y$ . A hipótese nula  $H_0$  é rejeitada pelo teste de Levene se a estatística  $W$  for maior que o valor crítico da distribuição  $F$  com  $\alpha$  de significância, e  $(k - 1)$  e  $(N - k)$  graus de liberdade, ou seja, se  $W > F_{\alpha, (k-1), (N-k)}$ .

Brown e Forsythe (1974) estenderam o teste de Levene substituindo a média  $\bar{Y}_i$  do subgrupo em  $Z_{ij}$  pela mediana ou pela média aparada<sup>3</sup>, tornando o teste mais robusto a valores extremos. De acordo com os autores,  $Z_{ij}$  pode ser redefinido como  $Z_{ij} = |X_{ij} - \tilde{X}_i|$ , em que  $\tilde{X}_i$  é a mediana  $i$ -ésimo subgrupo, ou  $Z_{ij} = |X_{ij} - \bar{X}'_i|$ , em que  $\bar{X}'_i$  é a média aparada de 10% do  $i$ -ésimo subgrupo. Essas substituições visam aumentar a robustez do teste, pois a mediana e a média aparada são menos sensíveis a *outliers* do que a média.

<sup>3</sup>A média aparada é calculada removendo-se uma proporção fixa de valores mais baixos e mais altos de um conjunto de dados antes de calcular a média.

Filliben e Heckert (2003), evidenciam que as opções para  $Z_{ij}$  determinam a robustez e o poder do teste de Levene, apontando que a robustez refere-se à capacidade do teste de evitar a detecção falsa de variações desiguais, mesmo quando o pressuposto de normalidade é violado e as variâncias são iguais, e o poder representa a habilidade do teste em identificar diferenças reais entre variâncias quando estas são de fato desiguais.

### 1.3 Transformações

As transformações de dados são ferramentas estatísticas que, ao modificarem a escala dos dados, visam melhorar a adequação dos dados a pressupostos como normalidade e homocedasticidade. Essa técnica possibilita que os dados transformados atendam aos critérios necessários para validação da análise, em situações que os dados originais violam os pressupostos da análise de variância tradicional. No contexto das ciências agrárias, em que a comparação de grupos experimentais é frequente, as transformações, em conjunto com a análise de variância, desempenham um papel crucial na verificação da existência de diferença entre múltiplos grupos experimentais.

Propostas por Edgeworth (1898), as transformações de dados surgiram da necessidade de relacionar distribuições observadas à distribuição normal. No século seguinte, com a análise de variância se tornando um método de comparação popular, por ser um teste poderoso e consolidado para comparações múltiplas, as transformações de dados passaram a ser utilizadas para validar as pressuposições para a realização da ANOVA. No entanto, o uso das transformações de dados pode se tornar problemático quando aplicado sem critério. De acordo com Nascimento *et al.* (2024a), o uso inadequado dessa técnica foi identificado em 94,28% da amostra de artigos científicos analisados na área das ciências agrárias.

Assis *et al.* (2023) afirmam que a transformação dos dados envolve substituir a medida da variável original por outra escala. Se os pressupostos não são adequados para a variável original, uma transformação apropriada pode ajudar a satisfazê-los. Esse procedimento pode ser aplicado em qualquer modelo de análise de variância ou regressão, seja em experimentos balanceados ou não balanceados, para amostras grandes ou pequenas. As estatísticas usuais e os graus de liberdade são mantidos, sem perda de precisão ou sensibilidade do teste. Embora uma análise descritiva e exploratória dos dados possa auxiliar na escolha da transformação, não existem regras gerais que garantam o sucesso dessa escolha.

Para Ramachandran e Tsokos (2021), o modelo básico para a ANOVA exige que as observações independentes provenham de populações normais com variâncias homogêneas. Essas exigências raramente são atendidas na prática, e o grau em que elas são violadas afeta a validade da inferência a respeito das variâncias por meio do teste  $F$ . Os autores indicam que é necessário examinar os dados para identificar desvios significativos dos pressupostos do modelo e, se necessário, aplicar uma transformação apropriada aos dados para alinhá-los às pressuposições básicas.

Assis *et al.* (2023) destacam que, quando os pressupostos da análise de variância paramétrica - como a normalidade dos erros, a homogeneidade das variâncias e a aditividade dos efeitos - não são

atendidos, o pesquisador pode utilizar técnicas de transformação de dados antes de recorrer a testes não-paramétricos. Os autores justificam essa recomendação, ressaltando que os testes paramétricos possuem maior poder estatístico em comparação aos não-paramétricos, desde que seus pressupostos sejam satisfatoriamente cumpridos.

Zar (2010) afirma que, quando os pressupostos paramétricos são atendidos, os testes paramétricos tendem a apresentar maior poder estatístico em comparação com os testes não-paramétricos. Isso se deve, ao fato de que os testes paramétricos aproveitam informações adicionais sobre a distribuição dos dados, como média e variância. Em contrapartida, os testes não-paramétricos, por dependerem menos dessas pressuposições, tornam-se mais robustos, mas geralmente apresentam menor poder sob condições de normalidade.

De acordo com Steel e Torrie (1980), não são todos os bancos de dados que podem ser transformados para corrigir a heterogeneidade dos erros em um experimento, pois essa heterogeneidade pode ser irregular ou regular. A heterogeneidade irregular ocorre quando alguns tratamentos têm maior variabilidade que outros, sem associação entre média e variância, e as transformações de dados geralmente não resolvem esse problema. Já a heterogeneidade regular geralmente resulta da falta de normalidade dos dados, onde há uma relação entre média e variância. Nesse caso, os dados podem ser transformados para obter distribuição normal e variâncias homogêneas. Portanto, é necessária a investigação adequada para determinar a viabilidade do uso da técnica de transformação de dados.

Assis *et al.* (2023) destacam que, uma vez normalizada e homogeneizada a distribuição dos dados amostrais por meio da transformação mais adequada, o pesquisador estará apto a utilizar testes paramétricos. No entanto, se, após a aplicação de diversas transformações, a distribuição dos valores, das variáveis ou dos erros continuar a apresentar um comportamento não normal, assimétrico, achatado, afilado, ou não homogêneo, ou ainda se os efeitos dos fatores do modelo matemático não forem aditivos, a alternativa restante é recorrer à estatística não paramétrica ou ao uso de modelos lineares generalizados.

A importância das transformações de dados para atender aos pressupostos de normalidade e homogeneidade de variâncias em análises estatísticas é amplamente reconhecida na literatura. Doncaster e Davey (2007) afirmam que violações dessas pressuposições são comuns e que as transformações são um procedimento eficaz para corrigi-las, ressaltando a necessidade de verificar os dados após a transformação para confirmar o efeito desejado. Christensen (2017) reforça que tais pressuposições são essenciais para a validade dos modelos estatísticos, e que transformações numéricas ajudam a aprimorar o seu cumprimento. Assis *et al.* (2023) também destacam que transformações adequadas podem resolver problemas de não normalidade e heterogeneidade de variâncias, tornando os dados mais apropriados para análise.

Mohr (2021) sugere que a forma preferida de lidar com assimetria e variâncias desiguais é buscar uma transformação de dados que mitigue esses problemas. Transformações que estabilizam a variância geralmente também reduzem a assimetria. O autor destaca que, em muitas situações, a dispersão dos grupos tende a seguir um padrão relacionado à média do grupo, o que pode ser tratado da seguinte forma:

- Se o desvio padrão ( $\sigma$ ) é proporcional à média, use o logaritmo dos dados.
- Se a variância ( $\sigma^2$ ) é proporcional à média, use a raiz quadrada dos dados.
- Se os dados são proporções ou percentagens, use a transformação arco seno.

Ele observa que as transformações logarítmica e raiz quadrada, não fazem sentido se alguns valores forem negativos, mas a adição de uma constante pode permitir que essas transformações sejam utilizadas. A estabilização das variâncias é crucial para a validade da análise de variância e para a correta interpretação dos resultados. Quando há grandes diferenças na dispersão entre grupos, a descrição de uma diferença como “grande” ou “pequena” se torna relativa.

As transformações logarítmica, raiz quadrada e arco seno são amplamente discutidas e aplicadas na comunidade acadêmica para homogeneizar variâncias. Assis *et al.* (2023) as apontam como as principais transformações de dados. Esse destaque é corroborado por Nascimento *et al.* (2024a), que identificaram essas três transformações como as mais utilizadas na amostra analisada de trabalhos da área de ciências agrárias que utilizaram transformações de dados para a validação da análise de variância. As três transformações e a transformação Box-Cox, serão abordadas com mais detalhes nas Subseções 1.3.1 a 1.3.4.

Um ponto chave do uso das transformações de dados, destacado por Mohr (2021), é que as inferências são baseadas nas médias dos valores transformados, e essas médias não são necessariamente iguais às médias transformadas dos valores originais. Isso significa que não é correto transformar as estatísticas calculadas a partir dos valores transformados de volta para a escala original. Quando as transformações não são úteis, o autor recomenda o uso do teste  $F$  de Welch, uma adaptação da análise de variância para casos de variâncias desiguais. Esse teste, implementado na maioria dos *softwares* estatísticos, é análogo ao teste  $t$  de Welch, usado em amostras independentes com variâncias desiguais. Se o problema for a não-normalidade das distribuições, o teste de Kruskal-Wallis é recomendado (Conover, 1999).

### 1.3.1 Transformação logarítmica

Para compreender a transformação logarítmica e seu comportamento em um conjunto de dados, é importante revisar primeiramente os conceitos de logaritmo e função logarítmica. O logaritmo e a potência, também conhecida como exponenciação, se relacionam de forma que cada elemento da progressão aritmética é o logaritmo do termo da progressão geométrica ao qual está associado.

Uma função vai além de uma operação. Enquanto uma operação produz um único resultado, uma função estabelece uma correspondência entre a operação e o resultado dessa operação para um conjunto de valores. A operação é realizada com os elementos do conjunto chamado domínio e seus respectivos resultados produz um outro conjunto, chamado imagem da função. Portanto, a função expressa uma relação entre conjuntos numéricos, que possuem características próprias e representação gráfica do comportamento dessa relação.

Conforme Iezzi *et al.* (2013), para os números  $a$  e  $b$ , reais e positivos, o logaritmo de  $b$  na base

$a$  é o expoente que se deve dar a base  $a$ , de modo que a potência obtida seja igual a  $b$ . Portanto,

$$\log_a b = x \Leftrightarrow a^x = b,$$

em que,  $x$  é o expoente da potência  $a^x = b$ .

Já a função logarítmica de base  $a$ , associa a cada  $x$ , real positivo sem zero ( $\mathbb{R}_+^*$ ), o número  $\log_a x$ .

$$\begin{aligned} f : \mathbb{R}_+^* &\rightarrow \mathbb{R} \\ x &\mapsto \log_a x \end{aligned}$$

O crescimento e decrescimento dessa função depende da base  $a$ . Se base  $a > 1$  a função logarítmica é crescente e se  $0 < a < 1$  ela é decrescente.

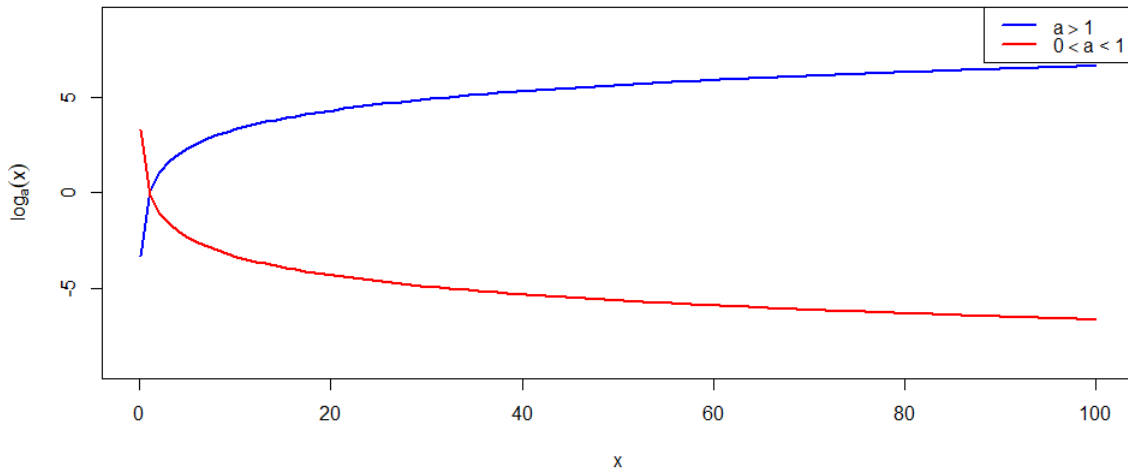


Figura 1.1: Curvas da função logarítmica para as bases 2 e 0,5.

A Figura 1.1 ilustra o comportamento da função logarítmica para valores maiores (em azul) e menores (em vermelho) que um. Observa-se que, para a função  $f(x) = \log_2(x)$ , o logaritmo de valores entre 1 e 100 são próximos de 5. De forma semelhante, para  $f(x) = \log_{0,5}(x)$ , os valores se aproximam de  $-5$ . Essa função reduz a amplitude de dados, aproximando números grandes e pequenos em termos de magnitude. Por isso, o logaritmo é utilizado como uma transformação de dados, pois a diminuição da amplitude dos dados faz com que a variabilidade também seja reduzida.

A transformação logarítmica é definida, por Mohr (2021), da seguinte forma:

$$\log_{10}(x+k) \text{ ou } \ln(x+k), \text{ para } k \geq 0,$$

em que  $x$  representa um dado valor da amostra e  $k$  uma constante que costuma ser 0,5 ou  $1^4$ , mas

<sup>4</sup>A escolha de  $k = 0,5$  ou  $k = 1$  é feita para evitar problemas ao calcular logaritmos de valores próximos ou iguais a zero, já que o logaritmo não é definido para números não positivos. Essas constantes deslocam os dados para uma escala positiva, garantindo a aplicabilidade da transformação.

pode ser qualquer valor real.

Conforme Assis *et al.* (2023), a transformação logarítmica, que na maioria das vezes utiliza o logaritmo natural ou logaritmo na base 10, estabiliza as variâncias de dados em que a média e a variância são positivamente correlacionadas. Uma correlação positiva indica que grupos com médias elevadas também apresentam grandes variâncias (na ANOVA) ou que a magnitude dos resíduos está correlacionada com a da variável independente (na regressão).

Assis *et al.* (2023) também indicam que essa transformação estabiliza as variâncias na situação em que as variâncias são proporcionais ao quadrado das médias dos tratamentos. Para dados univariados que são positivamente assimétricos (assimetria à direita), frequentemente contêm valores discrepantes elevados, com a transformação logarítmica, esses valores tendem a ser atraídos para a parte central da distribuição, tornando-a mais simétrica. Conjuntos de dados com médias e variâncias positivamente correlacionadas também tendem a apresentar resíduos positivamente assimétricos. A transformação logarítmica frequentemente resolve ambos os problemas simultaneamente.

Ramachandran e Tsokos (2021), indicam a transformação logarítmica quando a variância amostral for uma função da média, ou seja, quando for observada uma relação entre a média da amostra em relação à variância amostral. Quando há indícios de que a variância populacional é o produto da média por uma constante, os autores recomendam que cada observação  $x$  seja substituída pelo seu logaritmo de base 10 ou, quando há valores nulos, pelo logaritmo de  $(x + 1)$ , já que o logaritmo de 1 é zero.

Cochran (1947) ressalta que uma análise dos logaritmos das observações, em vez das próprias observações, é bem sucedida na conversão de dados para uma escala onde os erros seguem uma distribuição aproximadamente normal. Para Hoyle (1973), a transformação logarítmica pode ser utilizada quando se deseja uma redução da assimetria de uma distribuição, mas que essa redução seja menos severa do que a da transformação recíproca, que consiste em substituir cada observação  $x$  por cada observação  $\frac{1}{x}$ . Assis *et al.* (2023), destacam que é importante observar que a transformação logarítmica não pode ser aplicada a números negativos nem a zero. Por essa razão, dados que contêm zeros devem ser transformados usando a fórmula  $\log(x + k)$ , em que  $k$  assume os valores 0,5 ou 1.

Além disso, é importante destacar que ao transformar dados usando o logaritmo, as operações matemáticas, como somas e médias, mudam de significado. A soma dos logaritmos está relacionada ao produto dos valores originais, e a média dos logaritmos está relacionada à média geométrica dos valores originais, não à média aritmética. Portanto, é crucial ter cautela ao interpretar esses valores e ao tentar converter de volta para o espaço original usando exponencial. Após transformados, os dados devem ser interpretados na nova escala.

### 1.3.2 Transformação raiz quadrada

Para a compreensão da transformação raiz quadrada e seu comportamento em um conjunto de dados, vamos revisar os conceitos de radiciação. Assim como a subtração é a operação inversa

da soma e a divisão é a operação inversa da multiplicação, a radiciação é a operação inversa da potência.

Iezzi *et al.* (2013) definem  $b$  como a raiz  $n$ -ésima aritmética de  $a$ , indicado por  $\sqrt[n]{a} = b$ , de forma que  $b^n = a$ , em que  $a$  é chamado radicando e  $n$  índice. Portanto, o índice  $n$  da raiz corresponde ao expoente  $n$  da potência.

$$\sqrt[n]{a} = b, \text{ pois } b^n = a, \text{ para } a, b, n \in \mathbb{R}.$$

Para  $n = 2$ , teremos a raiz quadrada que se relaciona com a potência de exponte 2, ou seja, a raiz quadrada de  $a$  é um número  $b$  que, quando multiplicado por ele mesmo, resulta em  $a$ .

A função raiz quadrada é definida para todos os números reais não negativos, sendo dada por

$$\begin{aligned} f: \mathbb{R}_+ &\rightarrow \mathbb{R} \\ x &\mapsto \sqrt{x} \end{aligned}$$

A função raiz quadrada é crescente em todo o seu domínio e o seu comportamento, visto na Figura 1.2, se assemelha ao da função logarítmica com base  $a > 1$ . Ambas as funções crescem mais rapidamente para valores pequenos de  $x$  e desaceleram o crescimento à medida que  $x$  aumenta, e por isso a curva tende a se tornar menos inclinada em grandes valores de  $x$ .

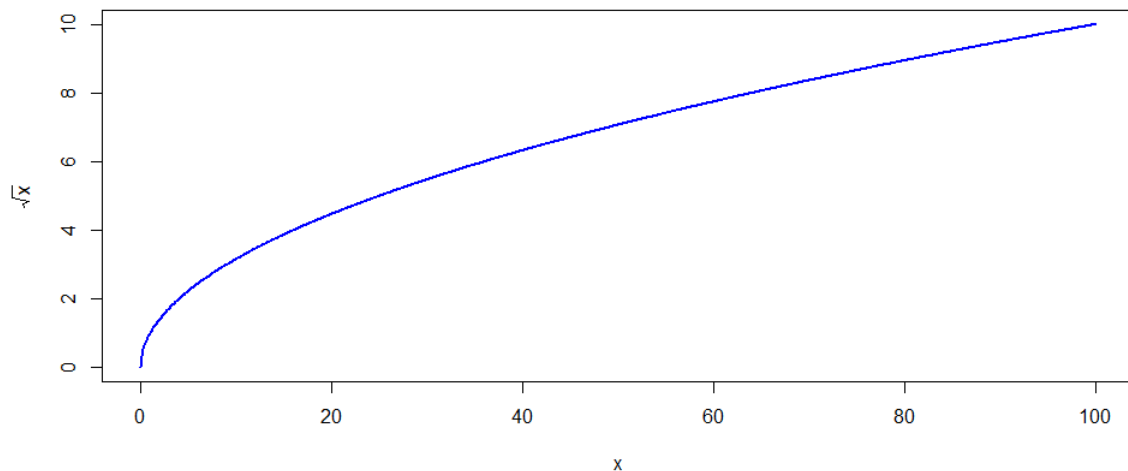


Figura 1.2: Curva da função raiz quadrada.

Apesar da semelhança entre as duas funções, a função raiz quadrada cresce mais rápido do que a função logarítmica para grandes valores de  $x$ . Por exemplo,  $\log_{10}(100) = 2$  enquanto  $\sqrt{100} = 10$ , fazendo com que a função logarítmica seja mais suave e a função raiz quadrada mais acentuada. Outra diferença entre essas funções é que a função logarítmica não é definida em zero e a função raiz quadrada é definida neste ponto. Todas essas características influenciam na escolha da raiz quadrada como transformação dos dados.

A transformação raiz quadrada é definida por Mohr (2021), da seguinte forma

$$\sqrt{x+k} \text{ para } k \geq 0,$$

o autor indica que essa transformação é frequentemente utilizada em dados provenientes de contagens.

Conforme Assis *et al.* (2023), a transformação raiz quadrada é indicada para estabilizar ou homogeneizar a variância quando existe correlação entre média e variância e a variável refere-se a uma contagem, como o número de plantas atacadas por uma doença ou o número de acidentes na produção de um material. Zar (2010) e Sokal e Rohlf (2012) destacam que essa transformação é especialmente útil para variáveis que seguem uma distribuição de Poisson. A transformação também é aplicável a qualquer tipo de contagem, como o número de frutos produzidos por planta, ou até mesmo para dados que não sejam de contagem, desde que as variâncias sejam proporcionais às médias dos tratamentos. Steel e Torrie (1980) recomendam, para contagens pequenas ou iguais a zero, somar 0,5 a cada valor antes de aplicar a raiz quadrada, o que resulta na transformação  $\sqrt{x+0,5}$ . Montgomery (2013) acrescenta que, nesses casos, a variável transformada pode ser considerada aproximadamente normal, atendendo melhor aos pressupostos de análise paramétrica.

De acordo com Bartlett (1947) e Hoyle (1973), a transformação raiz quadrada é recomendada quando os dados seguem uma distribuição de Poisson, na qual se conta o número de ocorrências de um determinado fenômeno durante um intervalo de tempo, especialmente quando há suspeita de heterogeneidade de variâncias. Mesmo que os dados não sigam exatamente uma distribuição de Poisson, esses autores ressaltam que a transformação ainda pode ser eficaz na estabilização das variâncias.

### 1.3.3 Transformação arco seno

Os conceitos de seno e arco seno são definidos no círculo trigonométrico<sup>5</sup>. Ao pegar um ponto qualquer nesse círculo, é possível medir a altura desse ponto em relação ao centro do círculo. Essa altura é o valor do seno para um certo ângulo. O arco seno é o processo de encontrar esse ângulo original, dado o valor do seno.

As operações seno e arco seno são inversas, assim como a subtração “desfaz” a soma, o arco seno “desfaz” o seno. O seno de um ângulo só pode resultar em valores de  $-1$  a  $1$ , e o arco seno só pode ser aplicado a números dentro desse intervalo. Para Iezzi *et al.* (2013),  $y$  é um arco cujo seno é  $x$ , ou seja, se o seno de  $\frac{\pi}{6}$  é igual a  $0,5$ , então o arco seno de  $0,5$  é  $\frac{\pi}{6}$ .

$$\arcsen(x) = y, \text{ se, e somente se, } \sen(y) = x.$$

Mohr (2021) indica o uso da transformação arco seno em conjuntos de dados que são medidas de proporções ou porcentagens.

---

<sup>5</sup>círculo com raio um

Para Assis *et al.* (2023), a transformação arco seno é aplicável a dados com distribuição binomial, expressos em frações ou porcentagens, ocorrendo estabilização da variância, como na proporção de sementes germinadas, registradas em um intervalo de 0 a 1. Além disso, Sokal e Rohlf (2012) observam que, para a transformação arco seno, se os dados estiverem entre 30% e 70%, a transformação pode ser dispensada, pois os dados já se aproximam de uma distribuição simétrica.

Conforme Sokal e Rohlf (2012), a transformação arco seno é dada por  $\arcsen(x)$  e a transformação  $\arcsen(\sqrt{x})$  é chamada transformação angular. Ambas são usadas para estabilizar as variâncias de variáveis que possuem distribuição binomial.

### 1.3.4 Transformação Box-Cox

A transformação Box-Cox é recomendada por Azevedo *et al.* (2016), para estabilizar ou reduzir a variabilidade dos dados, além de aproximá-los de uma distribuição normal. Segundo os autores, essa transformação melhora a homocedasticidade e a normalidade univariada e multivariada dos resíduos, sendo particularmente útil em situações em que há heterogeneidade de variâncias e distribuição assimétrica.

Conforme Box e Cox (1964), a família de transformações para  $x > 0$  é definida por

$$y(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \text{para } \lambda \neq 0 \\ \log(x), & \text{para } \lambda = 0, \end{cases} \quad (1.5)$$

em que,  $y(x)$  é a observação transformada,  $x$  representa os dados originais, e  $\lambda$  é o parâmetro de transformação.

Box e Cox (1964) ainda ressaltam que como uma análise de variância é inalterada por uma transformação linear, então a transformação (1.5) é equivalente a

$$y(x) = \begin{cases} x^\lambda, & \text{para } \lambda \neq 0 \\ \log(x), & \text{para } \lambda = 0. \end{cases}$$

Conforme Quinn e Keough (2002), as transformações raiz quadrada e logarítmica, fazem parte desta família de transformações.

### 1.3.5 Transformação potência

A família das transformações potência (power transformation) é reconhecida como uma das mais importantes em estatística, devido à sua eficácia em estabilizar variâncias e aproximar os dados de uma distribuição normal. Rupert (2001) destacou que essas transformações são ferramentas fundamentais para lidar com heterogeneidade de variância, tornando-as indispensáveis em análises como ANOVA e regressão. Hoyle (1973) definiu a transformação potência como:

$$\begin{cases} \frac{(x+w_1)^{w_2}-1}{w_2}, & \text{para } w_2 \neq 0 \\ \log(x+w_1), & \text{para } w_2 = 0, \end{cases} \quad (1.6)$$

Esta formulação, segundo Box e Cox (1964), é válida para  $x > -w_1$  e representa uma família de transformações flexíveis que podem ser ajustadas para atender às particularidades de diferentes conjuntos de dados. Ao fixar  $w_1 = 0$  e  $w_2 = \lambda$ , obtém-se a famosa transformação de Box-Cox, uma aplicação específica da transformação potência que tem sido utilizada em estudos de normalização de dados e redução de heterocedasticidade.

Além disso, Tukey (1957) já havia abordado conceitos relacionados à transformação potência em seu trabalho sobre transformação de dados, enfatizando a importância dessas técnicas para melhorar a interpretabilidade e a precisão dos modelos estatísticos. A transformação potência também é particularmente útil em contextos que a relação entre variáveis segue uma escala não linear, pois facilita a linearização e melhora a eficiência da estimativa dos parâmetros.

### 1.3.6 Transformação recíproca

Segundo Hoyle (1973), a transformação recíproca é geralmente utilizada para reduzir assimetria extrema, podendo fazer com que os dados satisfaçam as pressuposições de normalidade, homogeneidade de variâncias e independência. Esta transformação substitui os valores amostrais  $x$  pelo inverso multiplicativo de  $x$ , isto é,  $x^{-1}$ . Por sua definição matemática, esta transformação é restrita para valores diferentes de zero.

### 1.3.7 Transformação de Johnson

A transformação de Johnson é utilizada para a correção de não normalidade. Com base em Johnson (1949) e Bowman e Shenton (2006), os sistemas  $S_U$ ,  $S_L$  e  $S_B$  das distribuições de Johnson são baseados na transformação linear

$$Z = \gamma + \delta f(Y),$$

com respectivas funções de transformação  $f(Y)$  dadas por

$$i) f(Y) = \ln \left\{ Y + (1 + Y^2)^{\frac{1}{2}} \right\} = \sinh^{-1} Y, \quad -\infty < Y < +\infty;$$

$$ii) f(Y) = \ln(Y), \quad 0 < Y < +\infty;$$

$$iii) f(Y) = \ln \left( \frac{Y}{1-Y} \right), \quad 0 < Y < 1.$$

Todos estes sistemas levam à variável aleatória  $Z$  com distribuição normal padrão, ou seja, que possui média zero e variância um. O item *i*) corresponde ao sistema “*unbounded range*”, sistema com “alcance ilimitado” devido ao suporte da função que varia de menos infinito a infinito, o item *ii*) corresponde a família lognormal e o item *iii*) corresponde ao sistema “*bounded range*”, com “alcance limitado” de zero a um. Por isso os sistemas  $S_U$ ,  $S_L$  e  $S_B$ , são indexados com as letras U,

L e B. Segundo os autores, os parâmetros  $\gamma$  e  $\delta$  podem ser determinados a partir da equivalência de parâmetros de assimetria e curtose ou da equivalência de percentis e/ou momentos.

## 1.4 Revisão sistemática

Para Gough *et al.* (2012), uma revisão sistemática é um método rigoroso e estruturado para sintetizar e analisar de forma abrangente a evidência disponível sobre um determinado tópico de pesquisa. Ela envolve a busca sistemática por estudos relevantes, a seleção criteriosa dos estudos incluídos, a extração e a análise dos dados dos estudos selecionados. Tendo como principal objetivo fornecer uma visão abrangente e imparcial de estudos existentes, avaliando a consistência, a qualidade e a magnitude dos resultados, a fim de responder de maneira confiável a uma pergunta específica de pesquisa.

Segundo Khan (2020), uma revisão sistemática envolve a coleta e exame de trabalhos publicados para extrair estatísticas e informações relevantes dos estudos selecionados. Esse tipo de revisão é uma etapa da meta-análise e consiste na identificação de estudos que atendam aos critérios previamente definidos pelo pesquisador, de acordo com o objetivo do estudo. Para garantir a qualidade e rigor, a revisão sistemática deve seguir procedimentos e protocolos bem estabelecidos.

Moher *et al.* (2009) desenvolveram as diretrizes dos “Itens de relatório preferidos para revisões sistemáticas e meta-análises” (PRISMA), que descrevem procedimentos para o levantamento sistemático da literatura e realização de meta-análises baseados nas etapas de identificação, triagem, elegibilidade e inclusão. Na etapa de identificação, estudos relevantes são localizados em bases de dados por meio de estratégias de busca bem definidas. Em seguida, na triagem, os estudos são selecionados com base nos títulos e resumos, aplicando critérios de inclusão e exclusão previamente definidos. Na etapa de elegibilidade, os textos completos dos artigos selecionados são avaliados para garantir sua adequação ao objetivo da pesquisa. Finalmente, na etapa de inclusão, as informações relevantes dos estudos são extraídas e organizadas sistematicamente para análise qualitativa ou quantitativa, com o objetivo de sintetizar as evidências disponíveis. Essa abordagem padronizada assegura transparência, reprodutibilidade e rigor metodológico em revisões científicas, contribuindo para a confiabilidade dos resultados apresentados.

De acordo com Bernardes *et al.* (2022), realizar uma meta-análise pode ser uma tarefa desafiadora, pois envolve esforços intensivos da parte do pesquisador para revisar sistematicamente um grande volume de estudos publicados. Segundo Marshall e Wallace (2019), esta tarefa pode ser simplificada ao utilizar ferramentas de inteligência artificial, que possibilitam aplicar interativamente modelos de processamento de linguagem natural e *machine learning* (ML) para reduzir o esforço envolvido na revisão sistemática.

Para que os procedimentos interativos de *machine learning* minimizem o número de estudos a serem triados, priorizando os artigos com maior probabilidade de serem relevantes, Bernardes *et al.* (2022) interagiram com um modelo de *machine learning* em um *loop*, determinando critérios para classificar estudos do grupo de treino em relevantes e irrelevantes, fazendo com que o modelo

atualize as previsões feitas para o restante dos estudos. Van de Schoot *et al.* (2021) indicam que o estudo recém rotulado seja adicionado ao conjunto de dados de treinamento e outra rodada seja iterada.

Nascimento *et al.* (2024a) propuseram que pesquisadores que realizam revisões sistemáticas, especialmente quando a informação de interesse está na metodologia, podem otimizar a triagem dos trabalhos por meio da busca de termos específicos no texto completo dos artigos, geralmente disponíveis em formato PDF. A estratégia sugerida consiste em baixar todos os estudos identificados na etapa de identificação e, em seguida, aplicar técnicas automatizadas para separar os artigos que contêm os termos relevantes daqueles que não os contêm, economizando tempo e reduzindo o esforço manual. Em continuidade, Nascimento *et al.* (2024b) desenvolveram um aplicativo web dedicado a facilitar esse processo de triagem, permitindo que pesquisadores realizem essa etapa sem a necessidade de conhecimentos em programação.

A validade das evidências e decisões do estudo de interesse dependem diretamente da qualidade das fontes e da metodologia empregada para extraí-las e processá-las. Portanto, segundo Khan (2020), a implementação, avaliação e monitoramento dos critérios de seleção e exclusão, protocolos de estudo e avaliação de qualidade devem ser rigorosos para garantir que os resultados da síntese sejam reproduzíveis e válidos.

# Referências Bibliográficas

ACUTIS, M.; SCAGLIA, B.; CONFALONIERI, R. Perfunctory analysis of variance in agronomy, and its consequences in experimental results interpretation. *European Journal of Agronomy*, v. 43, p. 129–135, 2012.

ANDERSON, T. W.; DARLING, D. A. A Test of Goodness of Fit. *Journal of the American Statistical Association*, v. 49, n. 268, p. 765–769, 1954.

ASSIS, J. P.; DE LIMA, I. R. P.; FRANÇA, J. A.; DE SOUSA, R. P.; LINHARES, P. C. F.; CUSTÓDIO, T. N.; DE SOUSA, R. P.; RODRIGUES, W. M. *Transformação de dados aplicada à estatística*, 1st ed. Brasil: Pantanal Editora, 2023. 133 p.

AZEVEDO, A. M.; ANDRADE JÚNIOR, V. C.; FERNANDES, J. S. C. Transformação Box-Cox na homocedasticidade e normalidade uni e multivariada em experimentos de batata-doce. *Horticultura Brasileira*. v. 34, n. 1, p. 93–101, 2016.

BARTLETT, M. S. Properties of Sufficiency and Statistical Tests. *Journal of the Royal Statistical Society*, v. 160, n. 1 p. 268—282, 1937.

BARTLETT, M. S. The use of transformations. *Biometrics*, v. 3, p. 39–52, 1947.

BARTON, B.; PEAT, J. *Medical statistics a guide to SPSS, data analysis and critical appraisal*, 2nd ed. Chennai: BMJ Books, 2014. 408 p.

BOWMAN, K. O.; SHENTON, L. R. Johnson's system of distributions. In: KOTZ, S.; READ, C. B.; BALAKRISHNAN, N.; VIDA KOVIC, B. (Ed. 2). *Encyclopedia of statistical sciences*. New York: John Wiley & Sons, 2006. p. 3751-3762.

BERNARDES, R. C.; BOTINA, L. L.; ARAÚJO, R. S.; GUEDES, R. N. C.; MARTINS, G. F.; LIMA, M. A. P. Artificial intelligence-aided meta-analysis of toxicological assessment of agrochemicals in bees. *Frontiers in Ecology Evolution*, v. 10, p. 1—12, 2022.

BOLFARINE, H.; BUSSAB, W. O. *Elementos de Amostragem*, 2st ed. São Paulo: Blucher, 2005. 228 p.

BOX, G. E. P.; COX, D. R. An analysis of transformations. *Journal of the Royal Statistical Society*, v. 26, p. 211–252, 1964.

- BROWN, M. B.; FORSYTHE, A. B. Robust Tests for the Equality of Variances. *Journal of the American Statistical Association*, v. 69, n. 346, p. 364–367, 1974.
- CASELLA, G.; BERGER, R. L. *Statistical Inference*, 2nd ed. USA: Duxbury, 2002. 660 p.
- CHRISTENSEN, R. *Mathematical statistics with applications in R*. 2nd ed. CRC Press: Albuquerque, 2017. 620 p.
- CHRISTENSEN, R.; BEDRICK E. J. Testing the independence assumption in linear models. *Journal of the American Statistical Association*, v. 92, n. 439, p. 1006–1016, 1997.
- COCHRAN, W. G. The distribution of the largest of a set of estimated variances as a fraction of their total. *Annals of Eugenics*, v. 11, p. 47–52, 1941.
- COCHRAN, W. G. When the assumptions for the analysis of variance are not satisfied. *Biometrics*, v. 3, p. 22–38, 1947.
- COCHRAN, W. G.; COX, G. M. *Experimental design*. 2nd ed. New York: ISBN, 1992. 611 p.
- CONOVER, W. J. *Practical Nonparametric Statistics*. 3rd ed. USA: Wiley, 1999. 592 p.
- D'AGOSTINO, R. B.; BELANGER, A. A Suggestion for Using Powerful and Informative Tests of Normality. *The American Statistician*, v. 44, n. 4, p. 316—321, 1990.
- D'AGOSTINO, R. B.; PEARSON, E. S. Tests for departure from normality. Empirical results for the distributions of  $b_2$  and  $\sqrt{b_1}$ . *Biometrika*, v. 60, n. 3, p. 313—622, 1973.
- DONCASTER, C. P.; DAVEY, A. J. H. *Analysis of variance and covariance, how to choose and construct models for the life sciences*. 1st ed. New York: Cambridge University Press, 2007. 287 p.
- EDGEWORTH, F. Y. A suggestion for using powerful and informative tests of normality. *Journal of the Royal Statistical Society*, v. 61, n. 3, p. 534–544, 1898.
- EISENHART, C. The assumptions underlying the analysis of variance. *Biometrics*, v. 3, n. 1, p. 1–21, 1947.
- FERREIRA, D. F. *Estatística multivariada*. 3. ed. Lavras: UFLA, 2018. 652 p.
- FILLIBEN, J. J.; HECKERT, A. Exploratory data analysis. In: ZEY, C. 1st ed. *Engineering statistics handbook*. Maryland: NIST, 2003. p. 1-59.
- FISHER, R. A. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, v. 52, p. 399–433, 1919.
- FISHER, R. A. *Statistical methods for research workers*. 1st ed. London: Oliver and Boyd, 1925. 239 p.

- FISHER, R. A.; YATES, F. *Statistical tables for biological, agricultural and medical research*. 6th ed. Edinburgh: Oliver & Boyd, 1938. 112 p.
- GHASEMI, A.; ZAHEDIASL, S. Normality tests for statistical analysis: a guide for non-statisticians. *International Journal of Endocrinology & Metabolism*, v. 10, n. 2, p. 486–489, 2012.
- GOMEZ, K. A.; GOMEZ, A. A. *Statistical procedures for agricultural research*. 2nd ed. New York: Wiley, 1984. 704 p.
- GOUGH, D.; OLIVER, S.; THOMAS, J. *An Introduction to Systematic Reviews*. 2nd ed. Los Angeles: SAGE, 2012. 304 p.
- GREENE, J.; D'OLIVEIRA, M. *Learning to use statistical tests in psychology*. 3rd ed. New York: Open University Press, 2005. 256 p.
- HARTLEY, H. O. The maximum F-ratio as a short-cut test for heterogeneity of variance. *Biometrika*, v. 37, n. 3, p. 308–312, 1950.
- HERZOG, M. H.; FRANCIS, G.; CLARKE, A. *Understanding statistics and experimental design*. 1st ed. Switzerland: Springer, 2019. 146 p.
- HINTZE, G. K. *Power analysis and sample size system user's guide III*. 1st ed. Kaysville: NCSS, 2011. 350 p.
- HOLLANDER, M.; WOLFE, D. A.; CHICKEN, E. *Nonparametric Statistical Methods*. 3rd ed. New Jersey: Wiley, 2013. 848 p.
- HOYLE, M. H. Transformations: An introduction and a bibliography. *International Statistical Institute*, v. 41, n. 2, p. 203–223, 1973.
- IEZZI, G.; DOLCE, O.; MURAKAMI, C. *Fundamentos da matemática elementar - Logaritmos*, 10th ed. São Paulo: Atual, 2013. 288 p.
- JOHNSON, N. L. Systems of frequency curves generated by methods of translation. *Oxford University Press*, v. 36, n. 12, p. 149-176, 1949.
- KANJI, G. K. *100 statistical tests*. 3rd ed. London: SAGE Publications, 2006. 240 p.
- KISH, L. *Survey sampling*. 1st ed. New York: John Wiley & Sons, 1965. 664 p.
- KHAN, S. *Meta-analysis: methods for health and experimental studies*. 1st ed. Australia: Springer, 2020. 224 p.
- LEE, C.; PARK, S.; JEONG, J. Comprehensive comparison of normality tests: empirical study using many different types of data. *Journal of the Korean Data & Information Science Society*, v. 27, n. 5, p. 1399-1412, 2016.

- LEVENE, H. *Robust Tests for Equality of Variances*. In: OLKIN, I.; GHURYE, S. G.; HOEFFDING, W.; MADOW, W. G.; MANN, H. B. (Eds.). *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford: Stanford University Press, 1960. p. 278-292.
- LIM, T. S.; LOH, W. Y. A comparison of tests of equality of variances. *Computational Statistics & Data Analysis*, v. 22, p. 287–301, 1996.
- LILLIEFORS, H. W. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, v. 62, n. 318, p. 399–402, 1967.
- LIX, L. M.; KESELMAN, J. C.; KESELMAN, H. J. Consequences of assumption violations revisited: a quantitative review of alternatives to the one-way analysis of variance “F”. *American Educational Research Association*, v. 43, p. 129–135, 1996.
- MAYR, P.; EYE, A. *Cochran’s C test*. In: EVERITT, B.; HOWELL, D. *Encyclopedia of statistics in behavioral science*. 1st ed. Chichester: John Wiley & Sons, 2005. p. 349-350.
- MARSHALL, I. J.; WALLACE, B. C. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, v. 8, n. 163, p. 1–10, 2019.
- MARTINEZ, J.; IGLEWICZ, B. A test for departure from normality based on a biweight estimator of scale. *Biometrika*, v. 68, n. 1, p. 331–333, 1981.
- MENDEL, G. Versuche über pflanzenhybriden. *Verhandlungen des Naturforschenden Vereines in Brünn*, v. 4, p. 3-47, 1866.
- MILLER, S. *Experimental design and statistics*. 2nd ed. London: Routledge, 1984. 144 p.
- MOHR, D. L.; WILSON, W. J.; FREUND, R. J. *Statistical methods*. 4th ed. Chennai: Academic Press, 2021. 832 p.
- MONTGOMERY, D. C. *Design and Analysis of Experiments*. 8th ed. New York: John Wiley & Sons, 2013. 752 p.
- MURPHY, K. R.; MYORS, B. *Statistical power analysis*. 2nd ed. Mahwah: Lawrence Erlbaum Associates, 2004. 176 p.
- NASCIMENTO, J. S. N.; SILVA, J.; BERNARDES, R. C.; COSTA, G.; EMILIANO, P. C. Statistical data transformation in agrarian sciences for variance analysis: a systematic review. *F1000Research*, p. 1–12, 2024a.
- NASCIMENTO, J.; GONÇALVES, L. B.; BARBOSA, W. F.; BELO, L. P.; SILVA, J. F.; COSTA, G. S.; EMILIANO, P. C. Data transformations for ANOVA assumption violations in agrarian sciences subfields: a systematic review, simulation study, and practical guidelines. *DELLOS*, p. 1–27, 2024b.

- OLIVE, D. J. *Linear regression*. 1st ed. Illinois: Springer Nature, 2017.
- 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. 1, p. 171–176, 2006.
- QUINN, G. P.; KEOUGH, M. J. *Experimental design and data analysis for biologists*. 1st ed. New York: Cambridge University Press, 2002. 556 p.
- RAJAGOPALAN, V. *Select statistical tests*. 1. ed. New Delhi: New Age International Publishers, 2006. 324 p.
- RAMACHANDRAN, K. M.; TSOKOS, C. P. *Mathematical statistics with applications in R*. 3rd ed. London: Academic Press, 2021. 686 p.
- ROYSTON, P. Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and Computing*, v. 2, n. 3, p. 117-119, 1992.
- ROYSTON, P. A remark on algorithm as 181: the W-test for normality. *Royal Statistical Society*, v. 44, n. 4, p. 547-551, 1995.
- SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, v. 52, n. 3, p. 591-611, 1965.
- SNEDECOR, G. W. *Calculation and interpretation of analysis of variance and covariance*. 1st ed., Iowa: AMES, 1934. 96 p.
- SNEDECOR, G. W.; COCHRAN, W. G. *Statistical Methods*. 8th ed. Iowa: AMES, 1989. 503 p.
- SOKAL, R. R.; ROHLF, F. J. *Biometry*. 4th ed., W. H. Freeman, USA, 2012. 937 p.
- SPEED, T. P. What is an analysis of variance? *The Annals of Statistics*, v. 15, n. 3, p. 26–61, 1987.
- STEEL, R. G. D.; TORRIE, J. H. *Principles and Procedures of Statistics: A Biometrical Approach*. 2nd ed. USA: McGraw-Hill, 1980. 633 p.
- TAYLOR, J. R. *An introduction to error analysis: the study of uncertainties in physical measurements*. 2nd ed. California: University Science Books, 1997. 327 p.
- TUKEY, J. W. On the comparative anatomy of transformations. *Annals of Mathematical Statistics*, v. 28, n. 3, p. 602–632, 1957.
- TUKEY, J. W. *Exploratory Data Analysis*, Massachusetts: Addison-Wesley, 1977. 688 p.
- VAN DE SCHOOT, R.; SCHONBRODT, F. D.; VELDKAMP, S. A.; ROMANO, J. L. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, v. 3, n. 2, p. 125—133, 2021.

WANG, Y.; GIL, P. R.; CHEN, Y.; KROMREY, J. D.; KIM, E. S.; PHAM, T.; NGUYEN, D.; ROMANO, J. L. Comparing the performance of approaches for testing the homogeneity of variance assumption in one-factor ANOVA models. *Educational and Psychological Measurement*, v. 77, n. 2, p. 305–319, 2017.

WISHART, J. Statistics in agricultural research. *Wiley for Royal Statistical Society*, v. 1, n. 1, p. 26–61, 1934.

ZAR, J. H. *Biostatistical Analysis*. 5th ed. USA: Pearson Education, 2010. 944 p.

## Capítulo 2

**Publicado como artigo original na  
revista F1000Research**

# **Transformação de dados estatísticos nas ciências agrárias para análise de variância: uma revisão sistemática**

Jhennifer Nascimento<sup>1</sup>, Jonas Silva<sup>1</sup>, Rodrigo Cupertino Bernardes<sup>2</sup>, Guilherme Costa<sup>3</sup>,  
Paulo Emiliano<sup>1</sup>

<sup>1</sup>Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, Estado de Minas Gerais, Brasil

<sup>2</sup>Departamento de Biologia Geral, Universidade Federal de Viçosa, Viçosa, Estado de Minas Gerais, Brasil

<sup>3</sup>ICTP South American Institute for Fundamental Research e Instituto de Física Teórica, Universidade Estadual Paulista, São Paulo, Estado de São Paulo, Brasil

O presente capítulo apresenta a primeira revisão sistemática da nossa pesquisa, conduzida na área abrangente das ciências agrárias, para identificar as principais transformações de dados utilizadas na área para a validação dos pressupostos da análise de variância.

## Resumo

Em análises estatísticas, uma prática comum para aumentar a validade da análise de variância é a aplicação de transformação de dados para converter medições em uma escala matemática diferente. Essa técnica foi empregada pela primeira vez em 1898 por Edgeworth e continua relevante em publicações científicas atuais, apesar da proliferação de técnicas mais modernas e avançadas que eliminam a necessidade de certas suposições. Transformações de dados, quando usadas apropriadamente, podem fazer com que os termos de erro do modelo se aproximem de uma distribuição normal. Também é possível usar a técnica para corrigir a heterogeneidade de variâncias ou para renderizar um modelo aditivo, garantindo a validade da análise de variâncias. Dado que essa técnica pode ser aplicada às pressas, potencialmente levando a resultados errôneos ou inválidos, conduzimos uma revisão sistemática da literatura de estudos no campo das ciências agrárias que utilizaram transformações de dados para a validação da análise de variâncias. O objetivo foi verificar as transformações empregadas pela comunidade científica para identificar possíveis erros e inconsistências na aplicação da técnica em publicações. Apresentamos também um método adaptado para aqueles que precisam conduzir uma revisão sistemática quando as informações relevantes estão localizadas no texto principal do artigo, mas não no resumo ou no título. Neste estudo, identificamos deficiências e equívocos associados with the use of data transformation, e observamos utilização incompleta e inadequada da técnica em 94,28% da analyzed amostra, resultando em conclusões equivocadas e errôneas nos resultados da pesquisa científica.

Palavras-chave: *Manipulação de dados; revisão sistemática de literatura; ciências agrárias; ANOVA; análise descritiva.*

## 2.1 Introdução

A análise de variância (ANOVA) é um método utilizado para testar a hipótese de igualdade estatística entre um conjunto de médias populacionais, sendo uma opção superior ao teste t devido à maior probabilidade de se cometer um erro do tipo I ao utilizá-lo para comparar múltiplas médias, conforme afirma Mohr *et al.* (2021). Para validar a análise ANOVA, suposições como aditividade do tratamento e efeitos ambientais, independência, homocedasticidade das variâncias e normalidade dos erros experimentais precisam ser satisfeitas, conforme apontado por Cochran (1947). Transformações de dados podem ser aplicadas para validar a análise quando essas suposições não forem atendidas.

Inicialmente proposto por Edgeworth (1898) e conhecido como Método de Tradução, as transformações de dados surgiram da necessidade de reportar distribuições observadas à distribuição normal. De acordo com Bartlett (1947), o propósito usual da transformação é mudar a escala de medições para tornar a análise ANOVA válida. No entanto, uma transformação é bem-sucedida somente se ela satisfaz os requisitos de homogeneidade de variância, estrutura de aditividade e normalidade de erro através da mesma transformação, conforme declarado por Atkinson (1985).

Após qualquer transformação, é essencial verificar se houve melhora na distribuição dos dados, conforme destacado por Quinn e Keough (2002). Além disso, a transformação de dados fornece uma solução acessível para evitar distribuições de erros não normais, conforme afirmado por Pierre *et al.* (2018), permitindo análises fáceis aplicando modelos lineares. No entanto, os pesquisadores devem ter cautela ao usar transformações, pois elas podem ter efeitos contrários (RUPERT, 2001).

Apesar de frequentemente utilizar transformações de dados na literatura científica, elas recebem críticas de perspectivas teóricas e práticas (OLIVEIRA *et al.*, 2018). Uma das principais críticas às transformações de dados é o potencial de alterar a interpretação dos resultados (BOX e COX, 1964). Segundo esses autores, a transformação de dados pode mudar o foco do estudo, levando a conclusões diferentes daquelas obtidas a partir dos dados originais. Além disso, algumas transformações podem não ser prontamente interpretáveis, dificultando a comunicação dos resultados devido à escala transformada. Outra crítica é que as transformações podem afetar a robustez da ANOVA, resultando em falsos positivos ou negativos (HOCKING, 1976). Portanto, é crucial considerar cuidadosamente as implicações de qualquer transformação de dados antes de aplicá-la em uma análise de variância.

Devido ao grande volume de estudos em ciências agrárias que conduzem análises de variância e podem fazer uso de várias transformações de dados para validá-las, uma revisão sistemática da literatura é uma abordagem valiosa para identificar tendências de assunto e determinar se há utilização adequada de técnicas de transformação para garantir resultados confiáveis e válidos de análise de variância. De acordo com Khan (2020), uma revisão sistemática envolve coletar e examinar sistematicamente trabalhos publicados para extrair estatísticas e informações relevantes dos estudos selecionados. Por outro lado, a meta-análise é o aspecto quantitativo da síntese dos analyzedtrabalhos, permitindo um resumo numérico dos resultados observados.

Neste contexto, conduzimos uma revisão sistemática da literatura de trabalhos publicados em ciências agrárias que empregam transformações de dados para validar análises de variância. Realizamos uma análise detalhada de várias publicações que utilizaram esta técnica para identificar as transformações mais comuns adotadas pela comunidade científica and their motivations, providing an overview of data transformation practices in agrarian sciences. Este estudo também identificou deficiências e equívocos associados à aplicação deste método. Os resultados da analyzedamostra revelaram uma aplicação inadequada da técnica pela comunidade científica, destacando a necessidade de maior rigor metodológico em tais análises por pesquisadores.

No processo de estabelecer uma visão geral das práticas de transformação de dados em ciências agrárias por meio de uma revisão sistemática, adaptamos o método de triagem PRISMA para permitir buscas por palavras-chave dentro do texto completo dos artigos. Fornecemos descrições detalhadas de nossa adaptação para facilitar a pesquisa para aqueles que conduzem revisões sistemáticas, especialmente quando as informações relevantes estão localizadas dentro do texto principal do artigo, mas não no resumo ou título.

## 2.2 Métodos

### 2.2.1 Coleta de dados

O processo de revisão sistemática da literatura adotado neste estudo seguiu as diretrizes do PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), desenvolvido por Moher *et al.* (2021). Estas diretrizes descrevem os estágios de identificação, triagem, elegibilidade e inclusão, garantindo a validade e confiabilidade dos resultados.

Na etapa de identificação, os estudos de interesse são localizados nas bases de dados, seguidos pela seleção e classificação dos trabalhos com base nos títulos e resumos na etapa de triagem. Na etapa de elegibilidade, avaliamos o texto completo dos artigos selecionados da etapa anterior. Na etapa de inclusão, as informações relevantes são extraídas sistematicamente com base em tópicos predefinidos.

No entanto, nossa metodologia se adaptou a essas diretrizes na fase de triagem. Em vez de analisar apenas os resumos dos artigos, baixamos os artigos selecionados na etapa de identificação. Buscamos termos específicos no texto completo desses trabalhos para selecionar os artigos que prosseguiriam para as etapas de elegibilidade e inclusão. Essa estratégia foi empregada porque informações cruciais sobre as transformações às vezes estão presentes nos resumos, título e metodologia. Dessa forma, as informações de interesse seriam notadas.

#### 2.2.1.1 Etapa de identificação

Para identificar artigos no campo das ciências agrícolas que conduziram análises de variância, conduzimos buscas abrangentes em plataformas acadêmicas como Scopus e Web of Science, usando palavras-chave como as “variance analysis”, “analysis of variance” e “ANOVA”. Restringimos nossa busca a artigos publicados exclusivamente em inglês, a língua global da academia e da pesquisa. Essa abordagem garante uma base uniforme e acessível para os estudos escolhidos.

Devido à ausência de filtros específicos para ciências agrícolas nas plataformas utilizadas, empregamos diferentes abordagens de refinamento para cada plataforma. No Scopus, na aba “subject area”, filtramos por “Agricultural and Biological Sciences”, excluindo periódicos não relacionados a ciências agrícolas. No Web of Science, na aba “Web of Science Categories”, filtramos por “Agronomy”, “Agriculture Dairy Animal Science”, “Agriculture Multidisciplinary”, “Agricultural Engineering” e “Agricultural Economics Policy”. A última busca por cada fonte foi realizada em 14 de maio .

#### 2.2.1.2 Etapa de triagem

Devido à grande quantidade de arquivos a serem baixados e analyzed, trabalhamos apenas com arquivos de acesso aberto. Exportamos esses arquivos no formato RIS para o Rayyan, um aplicativo da web desenvolvido pelo *Qatar Computing Research Institute* para revisões sistemáticas. Organizamos os arquivos de acordo com o ano de publicação e excluímos manualmente as duplicatas para evitar downloads desnecessários de arquivos.

Utilizamos as bibliotecas “pyautogui”, “openly” e “pandas” para desenvolver uma automação Python que baixa os arquivos em formato pdf diretamente das plataformas *Scopus* e *Web of Science*. Os que não foram baixados por meio da automação foram obtidos manualmente.

Posteriormente, empregamos a biblioteca “pypdf4” para desenvolver um código de busca em Python para buscar os termos “data were transformed”, “transformation of the data” e “data transformation” nos textos completos dos trabalhos, visando reduzir o número de artigos necessários para leitura e análise.

Python é um software disponível gratuitamente em <https://www.python.org/>, e o código Python pode ser encontrado em <https://github.com/ghscosta/data-transform> para qualquer pessoa interessada em replicar os procedimentos descritos neste artigo.

### 2.2.1.3 Estapas de elegibilidade e inclusão

Na etapa de elegibilidade, dois revisores examinaram os textos completos dos artigos resultantes do processo de triagem e extraíram informações relevantes na fase de inclusão. Essas duas etapas foram realizadas simultaneamente. Dos artigos que passaram pelo processo de filtragem inicial, coletamos as seguintes informações: número de referência para identificação do artigo, título, periódico, ano de publicação, autores, utilização da análise de variância na metodologia, delineamento empregado, verificação dos pressupostos da ANOVA, tipo de transformação aplicada, justificativa para adoção da transformação, reavaliação dos pressupostos após transformação dos dados, escala de interpretação dos resultados, elemento de estudo no experimento e fator de impacto de cada periódico. A descrição de cada variável pode ser acessada na Tabela 1, disponível em <https://rpubs.com/JhenniferNascimento/table> e <https://zenodo.org/records/10757759>. Descartamos os artigos que não forneceram essas informações e consideramos apenas aqueles que continham os detalhes pertinentes para a análise descritiva.

### 2.2.2 Análise descritiva

Para apresentar os dados coletados e potenciais relacionamentos entre variáveis, empregamos gráficos de barras, gráficos de pizza e diagramas de Sankey para representar e analyzingas informações obtidas dos artigos amostrados. Nós tabulamos manualmente os dados, e um resumo conciso é fornecido em nossos dados estendidos, acessíveis na Tabela 2 em <https://rpubs.com/JhenniferNascimento/table> e <https://zenodo.org/records/10757759>.

Gráficos de pizza e barras foram usados para ilustrar os critérios de elegibilidade para determinar a inclusão de estudos em cada síntese, juntamente com as porcentagens de estudos descartados e incluídos. Diagramas de Sankey foram empregados para descrever variáveis-chave e elucidar potenciais relações entre elas.

## 2.3 Resultados

### 2.3.1 Coleta de dados

#### 2.3.1.1 Etapa de identificação

A busca pelos termos “variance analysis”, “analysis of variance” e “ANOVA”, que conduzimos nas plataformas *Web of Science* e *Scopus*, resultou na recuperação de 8.460 arquivos na primeira e 8.695 na segunda. Desse total, 3.615 arquivos da *Web of Science* e 3.529 da *Scopus* estavam disponíveis em acesso aberto. Consequentemente, durante a etapa de identificação, obtivemos um conjunto combinado de 7.144 artigos publicados entre 1961 e 2023 em ambas as plataformas.

#### 2.3.1.2 Etapa de triagem

Após a remoção de duplicatas, restaram 6.077 artigos para download. Baixamos 5.032 artigos por meio de automação, deixando 1.045 para download manual. Apesar de aplicar o filtro de acesso aberto em ambas as plataformas, não conseguimos acessar 52 dos artigos marcados para *download* manual, resultando em um total de 6.025 artigos baixados com sucesso.

Ao pesquisar os termos “data were transformed”, “transformation of the data” e “data transformation” no texto completo dos 6.025 artigos, reduzimos o número de artigos que exigem análise para 565 para análise nos próximos estágios.

#### 2.3.1.3 Estapas de elegibilidade e inclusão

Lemos e coletamos informações relevantes de 565 artigos, resultantes das etapas anteriores. Destes, 506 artigos foram descartados, com detalhes sobre os descartes apresentados na análise descritiva. Coletamos efetivamente informações de apenas 59 artigos relevantes para nossa busca.

Alguns dos artigos tabulados conduziram mais de um experimento em seu trabalho ou usaram mais de uma transformação, levando assim a mais de uma análise de variância. Cada experimento e transformação distintos foram tratados como observações separadas para garantir a contagem correta. Como resultado, a amostra de 59 artigos gerou 70 observações tabuladas. As características do estudo são apresentadas na Tabela 2, disponível em <https://rpubs.com/JhenniferNascimento/table>.

### 2.3.2 Análise descritiva

A Figura 2.1 apresenta as quantidades de artigos descartados e incluídos para tabulação, onde as barras empilhadas mostram visualmente o total de 506 e 59 artigos descartados e incluídos na amostra final, respectivamente. Por outro lado, o primeiro gráfico de pizza apresenta os motivos utilizados para o descarte, e o segundo mostra as transformações utilizadas para a validação da análise de variâncias.

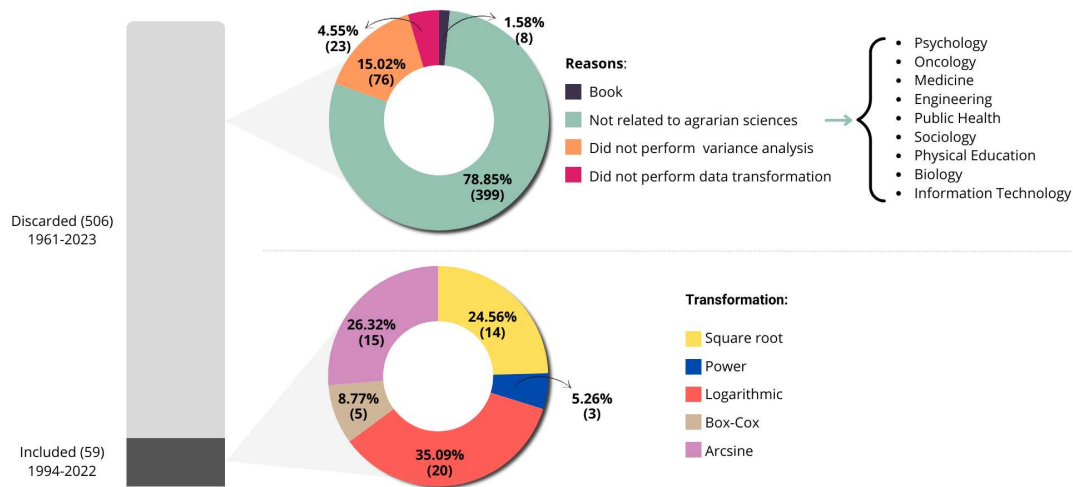


Figura 2.1: Análise descritiva dos artigos incluídos e descartados, considerando o campo de estudo das obras e os tipos de transformações nelas encontradas.

No primeiro setor circular, fica evidente que descartamos aproximadamente 78% dos artigos de áreas como medicina, engenharia, psicologia e educação física por estarem fora do campo das ciências agrárias. Descartamos aproximadamente 15% dos artigos por não conterem análise de variância, 4,55% por não utilizarem transformação de dados e pouco mais de 1% eram livros e, portanto, incompatíveis com os critérios de inclusão.

No segundo setor circular da Figura 2.1, que reporta artigos com informações relevantes para nosso estudo publicados entre 1994 e 2022, observamos cinco tipos de transformações. Entre elas, estão as transformações logarítmica, de raiz quadrada, Box-Cox, de potência e de arco seno, usadas para validar análises de variância em experimentos envolvendo culturas como milho, arroz, maracujá e outras, conforme ilustrado na Figura 2.2. Agrupamos categorias com apenas uma ocorrência como “other” para a variável “study element”.

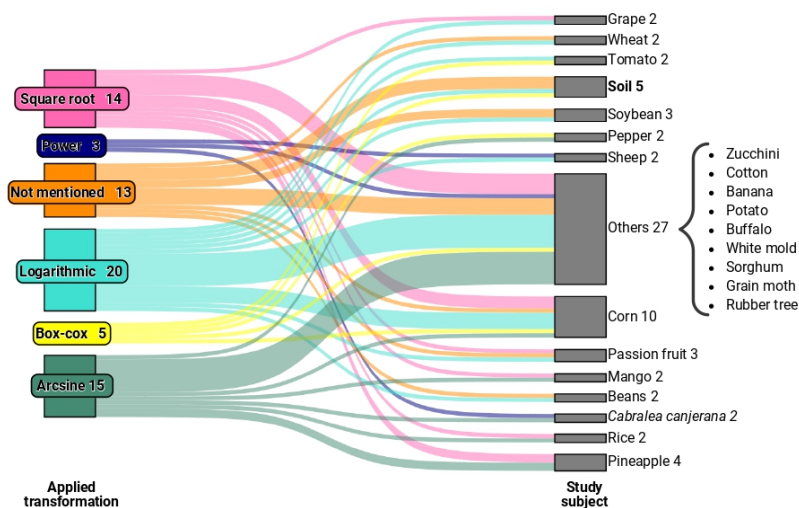


Figura 2.2: Diagrama de Sankey representando a relação entre as transformações usadas nos estudos avaliados e os elementos de estudo dos experimentos.

Além de examinar a relação entre os elementos de estudo do experimento e as transformações usadas, investigamos usando a Figura 2.3 para explorar as interações e conexões entre as transformações. Também verificamos suposições, avaliamos a motivação para empregar essas transformações, consideramos a escala de interpretação e reavaliamos as suposições da análise de variância após aplicar a transformação.

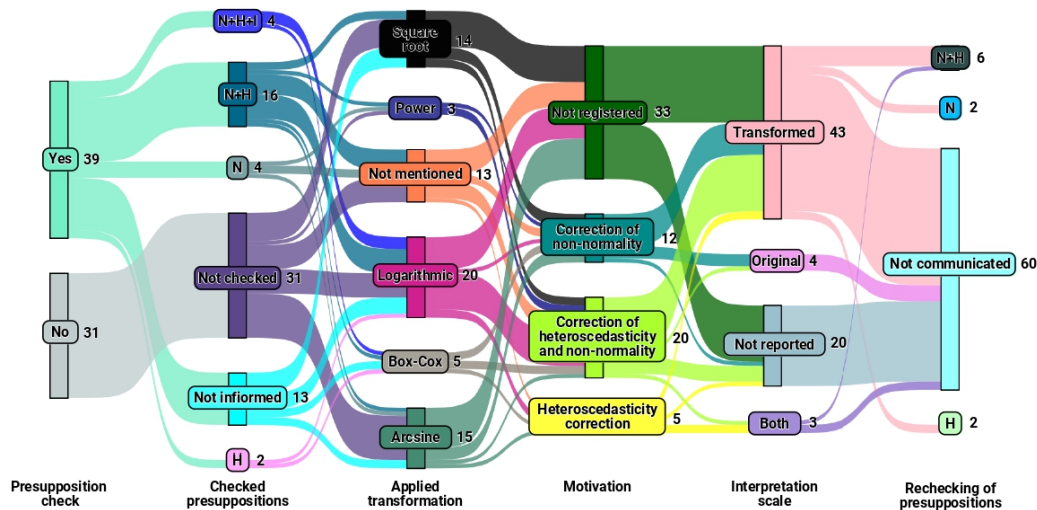


Figura 2.3: Diagrama de Sankey ilustrando as relações entre as transformações usadas, a verificação de suposições (N para normalidade, H para homogeneidade de variâncias, N+H para normalidade e homogeneidade de variâncias e N+H+I para normalidade, homogeneidade de variâncias e independência), a motivação para usar essas transformações, a escala de interpretação e a reverificação de suposições.

Como resultado, a Figura 3 destaca que as transformações foram aplicadas dentro da amostra final com (55,71%) e sem (44,28%) a verificação dos pressupostos da análise de variância. Entre os estudos amostrados que verificaram os pressupostos, 66,66% verificaram pelo menos um dos três pressupostos (normalidade, homogeneidade e independência dos resíduos) antes de aplicar a transformação, enquanto 33,33% mencionaram a verificação dos pressupostos, mas não especificaram qual o pressuposto verificado.

Dentre as transformações observadas na amostra (conforme ilustrado na Figura 2.3), as transformações logarítmica (28,57%), arco seno (21,42%) e raiz quadrada (20%) foram as mais utilizadas, independentemente de terem sido verificadas as hipóteses. Ao mesmo tempo, 18,57% não especificaram a transformação utilizada. Os 11,44% restantes referem-se a artigos que utilizaram transformações de potência e Box-Cox.

A Figura 2.3 também revela a presença de artigos que não verificaram pressupostos, não especificaram a transformação utilizada, não registraram o racional por trás da escolha da transformação e não relataram se os pressupostos medidos da análise de variância após a aplicação da transformação foram mantidos. Apenas 5,71% dos estudos realizaram verificação de normalidade, homogeneidade e independência dos pressupostos residuais e adotaram transformações com justificativas e motivações plausíveis, ilustrando a aplicação incompleta e potencialmente incorreta da técnica em 94,28% dos estudos considerados.

Por fim, a Figura 2.3 também fornece informações relevantes sobre a abordagem de interpretação adotada nos estudos. A maioria dos artigos interpretou seus resultados na escala transformada, considerando os dados após a aplicação das transformações. No entanto, alguns estudos mencionaram a transformação inversa para reverter os dados para a escala original antes da interpretação, o que é uma prática incorreta. Quanto à reavaliação de premissas após a implementação das transformações, a maioria dos estudos não mencionou essa etapa.

## 2.4 Discussão

### 2.4.1 Coleta de dados

A execução das etapas propostas pelo PRISMA (identificação, triagem, elegibilidade e inclusão), com as adaptações implementadas, apresenta uma abordagem inovadora para a busca e seleção de artigos para este estudo. A busca direta pelos termos de interesse no texto completo representou uma estratégia ousada, que pode potencialmente servir como uma técnica valiosa para futuras revisões sistemáticas, particularmente quando os pesquisadores buscam informações que podem não estar explicitamente declaradas no título ou no resumo, como foi o caso em nosso estudo.

A fase de triagem, implementada com download automatizado de artigos e busca por termos no texto completo, programada na linguagem Python, nos direcionou eficientemente para os artigos mais relevantes com esforço manual mínimo necessário. Embora alguns artigos exigissem download manual, reduzimos significativamente o esforço geral.

Durante as fases de elegibilidade e inclusão, identificamos e selecionamos artigos que relataram informações essenciais sobre a implementação de transformações de dados para validar a análise de variância. Percebemos que muitos desses estudos não documentaram, sem surpresa, várias informações relevantes.

### 2.4.2 Análise descritiva

Conforme ilustrado na Figura 2.1, grande parte dos artigos foi descartada, o que é comum em revisões sistemáticas de literatura, conforme apontado por Gerstner *et al.* (2017). Segundo os autores, informações essenciais são frequentemente omitidas em artigos científicos, levando à exclusão desses estudos durante a fase de revisão sistemática ou meta-análise. Em nosso estudo, apenas 10% da amostra forneceu as informações de interesse, com a maioria das exclusões ocorrendo porque não estavam relacionadas ao campo das ciências agrícolas. Esse resultado foi surpreendente, pois esperávamos obter uma amostra mais representativa do campo devido aos critérios de filtragem empregados nas plataformas.

Ao examinar as transformações empregadas nos estudos sobre os elementos analisados nos experimentos (como exemplificado na Figura 2.2), nenhum padrão específico emergiu que ligasse um tipo particular de elemento a uma transformação específica. Ao expandir a amostra, é concebível que possamos eventualmente discernir um padrão relevante entre os tipos de elementos usados em experimentos e sua interação com transformações específicas.

Em relação às interações e conexões das variáveis apresentadas na Figura 2.3, surgem preocupações sobre a validade das análises conduzidas em estudos onde 55% da amostra não verificaram pressupostos essenciais (normalidade, homogeneidade e independência dos resíduos) e optaram por transformações de dados. A prática de transformar dados sem antes determinar se é necessário para a análise compromete a interpretação e a confiabilidade dos resultados obtidos. Por outro lado, a constatação de que 26 dos 39 artigos que avaliaram efetivamente pelo menos uma das três premissas antes de aplicar as transformações (equivalente a 66,66%), representando apenas 37,14% da amostra total (de 70 artigos), levanta questões sobre a conscientização e o rigor estatístico dos autores no manuseio de tais análises de dados. Em contraste, os 33,33% dos artigos que mencionam a verificação de suposições sem especificar quais indicam uma necessidade de mais transparência nas práticas de análise estatística.

A distribuição de transformações de dados na amostra chama a atenção para a prevalência de transformações logarítmicas, de arco seno e de raiz quadrada, independentemente da verificação de suposições. No entanto, a omissão de casos de transformação levanta preocupações sobre a transparência e a reprodutibilidade da pesquisa, ressaltando a necessidade de documentação abrangente de métodos para garantir avaliação e replicabilidade precisas.

A descoberta de que apenas uma pequena parcela dos estudos adotou transformações com justificativas plausíveis após verificar as suposições de normalidade, homogeneidade e independência dos resíduos, somada à presença de estudos que não verificaram suposições, não mencionaram transformações utilizadas e não documentaram justificativas para suas escolhas, levanta questões sobre o rigor estatístico dos autores na condução de tais análises. Ela destaca a necessidade de maior rigor estatístico dentro da comunidade acadêmica e diretrizes mais explícitas para a condução de análises estatísticas robustas e transparentes, que garantam resultados confiáveis e significativos para o avanço da pesquisa.

Além disso, as descobertas destacadas também levantam considerações essenciais sobre o processo de revisão por pares e publicação de periódicos acadêmicos. O periódico com o maior fator de impacto na amostra inclui um dos artigos que não verificou suposições estatísticas, omitiu detalhes sobre transformações aplicadas e falhou em documentar a lógica por trás de suas escolhas. Aceitar trabalhos com tais características impacta a qualidade da literatura científica e compromete a integridade dos resultados apresentados e a interpretação correta das conclusões.

Portanto, os leitores desses trabalhos muitas vezes precisam de mais evidências para determinar a confiabilidade da análise executada e a veracidade do resultado, tendo que confiar no conteúdo apresentado sem a oportunidade de desenvolver suas conclusões com base nas análises. Isso mostra a importância de uma abordagem completa na descrição dos procedimentos utilizados em trabalhos científicos que são publicados e que os periódicos precisam ser mais cuidadosos, principalmente na hora de implementar essa metodologia tão antiga e utilizada.

Em relação aos aspectos relacionados à interpretação dos resultados e à abordagem metodológica adotada nos estudos da nossa amostra, a observação de que a maioria dos artigos optou por interpretar seus resultados na escala transformada aponta para uma prática comum e correta

na literatura. No entanto, retornar os dados à sua escala original antes da interpretação levanta uma discussão sobre a validade dessa abordagem, pois pode distorcer a interpretação correta dos resultados e potencialmente levar a conclusões errôneas. Para os artigos interpretados em ambas as escalas, aqueles fizeram um esforço desnecessário. A interpretação deve sempre ocorrer na escala transformada.

Por fim, outro ponto relevante é a ausência de menção à reavaliação de pressupostos após a aplicação de transformações. Isso levanta questões sobre a conscientização dos pesquisadores sobre a importância de verificar se os pressupostos da análise de variância são atendidos mesmo após a transformação dos dados. Omitir essa etapa pode comprometer a confiabilidade e a validade das conclusões. A documentação e a descrição detalhada dos métodos utilizados em toda pesquisa científica são cruciais, especialmente no campo das ciências agrícolas, que foi o foco do nosso estudo.

## 2.5 Conclusões

Introduzimos uma nova abordagem para triagem de revisões sistemáticas de literatura, que combina o download e a busca pelos termos de interesse diretamente dentro dos artigos de texto completo. Essa técnica se mostra valiosa, especialmente quando a pesquisa de interesse não está prontamente disponível em títulos ou resumos.

A análise descritiva das transformações usadas nos artigos revelou a prevalência de transformações logarítmicas, arco-seno e raiz quadrada. Tornou-se evidente que a verificação de suposições antes da aplicação de transformações foi conduzida consistentemente apenas em alguns casos, destacando a necessidade de uma abordagem mais detalhada e consistente nos procedimentos de planejamento e relato de pesquisa.

Interpretar resultados na escala transformada foi a abordagem mais comum, embora alguns estudos tenham adotado transformações reversas para retornar dados à escala original. No entanto, a omissão de detalhes da análise ressalta a necessidade de maior clareza e consistência na documentação dos procedimentos empregados.

Dado o nosso foco no campo das ciências agrícolas e as inconsistências identificadas na aplicação da técnica, surgem questões quanto ao uso adequado deste método em outras áreas do conhecimento.

## Disponibilidade de dados

### Dados subjacentes

Todos os dados subjacentes estão disponíveis como parte do artigo e não são necessários dados de origem adicionais.

### Dados estendidos

Zenodo : Dados estendidos para 'Transformação de dados estatísticos em ciências agrárias para análise de variância: uma revisão sistemática', <https://doi.org/10.5281/zenodo.10519177> (Nasci-

mento *et al.*, 2024).

### **Diretrizes de relatórios**

Zenodo : lista de verificação PRISMA e diagrama de fluxo para 'Transformação de dados estatísticos em ciências agrárias para análise de variância: uma revisão sistemática', <https://doi.org/10.5281/zenodo.10519177> (Nascimento *et al.*, 2024).

Os dados estão disponíveis sob os termos da Licença Pública Internacional Creative Commons Atribuição 4.0 (CC-BY 4.0 Internacional).

### **Disponibilidade do software**

Código fonte disponível em: <https://github.com/ghscosta/data-transform/tree/v0.2>. Software arquivado disponível em: <https://doi.org/10.5281/zenodo.10519177>. 15 Licença: GNU General Public Licença v3.0.

# Referências Bibliográficas

ATKINSON, A. C. *Plots, transformations, and regression*. 1<sup>a</sup> ed. USA: Oxford Science Publication, 1985. 320 p.

BARTLETT, M. S. The use of transformations. *Biometrics*, v. 3, n. 1, p. 39–52, 1947.

BOX, G. E. P.; COX, D. R. An analysis of transformations. *Journal of the Royal Statistical Society*, v. 26, n. 2, p. 211–243, 1964.

COCHRAN, W. G. When the assumptions for the analysis of variance are not satisfied. *Biometrics*, v. 3, n. 1, p. 22–38, 1947.

EDGEWORTH, F. Y. A suggestion for using powerful and informative tests of normality. *Royal Statistical Society*, v. 61, n. 3, p. 534–544, 1898.

GERSTNER, K.; MATEOS, D. M.; GUREVITCH, J.; *et al.* Will your paper be used in a meta-analysis? Make the reach of your research broader and longer lasting. *Methods in Ecology and Evolution*, v. 8, p. 777–784, 2017.

HOCKING, R. R. The Analysis and Selection of Variables in Linear Regression. *Biometrics*, v. 32, n. 1, p. 1–49, 1976.

KHAN, S. *Meta-analysis: methods for health and experimental studies*. 1<sup>a</sup> ed. Springer, 2020. 320 p.

MOHER, D.; LIBERATI, A.; TETZLAFF, J.; *et al.* Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *PLoS Medicine*, v. 6, p. e1000097, 2009.

MOHR, D. L.; WILSON, W. J.; FREUND, R. J. *Statistical methods*. 4<sup>a</sup> ed. Chennai: Academic Press, 2021. 720 p.

NASCIMENTO, J. S.; SILVA, J. F.; BERNARDES, R. C.; *et al.* Statistical data transformation in agrarian sciences for variance analysis: a systematic review [data and code]. 2024.

NASCIMENTO, J. S.; SILVA, J. F.; BERNARDES, R. C.; *et al.* Statistical data transformation in agrarian sciences for variance analysis: a systematic review [reporting guidelines]. 2024.

OLIVEIRA, J. P. R.; SANTANA, D. G.; PEREIRA, V. J.; *et al.* Data transformation: an underestimated tool by inappropriate use. *Acta Scientiarum*, v. 40, n. 1, p. 35300–35311, 2018.

PIERRE, A. P.; SHIKON, V.; SCHNEIDER, D. C. Count data in biology - Data transformation or model reformation? *Ecology and Evolution*, v. 8, n. 6, p. 3077–3085, 2018.

QUINN, G. P.; KEOUGH, M. J. *Experimental design and data analysis for biologists*. 1<sup>a</sup> ed. New York: Cambridge University Press, 2002. 556 p.

RUPERT, D. Statistical analysis, special problems of transformations of data. *Social and Behavioral Sciences*, v. 20, n. 1, p. 15007–15014, 2001.

## Capítulo 3

**Publicado como artigo original na  
revista DELOS**

# **Transformações de dados para violações das suposições de ANOVA em subáreas das ciências agrárias: uma revisão sistemática, estudo de simulação e diretrizes práticas**

Jhennifer dos Santos Nascimento<sup>1</sup>, Luciano Gonçalves Batista<sup>1</sup>, Wagner Faria Barbosa<sup>1</sup>,  
Lucas Pereira Belo<sup>1</sup>, Jonas Firmiano da Silva<sup>1</sup>, Guilherme Henrique da Silva Costa<sup>2</sup>,  
Paulo César Emiliano<sup>1</sup>

<sup>1</sup>Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, Estado de Minas Gerais, Brasil

<sup>2</sup>ICTP South American Institute for Fundamental Research e Instituto de Física Teórica, Universidade Estadual Paulista, São Paulo, Estado de São Paulo, Brasil

O presente capítulo apresenta a revisão sistemática realizada nos subcampos das ciências agrárias, incluindo entomologia, fitotecnia, ciências florestais e ciência do solo. Além disso, aborda um estudo de simulação que compara as transformações logarítmica, raiz quadrada, arco seno e Box-Cox em relação à sua eficácia na correção da não normalidade e heterogeneidade de variâncias. Por fim, são oferecidas diretrizes práticas para a execução e o reporte adequado da análise de variância em artigos científicos.

## Resumo

O uso de transformações de dados para a validação da análise de variâncias é uma técnica adotada pela comunidade científica desde o século XIX. Neste estudo, realizamos uma revisão sistemática abrangendo quatro subáreas das ciências agrárias (entomologia, fitotecnia, florestal e solos) para investigar como essa técnica tem sido empregada nessas áreas específicas. Além da revisão sistemática, conduzimos um estudo de simulação para avaliar a eficácia das transformações logarítmica, raiz quadrada, arco seno e Box-Cox na correção das violações dos pressupostos de normalidade e homogeneidade das variâncias, uma vez que essas foram as transformações mais frequentes na revisão sistemática. Embora algumas transformações sejam eficazes na homogeneização das variâncias, poucas corrigem simultaneamente a não normalidade e a heterogeneidade. Para facilitar a fase de triagem na revisão sistemática, desenvolvemos um aplicativo web como uma alternativa para usuários que não estão familiarizados ou não se sentem confortáveis com a interface do Python. Por fim, oferecemos diretrizes práticas para a aplicação correta da ANOVA com transformações, visando minimizar os erros e equívocos identificados na revisão sistemática.

Palavras-chave: *Análise de variância; método para triagem; download em massa; análise descritiva.*

### 3.1 Introdução

As transformações de dados surgiram como uma solução para relacionar distribuições assimétricas à distribuição normal, com Edgeworth (1898) sendo o pioneiro nessa abordagem. Essa prática tornou-se comum em análises de variâncias (ANOVA), que dependem de pressupostos como normalidade, homogeneidade das variâncias, independência dos erros e aditividade do modelo. Quando esses pressupostos não são atendidos, os resultados da ANOVA podem ser comprometidos, e as transformações de dados configuram uma ferramenta para garantir que os dados satisfaçam essas condições, permitindo dar continuidade à análise.

Ramachandran e Tsokos (2021) apontam que, na prática, é raro que os pressupostos de normalidade e homogeneidade das variâncias sejam completamente satisfeitos. A violação desses critérios compromete a validade do teste F, utilizado na ANOVA, tornando indispensável a verificação cuidadosa dos dados. No campo das ciências agrárias, onde a comparação de múltiplos grupos experimentais é frequente, Nascimento *et al.* (2024) destacam que as transformações de dados são utilizadas como ferramenta auxiliar para a execução da análise de variâncias.

Assis *et al.* (2023) afirmam que a transformação dos dados substitui a medida original da variável por uma nova escala. Quando os pressupostos não são atendidos na escala original, uma transformação apropriada pode auxiliar no cumprimento desses critérios. Esse procedimento pode ser aplicado em qualquer modelo de análise de variâncias ou regressão, independentemente de o experimento ser balanceado ou não, ou de o tamanho amostral ser grande ou pequeno. As estatísticas usuais e os graus de liberdade são mantidos, sem perda de precisão ou sensibilidade do teste. Contudo, embora análises descritivas e exploratórias possam guiar a escolha da transformação, não

existem regras gerais que garantam o sucesso dessa escolha.

Mohr *et al.* (2021) sugerem que, para lidar com assimetria e variâncias desiguais, a transformação dos dados é frequentemente a abordagem preferida para tornar a distribuição simétrica e variâncias homogêneas. Transformações que estabilizam variâncias também tendem a reduzir a assimetria, o que é fundamental para a validade da ANOVA e a interpretação correta dos resultados. No entanto, os autores também destacam que as inferências são baseadas nas médias dos valores transformados, que não correspondem necessariamente às médias dos valores originais. Isso significa que não é apropriado reverter as estatísticas para a escala original após a transformação. Quando as transformações não são eficazes, os autores recomendam, para os dados originais, o uso do teste F de Welch, nos casos de variâncias desiguais, ou o teste de Kruskal-Wallis, para a não-normalidade.

Transformações como logarítmica, raiz quadrada e arco-seno, são amplamente discutidas e aplicadas na comunidade acadêmica para homogeneizar variâncias. Assis *et al.* (2023) identificam-nas como as principais transformações utilizadas, enquanto Nascimento *et al.* (2024) corroboram essa posição, destacando que, além dessas três, a transformação de Box-Cox foi também utilizada com recorrência nos artigos da área de ciências agrárias para validação da ANOVA.

Dado o uso frequente da técnica pela comunidade científica, sua execução correta é crucial para a confiabilidade dos resultados de pesquisas que a adotam. Nascimento *et al.* (2024) abordaram as práticas inadequadas do uso de transformação de dados na área de ciências agrárias, indicando que uma análise mais aprofundada em seus subcampos poderia revelar padrões e problemas específicos na aplicação dessas metodologias, que não foram identificados na análise geral da área. O presente estudo aprofunda essa investigação, explorando especificamente como essas práticas estão sendo aplicadas em quatro diferentes subcampos das ciências agrárias.

Com este propósito, conduzimos uma revisão sistemática de literatura de artigos publicados em entomologia, fitotecnia, florestal e solos, subáreas de ciências agrárias, que empregam transformações de dados para validar pressupostos da análise de variâncias. Uma análise descritiva das transformações utilizadas nesses subcampos foi efetuada, com o objetivo de identificar padrões e problemas subjacentes em sua aplicação, contribuindo, assim, para a melhoria das práticas científicas de cada subcampo.

Para complementar a revisão sistemática, conduzimos um estudo de simulação com o objetivo de avaliar a eficácia das transformações mais frequentes, identificadas na literatura, na correção dos pressupostos de normalidade e homogeneidade das variâncias. Além disso, desenvolvemos diretrizes práticas para a execução adequada da análise de variâncias, com o objetivo de guiar pesquisadores na aplicação correta da técnica em artigos científicos, e evitar erros de execução e reporte da análise.

## 3.2 Revisão sistemática

### 3.2.1 Metodologia da revisão sistemática

A revisão sistemática de literatura conduzida neste estudo, para as subáreas de ciências agrárias, entomologia, fitotecnia, florestal e solos, foram orientadas pelas diretrizes PRISMA - *Preferred*

*Reporting Items for Systematic Reviews and Meta-Analyses*, conforme estabelecido por Moher *et al.* (2009). Essas diretrizes delimitam as fases de identificação, triagem, elegibilidade e inclusão, assegurando a confiabilidade e rigor da revisão sistemática.

Na fase de identificação, os estudos relevantes são localizados nas bases de dados. A seguir, na etapa de triagem, os artigos são selecionados e classificados com base nos títulos e resumos. Durante a fase de elegibilidade, procede-se à avaliação do texto completo dos artigos selecionados anteriormente. Finalmente, na etapa de inclusão, extrai-se sistematicamente as informações pertinentes conforme os tópicos previamente estabelecidos. Cada etapa será detalhada nas subseções 3.2.1.1 a 3.2.1.3.

### 3.2.1.1 Etapa de identificação

Na fase de identificação, escolhemos as bases de artigos científicos Scopus e Web of Science, que são plataformas internacionais consolidadas na comunidade científica, para a identificação de pesquisas que mencionaram o uso da análise de variâncias. Nossa revisão sistemática foca no uso de transformações de dados em conjunto com a ANOVA. Durante a fase de identificação, selecionamos estudos que aplicaram a ANOVA, e somente na triagem, filtraremos especificamente aqueles que utilizaram transformações de dados.

Assim como Nascimento *et al.* (2024), para identificar os artigos que mencionaram de alguma forma o uso da análise de variâncias, a busca em ambas as plataformas foi por meio dos termos (“variance analysis” OR “analysis of variance” OR “ANOVA”). Devido às diferenças intrínsecas às próprias plataformas, a seleção dos artigos por área foi distinta para cada uma, de acordo com as suas especificidades.

A plataforma *Web of Science*, possui filtros específicos para as subáreas entomologia, fitotecnia, florestal e solos, o que facilitou a busca. Dessa forma, foi necessário somente complementar a busca dos termos (“variance analysis” OR “analysis of variance” OR “ANOVA”), com o filtro “Limited to Article” em “Document type” para obter os artigos de interesse para o estudo.

Na plataforma *Scopus*, devido à ausência de filtros específicos para as subáreas de interesse, a busca foi realizada em mais passos. Após a busca dos termos (“variance analysis” OR “analysis of variance” OR “ANOVA”), em “Subject area” selecionamos a opção “Limited to Agricultural and Biological Sciences”, em “Document type” selecionamos a opção “Limited to Article”, e para identificar os artigos referentes à cada subárea, prosseguimos com a inclusão de palavras-chave adicionais, conforme descrito a seguir:

1. **Entomologia:** Após a busca inicial, adicionamos o conectivo AND e os termos (“entomology” OR “insect” OR “pest management”) para restringir os resultados aos artigos relacionados à entomologia.
2. **Fitotecnia:** Na subárea de fitotecnia, os termos (“plant science” OR “crop science” OR “botany”) foram adicionados.

3. **Florestal:** Para a área florestal, adicionamos o conectivo AND e utilizamos os termos (“forestry” OR “forest management” OR “silviculture”).
4. **Solos:** Finalmente, para a subárea de solos, utilizamos os termos complementares (“soil science” OR “soil fertility” OR “pedology”).

A combinação dessas palavras-chave com a busca inicial permitiu filtrar os artigos, focando nas publicações específicas de cada subárea. As listas com todos os artigos identificados nessa etapa foram exportadas por área no formato RIS, para que na etapa posterior possam ser realizados os procedimentos de triagem. A última busca em cada plataforma foi realizada em 02 de julho de 2024.

### 3.2.1.2 Etapa de triagem

Durante a fase de triagem, foram retiradas as duplicatas dos artigos selecionados na fase de identificação, feito o download desses artigos no formato pdf, e por fim realizada a busca de termos referentes ao uso de transformação de dados em seus respectivos textos completos.

Para a eliminação de duplicatas, utilizamos o Rayyan, uma aplicação web desenvolvida pelo *Qatar Computing Research Institute* especificamente para revisões sistemáticas. Os arquivos foram exportados no formato RIS para o Rayyan, onde foram organizados por ano de publicação. Optamos por utilizar o Rayyan devido à funcionalidade “Detect Duplicates”, que identifica automaticamente duplicatas ao comparar o percentual de similaridade entre os artigos, facilitando a eliminação de arquivos repetidos e evitando downloads desnecessários. Após a remoção das duplicatas, exportamos um arquivo CSV para cada área, contendo as informações dos artigos restantes.

Ao invés de analisar exclusivamente resumos, títulos e palavras-chave dos artigos, optamos por baixar os artigos selecionados na fase de identificação e realizar a triagem por meio de automação em Python para reduzir a quantidade de artigos a serem lidos completamente, seguindo a abordagem de Nascimento *et al.* (2024). Essa automação permite a busca de termos específicos no texto completo dos artigos, selecionando aqueles que contêm termos relacionados ao uso de transformações de dados.

Os arquivos CSV exportados do Rayyan foram importados para o gerenciador de referências open-source Zotero, para serem baixados utilizando a função “batch download”. Diferentemente e mais simples que o procedimento descrito por Nascimento *et al.* (2024), esse gerenciador permite fazer o download dos artigos que o usuário tem acesso, por serem de acesso aberto ou por meio do acesso institucional, utilizando o DOI dos artigos.

A maioria dos arquivos restantes, que não foram baixados automaticamente pelo Zotero, foram obtidos por meio de download manual. No entanto, ainda houve casos em que o acesso não foi possível, devido à restrição de acesso aberto ou à indisponibilidade para download institucional.

Após o download dos artigos, utilizamos o código de busca em Python, disponível em [https://github.com/JhenniferN/sistematic\\_review\\_R](https://github.com/JhenniferN/sistematic_review_R), para selecionar os artigos que mencionaram termos relacionados ao uso de transformações de dados. Enquanto Nascimento *et al.* (2024) buscaram nos textos completos os termos “data were transformed”, “transformation of the data” e “data

transformation”, nesse estudo optamos por buscar apenas o termo “transform”, que abrange os três mencionados, finalizando a etapa de triagem.

### 3.2.1.3 Etapas de elegibilidade e inclusão

As etapas de elegibilidade e inclusão foram conduzidas simultaneamente durante a avaliação dos artigos filtrados na triagem. A etapa de elegibilidade consiste em identificar se o artigo possui os critérios predeterminados, e a etapa de inclusão de coletar as informações de interesse, dado que o artigo é elegível. Quatro revisores avaliaram integralmente os artigos selecionados na triagem e extraíram as informações relevantes na fase de inclusão.

Artigos que não utilizaram análise de variâncias ou transformações de dados, foram automaticamente descartados. Os artigos que utilizaram ambas as técnicas foram selecionados e tiveram as seguintes variáveis coletadas: número de referência, título, periódico, ano de publicação, autores, uso da análise de variância na metodologia, delineamento experimental, verificação dos pressupostos da ANOVA, tipo de transformação aplicada, justificativa para a transformação, reavaliação dos pressupostos após a transformação, escala de interpretação dos resultados, elemento do estudo no experimento e fator de impacto do periódico.

Em [https://rpubs.com/JhenniferNascimento/interactive\\_data\\_from\\_systematic\\_review](https://rpubs.com/JhenniferNascimento/interactive_data_from_systematic_review), consta a descrição detalhada de cada variável.

### 3.2.1.4 Análise descritiva

A análise descritiva foi utilizada para descrever e apresentar as informações obtidas dos artigos sob estudo. Gráficos de barras, gráficos de setores e diagramas de Sankey foram utilizados para representar e analisar as variáveis coletadas na etapa de inclusão.

Utilizamos o software R (R Core Team, 2024) para fazer todos os gráficos e os códigos utilizados nas análises estão disponíveis em [https://github.com/JhenniferN/sistematic\\_review\\_R](https://github.com/JhenniferN/sistematic_review_R).

## 3.2.2 Resultados e discussão da revisão sistemática

### 3.2.2.1 Resultados e discussão da etapa de identificação

A busca feita na plataforma *Web of Science*, identificou 7.014 artigos, sendo 635 publicados de 1956 a 2024 na área de entomologia, 3.726 artigos de fitotecnia, publicados no período de 1951 a 2024, 1.658 da subárea florestal, de 1966 a 2024, e 995 de solos de 1977 a 2024.

Na *Scopus*, a busca inicial resultou em 45.074 artigos com o filtro “Agricultural and Biological Sciences”. Com os filtros adicionais, para cada subárea específica, identificamos 1.266 artigos relacionados à entomologia, 666 artigos que tratam de aspectos específicos de fitotecnia, 1.056 artigos relacionados à florestal e 319 artigos pertinentes à ciência dos solos, totalizando 3.307 artigos.

No total, foram identificados 1.901 artigos de entomologia, 4.392 de fitotecnia, 2.714 de ciências florestais e 1.314 de solos, somando os resultados das duas plataformas. É importante notar que alguns trabalhos podem ter sido duplicados, uma vez que o mesmo artigo pode ter sido encontrado

em ambas as bases de dados. Essas duplicatas serão eliminadas durante a etapa de triagem. A Tabela 3.1 apresenta a distribuição dos artigos por área e por plataforma.

Tabela 3.1: Quantidade de artigos identificados nas plataformas *Web of Science* e *Scopus* por subárea das ciências agrárias, incluindo entomologia, fitotecnia, florestal e solos.

PLATAFORMA/ÁREA	ENTOMOLOGIA	FITOTECNIA	FLORESTAL	SOLOS	TOTAL
<i>Web of Science</i>	635	3.726	1.658	995	7.014
<i>Scopus</i>	1.266	666	1.056	319	3.307
Total	1.901	4.392	2.714	1.314	10.321

Fonte: Elaborado pelos autores.

As palavras-chave escolhidas permitiram filtrar os artigos, focando nas publicações de cada subárea em ambas as plataformas. A diferença entre o número de artigos identificados na *Scopus* (3.307) e na *Web of Science* (7.014) pode ser atribuída às características intrínsecas de cada base de dados. Além disso, as metodologias de busca e os critérios de indexação foram distintas entre as plataformas, ocasionando diferentes quantidades de artigos identificados. A aplicação de palavras-chave específicas e filtros na *Scopus* restringiu ainda mais os resultados, permitindo um foco maior nas publicações relevantes para cada submapa. É importante ressaltar que a combinação dos periódicos de ambas as plataformas oferece uma visão mais completa e abrangente da literatura disponível.

### 3.2.2.2 Resultados e discussão da etapa triagem

O processo de triagem envolveu a remoção de artigos duplicados, feito por área no Rayyan, resultando na exclusão de 7,73% do total de registros. O *download* dos artigos, realizado utilizando o Zotero, foi responsável por baixar 83,15% do total de artigos, enquanto 14,96% foram baixados manualmente, restando 1,89% de artigos aos quais não tivemos acesso. Por fim, os artigos que mencionaram o termo “transform” em alguma parte do texto, foram filtrados por meio de automação em Python, resultando em 36,21% do total de artigos para as fases de elegibilidade e inclusão.

As quantidades de referências identificadas e exportadas para o Rayyan, as referências duplicadas detectadas, as duplicatas deletadas, as referências sem duplicatas, as referências a serem baixadas, as referências baixadas automaticamente, as referências baixadas manualmente, as referências que não foram baixadas e os artigos contendo o termo “transform” — que são os selecionados para as próximas fases da revisão sistemática — são todas apresentadas por área no Tabela 3.2.

Tabela 3.2: O quadro apresenta o processo de triagem das referências em quatro subáreas das ciências agrárias: entomologia, fitotecnia, florestal, e solos. Nela consta a contagem de referências importadas para o Rayyan, a quantidade de duplicatas detectadas e deletadas, a quantidade de referências restantes após a remoção de duplicatas, e o total de referências a serem baixadas. Também são mostradas as referências baixadas automaticamente e manualmente, além do número de artigos que não puderam ser baixados. Por fim, o quadro indica a quantidade de artigos em cada subárea que contém o termo “Transform”.

ETAPA DE TRIAGEM	ENTOMOLOGIA	FITOTECNIA	FLORESTAL	SOLOS	TOTAL
Referências identificadas	1.901	4.392	2.714	1.314	10.321
Duplicatas detectadas	312	202	912	186	1.612
Duplicatas deletadas	157	97	454	90	798
Referências sem duplicatas a serem baixadas	1.744	4.295	2.260	1.224	9.523
Referências baixadas pela automação	1.429	3.319	2.178	992	7.918
Referências baixadas manualmente	296	883	46	200	1.425
Artigos não baixados	19	93	36	32	180
Artigos com o termo “Transform”	764	1.347	846	488	3.445

Fonte: Elaborado pelos autores.

Sendo assim, por área, as taxas de download pela automação no Zotero para entomologia, fitotecnia, florestal e solos foram, respectivamente, 81,94%, 77,28%, 96,37% e 81,04% dos artigos. A taxa dos artigos baixados manualmente foi, respectivamente, 16,97%, 20,56%, 2,04% e 16,34%. As taxas dos artigos que não foram acessados foram, respectivamente, 1,09%, 2,17%, 1,59% e 2,61%. E as taxas dos artigos que contêm o termo “Transform” foram, respectivamente, 44,35%, 32,06%, 31,59% e 40,94%.

A utilização do Zotero facilitou o download em massa dos artigos, reduzindo em 83,15% o trabalho manual necessário para essa tarefa. Esse percentual variou conforme a subárea, mas, no total, menos de 15% dos artigos precisaram ser baixados manualmente, o que diminuiu consideravelmente o esforço necessário. Além disso, o uso do Zotero foi mais eficiente em comparação com a automação utilizada por Nascimento *et al.* (2024), pois reduziu o tempo e desempenho computacional exigido para o *download* dos artigos.

O uso da automação em Python para triar os artigos que utilizaram transformações de dados, com a busca do termo “transform” no texto completo dos artigos baixados, forneceu os estudos mais propensos a terem aplicado transformações de dados em sua metodologia, eliminando 63,82% dos artigos que, com certeza, não as utilizaram. Essa abordagem proporcionou uma expressiva redução no número de artigos que precisaram ser lidos e analisados nas etapas de elegibilidade e inclusão.

É importante ressaltar que a realização de revisões sistemáticas buscando os termos de interesse

diretamente no corpo dos artigos durante a fase de triagem é uma abordagem inovadora. Essa técnica permite revisões sistemáticas de temas que não são mencionados nos títulos, resumos e palavras-chave, como ocorrem nas revisões sistemáticas realizadas até o trabalho de Nascimento *et al.* (2024). Esse procedimento amplia as possibilidades de investigação de temas específicos que se encontram nas metodologias de trabalhos científicos.

### 3.2.2.3 Resultados e discussão das etapas de elegibilidade e inclusão

Na amostra final de artigos das áreas de entomologia, fitotecnia, florestal e solos, foram lidos e analisados 3.445 artigos nas fases de elegibilidade e inclusão, todos contendo o termo “transform” no corpo do texto. Destes, 1.226 foram selecionados para o estudo, pois utilizaram transformação de dados para a validação da ANOVA, resultando em 1.414 observações. Essa diferença se deve ao fato de que alguns estudos realizaram mais de um experimento ou aplicaram a ANOVA a diferentes variáveis com transformações distintas. Assim, cada transformação aplicada foi contabilizada como uma inclusão separada, mantendo-se a referência ao artigo original.

A Tabela 3.3 apresenta, por área, o número de artigos triados, os artigos selecionados na fase de inclusão e o total de observações geradas. A tabulação completa dos dados pode ser acessada em [https://rpubs.com/JhenniferNascimento/interactive\\_data\\_from\\_systematic\\_review](https://rpubs.com/JhenniferNascimento/interactive_data_from_systematic_review), que contém para cada área uma tabela interativa com todos os artigos avaliados, além de uma descrição detalhada das variáveis coletadas.

Tabela 3.3: Contagem dos artigos triados, artigos incluídos e observações geradas por área de estudo.

	ENTOMOLOGIA	FITOTECNIA	FLORESTAL	SOLOS	TOTAL
Artigos triados	764	1.347	846	488	3.445
Estudos incluídos	451	377	273	121	1.222
Total de observações selecionadas	525	430	321	139	1.415

Fonte: Elaborado pelos autores.

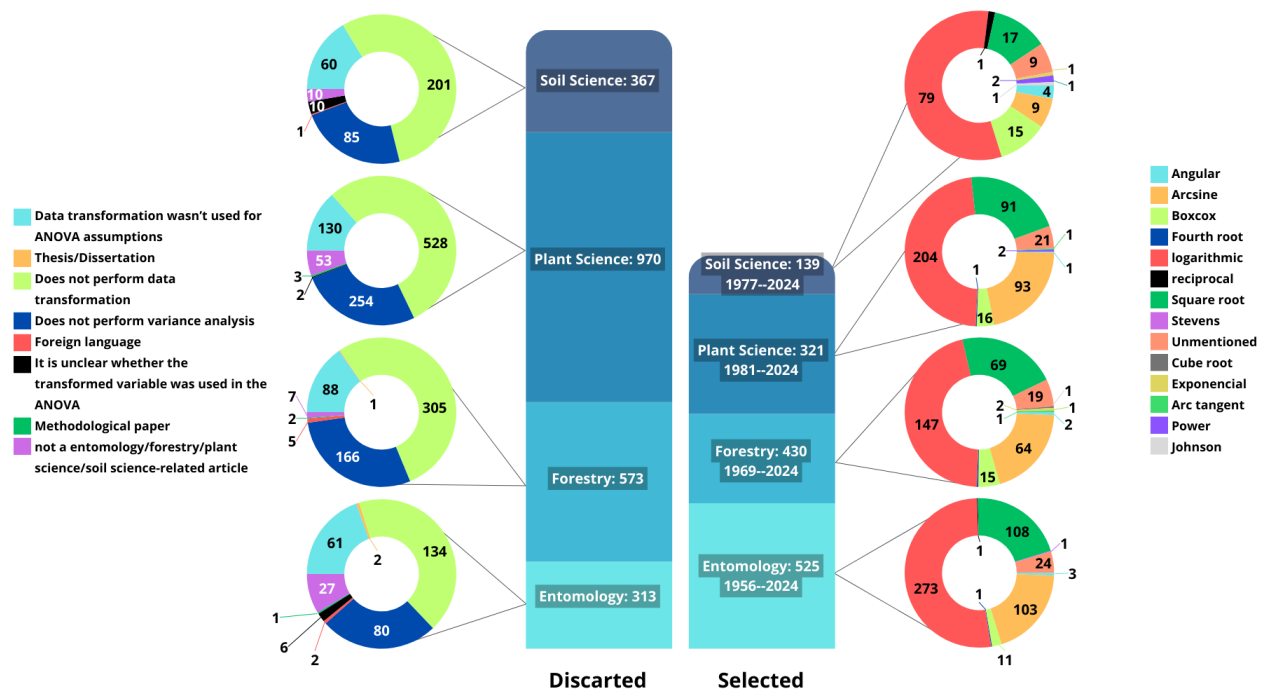
A busca pelo termo “transform” no corpo do texto trouxe artigos que poderiam ter sido excluídos caso tivesse sido utilizado um termo mais específico. No entanto, a escolha do termo de busca deve equilibrar a captura de estudos relevantes, pois termos mais específicos podem deixar de fora estudos importantes, dependendo da estratégia adotada pelo pesquisador, seja ela mais conservadora ou abrangente.

O processo de elegibilidade e inclusão de artigos em uma revisão sistemática exige um esforço considerável por parte dos pesquisadores envolvidos. Neste estudo, como a revisão abrangeu quatro subáreas distintas das ciências agrárias (entomologia, fitotecnia, ciências florestais e solos), o empenho necessário foi ainda maior. No entanto, esse trabalho foi facilitado pelas implementações propostas por Nascimento *et al.* (2024), além das atualizações técnicas que introduzimos neste artigo.

### 3.2.2.4 Resultados e discussão da análise descritiva dos estudos da revisão sistemática

A análise descritiva teve início com uma avaliação dos artigos descartados, mostrando os motivos para sua exclusão, bem como dos artigos incluídos, considerando o intervalo dos anos de publicação e as transformações utilizadas em cada área de estudo. As informações detalhadas estão representadas na Figura 3.1.

Figura 3.1: Visualização dos artigos incluídos e descartados na revisão sistemática, com indicação dos motivos de exclusão e análise do intervalo de anos de publicação, por área de estudo, e as transformações de dados utilizadas.



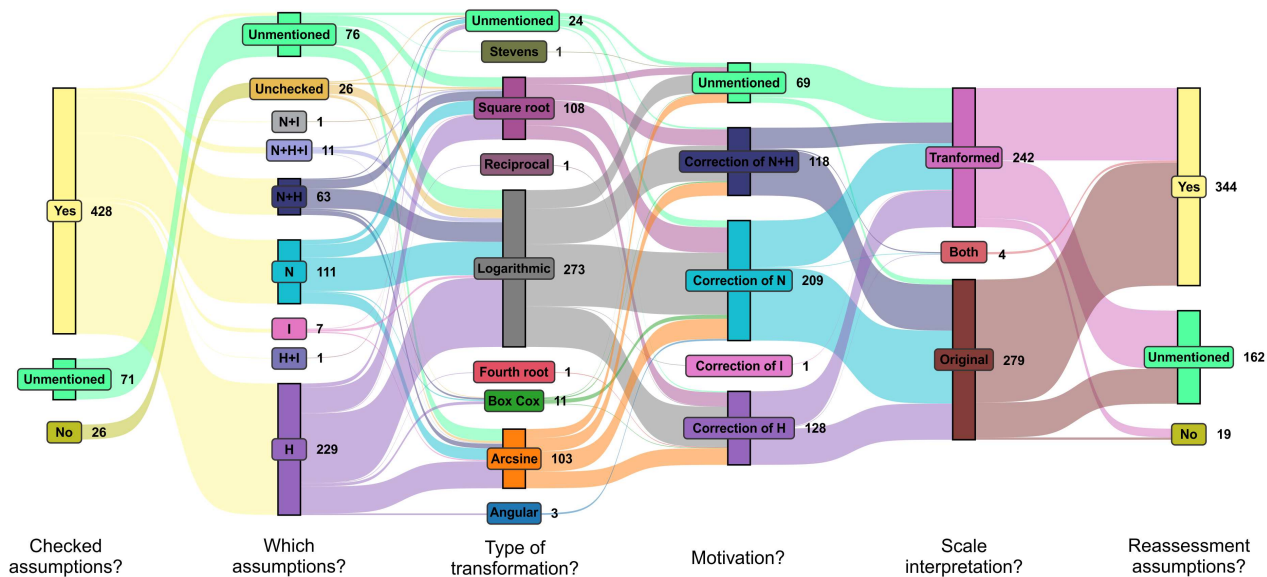
Fonte: Elaborado pelos autores.

Nos artigos selecionados, foram identificados estudos publicados entre 1956 e 2024, o que demonstra que o uso de transformações de dados para validação da ANOVA é uma técnica que permanece relevante até os dias atuais. Embora esperado que cada subárea apresentasse uma transformação predominante, as transformações logarítmica, raiz quadrada, arco seno e Box-Cox tiveram maiores frequências em todas as quatro subáreas analisadas, indicando uma tendência semelhante à encontrada por Nascimento *et al.* (2024), em sua revisão sistemática para a macro área de ciências agrárias.

A análise descritiva foi complementada com os gráficos de Sankey, apresentados nas Figuras 3.2 a 3.5, para as respectivas áreas de entomologia, fitotecnia, florestal e solos. Esses gráficos ilustram o fluxo de dados entre os diferentes aspectos do uso de transformações de dados para a validação da ANOVA, em cada área. Eles mostram como as transformações foram aplicadas para essa validação, e evidenciam as conexões entre as categorias das variáveis avaliadas, facilitando a visualização das principais tendências e relações dentro da amostra, permitindo a compreensão das

escolhas metodológicas predominantes em cada subárea pela comunidade científica.

Figura 3.2: Diagrama de Sankey para entomologia, ilustrando as relações entre as transformações utilizadas, a verificação dos pressupostos (N para normalidade, H para homogeneidade de variâncias, I para independência, N+H para normalidade e homogeneidade de variâncias, N+I para normalidade e independência, H+I para homogeneidade e independência, e N+H+I para normalidade, homogeneidade de variâncias e independência), a motivação por trás do uso dessas transformações, a escala de interpretação e a reavaliação dos pressupostos.



Fonte: Elaborado pelos autores.

Figura 3.3: Diagrama de Sankey para fitotecnia, ilustrando as relações entre as transformações utilizadas, a verificação dos pressupostos (N para normalidade, H para homogeneidade de variâncias, I para independência, N+H para normalidade e homogeneidade de variâncias, N+I para normalidade e independência, H+I para homogeneidade e independência, e N+H+I para normalidade, homogeneidade de variâncias e independência), a motivação por trás do uso dessas transformações, a escala de interpretação e a reavaliação dos pressupostos.

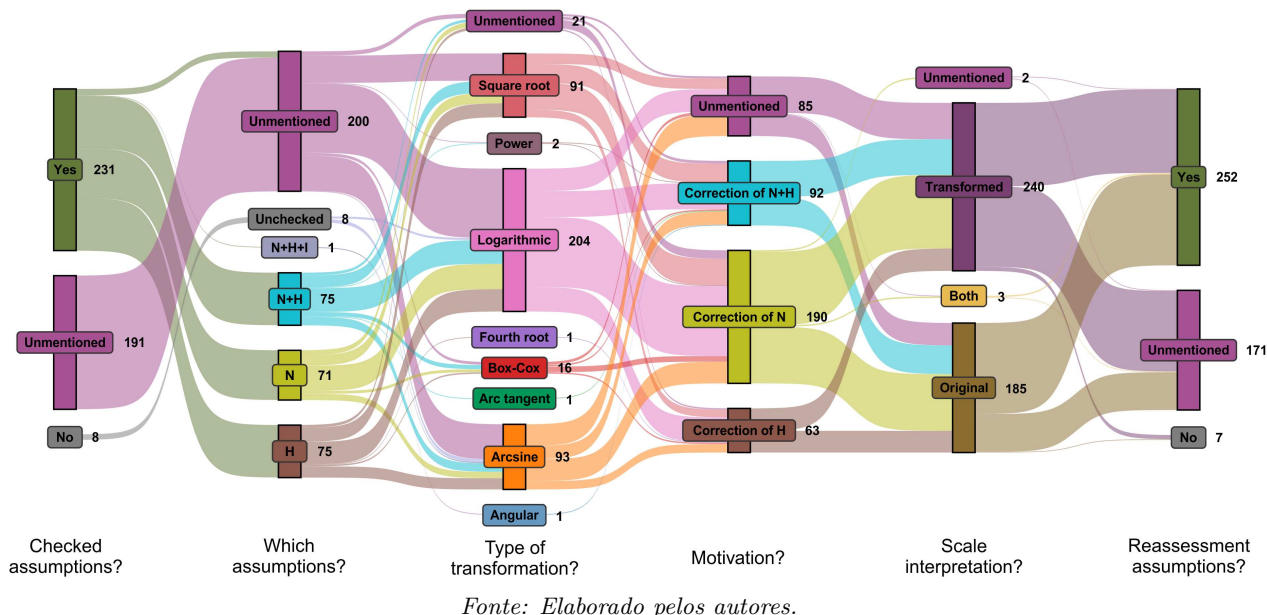


Figura 3.4: Diagrama de Sankey para ciências florestais, ilustrando as relações entre as transformações utilizadas, a verificação dos pressupostos (N para normalidade, H para homogeneidade de variâncias, I para independência, N+H para normalidade e homogeneidade de variâncias, N+I para normalidade e independência, H+I para homogeneidade e independência, e N+H+I para normalidade, homogeneidade de variâncias e independência), a motivação por trás do uso dessas transformações, a escala de interpretação e a reavaliação dos pressupostos.

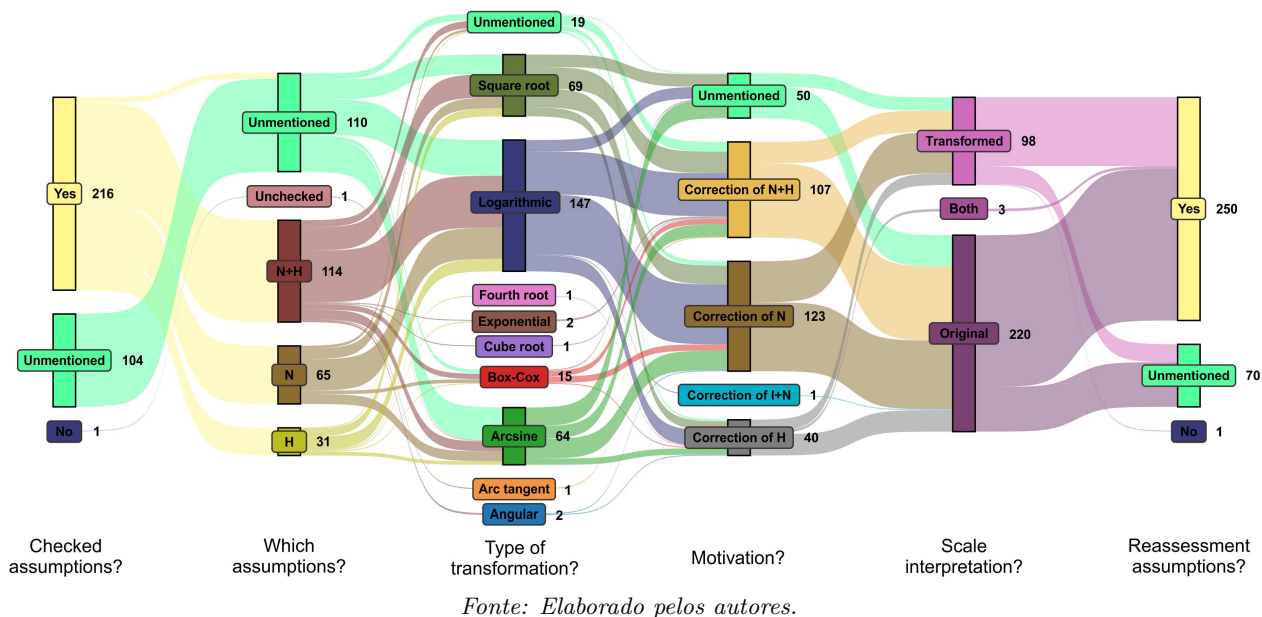
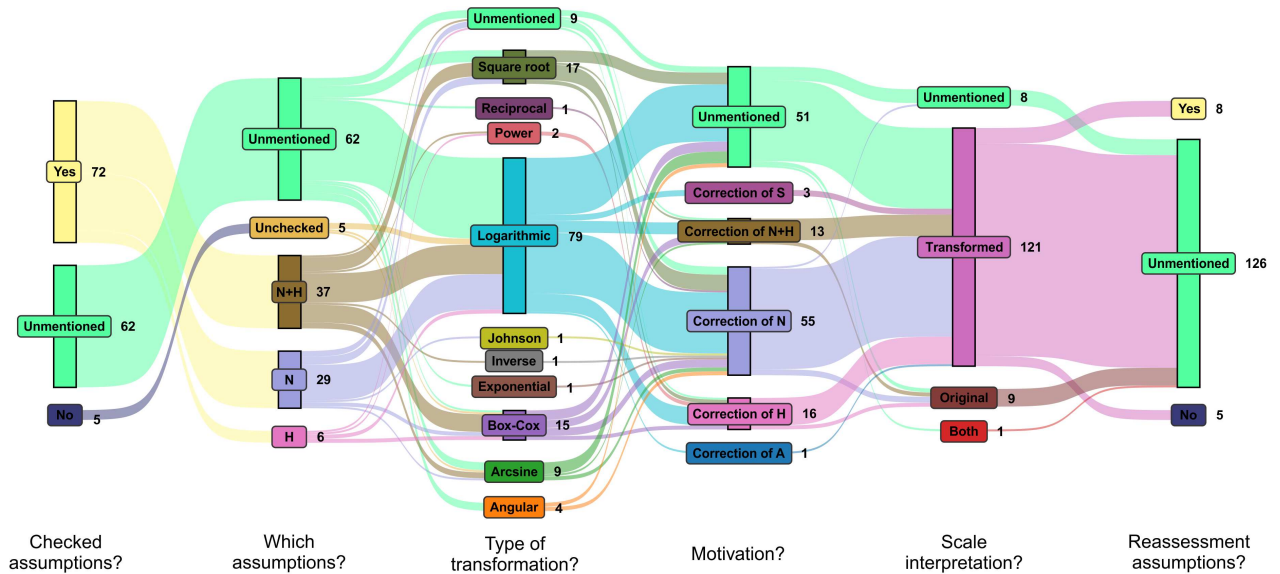


Figura 3.5: Diagrama de Sankey para ciência dos solos, ilustrando as relações entre as transformações utilizadas, a verificação dos pressupostos (N para normalidade, H para homogeneidade de variâncias, I para independência, N+H para normalidade e homogeneidade de variâncias, N+I para normalidade e independência, H+I para homogeneidade e independência, e N+H+I para normalidade, homogeneidade de variâncias e independência), a motivação por trás do uso dessas transformações, a escala de interpretação e a reavaliação dos pressupostos.



Embora os diagramas de Sankey expressem as particularidades de cada área, os mesmos problemas observados por Nascimento *et al.* (2024) também foram encontrados neste estudo. Em ambas as pesquisas, foi verificado que, em todas as variáveis analisadas, houve casos em que os procedimentos avaliados não foram mencionados ou não foram descritos de forma completa. Por exemplo, nas quatro subáreas analisadas, a maioria dos artigos verificou pelo menos um dos pressupostos da ANOVA, porém ainda foram identificados pesquisadores que, ou não verificaram, ou não reportaram a verificação desses pressupostos.

Na área de entomologia, a maioria dos estudos verificou apenas a heterogeneidade de variâncias, o que é válido, considerando que a ANOVA é robusta a desvios de normalidade (Sawyer, 2009). Já nas áreas de fitotecnia e solos, muitos artigos não mencionaram a verificação dos pressupostos, reportando apenas que realizaram transformações sem informar o pressuposto não atendido. Nos estudos de florestal, a quantidade de artigos que verificaram normalidade e heterogeneidade de variâncias foi semelhante àqueles que não mencionaram a verificação, o que reflete a diversidade de abordagens metodológicas adotadas pelos pesquisadores nessa área.

Quanto às transformações de dados utilizadas, houve uma variedade nas amostras, mas as que se destacaram foram as transformações logarítmica, arco seno, raiz quadrada e Box-Cox. Para essas transformações, realizamos um estudo de simulação, apresentado na Seção 3, a fim de avaliar sua capacidade de corrigir a não normalidade e heterogeneidade para a validação da ANOVA. A maior parte dos estudos interpretou os resultados na escala transformada. No entanto, em todas as áreas, houve casos em que a interpretação foi feita na escala original, o que constitui um erro

metodológico, já que, após a transformação dos dados, a interpretação deve ser conduzida na escala transformada, e não na original (Osborne, 2010).

Assim como em Nascimento *et al.* (2024), observamos que muitos estudos da presente revisão sistemática não documentaram diversas informações relevantes, conforme ilustrado pelas categorias “unmentioned” nas Figuras 3.2 a 3.5. Por isso, elaboramos diretrizes para o uso adequado de transformações de dados, apresentadas na Seção 3.4, que orientam os pesquisadores a evitar erros e equívocos no uso dessa técnica.

### 3.3 Estudo de simulação

O estudo de simulação foi conduzido com o objetivo de avaliar o desempenho de diferentes transformações de dados na correção de violações dos pressupostos da ANOVA, em particular a normalidade e a homogeneidade das variâncias dos resíduos. Sabendo que a ANOVA é amplamente utilizada em ciências agrárias, especialmente em experimentos que envolvem a comparação de múltiplos grupos, a correta aplicação de transformações de dados é fundamental para garantir a validade estatística dos resultados.

As transformações logarítmica, raiz quadrada, arco seno e Box-Cox foram selecionadas para o estudo, pois são as mais citadas nas literaturas avaliadas na revisão sistemática de literatura feita na Seção . O estudo foi conduzido utilizando um delineamento em blocos casualizados, com cinco tratamentos, três blocos e três repetições por tratamento em cada bloco, totalizando 45 observações. Essa estrutura experimental foi escolhida por ser representativa dos tipos de experimentos realizados nas ciências agrárias, permitindo a generalização dos resultados para contextos práticos.

#### 3.3.1 Metodologia do estudo de simulação

A metodologia foi organizada em seis etapas para aplicar sistematicamente as transformações logarítmica, raiz quadrada, arco seno e Box-Cox , avaliando seus efeitos sobre os pressupostos da ANOVA (normalidade e homogeneidade das variâncias dos resíduos), as etapas são descritas a seguir:

1. **Simulação do experimento:** Para cada simulação, os dados foram gerados com variâncias heterogêneas e resíduos não normais. Os dados de resposta foram obtidos somando valores fixos aos resíduos gerados a partir da distribuição qui-quadrado, escolhida por ser uma distribuição assimétrica. Diferentes graus de liberdade foram utilizados para simular variações nos tratamentos, refletindo a heterogeneidade esperada nos dados.
2. **Ajuste do modelo:** O modelo do delineamento em blocos casualizados foi ajustado utilizando a função `aov`, considerando os tratamentos e blocos como fatores. Posteriormente, os resíduos do modelo foram analisados para verificar a normalidade, utilizando o teste de Shapiro-Wilk, e a homogeneidade das variâncias, por meio do teste de Levene.

3. **Transformações dos dados:** Quando os pressupostos da ANOVA foram violados, ou seja, quando os testes indicaram ausência de normalidade ou heterogeneidade das variâncias residuais, foram aplicadas as transformações logarítmica, raiz quadrada, arco seno e Box-Cox aos dados.
4. **Reajuste do modelo pós-transformação:** Após a aplicação das transformações, o modelo de ANOVA foi novamente ajustado, e os pressupostos foram reavaliados com os testes de Shapiro-Wilk e Levene.
5. **Repetição e análise:** Esse processo foi repetido até que 1000 simulações fossem atingidas, nas quais os pressupostos da ANOVA fossem necessariamente violados. O sucesso de cada transformação foi avaliado com base na normalização dos resíduos, a homogeneização das variâncias e a taxa combinada (quando ambos os pressupostos foram corrigidos).
6. **Avaliação da eficácia:** Foram calculadas a frequência com que a transformação corrigiu ambos os pressupostos (normalidade e homogeneidade) e a frequência com que a transformação corrigiu apenas um dos pressupostos (normalidade ou homogeneidade).

Em [https://github.com/JhenniferN/sistematic\\_review\\_R](https://github.com/JhenniferN/sistematic_review_R), está disponível o código completo para reproduzir o estudo de simulação descrito.

### 3.3.2 Resultados e discussão do estudo de simulação

Para avaliar a eficácia das transformações logarítmica, raiz quadrada, arco seno e Box-Cox na correção dos pressupostos de normalidade e homogeneidade das variâncias residuais, calculamos as taxas de sucesso para a normalização, a homogeneização das variâncias e para ambos os pressupostos simultaneamente. Essas taxas representam a frequência com que cada transformação corrigiu os pressupostos, dividida pelo número total de simulações realizadas, conforme detalhado no Tabela 3.4.

Tabela 3.4: Taxas de sucesso das transformações logarítmica, raiz quadrada, arco seno e Box-Cox na correção dos pressupostos de normalidade e homogeneidade das variâncias residuais para 1000 simulações. As colunas indicam, respectivamente, a taxa de sucesso para a normalização dos resíduos, a homogeneização das variâncias e a correção simultânea de ambos os pressupostos (taxa combinada).

TRANSFORMAÇÃO	TAXA DE SUCESSO PARA NORMALIZAÇÃO	TAXA DE SUCESSO PARA HOMOGENEIZAÇÃO	TAXA COMBINADA PARA NORMALIZAÇÃO E HOMOGENEIZAÇÃO)
Logarítmica	0,308	0,911	0,237
Raiz quadrada	0,211	0,856	0,122
Arco seno	0,078	0,949	0,052
Box-Cox	0,525	0,971	0,504

*Fonte: Elaborado pelos autores.*

As transformações foram mais eficazes na homogeneização das variâncias do que na normalização dos resíduos. No entanto, para que uma transformação seja considerada eficaz, é necessário que ela

corrija ambos os pressupostos simultaneamente. Ao avaliar essa condição, a taxa de sucesso das quatro transformações mais utilizadas na revisão sistemática reduz-se de forma expressiva. As transformações logarítmica, raiz quadrada e arco seno corrigiram ambos os pressupostos em apenas 23,7%, 12,2% e 5,2% das simulações, respectivamente. Essas taxas são consideradas baixas porque o ideal seria que na maioria dos casos as transformações tivessem sucesso em corrigir ambos. A transformação Box-Cox corrigiu ambos os pressupostos em 50,4% das simulações, tornando-se a opção mais indicada. No entanto, mesmo com o uso dessa transformação, não há garantia de que os pressupostos serão plenamente atendidos, por isso a importância de reverificar os pressupostos pós transformação.

Quando o pesquisador precisa comparar múltiplas médias e se depara com a violação da homogeneidade das variâncias, todas as transformações corrigiram esse pressuposto em mais de 85% dos casos nas simulações realizadas. No caso de problemas relacionados apenas à normalidade, a transformação Box-Cox foi eficaz em aproximadamente 50% das simulações. No entanto, se for necessário corrigir tanto a normalidade quanto a homogeneidade das variâncias, o pesquisador deve estar ciente de que é menos provável que ambas as condições sejam atendidas simultaneamente.

Vale ressaltar que essas porcentagens refletem os resultados do estudo de simulação sob as condições consideradas e não podem ser diretamente generalizadas para todos os casos reais, embora indiquem uma tendência de maior efetividade na correção da homogeneidade das variâncias. Nesse caso, em vez de depender de transformações de dados, que podem não resolver o problema de forma adequada, é recomendável considerar o uso de métodos alternativos de análise, como técnicas não paramétricas ou modelo linear generalizado, que não exigem o cumprimento desses pressupostos.

Embora essas transformações tenham sido aplicadas principalmente para corrigir a normalidade na revisão sistemática em todas as quatro áreas, no estudo de simulação elas se mostraram mais eficientes na correção da heterogeneidade de variâncias. Isso sugere que pode haver um descompasso entre o objetivo da transformação e o problema a ser resolvido, já que algumas transformações estão sendo aplicadas para corrigir normalidade, quando são mais eficazes para heterogeneidade. Por isso, é de extrema importância a reverificação dos pressupostos após a aplicação das transformações nos dados, para garantir que a transformação tenha atingido seu objetivo específico.

### **3.4 Diretrizes para o uso de transformações de dados na validação da análise de variâncias**

Estas diretrizes têm como objetivo orientar pesquisadores na aplicação de transformações de dados para a validação dos pressupostos da análise de variância. O propósito é minimizar o uso incorreto da técnica e fornecer um procedimento claro para aqueles que optam por essa abordagem, mesmo com a ciência de que as transformações podem não resolver completamente a quebra de pressupostos.

A ANOVA é baseada em quatro pressupostos principais: aditividade, independência, homocedasticidade e normalidade dos resíduos. A violação desses pressupostos pode prejudicar a eficácia

da estimativa dos efeitos dos tratamentos, comprometer o poder do teste F e resultar em estimativas incorretas das médias dos tratamentos. A rigor, a não conformidade com esses pressupostos invalida os resultados da ANOVA (COCHRAN, 1947).

Dito isto, as transformações de dados são uma opção viável para quando os pressupostos da ANOVA não são atendidos. Embora análises não paramétricas sejam uma alternativa, o uso de transformações pode adequar os dados aos pressupostos, contribuindo para a validade dessa análise. A escolha da transformação deve ser cuidadosa, levando em conta o tipo de dados e a natureza das violações observadas. Seguindo os passos descritos, o pesquisador poderá executar a análise corretamente, minimizando a possibilidade de erros.

1. **Planejamento do experimento:** A primeira e mais crucial etapa na realização da análise de variâncias é o planejamento adequado do experimento. Embora os experimentos possam variar de uma pesquisa para outra, todos são regidos por três princípios básicos da experimentação: repetição, aleatorização e controle local (Cecon et al., 2012). Um planejamento experimental adequado assegura que as variáveis não controladas sejam atribuídas à aleatoriedade, resultando em informações mais precisas sobre os efeitos dos fatores em estudo (Ribeiro Júnior, 2013). Para um estudo mais aprofundado, a obra de Wu e Hamada (2021) oferece uma visão detalhada sobre técnicas de planejamento experimental.
2. **Verificação dos pressupostos:** Após coletar os dados, é essencial verificar os pressupostos da análise de variâncias. Este passo inclui a avaliação da normalidade dos resíduos e da homogeneidade das variâncias.
  - (a) **Análise preliminar dos dados:** Inicie a inspeção visual dos dados por meio de histogramas e gráficos Q-Q plot para verificar a independência dos resíduos e identificar possíveis assimetrias.

Tabela 3.5: Código em R e Python para a construção de histograma e Q-Q plot.

CÓDIGO	R	PYTHON
Histograma	<code>hist(residuos)</code>	<code>plt.hist(residuos, bins=20)</code>
Q-Q Plot	<code>qqnorm(residuos)</code>	<code>stats.probplot(residuos, dist="norm", plot=plt)</code>

*Fonte: Elaborado pelos autores.*

- (b) **Testes para normalidade:** Em seguida, é necessário realizar testes formais de normalidade dos resíduos. Entre os testes existentes, podem ser utilizados o Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling e D'Agostino. Esses testes podem ser realizados em R, Python ou outros softwares estatísticos.

Tabela 3.6: Código em R e Python para testes de normalidade residual.

TESTE	R	PYTHON
Shapiro-Wilk	<code>shapiro.test(dados)</code>	<code>stats.shapiro(residuos)</code>
Kolmogorov-Smirnov	<code>ks.test(residuos, "pnorm")</code>	<code>stats.kstest(residuos, 'norm')</code>
Anderson-Darling	<code>ad.test(residuos)</code>	<code>stats.anderson(residuos, dist='norm')</code>
D'Agostino	<code>agostino.test(residuos)</code>	<code>stats.normaltest(residuos)</code>

*Fonte: Elaborado pelos autores.*

- (c) **Testes para homogeneidade de variâncias:** Além da normalidade, é fundamental avaliar a homogeneidade das variâncias entre os grupos. Para isso, podem ser utilizados testes como Levene, Bartlett, Cochran e Hartley. Esses testes podem ser realizados em R (R Core Team, 2024) ou Python (Python Software Foundation, 2024).

Tabela 3.7: Código em R e Python para testes de homogeneidade residual.

TESTE	R	PYTHON
Levene	<code>leveneTest(residuos grupo)</code>	<code>levene(residuos_grupo1, residuos_grupo2)</code>
Bartlett	<code>bartlett.test(residuos grupo)</code>	<code>bartlett(residuos_grupo1, residuos_grupo2)</code>
Cochran	<code>CochranTest(residuos)</code>	<i>Não implementado em bibliotecas padrão do Python</i>
Hartley	<code>HartleyTest(residuos grupo)</code>	<i>Não implementado em bibliotecas padrão do Python</i>

*Fonte: Elaborado pelos autores.*

3. **Escolha e aplicação da transformação:** Se um ou ambos os pressupostos (normalidade ou homogeneidade de variâncias) forem rejeitados, uma solução possível é o uso de transformações de dados. A escolha da transformação mais adequada depende da natureza dos dados e do pressuposto que precisa ser corrigido. As transformações indicadas para cada caso específico estão descritas a seguir:

- (a) **Transformação logarítmica:** Indicada para dados assimétricos e para corrigir heterogeneidade de variâncias (Assis *et al.*, 2023).
- (b) **Transformação raiz quadrada:** Eficaz para dados de contagem ou variáveis discretas com assimetria, também indicada para a correção de heterogeneidade de variâncias (Assis *et al.*, 2023).
- (c) **Transformação arco seno:** Indicada para proporções e porcentagens, especialmente com valores próximos aos limites 0 ou 1, para a correção de heterogeneidade de variâncias e não normalidade (Mohr *et al.*, 2021).
- (d) **Transformação Box-Cox :** Indicada para dados que seguem distribuição assimétrica, essa transformação aproxima os dados da distribuição normal, sendo especialmente útil

para a correção de heterogeneidade de variâncias e aproximação à normalidade (Azevedo *et al.*, 2016).

4. **Reverificação dos pressupostos:** Após aplicar a transformação, refaça os testes de normalidade e homogeneidade para garantir que os pressupostos foram corrigidos adequadamente. Se a transformação não for eficaz, selecione outra transformação e repita o processo. É importante lembrar que a verificação dos pressupostos deve ser realizada sempre após cada nova transformação.
5. **Escala de interpretação dos resultados:** Após a aplicação da transformação, a interpretação dos resultados deve ser feita na nova escala transformada. É essencial que os dados transformados sejam compreensíveis e façam sentido dentro da escala em que foram analisados, garantindo que as conclusões refletem adequadamente o comportamento dos dados transformados.
6. **Análises alternativas:** Se, após as tentativas de correção, os pressupostos ainda não forem satisfeitos, considere utilizar uma das análises alternativas a seguir:
  - (a) **Teste F de Welch:** recomendado para situações com variâncias desiguais (Mohr *et al.*, 2021).
  - (b) **Kruskal-Wallis:** indicado para dados com distribuições não normais (Mohr *et al.*, 2021).
  - (c) **Modelo linear generalizado:** recomendado para lidar com dados que não atendem aos pressupostos da ANOVA tradicional (Agresti, 2015).
  - (d) **Modelo linear generalizado com link logarítmico:** que pode ser utilizado como alternativa à transformação logarítmica, permitindo a obtenção de estimativas não viesadas dos parâmetros (St-Pierre *et al.*, 2018).
7. **Tutorial para reportar a análise de variância em artigos científicos:** Para garantir que nenhum passo da análise seja omitido ou esquecido, recomenda-se utilizar o seguinte modelo de parágrafo para reportar a sua análise de variância em artigos científicos:
  - (a) **Caso em que não foi necessário o uso de transformações de dados:** A análise de variância (ANOVA) foi realizada para avaliar os efeitos dos [escreva os tratamentos da sua análise], utilizando o delineamento experimental [indique o delineamento utilizado no seu experimento e seus detalhes]. Antes da realização da ANOVA, os pressupostos de normalidade e homogeneidade das variâncias foram verificados e atendidos de acordo com os respectivos testes [indique o teste usado para verificar a normalidade] e [indique o teste usado para verificar a homogeneidade das variâncias].
  - (b) **Caso em que transformações de dados foram utilizadas:** A análise de variância (ANOVA) foi realizada para avaliar os efeitos dos [escreva os tratamentos da sua análise], utilizando o delineamento experimental [indique o delineamento utilizado no seu experimento e seus detalhes]. Antes da realização da ANOVA, os pressupostos de normalidade

e homogeneidade das variâncias foram verificados de acordo com os respectivos testes [indique o teste usado para verificar a normalidade] e [indique o teste usado para verificar a homogeneidade das variâncias]. No presente estudo, a transformação [indicar a transformação utilizada] foi aplicada às variáveis [nome das variáveis] para corrigir a violação de [indique o(s) pressuposto(s) violado(s)]. Após a transformação, os pressupostos foram reavaliados e considerados atendidos pelos mesmos testes. Os resultados da ANOVA foram interpretados utilizando a escala transformada para as variáveis transformadas e a escala original para as variáveis que não necessitaram de transformação.

### 3.5 Conclusão

A revisão sistemática de literatura nos proporcionou um panorama do uso de transformações de dados para a validação da ANOVA em artigos das áreas de entomologia, fitotecnia, florestal e solos. As técnicas aplicadas e as deficiências nos relatos científicos identificados nesses subcampos, foram similares àquelas observadas na revisão sistemática feita para a área de ciências agrárias. Constatamos que as transformações logarítmica, raiz quadrada, arco seno e Box Cox foram as mais utilizadas nas subáreas avaliadas. Verificamos no estudo de simulação que as taxas de sucesso para corrigir simultaneamente a não normalidade e a heterogeneidade residual foram baixas, o que sugere cautela no uso dessas transformações, especialmente devido às potenciais consequências negativas que a técnica pode ter na interpretabilidade dos dados. Também apresentamos alternativas de análises estatísticas não paramétricas para substituir a ANOVA, quando os pressupostos não são atendidos. No entanto, se o pesquisador optar pelo uso de transformações de dados para a validação dos pressupostos da ANOVA paramétrica, disponibilizamos diretrizes para o uso correto dessas transformações, com o objetivo de evitar erros e equívocos na execução da análise. Além disso, destacamos os pontos essenciais que devem ser reportados em futuros artigos, de modo a prevenir a repetição das falhas observadas por esta revisão. Por fim, a utilização do Zotero para o download em massa dos artigos na fase de identificação reduziu expressivamente o esforço manual e computacional envolvido nessa etapa.

# Referências Bibliográficas

AGRESTI, A. *Foundations of linear and generalized linear models*. New Jersey: John Wiley & Sons, 2015. 471 p.

ASSIS, J. P.; DE LIMA, I. R. P.; FRANÇA, J. A.; DE SOUSA, R. P.; LINHARES, P. C. F.; CUSTÓDIO, T. N.; *et al.* *Transformação de dados aplicada à estatística*. Mato Grosso: Pantanal Editora, 2023. 133 p.

AZEVEDO, A. M.; ANDRADE JÚNIOR, V. C.; FERNANDES, J. S. C. Transformação Box-Cox na homocedasticidade e normalidade uni e multivariada em experimentos de batata-doce. *Horticultura Brasileira*, v. 34, n. 1, p. 93–101, 2016.

CECON, P. R.; SILVA, F. F.; NASCIMENTO, M. *Métodos estatísticos*. 1<sup>a</sup> ed. Viçosa: Editora UFV, 2012. 377 p.

COCHRAN, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, n. 1, p. 22–38, 1947.

FILLIBEN, J. J.; HECKERT, A. Exploratory data analysis. In: ZEY, C. *Engineering statistics handbook*. Maryland: NIST, 2003. p. 1–636.

MOHER, D.; LIBERATI, A.; TETZLAFF, J.; ALTMAN, D. G. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *British Medical Journal*, n. 6, p. 332–336, 2009.

MOHR, D. L.; WILSON, W. J.; FREUND, R. J. *Statistical methods*. 4<sup>a</sup> ed. Chennai: Academic Press, 2021. 720 p.

NASCIMENTO, J. S.; SILVA, J.; BERNARDES, R. C.; COSTA, G.; EMILIANO, P. C. Statistical data transformation in agrarian sciences for variance analysis: a systematic review. *F1000Research*, n. 13, p. 1–12, 2024.

OSBORNE, J. W. Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research, and Evaluation*, n. 12, p. 1–9, 2010.

PYTHON Software Foundation. *Python Language Reference*, version 3.13. 2024. Disponível em: <https://www.python.org/>.

R CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. Disponível em: <https://www.r-project.org/>.

RAMACHANDRAN, K. M.; TSOKOS, C. P. *Mathematical statistics with applications in R*. 3<sup>a</sup> ed. London: Academic Press, 2021. 688 p.

RIBEIRO JÚNIOR, J. I. *Métodos estatísticos aplicados ao controle da qualidade*. 1<sup>a</sup> ed. Viçosa: Editora UFV, 2013. 254 p.

SAWYER, S. F. Analysis of variance: The fundamental concepts. *Journal of Manual & Manipulative Therapy*, n. 17, p. 27–38, 2009.

ST-PIERRE, A. P.; SHIKON, V.; SCHNEIDER, D. C. Count data in biology - Data transformation or model reformation? *Ecology and Evolution*, n. 8, p. 3077–3085, 2018.

WU, C. F. J.; HAMADA, M. S. *Experiments: planning, analysis and optimization*. 3<sup>a</sup> ed. Nova Jersey: Wiley, 2021. 649 p.

# CONCLUSÃO GERAL

O principal objetivo do presente estudo foi investigar o desempenho de diferentes transformações de dados para validação na ANOVA dentro das ciências agrárias. Como havia muitas transformações com essa finalidade na literatura, consideramos primeiramente realizar uma revisão sistemática de trabalhos que utilizavam essas técnicas, para identificar as transformações mais usadas e, posteriormente, avaliar seu desempenho. Um dos primeiros desafios enfrentados foi o acesso às informações de interesse, já que o reporte sobre o uso de transformações de dados costuma estar na seção de metodologia dos artigos, e não nos títulos, resumos ou palavras-chave, que são tradicionalmente as seções utilizadas nas fases de identificação e triagem da revisão sistemática. Diante desse desafio, decidimos que na fase de identificação seriam localizados nas bases os trabalhos que reportaram o uso da análise de variância, esses trabalhos seriam baixados e executada a triagem diretamente pela busca de termos relacionados ao uso de transformações de dados no texto completo dos artigos em PDF.

Essa abordagem levou a um segundo desafio: baixar em massa os artigos das bases Scopus e Web of Science, já que fazer essa tarefa manualmente seria inviável devido a grande quantidade de artigos encontrados na fase de identificação. Diante desse segundo desafio, uma automação em Python foi desenvolvida para simular um humano interagindo com a tela do computador para realizar o *download* dos trabalhos. Essa estratégia foi utilizada na revisão sistemática do primeiro artigo da tese (presente no Capítulo 2), mas exigiu supervisão, pois a automação era suscetível a erros. Para o segundo artigo (Capítulo 3), utilizamos o Zotero, que permite o *download* em massa a partir de uma lista de DOIs, economizando tempo e esforço, tanto pessoal quanto computacional, nessa etapa.

Após o *download* dos trabalhos, a triagem foi realizada por meio da busca de termos que indicassem o uso de transformações de dados nos PDFs, com o auxílio de um código elaborado em Python. Desse processo, foi desenvolvido um aplicativo *web* com o objetivo de facilitar a triagem para pesquisadores sem familiaridade com Python. Ou seja, na fase de identificação buscamos por trabalhos que realizaram análise de variância e na triagem buscamos nos PDFs termos relacionados ao uso de transformações de dados, filtrando dessa forma, os trabalhos mais prováveis de terem utilizado transformações para a validação da anova e analisamos como os pesquisadores aplicam a técnica.

As revisões sistemáticas ofereceram um panorama do uso de transformações de dados para a validação da ANOVA em artigos de ciências agrárias e nos subcampos de entomologia, fitotecnia,

florestal e solos. As práticas e as lacunas identificadas nesses subcampos foram consistentes com aquelas observadas para a área de ciências agrárias em geral. Constatamos que as transformações logarítmica, raiz quadrada, arco seno e Box-Cox foram as mais comuns e que a maioria dos trabalhos analisados apresentaram falhas na execução e no reporte dessa análise. Com essas constatações, realizamos finalmente o estudo de simulação para avaliar as taxas de correção da não-normalidade e heterogeneidade de variâncias para as quatro transformações mais frequentes e elaboramos um guia prático, com um passo a passo, para orientar a aplicação e o reporte adequado dessa técnica em trabalhos científicos.

O estudo de simulação mostrou que para o cenário avaliado, a transformação Box-Cox apresentou desempenho mais promissor, com as maiores taxas de correção da não normalidade (52,5%), da heterogeneidade (97,1%) e de ambos os pressupostos simultaneamente (50,5%). As outras três transformações apresentaram taxas de correção da heterogeneidade acima de 85% e para correção da não normalidade e dos pressupostos simultâneos elas não apresentaram um bom desempenho. O que sugere que o uso das transformações deve ser cauteloso, considerando o risco adicional de comprometer a interpretabilidade dos dados.

O guia para orientar a aplicação e o reporte adequado dessa técnica em trabalhos científicos minimizam os erros e equívocos de pesquisadores que o seguirem. Também destacamos os elementos essenciais que devem ser reportados em futuros estudos, com o objetivo de minimizar as falhas identificadas nas revisões feitas. Além disso, apresentamos alternativas de análises estatísticas não paramétricas como opções à ANOVA para situações em que seus pressupostos não são atendidos.

Por fim, nosso método de triagem oferece uma solução prática e acessível para pesquisadores que desejam fazer revisão sistemática de literatura e que as informações desejadas se encontram na metodologia dos trabalhos que precisam ser triados. Para pesquisadores que não estão à vontade com a interface do Python, o aplicativo *web* desenvolvido expande o acesso a nossa metodologia.

## Anexo

### Definições de conceitos estatísticos

- **Análise não paramétrica:** Conjunto de métodos estatísticos que não pressupõem distribuição específica para os dados, como normalidade. É útil para dados de natureza ordinal ou quando os pressupostos de testes paramétricos são violados. (Conover, 1999)
- **Análise paramétrica: vantagens e desvantagens:** Métodos baseados em pressupostos como normalidade e homogeneidade de variâncias. Vantagens: maior poder estatístico quando os pressupostos são atendidos. Desvantagens: sensibilidade a desvios desses pressupostos. (Montgomery, 2017)
- **ANOVA (Análise de Variância):** Método estatístico utilizado para comparar médias de três ou mais grupos e verificar se pelo menos um difere significativamente dos outros. (Cochran, 1947)
- **Amplitude:** Diferença entre o maior e o menor valor em um conjunto de dados, indicando a dispersão. (Moore & McCabe, 2006)
- **Caudas largas:** Distribuições que apresentam valores extremos com maior frequência do que a distribuição normal. Exemplo clássico: distribuição t de Student. (Lange, 2010)
- **Desvio padrão:** Medida de dispersão que indica a quantidade de variação ou dispersão de um conjunto de dados em relação à média. Fórmula:  $\sqrt{\sum(x_i - \bar{x})^2/n}$ . (Agresti, 2015)
- **Erro de tipo I:** Probabilidade de rejeitar a hipótese nula quando ela é verdadeira ( $\alpha$ ). Representa um falso positivo. (Montgomery, 2017)
- **Erro de tipo II:** Probabilidade de não rejeitar a hipótese nula quando ela é falsa ( $\beta$ ). Representa um falso negativo. (Montgomery, 2017)
- **Heterocedasticidade:** Condição onde a variância dos erros em um modelo de regressão não é constante. (Ramachandran, 2021)
- **Homogeneidade:** Propriedade onde as variâncias entre diferentes grupos ou condições são iguais. (Box & Cox, 1964)
- **Machine Learning:** Campo da inteligência artificial focado em criar sistemas que aprendem e melhoram automaticamente a partir de dados. (Hastie *et al.*, 2009)
- **Média:** Valor central calculado como a soma dos valores dividida pelo número de observações. (Moore & McCabe, 2006)
- **Mediana:** Valor que divide os dados ordenados em duas partes iguais. Insensível a valores extremos. (Moore & McCabe, 2006)

- **Moda:** Valor mais frequente em um conjunto de dados. (Moore & McCabe, 2006)
- **Outliers:** Observações que diferem substancialmente das demais, podendo distorcer resultados estatísticos. (Aggarwal, 2017)
- **Pressupostos da ANOVA:** Incluem normalidade dos resíduos, homogeneidade de variâncias e independência das observações. (Cochran, 1947)
- **Resíduos:** Diferença entre os valores observados e os valores ajustados por um modelo. (Ramachandran, 2021)
- **Revisão de literatura:** Síntese narrativa de estudos existentes em uma área para identificar tendências e lacunas. (Booth *et al.*, 2016)
- **Revisão sistemática:** Método rigoroso e replicável para reunir e analisar evidências de estudos relevantes para responder a uma questão de pesquisa. (Higgins & Green, 2011)
- **Robustez de um teste:** Capacidade de um teste estatístico produzir resultados confiáveis mesmo quando os pressupostos são levemente violados. (Montgomery, 2017)
- **Soma de quadrados:** Medida usada em análise de variância para particionar a variabilidade total dos dados em componentes atribuíveis a diferentes fontes. Fórmula geral:  $\sum(x_i - \bar{x})^2$ . (Montgomery, 2017)
- **Testes independentes:** Testes estatísticos aplicados a amostras independentes, como o teste t para amostras independentes. (Agresti, 2015)
- **Variância:** Medida de dispersão que expressa a média dos quadrados das diferenças dos valores em relação à média. Fórmula:  $\sigma^2 = \sum(x_i - \bar{x})^2/n$ . (Agresti, 2015)
- **Variâncias homocedásticas:** Condição onde as variâncias entre grupos são aproximadamente iguais, fundamental para ANOVA. (Cochran, 1947)
- **Variâncias homogêneas:** Sinônimo de variâncias homocedásticas, indicando igualdade de variâncias entre grupos. (Box & Cox, 1964)

# Referências Bibliográficas

- AGGARWAL, C. C. *Outlier analysis*. 2<sup>a</sup> ed. Nova York: Springer, 2017. 466 p.
- AGRESTI, A. *Foundations of linear and generalized linear models*. New Jersey: John Wiley & Sons, 2015. 471 p.
- BOOTH, A.; SUTTON, A.; PAPAIOANNOU, D. *Systematic approaches to a successful literature review*. Londres: Sage, 2016. 336 p.
- BOX, G. E. P.; COX, D. R. An analysis of transformations. *Journal of the Royal Statistical Society*, v. 26, n. 2, p. 211–243, 1964.
- COCHRAN, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, n. 1, p. 22–38, 1947.
- CONOVER, W. J. *Practical nonparametric statistics*. 3<sup>a</sup> ed. Nova York: John Wiley & Sons, 1999. 584 p.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning: data mining, inference, and prediction*. 2<sup>a</sup> ed. Nova York: Springer, 2009. 745 p.
- HIGGINS, J. P.; GREEN, S. *Cochrane handbook for systematic reviews of interventions*. 5<sup>a</sup> ed. Londres: Wiley, 2011. 672 p.
- LANGE, K. *Applied probability*. 2<sup>a</sup> ed. Nova York: Springer, 2010. 564 p.
- MONTGOMERY, D. C. *Design and analysis of experiments*. 9<sup>a</sup> ed. Nova York: John Wiley & Sons, 2017. 752 p.
- MOORE, D. S.; MCCABE, G. P. *Introduction to the practice of statistics*. 6<sup>a</sup> ed. Nova York: W. H. Freeman, 2006. 828 p.
- RAMACHANDRAN, K. M.; TSOKOS, C. P. *Mathematical statistics with applications in R*. 3<sup>a</sup> ed. Londres: Academic Press, 2021. 688 p.
- ST-PIERRE, A. P.; SHIKON, V.; SCHNEIDER, D. C. Count data in biology - Data transformation or model reformation? *Ecology and Evolution*, v. 8, n. 6, p. 3077–3085, 2018.