JANILSON PINHEIRO DE ASSIS
ISAAC REINALDO PINHEIRO DE LIMA
JOELMA DE ASSIS FRANÇA
ROBERTO PEQUENO DE SOUSA
PAULO CÉSAR FERREIRA LINHARES
TELDE NATEL CUSTÓDIO
ROBSON PEQUENO DE SOUSA
WALTER MARTINS RODRIGUES

TRANSFORMAÇÃO DE DADOS APLICADA À ESTATÍSTICA

$$Y = Log\ x$$

$$f_{\lambda}(x) = \frac{x^{\lambda} - 1}{\lambda}$$

$$T = \sqrt{X}$$

$$T = \sqrt{X + 0,50}$$

$$T = \sqrt{X + 1}$$

$$T = \sqrt[3]{X}$$

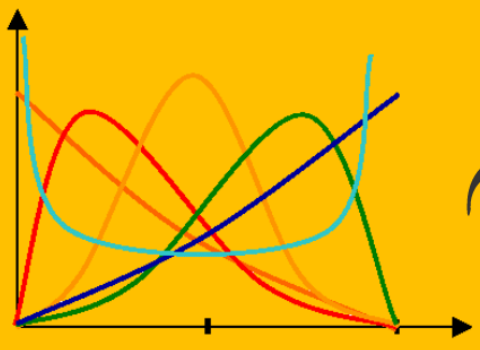
$$W = \sqrt{X + K}$$

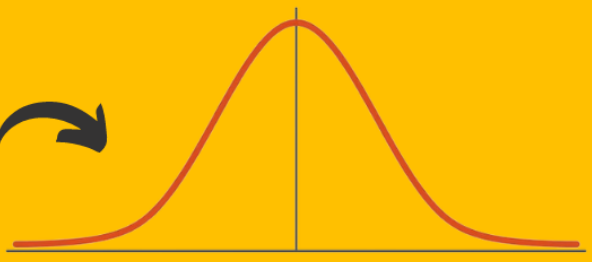
$$Z = Arc\ Sen\sqrt{X\%}$$

$$Log\,Y = a + b\,Log X$$

$$X_i = \frac{1}{x_i}$$

$$X_i = \frac{1}{x_i^2}$$





Pantanal Editora
2023

JANILSON PINHEIRO DE ASSIS
ISAAC REINALDO PINHEIRO DE LIMA
JOELMA DE ASSIS FRANÇA
ROBERTO PEQUENO DE SOUSA
PAULO CÉSAR FERREIRA LINHARES
TELDE NATEL CUSTÓDIO
ROBSON PEQUENO DE SOUSA
WALTER MARTINS RODRIGUES

TRANSFORMAÇÃO DE DADOS APLICADA À ESTATÍSTICA

Pantanal Editora

2023

Editor Chefe: Prof. Dr. Alan Mario Zuffo
Editores Executivos: Prof. Dr. Jorge González Aguilera e Prof. Dr. Bruno Rodrigues de Oliveira

Diagramação: A editora. **Diagramação e Arte:** A editora. **Imagens de capa e contracapa:** Canva.com. **Revisão:** O(s) autor(es), organizador(es) e a editora.

Ficha Catalográfica

Catalogação na publicação
Elaborada por Bibliotecária Janaina Ramos – CRB-8/9166

T772

Transformação de dados aplicada à estatística / Janilson Pinheiro de Assis, Isaac Reinaldo Pinheiro de Lima, Joelma de Assis França, et al. – Nova Xavantina-MT: Pantanal, 2023. 131p. ; il.

Outros autores: Roberto Pequeno de Sousa, Paulo César Ferreira Linhares, Telde Natel Custódio, Robson Pequeno de Sousa, Walter Martins Rodrigues.

Livro em PDF

ISBN 978-65-81460-95-2
DOI https://doi.org/10.46420/9786581460952

1. Estatística. I. Assis, Janilson Pinheiro de. II. Lima, Isaac Reinaldo Pinheiro de. III. França, Joelma de Assis. IV. Título.

CDD 310

Índice para catálogo sistemático

I. Estatística

Pantanal Editora

Rua Abaete, 83, Sala B, Centro. CEP: 78690-000.
Nova Xavantina – Mato Grosso – Brasil.
Telefone (66) 99682-4165 (Whatsapp).
https://www.editorapantanal.com.br
contato@editorapantanal.com.br

APRESENTAÇÃO

O assunto transformação de dados é pouco explorado na literatura, apesar de ser extremamente importante para auxiliar o pesquisador na aplicação correta de ferramentas de análise estatística dente elas a chamada análise de variância (ANAVA) aplicada em seus dados muitas vezes obtidos através de levantamentos amostrais ou em pesquisas planejadas, bem como no ajuste de modelos na regressão linear ou não linear, em técnicas univariada ou multivariada, na análise de sobrevivência, na aderência de distribuições teóricas ou especiais de probabilidade a conjuntos de variáveis aleatórias, etc.. Sendo assim este livro visa preencher uma lacuna na literatura especializada sobre o assunto, principalmente em textos escritos em língua portuguesa. Esperamos assim que o leitor encontre motivação para explorar este assunto tão palpitante na análise estatística de dados e resolva seus problemas e dúvidas sobre o assunto. Finalmente esta obra tem como objetivo fornecer a teoria dos mais diferentes métodos de transformações de dados, bem como mostrar através de exemplo numérico prático como este mecanismo de mudança nas variáveis afeta o conjunto de valores analisados, dando condições para o pesquisador aplicar técnicas de inferência estatística com validade ou obedecendo as pressuposições exigidas na construção de intervalos de confiança, bem como na aplicação de testes de hipóteses e no ajuste de equações em modelos matemáticos usuais na pesquisa científica conduzida nas ciências físicas, biológicas, sociais, na indústria, na genética, etc. Vale salientar que quaisquer erros que porventura sejam encontrados neste livro são de inteira responsabilidade dos autores.

Os autores
Mossoró, RN, Brasil, agosto de 2023

PREFÁCIO

A transformação de variáveis no planejamento experimental e nos levantamentos amostrais nas diversas áreas do conhecimento humano, é fator fundamental para proporcionar ao pesquisador o uso de técnicas de inferência estatística clássica tais como os testes de hipóteses e os intervalos de confiança, bem como o uso de técnicas econométricas e de regressão e correlação linear e não linear simples e múltipla, ampliando a amplitude do uso de tais técnicas e flexibilizando as analises estatísticas, aumentando o rol de alcance das conclusões obtidas, e do sucesso e êxito em suas conclusões racionais e elegantes, sem a necessidade de uso de técnicas mais complexas e sem a obrigação da adoção de pressuposições muitas vezes abstratas e difíceis de serem adotadas ou verificadas na prática da pesquisa científica.

Quando alguma das pressuposições da análise de variância paramétrica para o emprego da estatística clássica paramétrica, tanto unidimensional como multidimensional, tais como a normalidade da distribuição dos erros, a homogeneidade das variâncias, e a aditividade dos efeitos dos fatores de variação, não for verificada na prática da pesquisa científica e na estatística experimental ou levantamentos amostrais, o pesquisador pode ainda tentar o uso de técnicas de transformação dos dados, antes de adotar o uso de metodologia baseada na estatística não-paramétrica. É um meio alternativo e de suporte de análise que sempre deve ser tentado pelo investigador cientifico, porque a estatística paramétrica é sabidamente mais poderosa e de maior alcance que a estatística não-paramétrica. Na realidade a estatística não paramétrica somente foi desenvolvida como um recurso alternativo e de suporte complementar, destinado a cobrir a lacuna do uso de testes estatísticos nos casos em que alguma restrição impedia o uso da estatística paramétrica, ou quando a própria natureza das variáveis, muitas vezes categóricas ou qualitativas impedia o suo de tais técnicas.

As transformações mais comumente utilizadas de forma direta nos dados de observação, são a as transformações logarítmica, a logarítmica dos (dados +1), a raiz quadrada dos dados , a raiz quadrada dos (dados + 1, ou mais 1/2), a raiz cúbica dos dados, a transformação angular, a transformação hiperbólica de primeiro grau (ou o inverso dos dados) ou hiperbólica de segundo grau, a transformação percentual, e a transformação em valores de z, a transformação de Fisher, as de Box e Cox, etc. citadas em textos de diversos autores quando se referem aos testes para verificar a normalidade da distribuição dos erros amostrais.

Mesmo que o uso de tais transformações não seja adequado, o pesquisador pode adotar o uso dos modelos lineares generalizados os quais são o resumo e extensão notáveis de modelos de regressão familiares, como os modelos lineares. Essa técnica estatística é devida originalmente a Nelder and Wedderburn (1972) essa modelagem é especialmente adequada a aplicação de análises em dados de contagem, incluindo tabelas de contingência. Estes modelos são constituídos basicamente de Um

componente aleatório, onde neste componente especifica-se a distribuição condicional da variável resposta Y_i, para a i-ésima variável de n observações amostradas independentemente, dados os valores das variáveis explicativas no modelo, e também de um preditor linear que é uma função linear de regressores, sendo que como no modelo linear, os regressores $X_{\{ij\}}$ são funções pré-especificadas das variáveis explicativas e, portanto, podem incluir variáveis explicativas quantitativas, transformações de variáveis explicativas quantitativas, regressores polinomiais, regressores dummy, interações e outras, por isso que, uma das vantagens dos modelos lineares generalizados é que a estrutura do preditor linear é a estrutura familiar de um modelo linear, e finalmente por uma Uma função de ligação linearizante suave e invertível, que transforma a esperança da variável resposta, $\mu_i = E(Y_i)$, em um preditor linear. Sendo assim, o modelo linear generalizado pode ser pensado como um modelo linear para uma transformação da resposta esperada ou como um modelo de regressão não linear para a resposta.

Os dados que devem ser tratados com o uso do emprego dos modelos lineares generalizados são aqueles cujas distribuições pertencem a família exponencial, tais como gaussiana, binomial, Poisson, Gama e Gaussiana inversa.

Transformação de dados aplicado à estatística foi escrito com o objetivo de ser um livro texto inédito em disciplinas de Estatística para Cursos de Ciências Agrárias, Sociais e Humanas. A motivação para escrever este texto surgiu quando do início na década de 1980 como professor de estatística e estatística experimental na Escola Superior de Agricultura de Mossoró, e na Universidade Federal Rural do Semi-Árido, Bem como da experiência com o tratamento de dados por parte dos autores em áreas como educação, tecnologia da informação, ciências agrárias, ciências biológicas, ciências sociais e ciências humanas vividas pelos autores, e também através da curiosidade de alunos bolsistas quando se deparavam diante dos desafios das diferentes análises estatísticas de dados experimentais e oriundos de levantamentos amostrais em pesquisas nas mais diferentes áreas de atuação, e mostrando-lhes a necessidade e exigência de verificação das pressuposições no uso de técnicas estatísticas inferenciais. A motivação e o aumento do nível de conhecimento dos alunos, bem como a satisfação em prestar serviços de assessoramentos a muitos pesquisadores foram cada vez mais aumentando essa demanda, o que resultou na tomada de decisão do lançamento desta obra em forma de livro texto.

Este texto apresenta uma introdução à técnicas estatísticas de transformação de dados ou variáveis aleatórias de forma bastante aprofundada e orientada com enfoque didático, e com foco em uma leitura que deixe o leitor com uma visão ampla e profunda que lhe proporcione um assessoramento razoável no uso das técnicas de transformação de dados no dia a dia de suas atividades profissionais, na interpretação dos resultados, nas discussões científicas e na obtenção de conclusões válidas e abrangentes, fruto das análises obtidas em suas pesquisas, acompanhada ainda de uma orientação teórica de como planejar e conduzir uma pesquisa quantitativa e/ou qualitativa.

O livro tem como início uma abordagem geral das técnicas de transformação de dados e apresenta algumas ideias básicas e avançadas sobre o efeito nos resultados e discussão das análises de dados após

aplicação das transformações, bem como orientar no início do planejamento de uma pesquisa planejada em função de experimentos aleatórios com tratamentos e em levantamentos amostrais probabilísticos. O Capítulo final enfoca a questão de generalizar resultados da amostra para a população, através de exemplos numéricos práticos Finalmente, o último apresenta as considerações finais e conclusões deste importante procedimento estatístico para avaliar as exigências das pressuposições das análises de variância a priori, de regressão dentre outras, e assim como aplicar técnicas de inferência estatística clássica paramétrica para construir e utilizar modelos voltados para dezenas de tipos ou natureza de pesquisas cientificas, capacitando o leitor, professor, estudante, técnico ou qualquer outro indivíduo que utilize direta ou indiretamente as técnicas estatísticas aplicadas nas mais diferentes análises de dados de pesquisa científica.

Os Autores
Mossoró, RN, Brasil, agosto de 2023

SUMÁRIO

APRESENTAÇÃO **4**
PREFÁCIO **5**
CAPÍTULO 1 ..... **9**
CAPÍTULO 2 ..... **37**
CAPÍTULO 3 ..... **106**
Estimação do parâmetro λ 113
Método prático para estimar λ 114
REFERÊNCIAS **118**
APÊNDICE **126**
ÍNDICE REMISSIVO **127**
SOBRE OS AUTORES **129**

CAPÍTULO 1

Os dados oriundos de experimentos são a matéria prima para o pesquisador fazer inferências acerca dos seus resultados, no entanto muitas vezes ou nem sempre estas respostas obtidas se adequam aos requisitos pressupostos ou hipóteses básicas da chamada análise de variância paramétrica, tornando impraticável a construção de intervalos de confiança, aplicação de testes de hipótese bem como o ajuste de modelos de regressão e estimativas de correlação, sendo assim neste caso só resta duas alternativas, ou muda-se de escala, ou , seja se transformando os dados, ou muda-se de modelo, partindo-se para os testes não paramétricos, que são muito menos exigentes em condições para sua aplicação (SILVA; SILVA, 1999). Segundo Banzatto e Kronka (2006), todo delineamento experimental possui um modelo matemático e, para que se possa aplicar a análise de variância aos resultados de um experimento em um dado delineamento experimental, deve-se considerar o seu modelo matemático e aceitar algumas pressuposições básicas necessárias para a validade da estruturação desta análise de variância, bem como na obtenção de inferências estatísticas confiáveis. Uma das situações mais comuns de não verificação das pressuposições básicas na análise de variância (ANOVA) é aquela em que não existe homocedasticidade, isto é, a variância não é a mesma nos diferentes tratamentos. Isto origina o que se conhece por heterocedasticidade ou heterogeneidade dos erros, a qual pode ter origem na heterocedastidade irregular que ocorre quando certos tratamentos apresentam maior variabilidade que outros, ou ainda na heterocedasticidade regular que ocorre devido a falta de normalidade na distribuição empírica dos dados experimentais, existindo frequentemente, uma certa relação entre a média e a variância dos vários tratamentos ensaiados. Se a distribuição de probabilidade dos dados de observação for conhecida a relação dos tratamentos também o será e, os dados poderão ser transformados de tal maneira que passem a ter uma distribuição de probabilidade aproximadamente normal ou gaussiana e as médias e variâncias se tornem independentes, permitindo assim a aplicação da análise de variância na avaliação dos resultados do ensaio e a obtenção de inferências estatísticas confiáveis.

Um fator importante que antecede a análise de dados é a seleção de um método estatístico adequado que permita avaliar corretamente o comportamento dos efeitos dos tratamentos estudados e sua magnitude. Estas análises só devem ser realizadas após verificar se o modelo do delineamento experimental está correto e, se atende as pressuposições da análise de variância. Dentre estas são consideradas a distribuição dos erros dos resultados experimentais e a sua avaliação quando se afastam do conjunto de resposta. Num conjunto de dados se os erros têm distribuição assimétrica a variância residual é função da média (COCHRAN; COX 1957).

Outro fator a considerar são as observações discrepantes, as quais levam a perda na sensibilidade nos testes de significância e, deve haver critério para rejeitá-las (BUSTOS 1980, 1988, HOAGLIN et al.

1992). Na ausência de um teste estatístico clássico para avaliar a dispersão das variâncias dos tratamentos, da aditividade do modelo e da normalidade dos erros, os dados devem ser avaliados através de gráficos, estatísticas descritivas e clássicas para melhor identificar os pontos discrepantes que comprometem a significância do teste F (ZAR 1996).

Caso as observações discrepantes sejam detectadas através da análise exploratória recomenda-se a retirada deste valor, com critério, ou o uso da transformação que ermita validar os testes de significância e as estimativas dos limites de confiança (FINNEY 1960, BOX; COX 1964, BUSTOS 1988).

Por diversas vezes o pesquisador é conduzido a efetuar certas mudanças de variáveis, com o objetivo de tornar linear uma relação que não o era inicialmente ou com o fim de assegurar uma certa normalidade das distribuições ou uma certa igualdade das variâncias. Neste artigo será abordado o problema de uma forma mais sistemática, principalmente sob o aspecto de igualdade das variâncias.

Segundo Steel e Torrie (1981), a heterogeneidade dos erros em um experimento pode ser de dois tipos, a irregular e a regular. A do tipo irregular está associada ao fato de alguns tratamentos possuírem maior variabilidade que outros, sem qualquer aparente associação entre média e variância, e esta diferença pode ou não ser esperada, já que muitas vezes elas fazem parte de resposta do tratamento. Segundo Zimmermann (2004) um caso típico de diferença em variância esperada é o estudo dos efeitos de inseticidas no controle de insetos, pois a testemunha sem inseticida apresenta uma maior presença de insetos e de grande variabilidade quando comparada com os tratamentos com aplicação de inseticidas, onde a presença de insetos é bem reduzida. Casos semelhantes podem ser encontrados em experimentos com fungicidas, herbicidas, fertilizantes ou, ainda, com técnicas de irrigação. Neste caso, ou autores sugerem duas alternativas de solução, uma de eliminar da análise os tratamentos com variância muito discrepantes e outra, muito elegante, a de dividir o erro em componentes aplicáveis aos distintos tipos de comparações desejadas, uma vez que as transformações de dados, usualmente, não resolvem este problema. Já a heterogeneidade de variância do tipo regular surge quando ocorre algum tipo de falta de ajuste à normalidade, com a variabilidade dos tratamentos associada á sua média de alguma forma. De uma maneira geral, as diversas transformações de dados disponíveis estão associadas às distribuições de probabilidade e procuram tornar a média e a variância independentes.

Um fator importante que antecede a análise de dados é a seleção de um método estatístico adequado que permita avaliar corretamente o comportamento dos efeitos dos tratamentos estudados e sua magnitude. Estas análises só devem ser realizadas após verificar se o modelo do delineamento experimental está correto e, se atende as pressuposições da análise de variância. Dentre estas são consideradas a distribuição dos erros dos resultados experimentais e a sua avaliação quando se afastam do conjunto de resposta. Num conjunto de dados se os erros têm distribuição assimétrica a variância residual é função da média (COCHRAN; COX 1957). Outro fator a considerar são as observações discrepantes, as quais levam a perda na sensibilidade nos testes de significância e, deve haver critério para

rejeitá-las (BUSTOS 1980, BUSTOS 1988, HOAGLIN et al. 1992). Na ausência de um teste estatístico clássico para avaliar a dispersão das variâncias dos tratamentos, da aditividade do modelo e da normalidade dos erros, os dados devem ser avaliados através de gráficos, estatísticas descritivas e clássicas para melhor identificar os pontos discrepantes que comprometem a significância do teste F (ZAR 2010). Caso as observações discrepantes sejam detectadas através da análise exploratória recomenda-se a retirada deste valor, com critério, ou o uso da transformação que permita validar os testes de significância e as estimativas dos limites de confiança (FINNEY 1960, BOX; COX 1964, BUSTOS 1988).

As transformações de dados podem, em casos específicos, resolver os problemas de não normalidade, heterogeneidade de variâncias e não aditividade. Transformações apropriadas podem gerar dados com distribuição aproximadamente normal e com independência entre médias e variâncias, resultando em variâncias homogêneas. Os principais tipos de transformação são a logarítmica, a raiz quadrada e a arco-seno ou angular. A transformação logarítmica estabiliza a variância, na situação em que as variâncias são proporcionais ao quadrado das médias dos tratamentos. Em alguns casos, pode contribuir para a normalização dos dados e para a adequação do modelo aditivo linear.

Para uma variável Y com distribuição normal de média u e variância σ^2, o Log de Y tem variância aproximada de $\frac{\sigma^2}{u^2}$. A transformação raiz quadrada é indicada para estabilizar ou homogeneizar a variância quando existe correlação ente média e variância e a variável refere-se a uma contagem, com distribuição de Poisson (esta distribuição tem média igual a variância). Neste caso, a variável transformada pode ser considerada com distribuição normal. A transformação arco-seno é aplicável a dados com distribuição binomial, expressos em frações ou porcentagens, ocorrendo estabilização da variância. Em geral, quando todos os dados equivalem a porcentagens apenas na faixa de 30 a 70, a transformação provavelmente não seja necessária. Para dados discretos, em geral, recomenda-se verificar a existência ou não de correlação entre as médias de cada tratamento e suas variâncias. Se for constatada tal correlação, deve-se identificar a distribuição (Binomial ou Poisson) dos dados e aplicar a transformação recomendada. Uma maneira simples de decidir se uma transformação será efetiva consiste na verificação da proporção entre o maior e o menor valor do conjunto de dados. Se essa proporção for maior do que 20, a transformação será útil (Fernandez, 1992, citado por Resende, 2007). Um procedimento mais formal para detecção da necessidade de transformação dos dados e para indicação da transformação adequada a ser usada, refere-se ao método da transformação potência de Box e Cox (1964), citado por Resende (2007). Essa transformação é dada por $Y_t = Y^\gamma$, se $\gamma \neq 0$ e $Y_t = log\, Y$ se $\gamma = 0$. No caso, Y refere-se ao dado original e Y_t ao dado transformado. O parâmetro de transformação γ varia na amplitude de -2 a 2 e é determinado pela minimização da soma de quadrados residual. A transformação potência pode ser implementada segundo os seguintes passos: i) estimar as médias de tratamentos (M) e seus desvios padrões (S); ii) calcular os logaritmos de M e S; iii) plotar $\log(S)$ contra $\log(M)$ e verificar a linearidade

dessa relação. Uma forte relação não linear indica que a transformação potência não será apropriada ao conjunto de dados e, nesse caso, indica-se a utilização de testes não paramétricos baseados nos ranks das observações; iv) regressar $\log(S)$ em $\log(M)$ e testar a significância da relação linear. Se a regressão não for significativa (a 5%), não há necessidade de transformação dos dados. Se a regressão for significativa, usar a estimativa do coeficiente de regressão (β) para obtenção do parâmetro γ. v) obter γ por meio de $\gamma = 1 - \beta$. Por exemplo, se $\beta = 2$, o valor de γ será -1 e, portanto, a transformação ideal será a recíproca $\left(\frac{1}{Y}\right)$ dos dados. Alguns valores de γ conduzem às transformações comumente usadas, conforme abaixo (RESENDE, 2007).

β	γ	Transformação
2,00	-1	Recíproca
1,00	0	Logarítmica
0,66	0,33	Raiz Cúbica
0,50	0,50	Raiz Quadrada

Conforme ainda Resende (2007), quando a transformação potência é usada, perde-se um grau de liberdade no resíduo da análise, uma vez que o mesmo conjunto de dados foi usado para estimar o parâmetro γ e determinar a transformação apropriada. Os testes de significância e as comparações de médias devem ser realizado sobre os dados transformados. Mas as inferências e interpretações práticas devem ser realizadas na escala original, via transformação das médias e intervalos de confiança para a escala original. Em resumo, a determinação da transformação adequada, sem fazer a análise dos resíduos e/ou usar a transformação potência, nem sempre será efetiva.

Transformação significa uma troca de medida da variável original por uma outra escala. A idéia central é que, se para a variável original as suposições não são adequadas, pode existir uma transformação conveniente tal que, na nova escala, elas sejam razoavelmente satisfeitas. O uso de transformação é um procedimento que pode ser adotado para qualquer modelo de análise de variância e regressão, em experimentos balanceados ou não balanceados e para amostras grandes ou pequenas. As estatísticas utilizadas são exatamente as usuais e os graus de liberdade são mantidos. Portanto, não há a perda de precisão na análise e nem perda de sensibilidade do teste. Segundo Siqueira (1983), em análise de variância e análise de regressão, a transformação pode ser aplicada com um ou mais dos seguintes objetivos:

i. linearizar o modelo
ii. corrigir desvios das suposições do modelo;
iii. simplificar o modelo.

O objetivo i) restringe-se a modelos de regressão. Sua utilidade advém do fato de que procedimentos estatísticos, em geral, são mais complicados para relações não-lineares do que para as lineares.

Em geral, os fatores que causam maiores distúrbios na análise de variância são: presença de erros grosseiros (outliers), assimetria extrema, comportamento anormal de certos tratamentos ou parte do experimento (certos tratamentos apresentaram maior variabilidade que outros) e variâncias como função das médias.

Os principais métodos para contornar essas dificuldades, segundo Cochran (1947) são: omissão de certos tratamentos, observações ou repetições, subdivisão da variância residual e transformação de dados.

Como a falha de qualquer uma das hipóteses básicas resulta, geralmente, em heterogeneidade da variância dos erros e devido a grandeza das distorções que provoca no nível de significância dos testes, a transformação dos dados visa conseguir tal homogeneidade.

Vários fatores podem provocar a heterogeneidade de variância, e dentre eles, o mais importante é quando a variância da variável observada y é uma função da média, ou seja, $V(y) = \mu$ (NUNES, 1998).

Este assunto é pouco discutido na literatura em obras clássicas de estatística aplicada, no entanto dentre as raras exceções podem ser citados as obras como o livro de Li (1964 a e b); por outro lado, Thöni (1967) realizou um estudo mais geral sobre transformação de variáveis ou dados experimentais.

Em algumas situações a análise de variância fica inviabilizada por não ter suas premissas garantidas. Geralmente essas situações se denunciam por apresentarem coeficientes de variação muito altos. É possível encontrar uma resposta com distribuição normal, mas com alto coeficiente de variação. Neste caso, apenas a estratégia de aumento de amostragem, ou a instalação de delineamento especiais, para respostas de fluxo contínuo por exemplo, podem contornar as dificuldades nas comparações de médias. As variáveis que demandam transformação geralmente apesentam uma das alterações, ou os grupos experimentais revelam variâncias diversificadas, dependendo das suas respostas médias, ainda que suas distribuições continuem aparentemente normais, ou ainda pela natureza da resposta estudada é possível perceber que sua distribuição de frequência não é normal. Sendo assim a transformação de variáveis objetiva não só normalizar a resposta, mas também homogeneizar as variâncias dos grupos experimentais, mas nem sempre isto pode ser alcançado através das transformações, para estes casos que fogem à normalização mesmo após a transformação, existem técnicas não paramétricas de análise, as quais independem das premissas exigidas pela análise de variância (SAMPAIO, 2002).

Mischan e Pinho (1996), afirmam que uma maneira eficiente de se lidar com problemas de falta de normalidade dos erros e da heterogeneidade das variâncias nas exigências das pressuposições da análise de variância (ANAVA), é o uso da transformação de dados. Segundo estes autores o efeito da assimetria na distribuição normal consiste em se obter resultados significativos em excesso para o teste F. Este teste, no entanto, será pouco afetado se os desvios da normalidade forem pequenos. Se a distribuição se desviar da normal ela poderá aproximar-se, por exemplo, da Poisson, uma distribuição onde vale a relação variância = média, ou de log normal, onde se tem que a $variância = (média)^2$, ou ainda de uma

distribuição binomial, onde a variância é proporcional à quantidade $[média(1-média)]$. Nestes casos, a variância não mais se mantém constante, variando com as médias dos tratamentos, e o uso de uma estimativa conjunta de variância, S^2, não se justifica. Conforme ainda os autores, se a variância de uma variável X for função conhecida de sua média μ, isto é, $\sigma_X^2 = F(\mu)$, então podemos procurar uma transformação da variável X para outra, Y, isto é, $Y = f(X)$, tal, que a variância de Y seja constante isto é, $\sigma_Y^2 = k$. Escrevendo $Y = f(X)$ pela expansão de Taylor com um termo, $Y \cong f(\mu) + f'(\mu)(X-\mu)$, temos que a média é $E(Y) \cong f(\mu)$ e a variância, $\sigma_Y^2 = E[Y - E(Y)]^2 \cong E[f'(\mu)(X-\mu)]^2 = [f'(\mu)]^2 F(\mu)$. Por esta fórmula vê-se que para tornar σ_Y^2 aproximadamente constante, basta tomar $[f'(\mu)]^2$proporcional a $\frac{1}{F(\mu)}$. Então, se $[f'(\mu)]^2$=$\frac{1}{F(\mu)}$ e $f(\mu)$ =$\frac{1}{\sqrt{F(\mu)}}$, e para determinar $f(\mu)$, deve-se fazer, $f(\mu)d\mu = \frac{d\mu}{\sqrt{F(\mu)}}$, e integrando obtêm-se $f(\mu) = \int \frac{d\mu}{\sqrt{F(\mu)}} + C$.Segundo ainda estes autores na distribuição de Poisson, por exemplo, $\sigma_X^2 = F(\mu) = \mu$. Substituindo em $f(\mu) = \int \frac{d\mu}{\sqrt{F(\mu)}} + C$ tem-se, $f(\mu) = \int \frac{d\mu}{\sqrt{\mu}} + C = 2\sqrt{\mu} + C$. Verifica-se assim que $Y = \sqrt{X}$ é a transformação da variável X, com distribuição de Poisson, que estabiliza a variância na escala transformada, Y. Se ao invés disso for analisado a variável X, então será analisada Y, com variância aproximadamente constante.

Os autores relatam que outro exemplo é o caso da distribuição log normal, onde $\sigma_X^2 = F(\mu) = \mu^2$, onde então tem-se que, $f(\mu) = \int \frac{d\mu}{\sqrt{\mu}} + C = ln|\mu| + C$, e assim fica claro que para dados com distribuição log normal, a transformação logarítmica é a adequada.Já para o caso da variável com distribuição binomial, com $\sigma_X^2 = F(\mu) = k\mu(1-\mu)$, segundo o mesmo cálculo pode ser verificado que a função $Y = arc\ sen\sqrt{X}$é a transformação que estabiliza a variância. Estes autores descrevem alguns exemplos para a aplicação das transformações de dados como será visto a seguir. Transformação raiz quadrada: a transformação $Y = \sqrt{X}$ é recomendada principalmente quando tem-se dados de contagem de eventos raros, como por exemplo, o número de plantas atacadas por uma doença, número de acidentes na produção de um material, etc. Nestes casos a variável segue uma distribuição de Poisson. Também é utilizada para qualquer tipo de contagem, por exemplo, número de frutos produzidos por planta, ou mesmo de dados que não sejam de contagem, desde que se observe que as variâncias são proporcionais às médias dos tratamentos. Se alguns dados de contagem forem pequenos, ou iguais a zero, deve-se somar 0,5 a cada valor, antes de usar a raiz quadrada. Neste caso tem-se a transformação $Y = \sqrt{X + 0{,}5}$, sendo assim, deve-se observar que todos os dados serão acrescidos da constante 0,5, não só os pequenos. Um exemplo que os autores citam é o da produção de frutos, em número de frutos por 2 plantas. Já a transformação logarítmica, deve ser aplicada se os efeitos dos fatores no modelo matemático não forem aditivos, mas sim multiplicativos, e neste caso as observações estarão mais próximas da distribuição log normal, uma distribuição onde o logaritmo da variável é que tem distribuição normal. Já foi dito que na

distribuição log normal a variância é proporcional ao quadrado da média, ou seja, $\sigma^2 = k\mu^2$. A transformação $Y = ln\, X$é, portanto, a transformação adequada a dados onde se observa este tipo de relação entre média e variância, resolvendo tanto o problema de heterogeneidade de variâncias como o de falta de aditividade no modelo. Se ocorrerem valores de X menores que um (1) pode-se usar a transformação $Y = ln(X + 1)$, isto evita que se usem números negativos na análise, além de resolver o problema dos valores de X iguais a zero. Um exemplo que os autores citam é o do número de ovos e de ácaros em 4 variedades cítricas. Por fim, os autores descrevem a transformação arco seno, citando que esta transformação dada por $Y = arc\ sen\sqrt{\frac{X}{100}}$, também denominada transformação angular, é recomendada para estabilizar a variância para dados de porcentagens, principalmente se estiverem no intervalo de 0 a 30% e ou de 70 a 100%. Com dados de porcentagem existe um número máximo, n, de observações, que serve de base para o cálculo da porcentagem. Esse valor de n deve ser o mesmo, ou aproximadamente o mesmo, para cada unidade experimental. Neste caso os autores citam o exemplo de dados referentes as porcentagens de controle de herbicidas, aplicados em 3 diferentes épocas, sobre o crescimento da planta picão.

Segundo Calado e Montgomery (2003), dentre vários tipos de gráficos usados na análise de resíduos para se verificar a adequação do modelo da análise de variância nos experimentos, através do uso dos diversos tipos de delineamentos experimentais, estão aqueles que relacionam os resíduos com a sequência de tempo, com os valores previstos da variável dependente ou com os valores da variável independente. E conforme os autores existem padrões de comportamento característico, onde podem ser observados o tipo onde os pontos devem estar distribuídos de forma aleatória, de modo a caracterizar uma variância constante dos erros (Figura 1a). Pode ocorrer também o caso onde de um aumento na variância com o tempo ou com y_i ou x_i (Figura 1b). Para corrigir isso, conforme os autores, transformações podem ser feitas na variável dependente, tais como, $ln\, y\,, \sqrt{y}, \frac{1}{y}$. Pode ser verificado a desigualdade na variância (Figura 1c), ou ainda que termos de ordens mais elevadas devem ser adicionados ao modelo (Figura 1d).

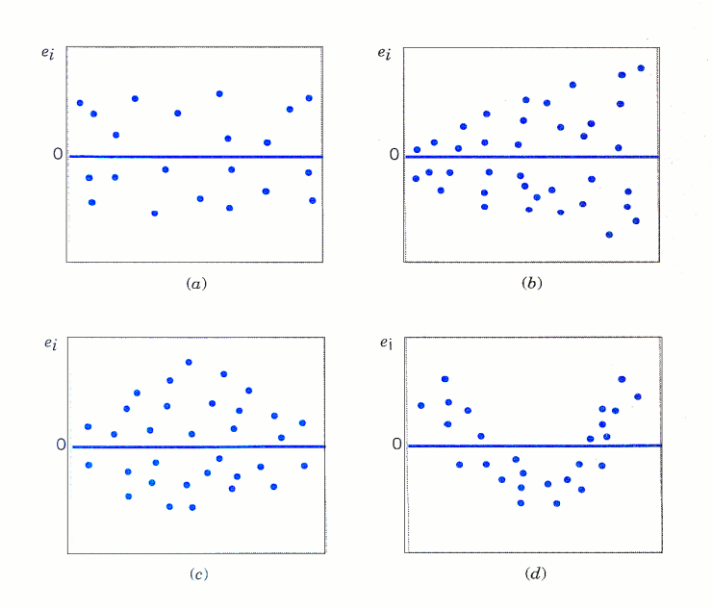


Figura 1. Padrões de comportamento dos gráficos dos resíduos. Mossoró, RN, 2023.

Vários procedimentos estatísticos são baseados na suposição de que os dados provêm de uma distribuição normal ou então mais ou menos simétrica. Mas, em muitas situações de interesse prático, a distribuição dos dados da amostra é assimétrica e pode conter valores atípicos. Se o pesquisador quiser utilizar tais procedimentos, o que se propõe é efetuar uma transformação das observações, de modo a se obter uma distribuição mais simétrica e próxima da normal. Uma família de transformações frequentemente explorada é:

$$x^{(p)} = \begin{cases} x^p, & se \quad p > 0 \\ ln(x) & se \quad p = 0 \\ -x^p, & se \quad p < 0 \end{cases}$$

Normalmente, o que se faz é experimentar valores de p na sequência.

$$\cdots, -3, -2, -1, -\frac{1}{2}, -\frac{1}{3}, -\frac{1}{4}, 0, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2, 3, \cdots$$

E para cada valor de p obtêm-se gráficos apropriados, como histogramas, desenhos esquemáticos, etc. para os dados originais e transformados, de modo a se escolher o valor mais adequado de p. Sabe-se que, para dados positivos, a distribuição dos dados é usualmente assimétrica à direita. Para essas distribuições, a transformação acima com $0 < p < 1$ é apropriada, pois valores grandes de x decrescem mais, relativamente a valores pequenos. Para distribuições assimétricas á esquerda, deve-se tomar $p > 1$. Os autores ainda afirmam que ao aplicar a transformação a dados de populações do Estado de São Paulo para p: $0, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}$, verificou que $\frac{1}{3}$ na transformação logarítmica e $p = \frac{1}{3}$ na transformação raiz cúbica forneceu distribuições mais próxima de uma distribuição simétrica (BUSSAB; MORETTIN, 2002).

A transformação dos dados ou variáveis em levantamentos amostrais e experimentais deve ser realizada baseada em razões próprias para esta transformação dos dados. Por exemplo, quando algum dos requisitos para o emprego da estatística clássica frequentista paramétrica não for atendido, tanto na

conhecida analise de variância (ANOVA) e no emprego de testes clássicos de comparações paramétricas, tias requisitos pode ser por exemplo, a obtenção da normalidade da distribuição dos erros, a prova da homogeneidade das variâncias, e o requisito da aditividade dos efeitos dos fatores de variação do modelo estatístico adotado, nestes casos pode-se adotar outras técnicas e regras para análise dos dados, como o uso de modelos lineares generalizados ou aplicação de técnicas de estatística não paramétrica. Ou seja, se mesmo assim o pesquisador não puder verificar ou provar que os seus dados da sua amostra experimental não preencheu tais requisitos, ai este pesquisador pode ainda tentar o recurso da transformação dos dados, erros ou variáveis do levantamento amostral ou obtidos em ensaios aleatorizados, antes de optar pela aplicação da estatística não paramétrica ou do emprego de modelos lineares generalizados. É um recurso que sempre vale a pena tentar, porque a estatística paramétrica é evidentemente mais poderosa, abrangente, flexível e de maior amplitude de alcance que a não paramétrica, apesar dos dados não terem distribuição de probabilidade livres. De fato, esta estatística de distribuição livre, somente foi desenvolvida como um recurso complementar, destinado a suprir a necessidade de testes estatísticos nos casos em que alguma restrição desaconselhava o uso da estatística paramétrica, ou quando a própria natureza dos dados, muitas vezes valores categóricos ou qualitativos, ou seja, nominais ou ordinais, isto, é, não exatamente numéricos, impedia a aplicação das técnicas paramétricas. As transformações mais comumente utilizadas diretas dos dados na estatística são: a logarítmica, a logarítmica $(log\ x)$, a $log\ x + 1$, a raiz quadrada dos dados $\left[\sqrt{X}\right]$, a raiz quadrada dos dados + 1$\left[\sqrt{X+1}\right]$ou mais 1/2$\left[\sqrt{X+\frac{1}{2}}\right]$, a raiz cúbica dos dados $\left[\sqrt[3]{X}\right]$, a transformação angular $\left[arc\ sen\sqrt{\frac{X\%}{100}}\right]$, a transformação hiperbólica de primeiro grau (ou o inverso dos dados $\left[X_i = \frac{1}{x_i}\right]$, ou hiperbólica de segundo grau $\left[X_i = \frac{1}{x_i^2}\right]$, a transformação percentual $\left[\sqrt{\frac{x\%}{100}}\right]$, e a transformação em valores de z nos testes paramétricos e na análise de regressão, para verificar a normalidade da distribuição dos erros amostrais por exemplo no uso dos testes de médias. Outra transformação adotada pelos pesquisadores é a transformação Box e Cox, que na estatística é vista como uma transformação de potência é uma família de funções que são aplicadas para criar a transformação monotônica de dados usando funções de potência. Esta é uma técnica de transformação de dados útil usada para estabilizar a variância, tornar os dados mais semelhantes à distribuição normal, melhorar a validade das medidas de associação, como por exemplo a correlação linear simples de Pearson entre as variáveis, bem como para outros procedimentos de estabilização de dados. Tanto a forma linear quanto a logarítmica são dois casos particulares de uma família mais extensa de transformações não-lineares. A transformação de potência é definida como uma função de variação contínua, em relação ao parâmetro de potência λ (lambda), ou seja, X^{λ}. Essa classe geral de transformação conhecida como transformação de Box-Cox é um dos tipos de transformação referida anteriormente a

qual é definida por: $f_\lambda(x) = \frac{x^\lambda - 1}{\lambda}$ para λ≠0 e $f_0(x) = log(x)$ para λ=0. A transformação de Box-Cox recebeu este nome em homenagem aos estatísticos que a formularam, George E. P. Box y David Cox, cujo trabalho foi publicado em artigo de 1964 ("An Analysis of Transformations"). É bastante conhecida nas aplicações em economia (econometria) e usada para resolver presenças do fenômeno de heterocedasticidade em modelos estatísticos de regressão, quando o modelo de hipótese matemático apresenta variâncias desiguais para Y e X $(X_1, X_2, X_3, \dots, X_n)$ para todas as observações) e/ou ausência de normalidade dos dados ou erros. A escolha do melhor valor de lambda pode ser automatizada. Na linguagem do programa, pacote ou software R, sua implementação pode ser feita através do pacote forecast, através da função BoxCox.lambda(). É recomendado sempre que possível deixar o valor de lambda restrito entre 0 e 1 (ou 0,5 e 1,0 se os dados incluíssem valores de zero), caso contrário podem resultar em previsões muito erráticas (principalmente quando os valores de lambda estiverem abaixo de 0,5).a transformação mais indicada para ser adotada pelo pesquisador, será baseada sempre numa justificativa objetiva, em geral bem definida matematicamente, para se optar por uma ou outra dessas transformações, sempre levando em consideração de como ou por que a distribuição amostral está se desviando da normalidade ou apresentando graus de assimetria e curtose maior ou menor. Os resultados práticos obtidos nas s do dia a dia é quem vai nortear o pesquisador a escolher e adotar qual a transformação mais indicada para cada caso particular. No entanto, com o uso frequente dos pacotes estatísticos ou softwares de análise estatística, as transformações se tornaram rotinas destes pacotes, e sua aplicação se tornou algo fácil e comum e de uso corrente nos laboratórios de análise. Sendo assim de forma rápida o pesquisador ou o estatístico pode aplicar e realizar as transformações de forma simultânea para ver qual a que produz o melhor resultado, gastando para isso pouco tempo, recursos, pessoal para se chegar a resultados precisos, rápidos e muito interessantes na sua análise estatística. O que o pesquisador deve ter em mente é que, a transformação mais indicada geralmente coincide com aquela que apresentar a ou provocar na distribuição dos dados a maior aproximação ou convergência para a distribuição normal ou gaussiana, desta forma para ele pouco importa saber a a justificativa matemática para tal transformação. Se a transformação não for adequada, a tendência é que a distribuição dos dados ou dos erros não converge para uma distribuição normal ou em forma de sino. A Interpretação dos resultados quando os dados já estão transformados, deve ser feita de maneira bastante criteriosa, pois, deve-se ter o cuidado, após transformar os dados experimentais, de usar o raciocínio em termos da natureza dos novos dados, por ocasião da discussão e da interpretação dos resultados. Por exemplo: algumas transformações invertem os valores dos dados, como é o caso da própria transformação inversa (ou hiperbólica de primeiro grau), na qual $\left[X_i = \frac{1}{x_i}\right]$, e da hiperbólica de segundo grau, em que $\left[X_i = \frac{1}{x_i^2}\right]$. Por outro lado, na transformação logarítmica. Não se deve esquecer, portanto, que, uma vez transformados os dados em logaritmos, a soma de dados logarítmicos não tem o mesmo valor que a

soma de seus antilogaritmos, mas representa o produto destes, de modo que a média dos logaritmos não corresponde ao logaritmo da média de seus antilogaritmos. Na verdade, o antilogaritmo da média dos logaritmos corresponde à média geométrica dos dados originais, e não à média aritmética destes. Por isso, no cálculo das médias, após a transformação logarítmica, não se pode esquecer de que os logaritmos passaram a ser tratados como simples dados numéricos, e não mais como logaritmos. Para fazer a conversão para os valores originais, as médias correspondentes às médias dos dados logarítmicos têm de ser calculadas a partir dos dados originais. A única exigência que é mantida nesses casos é a hierarquia dos dados, pois quando um dado original é maior do que outro, os seus logaritmos mantêm essa mesma ordenação hierárquica, ainda que os próprios valores numéricos passem a ser diferentes. Uma vez normalizada e homogeneizada a distribuição dos dados amostrais, por intermédio da transformação que se comprovar mais conveniente, o pesquisador estará apto a empregar os testes paramétricos. Contudo, se mesmo tendo adotado diversos tipos de transformações, e mesmo assim a distribuição dos valores, dos dados, das variáveis ou dos erros, continua apresentando um comportamento não normal, assimétrico, achatada, afilada, ou não homogênea, ou até mesmo com os efeitos dos fatores do modelo matemático não aditivo, não há outra alternativa senão utilizar a estatística não paramétrica ou o emprego dos modelos lineares generalizados (MLG).

Couto et al. (2009), utilizaram a metodologia Box-Cox (BOX e COX, 1964) para o conjunto de funções potências, em experimentos com abobrinha italiana em ambiente protegido para encontrar uma transformação para estabilizar ou reduzir a variabilidade existente entre os tratamentos simulados e normalizar os resíduos. Assim como para valores nulos, a família de transformações de Box-Cox fica restrita, avaliou-se a variável-resposta somada a uma constante. Adotou-se o procedimento proposto por Yamamura (1999), com $c = 0{,}5$, gerando as expressões e $f(y) = ln(y + 0{,}5)$, $\lambda = 0$, da família de transformações Box-Cox acrescida da constante $c = 0{,}5$, e assim após as análises concluíram que para o número e a fitomassa fresca de frutos da abobrinha italiana, a transformação indicada é a raiz quarta ($\sqrt[4]{X}$).

Lucio et al. (2011) estudando o uso de transformação Box e Cox em ensaios com pimentão, cujo objetivo foi definir uma transformação adequada para as variáveis observadas em colheitas por planta de experimentos com esta cultura realizados em ambiente protegido, e visando estabilizar a variabilidade gerada pela presença de valores zero nas múltiplas colheitas dos frutos, conduziram experimentos realizados nas estações sazonais verão/outono e inverno/primavera, utilizando a família de transformações de Box-Cox, com uma adaptação apresentada em Yamamura (1999), concluíram que apesar da redução na variabilidade e a normalidade dos resíduos, observadas em todos os experimentos, o uso do método não foi eficiente para tornar as variâncias homocedásticas. E assim afirmam que para dados de colheitas por planta de experimentos com pimentão, avaliando a fitomassa e o número de frutos, a transformação indicada é a inversa da raiz quarta.

Geralmente, dentro de sua área específica, o pesquisador sabe historicamente quais as variáveis que precisam ser transformadas. Na realidade, não é difícil perceber esta necessidade quando se consideram a natureza da variável estudada e os tratamentos testados. Por exemplo, se a resposta for infestação de helmintos em carneiros, dado pela contagem de ovos por grama de fezes, e há grupos experimentais envolvendo controle e vermífugos de médio e alto poder, pode-se antecipar que, em sendo aquela contagem uma variável muito instável, a variação entre os resultados obtidos no grupo controle será muito maior do que aquela observada no tratamento mais efetivo. Enquanto no grupo controle a média de ovos por grama de fezes (ogf) pode ser 3600, com uma variação livre em torno desse valor, o grupo de vermífugo mais poderoso poderia acusar uma média de apenas 80 ovos por grama de fezes com uma variação em torno desse valor necessariamente menor devido ao limite fixo e próximo de 0 (zero) ovos por grama de fezes. Neste caso pode-se prognosticar que os grupos com vermífugos mais efetivos terão menores variâncias em relação ao controle e a outros menos eficientes. Embora essa percepção possa ser feita antecipadamente, em caso de dúvida deve-se verificar a magnitude da variância dentro de cada grupo experimental. Vale lembrar que amostras reduzidas para cada grupo experimental, ou seja, menos de 5 observações podem casualmente sugerir variâncias diferentes sem que realmente isto seja verdade. É preciso estar atento para que a necessidade de transformação seja adequadamente evidenciada. (SAMPAIO, 2002).

Conforme Brito (2014), as transformações matemáticas dos resultados de experimentos aleatórios podem realizar-se para homogeneizar as variâncias e/ou normalizar a distribuição de variáveis respostas. As hipóteses são testadas nas variáveis transformadas através de testes não paramétricos, mas, se não for conveniente apresentar os dados na nova variável transformada, as médias podem ser transformadas de volta para a medida original. Segundo ainda o autor entre as transformações desenvolvidas para homogeneizar as variâncias, e que podem também conduzir a normalização da variável incluem-se as descritas a seguir.

Tabela 1. Tipos de transformações e exemplos de aplicações, Mossoró, RN, 2023.

Tipo de Transformação	Aplicar quando
$\sqrt{Y}$	Os valores dos Y_i forem contagens de números pequenos
$\sqrt{Y}+\sqrt{Y+1}$	Os valores dos Y_i forem contagens e alguns Y_i forem iguais a zero
$Log(Y)$	A dispersão dos Y_i é elevada, e as variâncias proporcionais às médias
$Log(Y+1)$	A dispersão dos Y_i é elevada e alguns valores dos $Y_i$forem iguais a zero
$\frac{1}{Y+1}$	Os valores dos Y_i forem muito próximos de zero
$Arcsen(\sqrt{Y})$	Os valores dos Y_i forem proporções ou percentagens dispersas
$Log\left[\frac{(1+Y)}{1-Y}\right]$	$-1 \leq Y \leq 1$
$(1-Y)^{\frac{1}{2}}-\frac{1}{3}(1-Y)^{\frac{3}{2}}$	$0 \leq Y \leq 1$

A não aditividade presente nos dados resulta em heterogeneidade das contribuições das observações para variância residual, não sendo esta, portanto, uma estimativa eficiente da variância comum. Em outras palavras, a expectância $V(e_i) = \sigma^2$ não se verifica. De fato, a variância do erro é geralmente maior do que a que seria obtida se o modelo fosse aditivo. O modelo linear aditivo pressupõe que os efeitos de tratamentos são aditivos, isto é, o valor de qualquer observação resulta da soma dos efeitos das várias causas de variação que afetam os dados. Uma forma comum de não aditividade ocorre quando os efeitos são multiplicativos. Suponha-se o caso de dois tratamentos repetidos em dois blocos. No modelo aditivo a diferença entre o bloco 1 e o bloco 2 é uma quantidade fixa, qualquer que seja o tratamento. Da mesma forma, a diferença entre o tratamento A e o tratamento B é também uma quantidade fixa, qualquer que seja o bloco. O modelo multiplicativo, a diferença entre os blocos 1 e 2 é uma percentagem fixa, independente do tratamento. Do mesmo modo, a diferença entre os tratamentos A e B é uma percentagem fixa, independentemente do bloco. O teste de Tukey para aditividade é empregado para vários objetivos, tais como auxiliar na decisão sobre a necessidade de transformação dos dados, sugerir a transformação mais conveniente e verificar se a transformação feita foi eficiente em produzir aditividade. O teste está teoricamente relacionado com uma transformação matemática dos dados experimentais para a forma $Y = X^P$, onde X representa dados na escala original e Y representa os mesmos dados na escala transformada. O problema consiste em encontrar o valor da potência P tal que os efeitos sejam aditivos na escala transformada $Y = X^P$. Se, por exemplo, $P = \frac{1}{2}$, isto representa a transformação dos dados em $\sqrt{X}$. Se $P = -1$, os X são transformados em $\frac{1}{X}$, o que implica em analisar os recíprocos de X em vez de X. Quando P = 0 deve-se interpretar como uma transformação logarítmica

porque a variável X^P se comporta como Log X quando P é muito pequeno. O argumento matemático do teste se baseia no cálculo integral. (NUNES, 1998).

Barbin (2003), afirma que a análise de resíduos deve ser usada para detectar heterogeneidade de variância, bem como outros fenômenos, apesar de ganhar ênfase em função dos pacotes estatísticos, na prática ainda é muito pouco usada. Segundo ainda o autor a experiência mostra que dados de produção geralmente satisfazem as exigências da análise da variância como a normalidade dos dados, pois sempre que possível trabalha-se com médias por parcela, por outro lado quando se tem dados de contagem ou de porcentagem, já se fazem o uso da transformação de dados. Ele comenta ainda que basta a não verificação de uma das exigências da análise da variância, para que esse tipo de análise não apresente validade, sendo assim recomenda o uso da transformação de dados, onde as mais usadas são: $\sqrt{X+K}$, onde $K \geq 0$ é uma constante, para dados de contagem, $arcoseno\sqrt{\frac{P}{100}}$, onde p = porcentagem, para dados de porcentagem, geralmente entre 0 e 30% ou entre 70 e 100 %, já a transformação $log(X+K)$, deve ser aplicada quando há proporcionalidade entre médias e desvios padrões. Pode – se, no entanto, segundo o autor, e de acordo com Box e Cox (1964) deve-se determinar analiticamente, que tipo de transformação pode ser usado. Estabele-se uma regressão linear entre o log S^2, como variável dependente e $log\ \hat{m}$, como independente. Determina-se $\hat{b}$ e, a seguir, $\lambda = 1 - \frac{\hat{b}}{2}$, e o valor de λ indica que tipo de transformação deve ser feito: se $\lambda \neq 0$ tem-se que $Y^* = Y^{\lambda}$e se $\lambda = 0$ tem-se que $Y^* = log\ Y$. Se por exemplo, o $\hat{b}$=1,1310, então $\lambda = 1 - \frac{1,1310}{2} = 1 - 0,5655 = 0,4345$. Isso permite que se use λ =0,5, por ser uma transformação de uso corrente, ou seja, $Y^* = Y^{0,5} = \sqrt{Y}$. Conforme ainda Barbin (2003), o resumo das várias transformações de dados é o seguinte:

Tabela 2. Tipos de transformações em função de valores diferentes para coeficientes de regressão, Mossoró, RN, 2023.

$\hat{b}$	λ	**Transformação**
0	1	nenhuma
1	$\frac{1}{2}$	$\sqrt{X}$ ou $\sqrt{Y}$
2	0	$log\ x$ ou $log\ y$
3	$-\frac{1}{2}$	$\frac{1}{\sqrt{X}}$ ou $\frac{1}{\sqrt{Y}}$
4	-1	$\frac{1}{X}$ ou $\frac{1}{Y}$

De acordo com Pagano e Gauvreau (2006), a análise dos resíduos em estudos de regressão pode sugerir falhas na suposição de homocedasticidade, isto porque este termo significa que o desvio padrão

dos resultados y, ou $\sigma_{y/x}$, é constante para todos os valores de x, e segundo este autor se o intervalo das grandezas dos resíduos ora aumenta ora diminui conforme $\hat{y}$ se torna maior produzindo uma dispersão com forma de leque, sendo assim isto implica que $\sigma_{y/x}$ não assume o mesmo valor para todos os valores de x, e neste caso a regressão linear simples não é apropriada para modelagem da relação entre x e y. O autor lembra ainda que se os resíduos não exibem uma dispersão aleatória, mas seguem uma tendência distinta, isto é, e_i aumenta conforme $\hat{y}_i$, por exemplo, sugere que a verdadeira relação entre x e y pode não ser linear, e neste caso uma transformação de x ou de y ou de ambas poderia ser apropriada. Conforme os autores ainda, quando a relação entre x e y não é linear, procura-se uma transformação da forma x^pou y^p, nas quais $P = \cdots -3, -2, -1, -\frac{1}{2}, ln, \frac{1}{2}, 1,2,3, \cdots$, deve-se lembrar que ln se refere ao logaritmo natural de x ou de y em vez de um expoente. Assim as transformações possíveis podem ser $ln(y), x^{\frac{1}{2}} = \sqrt{x},\ ou\ x^2$. O autor menciona também que o círculo de potências ou a escala de potências, como é chamada algumas vezes fornece uma diretriz geral para que o pesquisador escolha uma transformação. A estratégia está ilustrada na Figura 2, se os dados plotados se parecem com o padrão do quadrante I, por exemplo, uma transformação apropriada seria x para cima ou y para cima, em outras palavras, x ou y seria elevado a uma potência maior do que $P = 1$ os seja, quanto maior a curvatura nos dados, maior o valor de p necessário para se obter a linearidade, poderia por exemplo substituir x por x^2, ou ainda se um gráfico de dispersão bidimensional sugere que a relação entre y e x^2 seja linear, então deve-se ajustar um modelo da forma $\hat{y} = \hat{\alpha} + \hat{\beta}x^2$, em vez da forma usual $\hat{y} = \hat{\alpha} + \hat{\beta}x$. Por outro lado, se dos dados seguem a tendência do quadrante II, pode-se transformar y para cima ou x para baixo, ou seja, elevar x a uma potência menor do que 1 ou y a uma potência maior do que 1, portanto, poderia o pesquisador substituir x por $\sqrt{x}$ ou por $ln(x)$. Vale ressaltar que qualquer transformação que o investigador escolha, é necessário ele verificar a validade da suposição de homocedasticidade.

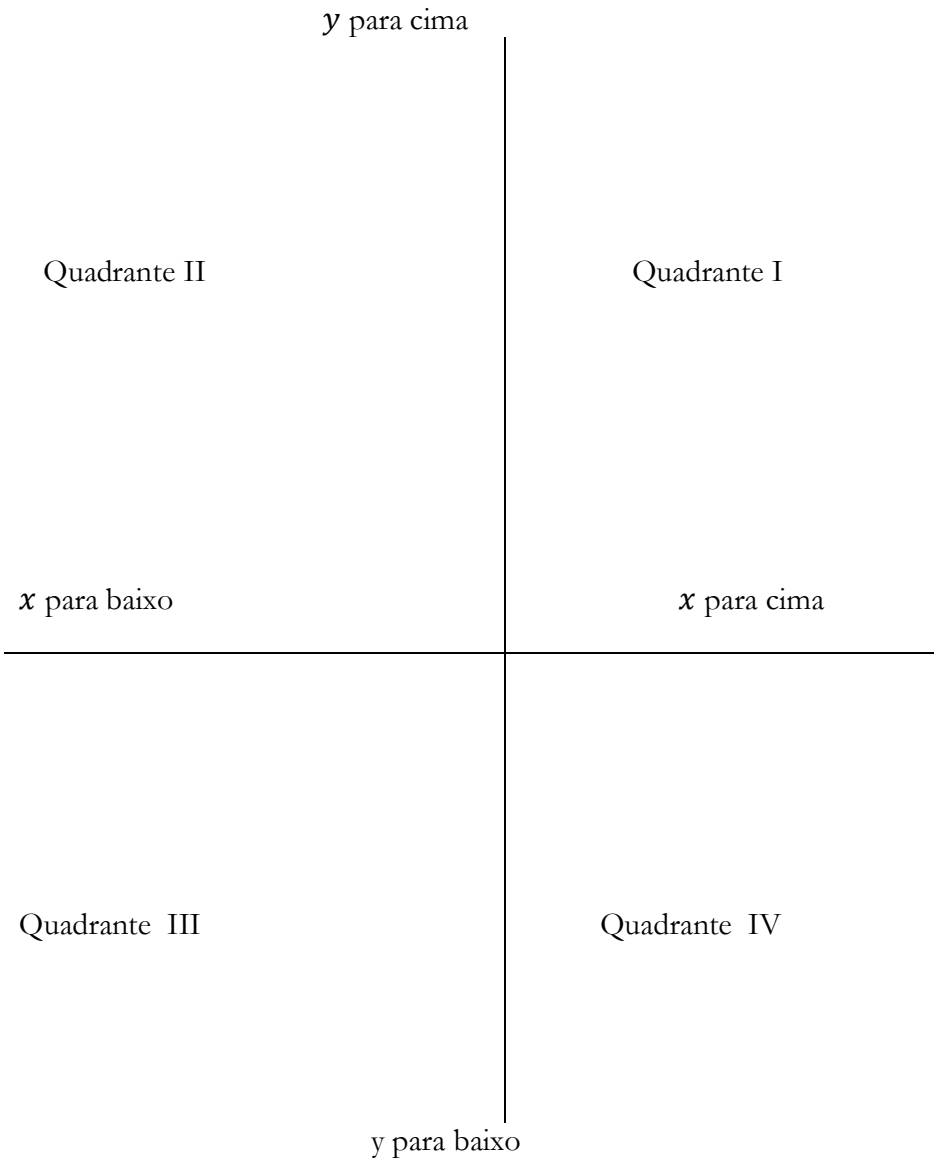


Figura 2. O círculo de potências, Mossoró, RN, 2023.

As transformações de dados podem, em casos específicos, resolver os problemas de não normalidade, heterogeneidade de variâncias e não aditividade. Conforme ainda o autor as transformações apropriadas podem gerar dados com distribuição aproximadamente normal e com independência entre médias e variâncias, resultando em variâncias homogêneas. Os principais tipos de transformação são a logarítmica, a raiz quadrada e a arco-seno ou angular. A transformação logarítmica estabiliza a variância, na situação em que as variâncias são proporcionais ao quadrado das médias dos tratamentos. Em alguns casos, pode contribuir para a normalização dos dados e para a adequação do modelo aditivo linear. A transformação raiz quadrada é indicada para estabilizar ou homogeneizar a variância quando existe correlação entre média e variâncias e a variável refere-se a uma contagem, com distribuição de Poisson, lembrando que esta distribuição tem média igual a variância. Neste caso, a variável transformada pode

ser considerada com distribuição normal. A transformação arco-seno é aplicável em dados com distribuição binomial, expresso em frações ou porcentagens, ocorrendo estabilização da variância. Em geral, quando todos os dados equivalem a porcentagens apenas na faixa de 30 a 70, a transformação provavelmente não seja necessária. Para dados discretos, em geral, recomenda-se verificar a existência ou não de correlação entre as médias de cada tratamento e suas variâncias. Se for constatada tal correlação, deve-se identificar a distribuição teórica de probabilidade que pode ser binomial ou Poisson, dos dados e aplicar a transformação recomendada (RESENDE, 2002).

Muitas vezes você irá ler em um artigo cientifico que os dados foram transformados antes das análises. Após a transformação, os dados atenderam aos pressupostos do teste estatístico utilizado e a análise foi executada. Em geral, pouca coisa é dita sobre as transformações de dados, e você pode perguntar-se porquê os dados foram transformados. Mas, primeiro, o que é uma transformação? Uma transformação significa simplesmente que uma função matemática foi aplicada a todas as observações de uma dada variável: $Y^* = f(Y)$. O Y representa a variável original; o Y^*, a variável transformada, e f é uma função matemática aplicada aos dados. A maioria das transformações corresponde a funções algébricas simples, sujeitas simplesmente à exigência de que sejam funções monótonas contínuas (Uma função contínua é uma função $f(X)$ tal que para quaisquer dois valores de uma variável aleatória X, x_i e x_j, que diferem por um número muito pequeno $(|x_i - x_j| < \delta)$, então $|f(x_i) - f(x_j)| < \varepsilon$, outro número muito pequeno. Uma função monótona é uma função $f(X)$ tal que para quaisquer dois valores da variável aleatória X, x_i e x_j, se $x_i < x_j$ então $f(x_i) < f(x_j)$. Uma função monótona contínua possui essas duas propriedades). Por serem monótonas, as transformações não alteram a ordenação dos dados, mas mudam o espaçamento relativo entre os pontos de dados e, portanto, afetam a variância e a forma da distribuição de probabilidades. Há duas razões legítimas para transformar seus dados antes das análises. Primeiro, as transformações podem ser úteis, porém, embora não estritamente necessárias, pois os padrões nos dados transformados podem ser mais fáceis de compreender e comunicar do que os nos dados brutos. As transformações podem ser necessárias para que a análise seja válida – este é o "atenderam aos pressupostos" usado com maior frequêncial em artigos científicos e discutidos nos livros de biometria Y^* (GOTELLI E ELLISON, 2010).

Transformações de dados como ferramenta cognitiva: As transformações muitas vezes são úteis para converter curvas em linhas retas. Conceitualmente, relações lineares são mais fáceis de entender e muitas vezes têm melhores propriedades estatísticas. Quando duas variáveis são relacionadas umas com as outras por funções multiplicativas ou exponenciais, a transformação logarítmica é uma das transformações de dados mais úteis. Um exemplo ecológico clássico é a relação espécie-área: a relação entre o número de espécies e a área de ilhas ou de amostras (PRESTON, 1962; MAcARTHUR; WILSON, 1967, citados por Gotelli e Ellison, 2010). Se medirmos o número de espécies em uma ilha e

plotarmos contra a área da ilha, os dados com frequência seguem uma função de potência simples: $S = cA^z$, Onde S é o número de espécies, A é a área da ilha, e c e z são constantes ajustadas aos dados. Por exemplo, o número de espécies de plantas encontradas em cada uma das ilhas de Galápagos (Preston, 1962) parece seguir uma relação assintótica. Se plotarmos os dados brutos, o número de espécies S em função da área A, veremos que a maioria dos pontos está agrupada à esquerda da figura (pois a maioria das ilhas são pequenas). Como um primeiro passo para a análise, poderíamos tentar ajustar uma linha reta para essa relação: $S = \beta_0 + \beta_1 A$.Neste caso, β_0 representa o intercepto da linha e β_1, sua inclinação. No entanto, a linha não se ajusta muito bem aos dados. Em particular, observe que a inclinação da linha de regressão parece ser dominada pelos dados de Albermarle, a maior ilha do conjunto de dados. O ajuste da linha aos dados não captura bem a relação entre a riqueza de espécies e a área das ilhas. Se a riqueza de espécies e a área das ilhas estão relacionadas exponencialmente (Equação $S = cA^z$), transformaremos essa equação tirando os logaritmos em ambos os lados: $log(S) = log(cA^z)$ e $log(S) = log(c) + z\,log(A)$. Essa transformação retira vantagem de duas propriedades dos logaritmos. Primeiro, o logaritmo de um produto de dois números é igual à soma de seus logaritmos: $log(ab) = log(a) + log(b)$. Segundo, o logaritmo de um número elevado a uma potência é igual à potência multiplicada pelo logaritmo do número:$log(a^b) = b\,log(a)$ Podemos reescrever a Equação $log(S) = log(cA^z)$, representando os valores transformados logaritmicamente como símbolo: $S^* = c^* + zA^*$. Assim, tomamos uma equação exponencial (Equação $S = cA^z$) e a transformamos em uma equação linear (Equação $S^* = c^* + zA^*$). Quando plotamos o logaritmo dos dados, a relação entre riqueza de espécies e a área das ilhas agora é muito mais clara e os coeficientes têm uma interpretação simples (GOTELLI E ELLISON, 2010).

Outras transformações podem ser usadas para converter relações não lineares em lineares. Por exemplo, a transformação pela raiz cúbica $\left(\sqrt[3]{Y}\right)$ é adequada para medidas de massa ou volume (Y^3), as quais relacionadas alometricamente com medidas lineares de tamanho ou comprimento corporal (Y). Em estudos que examinam as relações entre duas medidas de massa ou volume (Y^3), como a comparação das massas cerebral e corporal, tanto a variável X quanto Y são logaritmicamente transformadas. A transformação logarítmica reduz a variação dos dados, que pode variar em várias ordens de magnitude (GOTELLI E ELLISON, 2010).

Transformações de dados por demanda estatística: Todos os testes estatísticos requerem que os dados atendam a alguns pressupostos matemáticos. Por exemplo, os dados a serem analisados usando análise de variância devem satisfazer dois pressupostos: i) os dados devem ser homocedásticos, ou seja, as variâncias dos resíduos de todos os grupos de tratamentos devem ser semelhantes entre si;.ii) os resíduos, ou desvios da média de cada grupo, devem ser variáveis aleatórias normais. Da mesma forma, os dados a serem analisados usando regressão ou correlação também devem ter resíduos com distribuição

normal e que não sejam correlacionados com a variável independente. As transformações matemáticas dos dados podem ser usadas para atender esses pressupostos. Em geral, as transformações mais comuns visam a ambos os pressupostos de maneira simultânea. Em outras palavras, uma transformação que equaliza as variâncias (Pressuposto 1) frequentemente normaliza os resíduos (Pressuposto 2). Cinco transformações são muito usadas com dados ecológicos e ambientais: a logarítmica, a da raiz quadrada, a angular (ou arcoseno), a recíproca e a de Box-Cox (GOTELLI E ELLISON, 2010). A transformação logarítmica: A transformação logarítmica (ou transformação log) substitui o valor de cada observação pelo seu logaritmo: $Y^* = log(Y)$. A transformação logarítmica (na maioria das vezes usa o logaritmo natural ou logaritmo na base e, em geral equaliza as variâncias de dados em que a média e a variância são positivamente correlacionadas. Uma correlação positiva significa que grupos com médias grandes também terão variâncias grandes (na ANOVA) ou que a magnitude dos resíduos está correlacionada com a da variável independente (na regressão). Os dados univariados que são positivamente assimétricos (assimetria à direita) muitas vezes contêm alguns dados discrepantes grandes. Com uma transformação de log, esses dados geralmente são atraídos para a parte principal da distribuição, tornando-se mais simétrica. Os conjuntos de dados que apresentam médias e variâncias positivamente correlacionadas também tendem a ter dados discrepantes com resíduos positivamente assimétricos. A transformação logarítmica com frequência resolve os dois problemas ao mesmo tempo. Note que o logaritmo de 0 não é definido: independentemente da base b, não há nenhum número em que $a^b = 0$. Uma maneira de contornar esse problema é adicionar 1 a cada observação antes de tirar seu logaritmo (como o $log(1) = 0$, independente da base b). No entanto, essa não é uma solução útil ou adequada, se o conjunto de dados (ou especialmente se um grupo de tratamento) contém muitos zeros. A transformação da raiz quadrada: A transformação da raiz quadrada substitui o valor de cada observação pela sua raiz quadrada: $Y^* = \sqrt{Y}$. Essa transformação com frequência é usada com dados de contagem, como o número de lagartas por paineirinha ou o número de Rhexia por município. Mostramos que esses dados muitas vezes seguem uma distribuição de Poisson e salientamos que a média e a variância de uma variável aleatória de Poisson são iguais (ao parâmetro λ da razão de Poisson). Assim, para uma variável aleatória de Poisson, a média e a variância variam de igual forma. Calcular a raiz quadrada de variáveis aleatórias de Poisson produz uma variância que é independente da média. Como a raiz quadrada de 0 é igual a 0, a transformação da raiz quadrada não muda dados que são iguais a 0. Assim, para completar a transformação, você deve somar um número pequeno aos valores antes de calcular a raiz quadrada. Sokal; Rohlf (1995), citados por Gotelli e Ellison, 2010, sugerem adicionar 1/2 (0,5) a cada valor, enquanto Anscombe (1948), citado por Gotelli e Ellison, 2010 sugere 3/8 (0,325). A transformação do arcoseno ou arcoseno da raiz quadrada: A transformação do arcoseno, arcoseno da raiz quadrada ou transformação angular substitui o valor de cada observação pelo arcoseno da raiz quadrada: $Y^* = ar \rightleftarrows cos \rightleftarrows eno\sqrt{Y}$.Essa transformação é utilizada principalmente para proporções (e porcentagens), que são

distribuídas como variáveis aleatórias binomiais. observamos que a média de uma distribuição binomial $= np$ e sua variância $= np(1 - p)$, onde p é a probabilidade de sucesso e n é o número de tentativas. Assim, a variância é uma função direta da média (variância $= (1 - p)$ vezes a média). A transformação do arcoseno (que é o inverso da função seno) remove essa dependência. Como a função seno gera apenas valores entre -1 e $+1$, o inverso da função seno pode ser aplicado apenas a dados cujos valores estão entre -1 e $+1$: $-1 \leq Y_i \leq +1$. Portanto, essa transformação é apropriada apenas para dados que são expressos como proporções (como p, a proporção ou a probabilidade de sucesso em uma tentativa binomial). Duas advertências sobre a transformação arcoseno. Primeiro, se seus dados são porcentagens (escala de 0 a 100), devem ser convertidos para proporções (escala de 0 a 1,0). Segundo, a função arcoseno gera dados transformados com unidades em radianos, não em graus. A transformação inversa: a transformação inversa substitui o valor de cada observação pelo seu valor inverso: $Y^* = \frac{1}{Y}$. Normalmente é mais usada para dados que registram taxas, como o número de descendentes por fêmea. Em geral, dados de taxas parecem hiperbólicos quando plotados em uma função da variável no denominador. Por exemplo, se você plotar o número de descendentes por fêmea no eixo Y e o número de fêmeas na população no eixo X, a curva resultante pode parecer uma hipérbole, que decresce rapidamente no início, e depois conforme o X aumenta. A forma desses dados em geral é a $XY = 1$. (onde X é o número de fêmeas e Y é o número de descendentes por fêmea), que pode ser reescrita como uma hipérbole $\frac{1}{Y} = aX$. Transformar o Y no seu valor inverso, $\frac{1}{Y}$, resulta em uma nova relação $Y^* = \frac{1}{\left(\frac{1}{Y}\right)} = aX$ que é mais sujeita a uma regressão linear. A transformação Box-Cox: Usamos o logaritmo, a raiz quadrada, o valor in verso e outras transformações para reduzir a variância e a assimetria nos dados e criar uma série de dados transformados que possuam aproximadamente uma distribuição normal. A última é a transformação de Box - Cox, ou transformação de potência generalizada. Essa transformação, na verdade, é uma família expressa pela equação,

$$Y^* = \frac{\left(Y^\lambda - 1\right)}{\lambda} \quad (for \lambda \neq 0)$$

$$Y^* = log_e(Y) \quad (for \lambda = 0)$$

onde λ é o número que maximiza a função log da verossimilhança:

$$L = -\frac{v}{2} log_e(s_T^2) + (\lambda - 1)\frac{v}{n}\sum_{i=1}^{n} log_e Y$$

onde v são os graus de liberdade, n é o tamanho da amostra, e s_T^2é a variância dos valores de Y transformados (BOX; COX, 1964, citados por Gotelli e Ellison, 2010). O valor de λ, que resulta quando a equação $L = -\frac{v}{2} log_e(s_T^2) + (\lambda - 1)\frac{v}{n}\sum_{i=1}^{n} log_e Y$ é maximizada, é usado na equação

$$Y^* = \frac{(Y^\lambda - 1)}{\lambda} \quad (\text{para } \lambda \neq 0)$$

$$Y^* = log_e(Y) \quad (\text{para } \lambda = 0)$$

para fornecer o melhor ajuste dos dados transformados a uma distribuição normal. A Equação $L = -\frac{v}{2} log_e(s_T^2) + (\lambda - 1)\frac{v}{n}\sum_{i=1}^{n} log_e Y$ deve ser resolvida iterativamente (tentando diferentes valores de λ até L ser maximizado), usando programas de computador.

Determinados valores de λ correspondem às transformações já descritas. Quando $\lambda = 1$, a Equação resulta em uma transformação linear (operação de deslocamento); quando $\lambda = 1/2$, o resultado é a transformação da raiz quadrada; quando $\lambda = 0$, o resultado é a transformação logarítmica natural; e quando λ = – 1, o resultado é a transformação inversa. Antes de partir para o problema de maximizar a Equação $L = -\frac{v}{2} log_e(s_T^2) + (\lambda - 1)\frac{v}{n}\sum_{i=1}^{n} log_e Y$, você deve tentar transformar seus dados usando transformações aritméticas simples. Se seus dados têm assimetria à direita, tente usar as transformações mais familiares a partir da série $\frac{1}{\sqrt{Y}}, \sqrt{Y}, ln(Y), \frac{1}{Y}$. Se têm assimetria à esquerda, tente Y^2, Y^3, etc. (SOKAL; ROHLF, 1995, citados por Gotelli e Ellison, 2010). Apresentando resultados: transformados ou não? Embora você possa transformar os dados para análise, você deve relatar os resultados nas unidades originais. Por exemplo, os dados de espécie-área poderiam ser analisados usando aqueles transformados em log; porém, ao descrever o tamanho da ilha ou a riqueza de espécies, você deve apresentá-los em suas unidades originais, pela transformação reversa. Pode-se construir intervalos de confiança de maneira similar, tirando o antilog dos limites de confiança construídos e usando os erros-padrão das médias dos dados transformados. Isso normalmente resulta em intervalos de confiança assimétricos.

Em ensaios, onde é mensurada mais de uma variável resposta, a opção mais adequada é a análise de variância multivariada, no entanto poucos são os trabalhos aplicados, principalmente na área agrícola, cujos pesquisadores utilizam técnicas multivariadas e menores ainda o número de artigos que adotam análise de variância multivariada como técnica de análise estatística. Tal fato pode ser explicado pela maior complexidade dessas técnicas. As análises estatísticas de dados de experimentos desse tipo, em geral, são efetuadas para cada característica individual, o que leva a resultados próprios para cada característica, o que, às vezes, fica difícil de se chegar a uma conclusão geral. Uma alternativa, relativamente mais simples foi proposta por Pimentel - Gomes (2009) que utilizou a Função Discriminante Linear de Fisher – FDF, Fisher (1936) para a transformação de dados multivariados em uma nova variável, por meio da variável canônica principal, de maneira a atribuir a esta nova variável um valor máximo do teste F da análise de variância univariada, o que possibilita uma nova opção de análise de variância dos dados multivariados, a qual é dada por: $FDF = b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_iX_i$. Em que X_i, com $i = 1,2,\ldots,n$, representa

cada uma das características, e b_i com $i = 1,2,...n$, são os coeficientes ou pesos a ser determinados pelo método proposto.

Recentemente Campos et al. (2013), transformando os dados por meio da Função Discriminante Linear de Fisher, para posterior análise, concluiu que este tipo de transformação se mostrou como uma técnica viável para apurar ou detectar diferenças significativas. Ainda de acordo com estes autores, trabalhando com mudas de cafeeiro, estimaram a Função Discriminante Linear de Fisher como uma função linear nas características da qualidade de mudas, representada por: $FDF = b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7$, em que X_i, com $i = 1,2,3,4,5,6,7$ representa cada uma das sete características, e b_i com $i = 1,2,3,4,5,6,7$, são os coeficientes ou pesos a ser determinados pelo método proposto. Após a estimação dos coeficientes da FDF dada por $FDF = b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7$, substituiu-se os valores observados de cada característica para obtenção da nova variável (FDF), que explica grande parte das informações contida nas características avaliadas. Os valores obtidos dessa nova variável foram submetidos a análise de variância conforme esquema proposto para cada característica. Segundo ainda ou autores para construção da Função Discriminante Linear de Fisher foi utilizado o autovetor t, que é associado ao máximo autovalor e que maximiza a razão $\frac{t'Ht}{t'Rt}$, em que H e R são respectivamente as matrizes da soma de quadrados e da soma de produtos devidos aos efeitos dos tratamentos e dos resíduos, conforme análise multivariada apresentada por Padovani e Aragon (2005).

Conforme Resende (2007), no contexto dos modelos multivariados, as variáveis correlacionadas podem ser transformadas em variáveis não correlacionas, visando à realização de análises univariadas em vez de uma multivariada, porém sem perda de acurácia. Dentre os procedimentos de transformações, destacam-se: i) transformação canônica, que é adequada quando todos os caracteres são avaliados em todos os indivíduos, ou seja, quando as matrizes de incidência X e Z são as mesmas para todos os caracteres e ii) transformação Cholesky, quando observações de alguns caracteres são perdidas e a perda é sequencial. Segundo ainda este autor, devido à dificuldade de convergência e alto custo da análise sob modelo multivariado, o procedimento de transformação de variáveis originais em variáveis não correlacionadas geneticamente e residualmente, é muito interessante. Tal procedimento é denominado transformação canônica (Thompson, 1977, citado por Resende, 2007) e atua por meio da decomposição das matrizes de covariância genética e residual entre caracteres, usando matrizes de transformação. Sendo $Var(y) = Var(a) + Var(e) = G + R$e sendo G e R matrizes simétricas e positivas definidas, existe uma matriz de transformação T tal que $TRT' = I$ e $TGT' = H$, em que I é uma matriz identidade e H é uma matriz diagonal. Assim, fazendo-se essas multiplicações, R se torna igual a identidade e G se torna diagonal, ou seja, não existem mais covariâncias explicitas em R e G. E multiplicando os dados originais por T, obtém-se o vetor $y*$ das variáveis não correlacionas. Análises univariadas são então aplicadas

sobre $y*$ e os resultados (u) são convertidos para a escala original, produzindo valores genéticos idênticos àqueles obtidos sob o modelo multivariado (m). As expressões de conversão são dadas por $\hat{b}_m = T^{-1}\hat{b}_u^*$ e $\hat{a}_m = T^{-1}\hat{a}_u^*$. É importante mencionar que, para a realização das análises univariadas sobre $y*$, a herdabilidade é dada por $h_a^2 = \frac{\sigma_a^2}{(\sigma_a^2+1)}$ e o coeficiente lambda associado ao fator shrinkage nas equações de modelo misto é dado por $\lambda_1 = \frac{1}{\sigma_a^2}$, visto que a variável transformada apresenta variância residual unitária. Segundo ainda Resende (2007), outro tipo de transformação que pode ser usada é a transformação Cholesky. Tal transformação produz variáveis com correlação residual nula e variância unitária. A matriz de transformação $L^{\{-1\}}$ é obtida por meio da decomposição Cholesky de R. Tal decomposição é dada por $R = LL'$, em que L é uma matriz triangular inferior. A matriz de transformação é a inversa de L. A decomposição de Cholesky é uma especialização da decomposição LU (em matriz triangular inferior L e superior U), a qual é uma consequência do método de eliminação de Gauss ou de absorção de uma linha da matriz por vez, via operações elementares sobre linhas da matriz, em processo que culmina com a triangularização. No processo de eliminação de Gauss, os elementos diagonais são denominados pivôs. A transformação de variáveis via decomposição Cholesky é equivalente ao processo de condensação pivotal (Aitken, 1937, citado por Resende, 2007) usado na obtenção de variáveis canônicas.

No processamento analítico de dados experimentais ou não se utiliza, com frequência, a técnica de análise de variância. Para tanto, adota-se um modelo estatístico linear, tal como $Y_i = \mu + \varepsilon_i \quad (i = 1,2,\cdots,N)$, em que Y_i representa o valor observado do elemento i da população estudada; μ representa a média populacional e ε_i representa o erro ou desvio do valor observado i em relação a média, o qual assume premissas básicas, a saber; i) a aditividade que é a condição imposta pelo modelo, em que os diversos efeitos se somam; ii) a independência dos erros que assegura que a probabilidade do erro de uma observação qualquer assumir um dado valor não deve depender dos valores dos outros erros; iii) a homogeneidade de variância que é sinônimo de homocedasticidade, que assume que eles tenham todos a mesma variância, para assim validar os testes F e de comparações de médias e o processo de estimação de componentes de variância; iv) a normalidade que assegura que os erros devem ser normalmente distribuídos, a qual é importante para validar os testes de hipóteses. Ocorrendo desvios importantes dessas pressuposições devem-se adotar um dos dois expedientes abaixo: i) emprego de técnicas analíticas que não exijam atendimento a essas premissas, tais como as não paramétricas, os mínimos quadrados ponderados, entre outras; ii) o emprego de funções de transformação dos dados originais, de modo que as pressuposições sejam atendidas (DIAS; BARROS, 2009). Sendo assim conforme ainda estes autores com frequência surgem conjuntos de dados em que a média e a variância são relacionadas. Esse fato é o principal responsável pela heterogeneidade de variância, também chamada heterocedasticidade. Logo,

essa relação tem que ser quebrada para restituir a normalidade aos dados. Isso é possível encontrando-se uma função logarítmica, raiz quadrada, angular ou recíproca, capaz de desacoplar média e variância. Tem-se então: i) transformação logarítmica que segundo Steel et al. (1997) citado por Dias e Barros (2009) é aquela que é útil quando o efeito de um tratamento é a mudança percentual, instalando um modelo multiplicativo ao invés do aditivo, como exemplo os autores citam o conjunto de dados de produção de frutos de cacaueiros tomados ao acaso e deles acrescentam 50%, nas escalas normal e logarítmica, onde se observa que os dois conjuntos de dados têm a mesma variância na escala logarítmica e que a média do controle, nesta mesma escala, corresponde a média geométrica dos dados originais e não propriamente ao log de 46,4, que é a média original ou de controle. Basicamente a transformação logarítmica comum ou natural estabiliza a variância e remove a assimetria da distribuição dos dados. Em suma, se $X = log\ (Y)$ tem distribuição normal, então Y é dito ter distribuição lognormal. Atenção deve ser dispensada ao fato de que a transformação logarítmica não pode ser aplicada a números negativos e nem a zero. Por essa razão, os dados contendo zeros devem ser transformados para $log\ (Y + C)$, onde C assume valores 0,50 ou 1; ii) transformação raiz quadrada: este tipo de função de transformação $X = \sqrt{Y}$ou $X = \sqrt{Y + C}$é recomendada para dados oriundos de contagem, supostamente distribuídos conforme Poisson. Algebricamente, tem-se $E(Y) = V(Y) = \mu$. A transformação raiz quadrada estabiliza a variância, independentemente da média. A soma de $C = 0{,}5$ ou $1{,}0$ à raiz quadrada do dado original se faz necessária em presença de zeros; iii) transformação angular: A transformação angular ou seno inversa ou arco seno, onde $X = seno^{-1}\sqrt{Y}$ou $X = arccosseno\sqrt{Y}$, foi proposta para tratar dados de proporção, supostamente distribuídos conforme a binomial. Exemplo disso é a proporção de sementes germinadas, registradas em um intervalo de 0 a 1. Neste caso, o intervalo de proporções entre 0,3 e 0,7 dispensa transformação; iv) transformação recíproca; trata-se de uma função de transformação adequada quando Y refere-se à taxa de sobrevivência. Por conseguinte, $\frac{1}{Y}$ refere-se à taxa de mortalidade.

FERREIRA (2000), afirma que além de aprender as regras para levar a cabo uma análise de variância, todo pesquisador deve buscar o domínio e a compreensão dos princípios inerentes a mesma, para não se defrontar com sérios problemas, como por exemplo, chegar a conclusões que não têm justificativas ou não alcançar conclusões importantes porque os dados não foram analisados adequadamente. Desse modo para que a análise de variância possa ter validade, o pesquisador deve atender às seguintes pressuposições: i) os efeitos principais devem ser aditivos, para tanto nos experimentos, os vários efeitos devem ser aditivos e para isso existe em cada delineamento estatístico um modelo linear aditivo, e caso o que foi exposto acima não se verifique, é necessário transformar os dados experimentais para justá-los ao modelo aditivo; ii) os erros de observação devem ser independentes, ou seja, cada observação possui um erro que deve ser independente dos demais. O princípio da casualização assegura a validade da estimativa do erro experimental, pois permite uma distribuição independente do

mesmo. A casualização evita que todas as parcelas que recebem o mesmo tratamento ocupem posições adjacentes na área experimental; iii) os erros de observação devem ser normalmente distribuídos, pois a única fonte de variação dentro de amostragens são os erros aleatórios, felizmente, as variações da suposição de normalidade não afetam muito seriamente a validade da análise de variância. A normalidade dos dados pode ser verificada por um teste de normalidade como, por exemplo, o teste do Qui-quadrado, desde que o número de amostras com as quais estamos trabalhando seja definitivamente grande. Quando se verifica que falta normalidade aos dados, usamos as transformações para que os mesmos sejam normalmente distribuídos. De modo geral dados médios de parcelas têm distribuição normal; iv) as variâncias das diferentes amostras devem ser homogêneas, pois na análise de variância, o valor do quadrado médio do resíduo que corresponde a estimativa da variância do erro experimental, é utilizada nas fórmulas matemáticas dos testes de hipóteses. Tais testes são utilizados para verificar se existe ou não diferença significativa entre tratamentos avaliados. O quadrado médio do resíduo nada mais é que a média das variâncias de cada tratamento (amostra). Assim sendo, é importante que as variâncias das diferentes amostras sejam homogêneas, de modo que os resultados obtidos dos testes de hipóteses tenham validade. Entre os vários testes estatísticos utilizados para verificar a homogeneidade de variâncias, temos o teste F máximo proposto por Hartley. Uma regra prática e rápida para verificar a homogeneidade de variâncias é que a relação entre a maior e a menor delas não pode ser superior a mais de quatro vezes para que elas sejam homogêneas. Quando as variâncias das diferentes amostras não são homogêneas, temos diversos cursos que podemos seguir. Primeiro, podemos separar as amostras em grupos, de modo que as variâncias dentro de cada grupo sejam homogêneas. Assim, a análise de variância poderá ser efetuada para cada grupo. Segundo, podemos utilizar um método descrito em textos mais avançados de estatística, o qual contempla um procedimento bastante complicado para ponderar médias de acordo com suas variâncias. Terceiro, podemos transformar os dados de tal forma que eles fiquem homogêneos, sendo este último método o mais utilizado na prática. Segundo ainda o autor na transformação de dados deve ser observado o seguinte: Como visto, na análise de variância, algumas condições são exigidas para que os testes de hipóteses tenham validade. Contudo, como tais condições raramente são verificadas na prática, vários procedimentos são utilizados com o fim de reparar pelo menos aproximadamente, a falta de verificação dessas condições. Dentre os procedimentos, geralmente utilizam-se transformações de dados. Uma transformação é qualquer alteração sistemática num conjunto de dados onde certas características são mudadas e outras permanecem inalteradas. Conforme ainda o autor as principais transformações são: i) raiz quadrada: própria para certos tipos de dados em que a média é aproximadamente igual a variância, ou seja, para dados oriundos de uma distribuição de Poisson, que é um tipo de distribuição em que os dados apresentam uma probabilidade muito baixa de ocorrência em qualquer indivíduo. Os fenômenos naturais são os exemplos mais óbvios desse tipo de ocorrência. Tais tipos de dados ocorrem quando as variáveis são oriundas de contagem como por exemplo, sementes

por parcela, período de enraizamento de bulbos, insetos por planta, carrapatos por animal, etc. Os dados provenientes de uma escala de notas também devem ser transformados através da raiz quadrada. Também os dados de porcentagens, referentes às contagens, quando variam de 0 a 20 % ou de 80 a 100 %, podem ser transformados através da raiz quadrada. Neste caso, as porcentagens entre 80 e 100 % devem ser, de preferência, subtraídas de 100, antes de se fazer a transformação. A transformação da raiz quadrada é, ainda, indicada no caso de porcentagens, fora dos limites acima considerados, quando as observações estão claramente numa escala contínua, neste caso temos $\sqrt{X}$. Quando neste tipo de transformação os dados variam de 0 a 10, nós trabalhamos com $\sqrt{X + 0{,}5}$ ou $\sqrt{X + 1}$, em lugar de $\sqrt{X}$. ii) logarítmica: é usada sempre que temos dados em que os desvios padrões das amostras são aproximadamente proporcionais às médias, ou seja, todas as amostras apresentam o mesmo coeficiente de variação. Também quando os efeitos principais são multiplicativos, em vez de aditivos, os dados devem ser transformados através desse tipo de transformação. Essas transformações são satisfatórias quando os dados se referem à contagem de bactérias, de esporos, de grãos de pólen, etc. Dados provenientes de adição de vitaminas em animais também devem ser transformados através da transformação logarítmica. É utilizada, ainda, quando os dados são apresentados por porcentagens que abrangem uma grande amplitude de variação. Nesse caso temos: $log\, X$. Na transformação logarítmica, quando a amostra possui dados iguais a zero ou muito próximos de zero, nós trabalhamos com $log(X + 1)$. Essa transformação deve ser usada quando as variâncias de cada amostra possuem, no mínimo, 12 observações. iii) arcoseno ou angular: própria para dados em que a média é proporcional à variância, ou seja, para dados oriundos de uma distribuição binomial, que é o tipo de distribuição em que os dados apresentam uma probabilidade calculável de ocorrência ou não em qualquer indivíduo. Tais tipos de dados ocorrem quando as variáveis são oriundas de proporção como: porcentagem de germinação de sementes, porcentagem de mortalidade de plantas infectadas com vírus, porcentagem de sobrevivência de bezerros da raça Nelore, etc. Neste caso temos: arco seno $\sqrt{X\%}$. Na transformação arco seno, quando todos os dados estão entre 30 e 70 % não precisa usar a transformação. Se os dados extrapolam esta amplitude, usa-se então a transformação. Quando o número de observações for menor que $50\ (N < 50)$, a proporção 0% deve ser substituída por $\frac{1}{4}N$ e a proporção 100% para $100 - \frac{1}{4}N$, antes de transformar os dados em arco seno $\sqrt{X\%}$. Segundo ainda Ferreira (2000), na escolha da melhor transformação deve-se observar o seguinte: Em alguns casos ficamos sem saber qual seria a transformação mais adequada. Quando nos defrontamos com tais situações, dispomos de várias maneiras para escolher a melhor transformação. Entre as várias maneiras, uma das mais simples é por meio de gráficos, onde se coloca no eixo dos x e y as médias e variâncias respectivas de cada amostra para cada transformação e seleciona-se a que apresentar menor dispersão. Outro procedimento é aplicar cada transformação para o maior e o menor dado de cada amostra. A amplitude dentro de cada amostra é determinada e a razão entre a maior e a menor amplitude

é calculada. A transformação que produz a menor razão é a selecionada. Segundo ainda este autor, pode ser usado o coeficiente de variação como indicativo para o uso de transformações: uma indicação razoável do efeito favorável das transformações de dados é o coeficiente de variação (CV). Quando o valor do CV dos dados transformados for menor que o valor de CV dos dados originais ou não transformados, indica que a transformação foi válida. Em caso contrário, não se justifica o seu uso. O autor mostra resultados exemplificando e considerando os dados de um exemplo em sua obra que é referente ao período de enraizamento em dias de cultivares de cebola (allium cepa L.) de dias curtos. Piracicaba, SP. Ferreira, 1982. Temos que:

Dados originais	CV = 38,26 %
Dados transformados em $\sqrt{X}$	CV = 21,35 %
Dados transformados em log X	CV = 32,49 %

Realmente, as transformações de dados foram válidas, pois houve uma redução muito significativa nos coeficientes de variação em relação aos dados originais, indicando que os dados experimentais foram ajustados de acordo com as exigências da análise de variância. Contudo, a transformação da raiz quadrada foi novamente confirmada como sendo a melhor transformação para tias dados.

Ferreira (2000) descreve ainda algumas considerações sobre transformações de dados, tais como mostrado a seguir: quando é utilizada uma transformação de dados, todas as comparações entre médias de tratamentos são feitas na escala transformada. Quando se achar preferível não apresentar os resultados na escala transformada, os dados finais devem ser transformados novamente para a escala original. Isto é feito elevando-se ao quadrado, no caso de $\sqrt{X}$, achando o antilogarítmo no caso de $log\ X$, e procurando o valor correspondente na tabela de arco seno $\sqrt{X\%}$, no caso de transformação angular. Em certos casos, nenhuma transformação existe que possibilite o uso da análise de variância. Isto ocorre quando: i) as médias são aproximadamente iguais e as variâncias heterogêneas; ii) as variâncias são homogêneas porém os níveis dos tratamentos são heterogêneos em forma; iii) as médias variam independentemente das variâncias. Se alguns destes casos ocorrem, a análise dos dados é feita através de métodos não paramétricos.

Diante do exposto, o presente livro tem como objetivo discutir a demanda de transformações de dados em variáveis respostas de ensaios aleatórios, não aleatórios e quase experimentais em estudos descritivos, em levantamentos amostrais, em estudos de caso controle e coorte na epidemiologia e na área de saúde pública, em análise de regressão e correlação linear e não linear, na análise multivariada, em análise não paramétrica, no ajuste ou aderência de modelos matemáticos de funções de probabilidade e funções densidade de probabilidade à séries de valores ou a uma distribuição empírica de valores de uma série histórica obtidas nas ciências atmosféricas, da terra e do ambiente, nas ciências físicas, na biologia, nas ciências agrárias, nas ciências sociais, na medicina, na paleontologia, arqueologia, botânica, botânica

sistemática, botânica de anatomia de plantas, ecologia, zoologia, aquicultura, piscicultura, engenharia de pesca, no comércio, na indústria, em estudos de simulação e em outros tipos de análise estatística e aplicação de técnicas de inferência estatística de uma maneira geral, bem como sugerir diferentes tipos de transformação para os dados nestes experimentos planejados e conduzidos nas mais diferentes áreas do conhecimento humano, bem como apresentar um estudo de caso.

CAPÍTULO 2

Conforme Ayres et al. (2007), as transformações são procedimentos estatísticos de mudanças de escalas, com a finalidade de obter a normalidade da distribuição dos escores e a estabilização da variância. Na distribuição normal os escores amostrais são transformados de modo linear em z escores, cuja média é igual a zero (0) e o desvio padrão igual a uma unidade (1). Outras transformações não - lineares podem ser efetuadas para atender ao modelo de distribuição normal, sobretudo quando os escores brutos apresentam acentuada assimetria ou quando médias muito grandes são acompanhadas por variâncias também muito elevadas, com perda da normalidade, e da homocedasticidade.

As principais transformações de variáveis oriundas de experimentos agrícolas conforme Dagnelie (1973) são as seguintes:

As transformações lineares do tipo $\hat{y}_i = \hat{a} + \hat{b}x_i$, permitem simplificar em certos casos, a redução dos dados relativos a uma ou a várias variáveis e também a inferência estatística relativa as médias e às variâncias, incluindo a análise da variância (anova). No entanto estas transformações não apresentam interesse no que se refere à aplicabilidade dos métodos estatísticos, pois não modificam a normalidade nem a não normalidade das distribuições, nem a igualdade ou desigualdade das variâncias, nem a linearidade ou não linearidade de uma eventual regressão.

Outra transformação é a raiz quadrada dada por $y = \sqrt{x}$, permite estabilizar as variâncias sempre que há proporcionalidade entre a variância e a média da variável inicial. Este é um caso particular em que os dados tem distribuição de Poisson. Sendo assim nestas condições, temos o seguinte: $\sigma_X^2 = k\ m_X$, de tal forma que $\sigma_Y^2 \approx k\ m_X \left(\frac{d\sqrt{X}}{dX}\right)^2_{m_X} = k\ m_X \left(\frac{1}{2\sqrt{m_X}}\right)^2 = \frac{k}{4}$. Esta transformação é útil sobretudo para as variáveis aleatórias discretas ou descontínuas, possuindo distribuições semelhantes às distribuições de Poisson. Quando os valores observados são no conjunto, relativamente baixos, a prática mostra que é preferível utilizar a transformação $y = \sqrt{x + \frac{1}{2}}$, ou até, segundo alguns autores como Anscombe (1948) e De Munter (1958) a fórmula pode ser $y = \sqrt{x + \frac{3}{8}}$. Vale lembrar que é por intermédio de uma transformação raiz quadrada que a aproximação para a distribuição normal é utilizada no caso das distribuições Qui-quadrado (χ^2) com mais de 30 graus de liberdade. Frequentemente esta transformação é utilizada para dados de contagem como o número de insetos mortos ou capturados em armadilhas luminosas, número de plantas com uma determinada sintomatologia, número de frutos com determinada característica ou sintoma de uma doença, número de ovos de determinado tipo, etc. Esses dados discretos, normalmente seguem a distribuição teórica de probabilidade de Poisson, na qual a média é igual a variância. Nesse caso a transformação adequada para homogeneizar as variâncias dos tratamentos, de

tal sorte que os dados passem a apresentar uma distribuição normal é a raiz quadrada. A transformação básica para os dados nas condições especificadas consiste simplesmente na extração da raiz quadrada dos dados, e em seguida procede-se a análise da variância. Se os dados apresentarem valores zero, recomenda-se acrescentar ao valor genérico X que representa a parcela, uma constante k, que normalmente assume 0,5 ou 1,0 ($\sqrt{x+0{,}5}$ou $\sqrt{x+1{,}0}$), podendo, entretanto, assumir outros valores, desde que, logo após a transformação, se aplique os testes que comprovem a melhor homogeneidade dos tratamentos, além da redução do valor do coeficiente de variação (CV) (SILVA; SILVA, 1999).

Segundo Ayres et al. (2007), a transformação raiz quadrada é sugerida em variáveis referentes a medidas de superfície ou em variáveis discretas tais como o número de leucócitos por mm^3, número de filhos por casal, etc. Emprega-se, também, quando o modelo de distribuição dos escores brutos é o de Poisson, no qual a média e a variância têm o mesmo valor ($\lambda = np$). Com essa transformação a variância e a média tornam-se independentes. Quando o valor da variável for zero, acrescenta 0.5 a todos os valores.

Segundo Zimmermann (2004), alguns exemplos de dados de observação que seguem a distribuição raiz quadrada dada por $\left(\sqrt{X+K}\right)$ são as contagens de números de plantas daninhas ou de insetos de uma dada espécie por unidade de área, por exemplo, onde as contagens são sempre em números inteiros ou baixos. Dados em porcentagem baseados em contagem e com denominador comum, situados entre 0 % e 20% ou entre 80% e 100%, também podem ser transformados por raiz quadrada. Segundo este autor a constante k adicionada ao dado pode, usualmente, assumir três valores, 0, $0{,}5$ e 1. Quando os valores são muitos pequenos e, principalmente, se zeros estão presentes, a simples raiz quadrada ($k = 0$) tende a supercorrigir e, portanto, $k = 0{,}5$ ou $k = 1$ devem ser empregados.

Segundo Sampaio (2002), algumas situações envolvem respostas discretas correspondentes a contagem. Uma variável discreta pode apresentar aproximadamente uma distribuição normal, mas se isto não ocorrer, haverá necessidade de transformação, principalmente se o evento estudado for difícil de ser detectado nas amostras experimentais. Neste caso, o fenômeno é dito raro e sua distribuição aponta altas frequências pra contagem nulas ou baixas e poucas frequências de contagens altas. Este tipo de distribuição, chamada de distribuição de Poisson, se caracteriza por ter um valor médio equivalente ou proporcional à variância. Conforme ainda este autor o número de peixes capturados por armadilhas, quando o peixamento do local tenha sido recente, a frequência de uma espécie vegetal ou animal ameaçada de extinção por unidade de área e o número de colônias bacterianas por placa semeada são exemplos clássicos de variáveis que demandam a transformação radicial, uma das que mais drasticamente controlam a variação original observada, e que consiste em substituir a resposta X por $\sqrt{X}$ou $\sqrt{X+1}$, neste último caso se houver muitos valores nulos, para assim alcançar as premissas de uma análise de variância.

A transformação logarítmica que é do tipo $y = log_e x$, é bastante usada em dados com distribuições log – normais e em estudo quando a equação de regressão é uma função exponencial ou modelo potência, mas o emprego desta transformação também pode ser recomendado, com o objetivo de estabilizar as variâncias, sempre que há proporcionalidade entre médias e os desvios padrões. Nestas condições, pode ser escrito que: $\sigma_X^2 = k\ m_X^2$ e $\sigma_Y^2 \approx k\ m_X^2 \left(\frac{d\ log_e X}{dX}\right)_{m_X}^2 = k\ m_X^2 \left(\frac{1}{m_X}\right)^2 = k$. Esta situação é bastante comum em dados oriundos de material biológico, quando estão sendo avaliados resultados de respostas em processo de crescimento ou de multiplicação. No que se refere ao desvio padrão, a variabilidade observada é então muitas vezes proporcional à média, o que implica, evidentemente, que o coeficiente de variação correspondente é constante. Na prática, os pesquisadores geralmente utilizam a relação $y = log_{10} x$, ou ainda $y = log_{10}(x + 1)$. O uso desta segunda relação deve ser aplicado quando os valores observados são números inteiros cuja média ou médias se aproximam de 0. De resto, neste caso, são relativamente frequentes os valores observados nulos, o que impede a utilização da primeira transformação. Esta transformação também é aplicável ao caso, relativamente raro, em que é necessário estabilizar a variância de uma série de variâncias observadas. Para uma população normal, o erro padrão da variância é, com efeito, proporcional a média, uma vez que $E[S^2] = \frac{(n-1)\sigma^2}{n}$ e $Var[S^2] = \frac{2(n-1)\sigma^4}{n^2}$. Esta transformação é utilizada quando se constata a proporcionalidade entre as médias e os desvios padrões dos diversos tratamentos, que acarreta coeficientes de variação constantes ou de valores muitos próximos, de tratamento para tratamento. A transformação $log\ X$ ou $ln\ X$, se aplica em casos onde a população de insetos é grande, o que implica normalmente em contagens também grandes para a testemunha e para os tratamentos pouco eficientes, ou seja, numa faixa de variação de 100 pra 10000, por exemplo, enquanto que para os tratamentos que controlam melhor a praga, a faixa de variação é baixa, de 5 a 50 insetos, por exemplo, (SILVA; SILVA, 1999).

Segundo Demétrio (1978), a proporcionalidade entre a média e o desvio padrão é encontrado geralmente quando os efeitos são multiplicativos em lugar de aditivos. Nessa situação, a transformação $log\ x$ ou $ln\ x$, além de estabilizar a variância produz aditividade nos efeitos, e tende a normalizar a distribuição dos erros. Ainda segundo Demétrio, (1978), a base 10 para o logaritmo é mais usada, por conveniência, contudo qualquer base é satisfatória, e recomendada que a transformação seja usada para números inteiros positivos que cobrem uma grande amplitude, sendo que, não pode ser usado quando os valores estiverem numa faixa abaixo de 10. Nesse caso, a transformação mais indicada é $log\ (X\ + 1)$.

De acordo com Ayres et al. (2007), a transformação logarítmica, por exemplo, é indicada quando os valores numa distribuição linear são multiplicativos, como crescimento bacteriano, ou quando a distribuição é muito assimétrica, sendo indicada também em dados de pH. Esta transformação não

admite, obviamente, valores negativos ou nulos. Quando o escore for zero, o programa adiciona um (1) a todos os valores.

Almeida (2005), afirma que quando se quer usar uma regressão curvilínea a forma mais direta é fazer uma inspeção visual nos dados. Fazer um gráfico de pontos, se o gráfico não se apresenta como linear (alguma indicação de linearidade), use então uma regressão curvilinear. Ou então quando há outras razões para suspeitar que as relações não são lineares, como por exemplo, fenômenos claramente modelados por leis de potência, ou lei exponencial, etc, então as Relações devem ser convertidas para formas lineares. Segundo o autor existem muitos tipos possíveis de regressão curvilínea, baseados numa variedade de relações entre as variáveis: $y = bx^a, y = a + b/x, y = ab^x$. Existem várias outras possibilidades. Transformações para relações lineares podem ser usadas assim: Use qualquer transformação que leve a representar a relação através de funções de forma linear, como: logaritmos, multiplicação, divisão, etc. Exemplo quer se obter $y = a + b^{\{*\}}$, ou uma forma similar. Na transformação, usar alguma função da variável resposta y em lugar do próprio y. Regressão curvilinear é um exemplo dessa transformação.

$$Y = \frac{a}{b}x^{a+1}, \ln Y = \ln\left(\frac{a}{b}\right) + (a+1)\ln x \text{ e } Y' = A + Bx'$$

Conforme ainda Almeida (2005), as técnicas tem aplicação mais geral como outras transformações lineares sendo não linear em linear. $Y = a + \frac{b}{x} \Rightarrow Y = a + b\left(\frac{1}{x}\right)$; $Y = \frac{1}{(a+bx)} \Rightarrow \frac{1}{Y} = a + bx$; $Y = \frac{x}{(a+bx)} \Rightarrow \left(\frac{x}{y}\right) = a + bx$; ou ainda outros caos tais como, $Y = axb^x \Rightarrow \ln Y = \ln a + x \ln b$, ou ainda, $Y = a + bx^n \Rightarrow Y = a + b(x^n)$. O autor responde a pergunta quando transformar? Assim: Quando as propriedades conhecidas do sistema medido sugerem uma transformação. Quando o intervalo dos dados medidos cobre várias ordens de grandeza. Quando a hipótese de uma variância homogênea dos resíduos é violada. A transformação devido a homocedasticidade deve ser feita: Se em um gráfico de pontos a diferença dos resíduos versus a resposta prevista não é homogênea. Então os resíduos são ainda uma função das variáveis previsoras. A transformação da resposta pode resolver o problema. Então qual transformação deve ser usada, e assim segundo o autor verificar o seguinte: Calcule o desvio padrão dos resíduos. Coloque num gráfico de pontos esses desvios como função da média das observações. Considere múltiplos experimentos para um conjunto de valores previsores. Verifique a linearidade, se há essa linearidade, então use a transformação logarítmica. Outros testes podem ser usados para verificar a necessidade de transformação: Se a variância versus a média das observações medidas é linear, use uma transformação de raiz quadrada. Se o desvio padrão versus o quadrado médio é linear, use uma transformação inversa. Se o desvio padrão versus a média elevada a uma potência é linear use uma transformação de potência. O princípio geral das transformações é: Para uma função observada $S = g(\bar{y})$ se $h(y) = \int \frac{1}{g(y)} dy$ transforme para $w = h(y)$. Por exemplo, uma transformação logarítmica: Se

o desvio padrão versus a média é linear, então $g(y) = ay$, assim $h(y) = \int \frac{1}{ay} dy = a \ln y$. Em estudos de regressão segundo este autor na verificação de não linearidade desenhe o gráfico de pontos, se não for linear, verifique as possibilidades curvilineares e suas transformações. O uso de uma regressão linear quando as relações entre resposta e previsores não são lineares é um erro.

As funções linearizáveis podem ser usadas no ajuste de modelos de regressão quando os modelos lineares não são adequados. O modelo de regressão linear simples é útil em muitas situações reais. Um caso que merece menção e que recai naquele modelo é o das funções linearizáveis. Certas funções, mediante transformações convenientes, linearizam-se, o que torna simples a solução do problema de regressão. Assim, por exemplo, se admitirmos que a função de regressão seja uma função exponencial do tipo $y = \alpha\beta^x$, a aplicação de logaritmos promove a linearização da função na forma $logy = log\alpha + xlog\beta$. Um artifício simples para se saber se a transformação logarítmica promove uma boa linearização consiste em usar papéis monologarítmicos ou dilogarítmicos. No presente caso, a adequabilidade da transformação vista seria evidenciada se, plotando-se os valores de x segundo escala linear e os de y segundo a escala logarítmica de um papel monologarítmico, os pontos observados se aproximassem de uma reta. Chamando $z = log\ y$, $A = log\alpha$ e $B = log\beta$ e, passamos a ter o problema de estimar os parâmetros da reta $z = A + Bx$, o qual sabemos resolver. Para tanto, basta trabalhar com os valores x_i versus $z_i = logy_i$, obtendo as estimativas de A e B, cujos antilogaritmos serão as estimativas de α e β. Analogamente, se tivermos $y = \alpha x^\beta$, o problema se resolverá trabalhando-se com $\log y_i$ versus $\log x_i$. Outros casos de fácil linearização podem também ser encontrados, como, por exemplo, se $y = a + bx^2$, $y = (a + bx)^{-1}$, etc. Cuidado especial, entretanto, deve ser tomado quanto o processo de linearização envolve transformações na variável dependente Y. O autor ainda comenta que n análise de variância deve ser feita a análise de resíduos, pois a análise de variância considera que as observações analisadas sigam distribuição, com variância constante. Isto, em geral, é verificado através da análise dos resíduos (ou erros) calculados entre as observações e a média do grupo ao qual elas pertencem. Um resíduo é a diferença entre uma observação $y_{i,j}$ e o seu valor estimado (ou ajustado) a partir do modelo estatístico estudado, $\hat{y}_{i,j}$. Para que o modelo de análise seja válido, os resíduos devem ser aleatórios, normais, independentes (não correlacionados) e identicamente distribuídos. Se um modelo está correto, os resíduos devem estar desestruturados; em particular, eles não devem estar correlacionados com qualquer variável incluída na resposta predita. Uma anomalia que ocasionalmente ocorre neste tipo de gráfico é o da Variação não constante. Às vezes, a variância das observações aumenta com o incremento da magnitude das observações. Isto viola o princípio da homogeneidade da variância. Os gráficos que representam a heterocedasticidade apresentarão resíduos dispostos em forma funil ou de borboleta, o que caracteriza um padrão de tendência. Variância não constante também ocorre quando os dados não são normalmente distribuídos, seguindo, portanto, distribuições assimétricas. Nesse tipo de distribuição a variância tende a

ser função da média. Os resíduos devem apresentar comportamentos aleatórios quando dispostos em uma ordem temporal. Se qualquer padrão não-aleatório for identificado nestes tipos de gráficos, uma transformação de dados (Log, Raiz Quadrada, Box-Cox) na resposta Y deve ser realizada. A tendência de se ter resíduos positivos ou negativos indicam uma correlação positiva, o que implicaria na violação da hipótese de independência dos resíduos. Uma aleatorização apropriada do experimento é um bom procedimento para a garantia dessa independência. Algumas vezes, a habilidade do experimentador em executar o experimento pode mudar com o progresso da experimentação, ou, conquanto o processo esteja sendo estudado constantemente, pode se tornar mais inconsistente. Isto sempre resultará em uma mudança no erro da variância ao longo do tempo. O gráfico que surge dessa situação revela uma relação resíduo versus ordem com mais variação do que as outras. A utilização dos Resíduos Padronizados gera mais informação sobre a qualidade da resposta do que os resíduos ordinários. Os Resíduos Padronizados (d) podem ser obtidos dividindo-se os resíduos das observações em relação à média das réplicas (FIT) pelo desvio padrão das réplicas (SE Fit). Deste modo, tais resíduos possuirão média igual a zero e desvio padrão unitário. Consequentemente, eles serão úteis na identificação de outliers. A maior parte dos resíduos padronizados deve encontrar-se no intervalo $-3 \leq d \leq 3$, e qualquer observação com resíduo padronizado fora deste intervalo pode representar uma resposta observada potencialmente não usual ou incorreta. Estes outliers devem ser examinados com cuidado: tanto pode indicar erros de registro, quanto algo mais preocupante. Uma outra maneira de verificar a normalidade dos resíduos é proceder ao teste de normalidade, disponível em muitos pacotes estatísticos computacionais. Um valor de p-Value maior do que o nível de significância de 0,05 (5%) indica que os resíduos são normais (BALESTRASSI; PAIVA 2007).

Matos, (1995), Descreve a importância do uso de transformação de dados afirmando que em alternativa ao uso das variáveis originais ("raw"), podem ser usadas variáveis centradas ("centered"), estandardizadas ("standardized") ou com norma unitária ("unit lenght"), obtidas através das transformações indicadas a seguir. Todos estes procedimentos visam compatibilizar, de algum modo, variáveis que podem ter escalas e dispersões muito diferentes. Em particular, a comparação da influência relativa das diversas variáveis, com base nos parâmetros estimados, só faz sentido se as variáveis forem normalizadas. Como se verá noutro local deste texto, os resultados obtidos depois de qualquer das transformações que se descrevem a seguir são sempre iguais aos da versão com os dados originais. Também os parâmetros têm relações simples entre si, permitindo passar facilmente de uma formulação a outra. Por exemplo, na centragem uma transformação simples consiste em centrar cada variável em relação à sua média. A variável transformada M_k obtém-se de X_k através de: $m_{ik} = x_{ik} - \bar{X}_k$. Semelhantemente ao que se fez para X, também aqui se define $M = [m_1, m_2, \ldots m_p]$. Estandartização: A estandardização corresponde a uma transformação para média nula e desvio padrão unitário de cada

variável original X_k. A nova variável Z_k é obtida através de: $z_{ik} = \frac{x_{ik} - \bar{X}_k}{s_k}$. Neste caso, define-se $Z = [z_1, z_2, \ldots z_p]$. Norma unitária: Esta transformação substitui os valores de cada variável X_k por uma nova variável W_k, obtida pela seguinte regra: $w_{ik} = \frac{x_{ik} - \bar{X}_k}{d_k}$. Definindo aqui também $W = [w_1, w_2, \ldots w_p]$, verifica-se que a matriz $W' \cdot W$ apresenta diagonal unitária (daí o nome da transformação). Os restantes elementos $(W' \cdot W)_{uv}$ correspondem à correlação entre X_u e X_v. Note-se ainda que $Z' \cdot Z = (n-1)W' \cdot W$. A Estimação de parâmetros: A estimativa não tendenciosa de b pelo método dos mínimos quadrados é dada por: $\hat{b} = (X'_a . X_a)^{-1} X'_a y$. No caso de variáveis centradas, estandardizadas ou de norma unitária, o processo de obtenção da estimativa dos parâmetros b_0 utiliza uma expressão análoga à anterior, substituindo-se X_a respectivamente por M, Z ou W. A estimativa de a é, em todos esses casos, igual à média de Y. Os valores de $\hat{b}_k$ obtidos se as variáveis forem centradas são iguais aos do caso geral. Para variáveis estandardizadas e de norma unitária, cada $\hat{b}_k$ vem multiplicado respectivamente por s_k e d_k em relação ao caso geral. A menos de erros de arredondamento, os valores estimados com qualquer dos modelos são rigorosamente correspondentes. Segundo ainda o autor na análise de resíduos para verificar a adequação do ajuste do modelo de regressão o pesquisador tem que verificar que de acordo com os pressupostos da regressão, os resíduos devem distribuir-se aleatoriamente em torno de 0, tanto no modelo global como em relação a cada variável. Caso tal não se verifique, será normalmente necessário alterar o modelo, incluindo ou retirando variáveis, ou realizando alguma transformação que adeque melhor o modelo aos dados (por exemplo, X_k^2 em vez de X_k). Verificação de pressupostos: Apresentam-se, a seguir, alguns testes que permitem verificar se os pressupostos em relação aos erros do modelo são verificados pelos resíduos. Trata-se de verificações a posteriori que poderão levar à revisão do modelo. Aleatoriedade: Uma forma corrente de verificar a aleatoriedade dos resíduos é o teste às sequências de sinais dos resíduos, através do "runs test" (teste de corridas), importante sobretudo quando as observações dependem do tempo. Considerando apenas os sinais (+ ou -) dos resíduos, pela ordem em que foram recolhidos, haverá n_1 sinais (+), n_2 sinais (-) e r corridas (sequências máximas de sinais iguais seguidos). Na sequência (+ - - + + + + - - - + + -), por exemplo, será $n_1 = 7$, $n_2 = 6$ e $r = 6$. Usando em seguida tabelas para o "runs test", determinam se valores críticos que ajudam a determinar, com nível de significância 5%, se a sequência é ou não aleatória. Em função de n_1 e n_2, as tabelas dão dois valores (inferior e superior) que terão que enquadrar o valor de r. Caso contrário, suspeita-se de não-aleatoriedade. No caso do exemplo, os dois valores são 3 e 12, concluindo-se pela aleatoriedade, uma vez que $3 \leq r \leq 12$. Correlação sucessiva: A verificação de independência é usualmente feita através do teste de Durbin-Watson à correlação entre resíduos sucessivos. O teste é útil, sobretudo em dados dependentes do tempo. Heteroscedaticidade: A detecção de desigualdades de variância dos erros pode ser realizada a partir de um gráfico dos resíduos r_i em função dos $\hat{y}_i$. Se o aspecto não for uma mancha

de largura uniforme, por exemplo alargando com o aumento de $\hat{y}_i$, poderá ser necessário transformar Y (lnY, $\frac{1}{Y}$, etc.) ou alterar o modelo. Um gráfico semelhante, mas dos quadrados dos resíduos, pode confirmar suspeitas e ajudar a detectar isolados. Normalidade: A verificação visual da normalidade é feita ordenando os resíduos de forma crescente, e desenhando-os em papel de distribuição normal. Se a presunção de normalidade se verificar, os resíduos deverão estar aproximadamente em linha reta. Expressão do modelo: São úteis alguns gráficos de resíduos em relação a variáveis, para verificação visual da correção da expressão do modelo. Os gráficos potencialmente mais interessantes são: Resíduos em função das variáveis. Permitem verificar se é necessário transformar as variáveis (lnX, $\sqrt{X}$, etc.). Resíduos em função de produtos de variáveis. No caso de ser detectado um padrão, deve ser incluído no modelo um novo termo com o produto em causa ($X_u \cdot X_v$, por exemplo); Resíduos parciais. Gráfico dos resíduos obtidos sem incluir X_k, em função de X_k. Permitem detectar não-linearidades que levem à transformação de X_k. Se o ajuste for bom, o gráfico tem o aspecto de uma reta com inclinação igual ao parâmetro da variável na regressão.

Conforme Nunes (1998) a validade de testes de significância na análise da variância requer o atendimento de alguns pressupostos, ou seja, os erros experimentais devem ser normalmente distribuídos com média zero e variância comum e serem independentes entre si e dos efeitos dos tratamentos, e que adicionalmente, a escala de mensuração deve ser tal que as observações possam ser representadas por um modelo aditivo. Segundo ainda o autor, as transformações de dados são alterações sistemáticas em um conjunto de dados, as quais têm por objetivo, obter a homogeneidade das variâncias, obter aditividade dos efeitos de tratamentos, obter a normalidade dos erros de observação e aumentar a eficiência da estimativa da média. O autor afirma ainda que esses objetivos nem sempre são alcançados, pois nem sempre é possível determinar qual a transformação apropriada, e que nestes casos, outros tipos de análise devem ser tentadas. De acordo com o autor as principais transformações são as seguintes: A transformação raiz quadrada$\left(\sqrt{X}\right)$ cujo emprego serve para dados experimentais consistindo de pequenos números inteiros, tais como número de plantas daninhas em parcelas tratadas com herbicidas, número de insetos de determinada espécie em certa unidade de área, número de colônias de bactérias em uma placa de Petri, que seguem geralmente a distribuição de Poisson. Nesta distribuição, a variância é teoricamente igual a média. Em tais circunstâncias, o erro experimental é muito grande, prejudicando a sensibilidade dos testes de significância. Diferenças reais, quando da comparação entre médias dos tratamentos, deixam de ser destacadas. Em casos como esses, a análise da variância se torna mais eficiente tomando-se antes a raiz quadrada dos números. Pode ainda ser empregada em dados de percentagem calculadas a partir de um mesmo número total, quando os valores podem ser agrupados entre os intervalos 0 a 20% ou entre 80 a 100%, mas não ambos. As percentagens entre 80 e 100% devem ser subtraídas de 100 antes da transformação. Pode ainda ser utilizada em dados de percentagens dentro dos

mesmos limites quando os dados estão em uma escala contínua. É conveniente observar que, se os valores envolvidos são muito pequenos, $\sqrt{X}$ tende a super corrigir a distribuição dos dados, resultando que os intervalos dos valores pequenos transformados se tornem maiores do que o dos valores transformados de médias maiores. Por isso, quando alguns valores envolvidos estão abaixo de 10, e, em especial, se zeros estão presentes, recomenda-se a sua transformação para $\sqrt{X+\frac{1}{2}}$. As médias dos tratamentos devem ser expressas na escala original devendo-se, para isso, elevar ao quadrado as médias de tratamentos computadas a partir de $\sqrt{X}$. Lembrado, finalmente, que os dados transformados devem satisfazer ao modelo aditivo. Ainda aquele autor afirma que a transformação logarítmica ($log\ X$) é empregada quando as variâncias são proporcionais aos quadrados das médias dos tratamentos, e assim a transformação para Log X equaliza as variâncias. Efeitos que são multiplicativos na escala original tornam-se aditivos na escala logarítmica. Em geral, usam-se logaritmos decimais. Todavia, qualquer base é satisfatória. O seu emprego é também feito em dados que consistem de números inteiros entre amplos limites de magnitude. Também em alguns tipos de trabalhos experimentais em que as variáveis a serem submetidas à análise da variância são variâncias, ou seja, analisam-se l$og\ S^2$. A transformação logarítmica não pode ser usada para valores zero. Quando os valores são inferiores a 10, em geral, é desejável uma transformação que produza o resultado de $\sqrt{X}$ para valores pequenos e, ao mesmo tempo, se comporte como logaritmos para valores grandes. $log\ (X+1)$ se comporta como $\sqrt{X}$ até $X = 10$ e pouco difere de $log\ X$ a partir daí. Finalmente o autor afirma ainda que no caso da transformação angular ou arco seno ($\text{arcseno}\sqrt{X}$) esta foi desenvolvida para emprego em proporções binomiais. Se um sucesso a_{ij} entre n sucessos possíveis é obtido na repetição j do tratamento i, a proporção $\hat{p}_{i,j} = \frac{a_{i,j}}{n}$ tem por variância $\frac{p_{i,j}(1-p)}{n}$. Tabelas especiais construídas por Bliss (1934) e Bliss (1938) substituem $\hat{p}_{i,j}$ por ângulos cujo seno é $\sqrt{\hat{p}_{i,j}}$. Nesta escala angular, as proporções próximas de 0 ou 1 são ampliadas de forma a aumentar suas variâncias. Se todas as variâncias do erro forem binomiais, a variância do erro na escala angular será uma constante aproximadamente igual a $\frac{821}{n}$. Se n varia, uma análise ponderada na escala angular é recomendada uma vez que a transformação não elimina inequalidade de variâncias decorrentes de diferentes valores de n. Para amostras com $n < 50$, uma proporção zero deve ser considerada igual a $\frac{1}{4n}$ antes de ser feita a transformação angular, assim como uma proporção 100% deve ser considerada igual a $\left(n-\frac{1}{4}\right)n$. A transformação angular pode ser usada, também, para proporções sujeitas a outros tipos de variação que não a binomial, desde que haja motivos para se pensar que $\hat{p}_{i,j}$é algum múltiplo de $p_{i,j}(1-p_{i,j})$.

Segundo Zimmermann (2004), a transformação logarítmica obtida pela equação $[log(X+K)]$, é adequada em três situações, onde as variâncias são proporcionais ao quadrado das médias, os efeitos multiplicativos na escala original que, portanto, se tornam aditivos com a transformação, e medidas em valores inteiros que cobrem uma amplitude muito grande. Conforme ainda o autor teoricamente qualquer base logarítmica pode ser adotada, e até o surgimento dos computadores, a base 10 era a mais usada por conveniência ou tabelas mais facilmente disponíveis. Neste tipo de transformação, a constante k pode assumir, no geral, dois valores 0 (zero) e 1 (um). O valor 1(um) é empregado obrigatoriamente quando da ocorrência de zeros ou optativamente coma ocorrência de valores menores que 10.

De acordo com Sampaio (2002), quando uma resposta muito instável é medida sob diferentes tratamentos, é comum observar-se um aumento de instabilidade à medida que o valor médio observado no tratamento aumenta. Neste caso observa-se uma proporcionalidade entre a média do grupo experimental e seu respectivo desvio padrão. Quando esta relação for observada, a transformação logarítmica será a recomendada, e se X for a resposta medida, ela deverá ser analisada como $log\ (X)$ ou $log\ (X+1)$. Pode ser usada qualquer base, ou seja, a decimal ou neperiana e se houver algum resultado zerado, deve-se somar uma unidade a todos os valores X pois $log\ (0)$ é indeterminado. Segundo aquele autor ainda, na experimentação animal, variáveis como a contagem de ovos por grama de fezes, usada em parasitologia, de ovos por postura enumerado na malacologia, número de carrapatos por animal contados em epidemiologia, número de células apoptóticas por campo contados na histologia, número de insetos capturados por armadilha avaliados na entomologia, são todas sujeitas a este tipo de transformação.

Uma das finalidades da transformação, além de normalizar a distribuição dos erros e produzir aditividade entre os efeitos dos fatores, em função do modelo matemático, é estabilizar as variâncias dos tratamentos, se a heterocedasticidade for regular, isto, é, se esta é devida a falta de normalidade dos erros experimentais. Neste caso, um teste bastante simples que deve-se aplicar, antes e depois de transformados os dados, é o de Hartley, ou o teste da razão máxima, estatística que baseia-se no quociente entre a variância máxima e a mínima, entre os tratamentos estudados, como se segue: $H_C = \frac{S^2_{max}}{S^2_{min}}$, onde, S^2_{max}= maior variância; e S^2_{min}= menor variância. As hipóteses estatísticas para esse teste, cujos valores críticos ou tabelados de $H_{(g;r-1)\alpha}$, estão na tabela 31 de Pearson e Hartley (1970) ou Tabela 8 de Banzatto e Kronka (2006), onde, g é o número de grupos ou tratamentos e r é o número de repetições, se iguais para todos os tratamentos, ou o quociente entre o total das repetições de todos os tratamentos e o número de tratamentos g. Sendo assim se $H_C < H_{(g;r-1)\alpha}$, aceita-se a hipótese de homocedasticidade ao nível α de probabilidade, e conclui-se que a esse nível de significância os grupos ou os tratamentos são homogêneos, quanto as variâncias. Por outro lado, se $H_C \geq H_{(g;r-1)\alpha}$, rejeita-se a hipótese de homocedasticidade ao

nível α de probabilidade, e conclui-se que a esse nível de significância os grupos ou os tratamentos não são homogêneos, quanto as variâncias (SILVA; SILVA, 1999).

Conforme Radtek (2015) a transformação de dados em algumas situações é necessária, pois o pesquisador ao realizar a transformação dos dados amostrais, tem como objetivo atingir determinadas exigências de certos testes estatísticos (pressupostos). a transformação obtida geralmente melhora a aproximação dos dados à distribuição normal. a normalidade dos dados é uma exigência comum para a aplicação de testes de hipótese. se a suposição de normalidade dos dados não é aceitável, podemos adotar a estratégia de transformação da variável. Transformações são nada mais do que uma forma de reescrever os dados numa unidade diferente. em muitas situações práticas a escolha da transformação para melhorar a aproximação à distribuição normal não é óbvia. Segue abaixo algumas transformações comumente utilizadas:

Para dados de contagens: $\sqrt{X}$ Contribui para tornar as variâncias muito menores e desta forma mais facilmente obter homocedasticidade (variâncias iguais).

Para dados de proporções: $\frac{1}{2} log\left(\frac{X}{1-X}\right)$ou arc sen ($\sqrt{X}$) - Contribuem para alterar a forma da distribuição dos dados.

Para estudos de correlações: Fisher: $z(x) = \frac{1}{2} log\left(\frac{1+x}{1-x}\right)$

Para dados de concentrações: $log(x)$ ou $ln(x)$. Contribui para tornar as variâncias muito menores e desta forma mais facilmente obter homocedasticidade.

A transformação angular ou arco seno é muito utilizada em problemas relativos a proporções, a qual é dada por $y = 2\ arc\ sen\ \sqrt{\frac{x}{n}}$, ou ainda, $y = 2\ arc\ sen\ \sqrt{x'}$. O seu emprego justifica-se nas variáveis que possuem distribuições binomiais, sempre devido a mesma propriedade. Pode-se generalizar ainda o uso desta transformação nas razões compreendidas entre 0 e 1, mesmo quando se trata de variáveis continuas. Vale salientar que o emprego dessa transformação só pode ser recomendado quando o valor de n ou o denominador da razão é constante ou sensivelmente constante. O uso dessa transformação para tamanhos variáveis é abordado principalmente por Gabriel (1963). Na prática deve ser feito o uso de correções nos casos extremos tais como $\frac{x}{n} = 0$ e $\frac{x}{n} = 1$, sendo que estes valores podem ser substituídos respectivamente por, $\frac{1}{4n}$ e $1 - \frac{1}{4n}$. Alguns autores sugerem também que se utilize o fator $\frac{3}{8}$, em vez de $\frac{1}{4}$, como Anscombe (1948) e De Munter (1958). Esta transformação é normalmente recomendada quando os dados são expressos em porcentagem, como por exemplo, porcentagem de plantas com um determinado sintoma de doença, porcentagem de estacas enraizadas, porcentagem de sementes germinadas, porcentagem de raízes com sintomas de nematoides, porcentagem de pintos dentro de uma faixa de peso, porcentagem de plantas com deficiência de um nutriente, etc. Quando os dados

são dessa natureza, normalmente têm distribuição teórica de probabilidade binomial, e se justifica a transformação aplicando-se a expressão $arc\ sen\sqrt{\frac{X}{100}}$, ou aplicando-se diretamente tabelas onde com o valor da porcentagem se obtém a transformação correspondente, tais como as tabelas publicadas por Fisher e Yates, denominadas de Tabelas estatísticas para biologia, medicina e agricultura. Se as porcentagens estiverem numa faixa entre 30 % e 70 %, torna-se desnecessária a transformação, podendo os dados originais serem analisados diretamente. Isto é, a transformação $arcoseno$ ou angular produzirá sensíveis alterações se as porcentagens estiverem entre 0 % e 20 %, ou entre 70 % e 100 %, neste caso a transformação da raiz quadrada é indicada, sendo que nessa situação os dados devem ser subtraídos de 100 antes da transformação. Além do caso das porcentagens estarem entre 30 % e 70 %, é também, desnecessária a transformação, quando as porcentagens são resultantes da divisão dos valores observados nas parcelas por um valor constante, como, por exemplo, a média do tratamento testemunha, ou quando o denominador é a observação maior, o que dá para esses dados 100%, e as demais porcentagens abaixo disso, ou quando são representativos de concentração, como o teor de nitrogênio (N), o teor de proteína do trigo, o teor de sacarose na cana - de - açúcar, etc. Na realidade a transformação angular deve ser aplicada as porcentagens obtidas de quocientes entre variáveis discretas, como o número de sementes germinadas sobre o total de sementes plantadas, o número de plantas com sintomas típicos da doença sobre o número total de plantas observadas, etc. e preferencialmente, quando as porcentagens forem calculadas, para todos os dados, sobre o mesmo denominador (SILVA; SILVA, 1999).

Conforme Zimmermann (2004), a transformação $arcoseno$ determinada pela fórmula $(arcsen\sqrt{X})$, é especialmente recomendada para dados expressos em porcentagem, e principalmente para casos em que os dados estejam nas faixas entre 0 % e 20 % ou entre 80 % e 100 %. Os valores transformados tanto podem ser expressos em graus como em radianos. Se os valores em porcentagem estiverem todos entre 20 % e 80 %, pouco ou nada se ganhará ao fazer-se a transformação. A variância dos novos valores é aproximadamente constante, sendo igual a $\frac{821}{n}$, quando expressos em graus, e igual a $\frac{0,25}{n}$, se expressos em radianos (STEEL; TORRIE, 1981). Isto mostra que se pressupõem um denominador constante, mas esta transformação é comumente empregada quando os denominadores são diferentes.

Segundo Sampaio (2002), as respostas percentuais possuem um universo restrito aos limites de 0 a 100 %. Pode-se perceber que se uma população apresentar um valor médio percentual entre 30 e 70 %, haverá maior chance de, observado o valor de seu desvio s, encontrar 95 % dos valores desta população distribuídos simetricamente em torno do valor médio central. Por exemplo, se uma resposta percentual tal como motilidade do sêmen bovino apresentar um valor médio de 83% e um desvio de 8 %, o intervalo de respostas típicas estará teoricamente entre os valores 83 ± 1,96 (8), ou seja, de 67,3 a 98,6 %. Este

intervalo não viola o limite superior possível de motilidade (100 %) e, portanto, há uma grande chance de existir uma distribuição normal de valores naquele intervalo. Por outro lado, se a percentagem de defeitos de peça intermediária dos espermatozoides fosse estudada nas mesmas amostras, poderíamos encontrar um valor médio de 6 % com um desvio de 10 %. O intervalo de respostas típicas deveria ser, portanto 6 ± 1,96 (10), ou seja, de – 13,6 a 25,6 %. Como o limite inferior para dados percentuais é 0 (zero), conclui-se que uma distribuição que teoricamente seria simétrica entre -13,6 e 25,6 % não poderá manter a simetria quando os limites típicos observados forem 0 e 25,6%. Sendo uma distribuição assimétrica cai uma das premissas para garantir a análise de variância. Deve-se considerar que, se em um ensaio, pelo menos um dos grupos experimentais violar o intervalo real de 0 a 100% ao ser definido seu intervalo de valores típicos, então deve-se fazer uma transformação angular para cada observação X, dada por: $arcsen\ \sqrt{X}$, onde X precisa ser expresso em $\frac{percentual}{100}$, ou seja, X ≤ 1. Por exemplo, se uma observação for 9 % de defeito de peça intermediária, seu valor transformado seria $arcsen\ \sqrt{0{,}09} = arcsen(0{,}30) = 17{,}4576 \approx 17{,}46.$ Nesta situação se está utilizando a transformação em graus, embora possamos utilizar radianos. A leitura acima indica que o seno do ângulo de 17,46° é igual a 0,30. A expressão matemática em $arcsen\ \sqrt{X}$ significa que o arco ou o ângulo cujo seno é igual a 0,30 é o de 17,46°. Como foi efetuada a raiz quadrada do valor percentual, que é uma operação de grande poder de redução na variabilidade observada, os valores dos ângulos então analisados dificilmente violarão os limites de 0 (zero) a 90°, que correspondem aos senos de 0 (zero) a 19um), quando for definido um intervalo de respostas típicas. Ainda de acordo com Sampaio (2002), é preciso esclarecer que quando as respostas percentuais são medidas, a amostra investigada que originará o percentual deverá ser substancial, por exemplo, $n \geq 30$. Variáveis percentuais obtidas de baixa amostragem, por exemplo, percentagem de natimortalidade de leitões por leitegada, onde uma leitegada pode conter até 16 leitões, certamente apresentará uma alta instabilidade e a técnica de transformação angular poderá não resultar em boas comparações. Para estes casos, outras estratégias de análise podem ser mais eficientes, como o estudo de tabelas de contingência. Segundo este autor para que se possa controlar variações extremas, quando o contingente amostral (n) que define a percentagem for menor que 50 observações, todo percentual com o valor zero deve substituído por $\frac{1}{4}n$ e os iguais a 100% pelo valor $\left(\frac{n-\frac{1}{4}}{n}\right)$, antes de ser executada a transformação. A transformação só será eficiente quando for observada a violação dos valores limites 0 e 100% ao ser definida o intervalo de respostas típicas, de pelo menos um grupo experimental. A mortalidade de pintos de um dia em linhagens comerciais é geralmente baixa, por exemplo, 8%. Aparentemente há maiores chances de violar o limite 0% pelo fato da média estar tão próxima dele. Entretanto, essa variável é muito pouco instável, com um desvio padrão de 3% e como o

intervalo 8 ± 1,96 (3) não alcançará o valor limite zero então, não será necessário transformar os resultados de mortalidade, desde que este fato ocorra em todos os grupos experimentais.

A transformação argumento tangente hiperbólica é dada por: $z = arg\ \ th\ r = \frac{1}{2} log_e \frac{1+r}{1-r}$, a qual permite normalizar a distribuição de amostragem do coeficiente de correlação linear simples de Pearson r, bem como estabilizar a sua variância, e assim poder aplicar inferências tais como construir intervalos de confiança para o coeficiente de correlação linear populacional rô (ρ), bem como testar a hipótese nula de que este coeficiente é nulo ($H_0: \rho = 0$) .

A transformação argumento seno hiperbólico são empregadas também mas em um tipo particular de famílias de distribuições teóricas de probabilidades como variáveis que possuem distribuição binomial negativa, outras, pelo contrário, são mais gerais, é o que ocorre com as transformações $y = \sqrt{x + a}$; $y = x^a$e $y = log(x + a)$, as quais são casos gerais de outras transformações já comentadas anteriormente como; $y = \sqrt{x}$ e $y = log\, x$. Na literatura especializada são encontradas diversas discussões aprofundadas sobre este assunto bem como fornecem informações complementares importantes (BARTLETT, 1947; GRIMM, 1960; HEALY; TAYLOR, 1962; KLECZKOWSKI, 1949; TAYLOR, 1961; TUKEY, 1957). Conforme Lucio et al. (2012), esta é a transformação que possibilita maior proporção de atendimento das variáveis produtivas e morfológicas aos pressupostos de normalidade, homogeneidade e aleatoriedade dos erros, em dados obtidos de experimentos com tomateiros em túnel plástico e em campo.

De acordo com Bussab (1986) e Weisberg (1980), em estudos de ajuste de equações de regressão, logo após o pesquisador identificar a ausência de homogeneidade nos dados, se faz necessário o uso de métodos alternativos que atendam a esta demanda, e que felizmente, na maioria dos casos a variabilidade segue algum padrão identificável. Com grande frequência esta é uma função da variável preditora, ou de uma outra variável independente, e para resolver esta questão o investigador tem que procurar remover a heterocedasticidade através de transformação da variável resposta y, ou da variável preditora x, ou então de ambas, são as chamadas transformações estabilizadoras da variância. A seguir será apresentado algumas transformações mais usadas nestes casos segundo aquele autor, bem como algumas recomendações de quando utilizá-las. A transformação raiz quadrada ($\sqrt{y}$) é recomendada quando a variância do erro isto é $Var(e_i)$ cresce proporcionalmente em relação a variável independente x_i. Por outro lado a transformação logarítmica ($log\, y$), é sugerida o seu uso quando o crescimento da variância do erro $Var(e_i)$ é mais acentuado do que o caso anterior, isto é, a variância cresce proporcional ao valor x_i^2, e por último a transformação $\text{arcsen}\sqrt{y}$, é recomendada quando a variável resposta y é do tipo proporção, isto é, $0 \leq y \leq 1$.

Na transformação $logit$ utiliza-se uma propriedade interessante da função logística a qual é a possibilidade dela poder ser linearizada. Denotando-se $E(Y)$ por π, pois a resposta média é a

probabilidade quando a variável resposta é binária. Fazendo-se a transformação: $\pi' = log_e\left(\frac{\pi}{1-\pi}\right)$, então, obtêm-se $\pi' = \beta_0 + \beta_1 X$. Esta transformação é chamada de transformação logit da probabilidade π. A razão $\pi/(1-\pi)$ na transformação $logit$ é chamada de Odds (Chance). A função resposta transformada $\pi' = \beta_0 + \beta_1 X$ é denominada como função resposta $logit$, e π' é denominada de resposta média $logit$. Observe em $\pi' = \beta_0 + \beta_1 X$ que: $-\infty \leq \pi' \leq \infty$ para $-\infty \leq X \leq \infty$ (AYRES et al., 2007).

Para obter a transformação $probit$ considere que uma função de resposta curvilínea com a mesma forma da função logística $E(Y) = \frac{exp(\beta_0+\beta_1 X)}{1+exp(\beta_0+\beta_1 X)}$ ou equivalente a $E(Y) = [1 + exp(-\beta_0 - \beta_1 X)]^{-1}$, é obtida transformando π por meio da distribuição normal acumulada. Esta transformação é chamada de transformação $probit$. O modelo de regressão $probit$ é menos flexível do que a regressão logística pois não pode ser diretamente aplicada com mais de uma variável preditora. A distribuição de probabilidade acumulada é dada por: $P(T < X) = \Phi(\beta_0 + \beta_1 X)$. As transformações $logit$ e $probit$ são muito utilizadas em inferência estatística para duas dimensões na interpretação dos resultados de ensaios biológicos como, por exemplo, em ajuste de modelos de regressão (AYRES et al., 2007).

Outra transformação bastante aplicada é aquela onde uma outra função de resposta curvilínea é usada e ela é denominada de transformação complemento $log\ log$ da probabilidade π dada por: $log_e(-log_e(1-\pi))$. Diferentemente das transformações logit e probit esta transformação não é simétrica em torno de $\pi = 0{,}5$ (AYRES et al., 2007).

Já a transformação de Box–Cox (1964), conforme Ayres et al., (2007), é usada quando a distribuição normal não se adequa aos dados, muitas vezes é útil aplicar a transformação de Box - Cox para obtermos a normalidade. Considerando $X_1, \ldots, X_n$ os dados originais, a transformação de Box - Cox consiste em encontrar um λ tal que os dados transformados $Y_1, \ldots, Y_n$ se aproximem de uma distribuição normal. Esta transformação é dada por: $Y_i = \begin{cases} ln(X_i), & se \quad \lambda = 0 \\ \frac{X_i^{\lambda}-1}{\lambda}, & se\ \lambda \neq 0 \end{cases}$.

Ainda de acordo com Ayres et al. (2007), Box e Cox desenvolveram este procedimento a fim de estimar a melhor transformação para obter a normalidade de um conjunto de escores. Trata-se de um modelo de função de máxima-verossimilhança, podendo o pesquisador, por esse método, verificar qual a transformação sugerida para seus dados. Recomenda-se, todavia, ponderar as vantagens e as desvantagens da transformação de dados. Quando ocorrer dúvida, é preferível manter os escores originais e utilizar testes não-paramétricos.

A transformação Box – Cox pode ser particularmente útil quando uma variável de resposta não cumpre com os pressupostos de normalidade e/ou homoscedasticidade. Em Economia é comum termos variáveis dependentes que só apresentam uma distribuição (aproximadamente) normal após uma transformação logarítmica ou, menos frequentemente, uma transformação inversa. Esses são casos

particulares da transformação Box–Cox: o primeiro corresponde à fórmula $y^{(\lambda)} = log\, y$ para $\lambda = 0$, e o segundo é um caso particular da fórmula $\begin{cases} y^{\lambda} - 1 \\ y^{(\lambda)} \quad \lambda \neq 0 \cdot \\ \lambda \end{cases}$

A hipótese de trabalho por trás da transformação é equivalente à afirmativa de que existe algum valor de λ para o qual y^{λ} tem distribuição (aproximadamente) normal com média dada pelo produto vetorial $\beta' X$ e variância constante. Nas fórmulas anteriores observa-se que, quando $\lambda = 1$, a transformação deixa a variável praticamente inalterada (a não ser pela subtração da unidade de cada observação, cuja consequência é o deslocamento de toda a distribuição para esquerda nessa quantia, mas mantendo a mesma variância). Por sua vez, como já foi mencionado, λ = 0 implica na transformação logarítmica da variável. Por último, para $\lambda = -1$, temos que $y^{-1} = 1 - y^{-1}$.

Quando as respostas são constituídas de valores nulos, a família de transformações de Box-Cox fica restrita, e o pesquisador pode utilizar a variável somada a uma constante. Este procedimento proposto por Yamamura (1999), com o parâmetro c é o valor constante, gera assim as expressões: $f(y) = \frac{(y+0{,}5)^{\lambda}-1}{\lambda}, \lambda \neq 0$ e $f(y) = ln(y + 0{,}5)\,, \lambda = 0.$

Outros princípios alternativos foram propostos por Hill (1963), Griffiths (1980) e Berry (1987) para determinar o valor dessa constante c. Os procedimentos propostos por esses autores são aplicáveis em um determinado conjunto de pressupostos e requerem cálculos complexos.

Após aplicarmos essa transformação aos dados, as especificações e os parâmetros do processo (média, variabilidade inerente e total) são obtidos para os dados transformados, aplicando a análise via dados normais. Da mesma forma, os índices são calculados para os dados transformados com a distribuição normal. Para verificarmos se a transformação foi eficiente basta analisarmos a normalidade dos dados transformados através do histograma e polígono de frequências, papel de probabilidade normal ou teste não paramétrico de normalidade de Kolmogorov - Smirnov, Qui – quadrado, Anderson-Darling ou outro.

Segundo Ayres et al. (2007), a transformação ordinal (rank) é designada para distribuições muito assimétricas, sendo largamente usada nos testes não-paramétricos-correlação de Spearman, análise de variância de Kruskal-Wallis, etc. Pode-se efetuar a transformação ordinal (rank) pela ordem numérica ou alfabética, de modo crescente ou decrescente. Algumas vezes há necessidade de ordenar um conjunto de colunas, sendo a primeira àquela que conduzirá os respectivos valores das demais. Ainda segundo o autor a transformação recíproca$\left[\frac{1}{X} = X^{-1}\right]$é empregada em pesquisas cujos escores envolvem diluições, como ocorre em titulações sorológicas. Quando houver zeros entre os dados, não deverá ser usada essa transformação.

A escolha de uma mudança de variável pode fazer-se com base em considerações teóricas ou em bases puramente empíricas, no entanto é difícil fornecer regras absolutamente gerias sobre este assunto. Do ponto de vista teórico, é lógico, por exemplo, aplicar uma transformação angular às variáveis de natureza binomial e, por extensão, às proporções e percentagens compreendidas entre 0 (zero) e 1(um) ou entre 0 (zero) e 100, desde que estes valores sejam calculados a partir do mesmo número n de indivíduos ou, de uma maneira mais geral, a partir do mesmo número n de indivíduos ou, de uma maneira mais geral, a partir de denominadores iguais ou quase iguais. Do mesmo modo, os números inteiros resultando de simples enumerações, são muitas vezes valores observados de variáveis de Poisson ou de variáveis com distribuições vizinhas das de Poisson. Geralmente, pode- se aplicar com vantagem uma transformação raiz quadrada. Finalmente, em muitos estudos de crescimento, verificamos que os pesos, incluindo também os rendimentos expressos em pesos, possuem coeficientes de variação sensivelmente constantes, para indivíduos que vivam em condições análogas, e neste caso é lógico, portanto, o uso de uma transformação logarítmica. No entanto seria perigoso aplicar estes escassos princípios sem discernimento, e assim não constitui uma exceção, por exemplo, que a transformação raiz quadrada seja insuficiente para estabilizar a variância dos resultados de certas enumerações, sobretudo no caso das distribuições agregativas. Nestas situações, a transformação logarítmica é muitas vezes mais adequada do que a transformação raiz quadrada. (DAGNELIE, 1973).

Por outro lado, muitas vezes não dispomos de nenhuma razão teórica que permita justificar a priori a utilização de determinada transformação, nestas circunstâncias o pesquisador é obrigado a efetuar uma escolha empírica. É possível orientar esta escolha estabelecendo um diagrama de dispersão das médias e das variâncias, de preferência com escalas logarítmicas. Quando as médias estão marcadas no eixo das abscissas e as variâncias no eixo das ordenadas, podem apresentar-se dois casos típicos. Por um lado, a nuvem de pontos pode estar orientada paralelamente à bissetriz do diagrama, o que mostra que a relação existente entre as médias e as variâncias é do tipo $log\,\sigma_x^2 = log\ \ m_X + log\,k$, ou então $\sigma_X^2 = km_X$ (DAGNELIE, 1973).

A transformação a ser utilizada é então a transformação raiz quadrada. Por outro lado, a nuvem de pontos pode ser orientada paralelamente a uma reta de declive 2, sendo a relação médias – variâncias do tipo: $log\,\sigma_x^2 = 2\,log\ \ m_X + log\,k$ ou então, $\sigma_X^2 = km_X^2$.Neste caso, deve ser empregada como é evidente, a transformação logarítmica (DAGNELIE, 1973).

Portanto é a inclinação da nuvem de pontos que permite orientar a escolha da transformação. Esta inclinação pode ser estimada através do coeficiente de regressão dos logaritmos das varâncias em função dos logaritmos das médias ou, mais simplesmente, pelo coeficiente da reta dos mínimos quadrados, que não é mais do que a razão entre o desvio padrão dos logaritmos das variâncias e o desvio padrão dos logaritmos das médias. Quando as amostras são do mesmo tamanho, pode-se substituir as médias e as variâncias, respectivamente, pelas somas e pelas somas dos quadrados dos desvios. Esta

substituição não altera em nada a inclinação da nuvem de pontos no caso do gráfico de escalas logarítmicas (DAGNELIE, 1973).

Conforme Heath (1981), o emprego de uma estimativa única de variância residual (ou de "erro"), numa análise de variância, pressupõe que os desvios residuais R, para todas as unidades, se distribuam normalmente, tendo a mesma variância "verdadeira" σ^2. A consequência mais importante disso é o fato de que as variâncias residuais, nos diversos tratamentos, embora sujeitas a variações aleatórias, devem independer da magnitude das médias dos tratamentos. Esse, aliás, é o objetivo imediato de todas as transformações; o segundo objetivo é aproximar a distribuição da normal. Se os efeitos dos tratamentos são muito amplos (como ocorre com os efeitos da idade sobre o tamanho de plantas que crescem com rapidez ou com os efeitos da nutrição, nas culturas feitas em água ou em areia), os tratamentos que apresentam médias maiores têm, em geral, as maiores variâncias. Para os dados, relativos ao peso de matéria seca, as estimativas $\sum(x - \bar{x})^2/3$ (que incluíam variações devidas a diferenças entre blocos, em cada caso) foram comparadas, aos pares, por meio do teste; a diferença relativa a NPK foi de longe a maior e significantemente maior do que todas as outras, exceto a diferença relativa a NP. A transformação logarítmica tornou as variâncias muito mais homogêneas e, aparentemente, independentes da magnitude das médias. Essa transformação é apropriada quando a variância para os dados puros é aproximadamente proporcional ao quadrado da média.

Dados que assumem a forma de números inteiros tendem, com frequência, a distribuir-se de acordo com a distribuição de Poisson. Exemplos: números de colônias de bactérias, em cultura em placas de Petri; número de sementes ou insetos, de uma determinada espécie, em espaços quadrados ou em pequenas parcelas; são condições necessárias, em tais casos, que os indivíduos se apresentem aleatoriamente em uma determinada área, independentemente de nessa área existirem ou não outros indivíduos (ou seja, independentemente de atrações ou repulsões mútuas) e é necessário, ainda, que haja um número suficientemente grande de indivíduos para que muitas das áreas examinadas estejam ocupadas por um ou mais indivíduos. Nessa distribuição, a variância é igual a média, de modo que é proporcional a ela. A transformação adequada consiste em utilizar as raízes quadradas dos dados originais apresentados para análise, caso os números (x), relativos a cada área, fiquem situados entre 10 e 100; ou utilizar $\sqrt{(x + 1/2)}$, caos x tenha geralmente valor inferior a 10. Se a maior parte dos números é superior a 100, as transformações serão, provavelmente, desnecessárias. (HEATH, 1981).

Frações ou porcentagens (baseadas em números inteiros), representando a proporção de indivíduos, considerados a partir de um total fixo, pertencentes a uma determinada categoria, tendem a distribuir-se de acordo com a distribuição binomial. A variância, nesse caso, é máxima em 50 por cento, decrescendo simetricamente para zero, quer na direção de zero por cento, quer na direção de cem, por cento. Exemplificando: porcentagens de plantas que florescem como consequência de um tratamento fotoperiódico, em experimento realizado em estufas. Em tais casos, a transformação apropriada é a

angular (arco-seno). A análise é feita com base em ângulos, entre zero e noventa graus, cujos senos são iguais às raízes quadradas das frações. Estes foram tabulados por Fisher e Yates (1963). Todavia, quando as porcentagens (p) ficam situadas entre 30 e 70, a variância é praticamente independente da média e não é necessário efetuar qualquer transformação; quando as porcentagens são todas inferiores a 20, a variância é quase proporcional à média, podendo-se, então, aplicar uma transformação de raiz quadrada, o que também se usa com (100-p), se todos esses valores são maiores do que 80. Outras informações sobre transformações podem ser obtidas em Cochran (1938) e Bartlett (1947) (HEATH, 1981).

Segundo Lima e Lima (2014), uma transformação adequada aos dados é aquela em que a variância da variável transformada não é afetada por mudanças do valor médio. A variável transformada é normalmente distribuída; A escala de transformação é tal que a média aritmética estime imparcialmente a média verdadeira e que a escala de transformação é tal que os efeitos reais são lineares e aditivos. Quando uma transformação de dados é feita, todas as comparações e estimativas de intervalo de confiança devem ser determinadas na nova escala, sendo que as médias podem ser transformadas para a escala original. A mudança exata da escala é, em geral, difícil e a escolha de uma transformação adequada depende, em parte, da experiência do estatístico. Devemos lembrar-nos de verificar se as hipóteses fundamentais foram satisfeitas após a escolha e aplicação de uma transformação de dados. O estudo das relações entre médias e variância de tratamentos pode sugerir uma transformação apropriada, que segundo estes autores pode ser visto em Box e Cox (1964).

Segundo Ribeiro Júnior (2001), as transformações de dados são necessárias quando pelo menos uma das condições da análise de variância não forem satisfeitas. Quando uma transformação for feita, deve-se verificar novamente as condições para a análise e todas as comparações devem ser realizadas na nova escala. Sendo assim, conforme este autor as transformações podem ser as seguintes: i) Transformação raiz quadrada: A transformação raiz quadrada $\sqrt{X}$é feita quando os dados observados de uma variável X seguem distribuição de Poisson, na qual a média e a variância são iguais. Esta distribuição se refere à contagem do número de vezes que ocorre um determinado evento por unidade de tempo ou por uma unidade de medida. Pode também ser usada quando a variância de X é proporcional à média de X e para dados de porcentagem baseados em contagens, sendo a amplitude de 0 a 20% ou de 80 a 100%, mas não ambas. Quando os dados estão situados entre 80 e 100%, eles devem ser subtraídos de 100 antes da transformação. Quando entre os dados ocorrem valores pequenos inferiores a 10 e, principalmente zeros, as transformações recomendadas são $\sqrt{X+0{,}5}$, $\sqrt{X+1}$ ou $\sqrt{X}+\sqrt{X+1}$, como exemplo de variável que pode sofrer transformação raiz quadrada pode ser citado o número de plantas atacadas; ii) transformação logarítmica: a transformação $log\ X$ ou $ln\ X$ é utilizada quando os desvios padrões variam diretamente com as médias dos diversos tratamentos, ou seja, quando o coeficiente de variação é constante de tratamento para tratamento. Esse tipo de relação entre a média e o desvio padrão é encontrado geralmente quando os efeitos são multiplicativos em lugar de aditivos. Essa transformação é

indicada para observações com números inteiros positivos que cobrem uma grande amplitude, sendo que não pode ser usada diretamente quando ocorrem zeros ou quando alguns dos valores são menores que 10. Nesta última, a transformação $log\ X$ mais um, isto, é, $log\ (X\ +\ 1)$ é a mais indicada, como exemplo de variável que pode sofrer transformação logarítmica pode ser citado o número de insetos coletados; iii) transformação angular: este tipo de transformação é recomendável para dados expressos em porcentagens, que geralmente seguem distribuição binomial, ou seja, para aquelas variáveis que apresentam somente dois resultados possíveis em cada avaliação. Porém, se as porcentagens estiverem entre 30 e 70 %, a transformação angular não será necessária. A transformação também será desnecessária quando as porcentagens forem resultantes da divisão dos dados observados por um valor constante ou quando são representativas de concentração. Para uma variável X, esta transformação é dada por $arcsen\sqrt{\frac{X}{100}}$, como exemplo de variável que pode sofrer transformação angular ou arco seno, pode ser citado a porcentagem de danos em plantas.

As principais transformações utilizadas nas análises estatísticas são: Raiz Quadrada, Logarítmica e Angular. Transformação Raiz Quadrada - Esta transformação é utilizada para dados provenientes de contagens como: número de bactérias em uma placa, número de plantas ou insetos em uma dada área, número de defeitos ou acidentes. Geralmente eles se distribuem de acordo com a distribuição de Poisson, em que a média e a variância são iguais. Neste caso, a transformação raiz quadrada dos dados estabiliza a variância, além de torná-la independente da média. A transformação raiz quadrada pode também ser usada com dados de contagens em que a variância de X é proporcional à média de X, ou seja, $\sigma_x^2 = K\bar{X}$. Para a distribuição de Poisson tem-se $K\ =\ 1$ mas, frequentemente, encontra-se $K\ >\ 1$, o que indica que a distribuição dos erros tem uma variância maior que aquela de Poisson. Dados de porcentagem baseados em contagens com um denominador comum, sendo a amplitude de 0% a 20% ou de 80% a 100%, mas não ambas, podem também ser analisados utilizando-se a transformação raiz quadrada. Quando os dados estão situados entre 80% e 100%, devem ser subtraídos de 100 antes da transformação. A mesma transformação é útil para porcentagens na mesma amplitude quando as observações provêm de uma mesma escala contínua, desde que médias e variâncias sejam aproximadamente iguais. Quando entre dados ocorrem valores pequenos, inferiores a 10 e, principalmente, zeros, as transformações $\sqrt{X+1/2}$, $\sqrt{X+1}$ ou $\sqrt{X}+\sqrt{X+1}$ estabilizam a variância mais efetivamente que $\sqrt{X}$, sendo X o valor observado. A transformação raiz quadrada afeta o tipo de achatamento de distribuição de frequência dos erros e a medida de aditividade. Assim, se os efeitos de blocos e tratamentos são aditivos na escala original, geralmente não o serão na escala raiz quadrada ou vice versa. Contudo, a menos que efeitos de blocos e tratamentos sejam ambos grandes, efeitos que são aditivos em uma escala serão aproximadamente aditivos na escala raiz quadrada. As médias obtidas com os dados transformados são reconvertidas para a escala original, utilizando-se a operação inversa, ou seja, sendo elevadas ao quadrado. Os valores

obtidos, geralmente são ligeiramente menores que as médias originais, porque a média de uma série de raízes quadradas é menor que a raiz quadrada da média original (LIMA; LIMA, 2014).

Por outro lado, a transformação logarítmica estabiliza a variância quando o desvio padrão na escala original varia diretamente com a média, ou seja, o coeficiente de variação é constante de tratamento para tratamento. Esse tipo de ralação entre média e desvio padrão é encontrado geralmente quando os efeitos são multiplicados em lugar de aditivos. Nessa situação, tal transformação, além de estabilizar a variância, produz aditividade nos efeitos e tende a normalizar a distribuição dos erros. A base 10 para o logaritmo é a mais usada, por conveniência, contudo, qualquer base é satisfatória. Essa transformação é usada para números inteiros positivos que cobrem uma grande amplitude, sendo que não pode ser usada diretamente quando ocorrem zeros ou quando alguns dos valores são menores que 10. Neste caso, é necessário ter-se uma transformação que equivale à transformação $\sqrt{X}$ para valores pequenos e $log\, X$ para valores grandes de X. A transformação $log(X + 1)$ é a que mais se aproxima da desejada. As médias obtidas na escala logarítmica são convertidas na escala original através da operação inversa, ou seja, utilizando-se antilogaritmos dos valores obtidos para essas médias estando, porém, afetadas de um erro. A transformação Angular ou $arcosen\sqrt{\mathrm{p}/100}$ é utilizada para homogeneizar a variância residual dos dados de proporção X/N, ou porcentagens $100(X/N)$, correspondentes a indivíduos portadores de um dado atributo, em uma amostra de tamanho N é especialmente recomendada quando as porcentagens cobrem uma grande amplitude de valores. Admite-se que as proporções têm distribuição binomial com média igual a μ e a variância igual a $\mu(1-\mu)/N$. Desde que as proporções têm distribuição binomial, essa variância será máxima para $\rho = 0{,}5$. As proporções igualmente afastadas de 0,5 terão variâncias iguais e quanto mais afastadas de 0,5, valores menores. A transformação irá, pois, alterar as porcentagens extremas, ou seja, aquelas de menores variâncias. Snedecor e Cochran (1967), dizem que essa transformação também pode ser usada para proporções que estão sujeitas a outra causa de variação que não a binomial, sendo, porém, que as variâncias dessas proporções devem ser um múltiplo de $\mu(1 - \mu)$. Como, porém, esse produto varia pouco se as porcentagens estiverem todas entre 30% e 70%, a transformação angular será desnecessária. Essa transformação produzirá sensíveis alterações nos valores das porcentagens se estiverem entre 0% e 30% ou 70% e 100%. A transformação $arcosen\sqrt{\%}$ dará melhores resultados quando todas as porcentagens forem baseadas em denominadores iguais, porém, tem sido frequentemente usada quando são diferentes, especialmente, se são aproximadamente iguais. Pode acontecer que a variável $\frac{n}{X}$ não tenha distribuição binomial e que a transformação angular não atinja seu objetivo, como é o caso, muitas vezes, de dados de controle de pragas e moléstias no campo. Neste caso, deve-se considerar o numerador da proporção como a variável aleatória, podendo ser analisada utilizando-se uma das transformações citadas anteriormente. A transformação raiz quadrada é

recomendada para porcentagens entre 0% e 20% ou 80% e 100% sendo subtraídos de 100 antes da transformação. (LIMA; LIMA, 2014).

De acordo com Heath (1981), o emprego de dados transformados altera o significado da ausência de interação que, para os dados brutos, foi definida como independência e aditividade dos efeitos dos tratamentos. Na transformação logarítmica, a ausência de interação faz com que sejam independentes e aditivos os efeitos dos tratamentos sobre os logaritmos das colheitas; por conseguinte, e tendo em conta os dados brutos, existe uma interação multiplicativa, ou seja, um determinado aumento no nível de um dos fatores (ou tratamentos) multiplica a colheita por uma constante, sejam quais forem os níveis dos demais fatores. Para alguns sistemas, essa poderia ser a definição mais apropriada de independência de fatores (mais apropriada do que a definição em termos de aditividade dos efeitos desses fatores, tendo em conta os dados brutos). Por exemplo, é de se esperar que os efeitos de fatores sobre o número de bactérias sejam proporcionais ao número de células presentes e que sejam, portanto, aditivos numa escala logarítmica. Na prática, entretanto, não faz muita diferença qual das duas definições se aceita, a não ser que as respostas aos dois fatores considerados sejam ambas proporcionalmente grandes. O autor ainda afirma que não viu nenhuma interpretação satisfatória das implicações biológicas da falta de interação, em termos de ângulos.

Toda análise da variância de um experimento pressupõe um modelo matemático e a aceitação de algumas hipóteses básicas. Se tomarmos como exemplo um experimento em blocos ao acaso, teremos como modelo matemático $Y_{ij} = m + t_i + b_j + e_{ij}$, onde Y_{ij}é o valor observado relativo à parcela que recebeu o tratamento i no bloco j; m é a média geral; t_i mede o efeito do tratamento i; b_j mede o efeito do bloco j; e e_{ij} é a contribuição do acaso, isto é, a parte da variação devida a fatores não controlados. Na análise admitimos as seguintes hipóteses: a) Que os diversos efeitos são aditivos, como se vê no modelo matemático adotado e não, por exemplo, multiplicativo, como seria o caso num modelo $Y_{ij} = mt_i b_j e_{ij}$; b) Que os erros ou desvios e_{ij} são independentes, de onde resulta que não são correlacionados; c) Que os erros e_{ij} têm todos a mesma variância σ^2; d) Que os erros e_{ij} têm distribuição normal. Estas hipóteses parecem muito restritivas, mas não são tanto assim, pois em geral não há grande importância em que se verifiquem apenas aproximadamente. Por exemplo, os testes mais em uso como o teste t de Student, o teste F de Fisher–Snedecor e o teste de F. G. Brieger (t e F ou v) não se alteram muito se a distribuição for apenas aproximadamente normal, ou mesmo que a distribuição se afaste bastante da normalidade. Saliente-se, aliás, que a normalidade dos erros jamais se verifique nos experimentos, pois, como primeira condição, deveria haver possibilidade de observações desde $-\infty$ até $+\infty$, o que não ocorre. Mas é suficiente que haja uma aproximação razoável, principalmente quando há repetições para todos os tratamentos (PIMETEL GOMES, 1987; SNEDECOR e COCHRAN, 1967). A desigualdade de variâncias traz problemas mais sérios, mas não deve ser encarada com excessivo rigor, pois é suficiente

que as diferenças entre elas não sejam muito grandes (BOX, 1954; CONAGIN et al., 1993). Para avaliar a heterogeneidade das variâncias, usam-se, em geral, o teste de Bartlett e o F máximo, quociente do maior quadrado médio sobre o menor. Lamentavelmente, ambos estes testes são muito sensíveis à falta de normalidade (SNEDECOR e COCHRAN, 1967). Não havendo suspeita de grave afastamento da normalidade, uma solução recomendável por Pimentel Gomes, 2000, a qual permite a aplicação sem susto da análise da variância, juntamente com os testes F, de Tukey, de Duncan, de Bonferroni e t de Student. (PIMENTEL GOMES, 2000).

Quando são excessivamente heterogêneas as variâncias, pode-se tentar a transformação da variável em estudo. A transformação recomendável frequentemente pode ser obtida com o auxílio da equação de regressão $V = A\hat{m}^b$, onde $\hat{m}$ é a estimativa da média de sua variância. A transformação indicada é então: $y = (x + k)^{1-\frac{b}{2}}$, para $b \neq 2$, $y = log(x + k)$, para $b = 2$. Onde k é uma constante positiva ou nula (THÖNI, 1978) e o logaritmo pode ser decimal ou neperiano, à vontade. Há casos, porém, em que, independentemente de estudo da heterogeneidade das variâncias, é indicada a transformação dos dados. Tal ocorre, por exemplo, no caso de contagens com números relativamente baixos. Admite-se então, em geral, que se trate da distribuição de Poisson, para a qual o valor de $b = 1$ e, pois, a transformação indicada é: $y = (x + k)^{1-\frac{b}{2}} = (x + k)^{\frac{1}{2}} = \sqrt{x + k}$. Quando se incluem valores de x superiores a 15, convém tomar $k = 0{,}5$ ou $k = \frac{3}{8}$. É o que ocorre comumente com experimentos de insetos em laboratórios. Quando todos os valores de x são maiores do que 15 pode-se tomar k = 0. Outro caso a considerar é o que trata de porcentagens $p = \frac{x}{n} \cdot 100$, relativas a n observações por parcela. Em tais condições, os dados têm em geral, distribuição binomial e a transformação indicada é $y = arcsen\sqrt{\frac{x}{n}}$,que pode ser comodamente aplicada com o auxílio de uma calculadora, com y expresso em graus ou em radianos, indiferentemente. Mas só se faz necessária a transformação quando a porcentagem cai abaixo de 15% ou excede 85%. Assim, se todos os dados estiverem no intervalo [15%, 85%], a transformação não é necessária. Além disso, no caso de $x = 0$, o valor $\frac{0}{n}$ deve ser substituído por $\frac{1}{4n}$, e, no caso de $x = n$, o valor $\frac{n}{n}$ será substituído por $1 - \frac{1}{4n} = \frac{4n-1}{4n}$.

Segundo Silva (2013), as pressuposições nas quais a análise de variância está fundamentada são as seguintes: efeitos aditivos dos efeitos dos tratamentos e dos efeitos das outras fontes de variação, independência dos erros experimentais, homogeneidade da variância dos erros e distribuição normal dos erros. Essas suposições, aparentemente muito rigorosas, são geralmente encontradas, ao menos aproximadamente, nos experimentos agronômicos. Quando elas não são satisfeitas, alguns procedimentos devem ser tomados para que uma análise de variância adequada seja feita. Conforme o autor são aplicadas correções como as descritas a seguir: Falta de aditividade: Se os efeitos não são

aditivos, mas multiplicativos, a transformação logarítmica conduz à aditividade. Uma comparação dos modelos aditivo e multiplicativo é dada na Tabela 3 que considera um experimento com dois tratamentos, em blocos ao acaso com duas repetições, que ignora os efeitos de erros experimentais. Para o modelo aditivo, o aumento do bloco 1 para o bloco 2 é uma quantidade fixa (10), independentemente do tratamento. O mesmo é verdadeiro para tratamentos (20). No modelo multiplicativo, os efeitos são aditivos apenas quando expressos em percentagens, ou seja, a redução do bloco 1 para o 2 é uma percentagem fixa (50%), independentemente do tratamento. O mesmo é verdadeiro para os tratamentos (33,3%). Quando os efeitos são multiplicativos, os logaritmos dos dados exibem os efeitos em uma feição aditiva e uma análise de variância dos logaritmos é apropriada. Note que os efeitos do modelo multiplicativo se tornam aditivos, após a transformação logarítmica.

Tabela 3. Dados fictícios para os modelos aditivo e multiplicativo de um experimento com dois tratamentos e duas repetições. Mossoró-RN, 2023.

Modelos	Tratamentos	Blocos		Efeitos dos blocos (Bloco 1 Bloco 2)
		1	2	
Aditivo	A	40	30	10
	B	20	10	10
	Efeitos dos tratamentos (A-B)	20	20	-
Multiplicativo	A	60	30	30
	B	20	10	10
	Efeitos dos tratamentos (A-B)	40	20	-
Aplicação de Log_{10} aos dados do modelo multiplicativo	A	1,78	1,48	0,30
	B	1,30	1,00	0,30
	Efeitos dos tratamentos (A-B)	0,48	0,48	-

Falta de independência dos erros experimentais: a independência dos erros experimentais, isto é, a não correlação dos erros, é geralmente obtida por meio de uma casualização apropriada. Em certos experimentos, onde não se pode fazer casualização, os tratamentos são dispostos sistematicamente nas parcelas, ao invés de modo aleatório. Os efeitos da profundidade do solo sobre os atributos químicos dele constituem um exemplo. Nesse caso, as profundidades poderiam ser (em cm): 0 – 10 e 10 – 20, 20 – 30 e 30 – 40. Neste caso, os erros associados às profundidades 0 – 10 e 10 – 20 tenderiam a ser correlacionados, pois essas camadas estariam sempre próximas. O mesmo ocorreria com as camadas 20 – 30 e 30 – 40.

Falta de homogeneidade da variância dos erros experimentais: Quando os erros não são homogêneos, uma solução é dividir o termo erro em componentes homogêneos para testar comparações de tratamentos específicos. Algumas vezes, se as médias de um ou dois tratamentos são muito maiores do que as dos outros e tem variação significativa maior, estes tratamentos podem ser excluídos da análise. Outra solução é buscar uma transformação adequada dos dados experimentais.

Falta de normalidade da distribuição dos erros experimentais: Se existem razões para que se suspeite de uma não normalidade da distribuição dos erros, é aconselhável fazer-se uma transformação das variáveis. Essas transformações são feitas para tornarem homogêneas as variâncias dos erros. Em geral também, tendem a aproximar os dados da normalidade.

Quando as médias dos tratamentos e as variâncias respectivas estão relacionadas, isto é, não são homogêneas, o correto não é analisar os dados originais, mas realizar uma transformação apropriada desses dados e depois fazer uma análise com os dados transformados. Em outras palavras, são analisados valores dependentes dos dados originais, mas que igualam as variâncias dos tratamentos, independente das médias desses tratamentos. Geralmente, a transformação também serve para tornar a distribuição mais próxima da normal, de tal modo que os testes para comparações de médias e os das razões de variâncias podem ser aplicados sem qualquer dificuldade (SILVA, 2013).

A heterogeneidade da variância ocasionada pela não normalidade da distribuição dos erros é corrigida através de transformações da variável. Se X tem distribuição normal, a médias $m = E(X)$ é independente da variância $V(X)$. Mas, se a variância é função da média, isto é, $V(X) = f(m)$, ela somente ficará estabilizada por meio de uma mudança conveniente na escala de mensuração. Assim, o problema consiste em determinar uma nova variável y, em função de X, tal que a variância seja independente (ou aproximadamente independente) da média. Uma vez encontrada a transformação, a análise estatística é feita com as variáveis transformadas (SILVA, 2013).

Já a transformação angular é apropriada para dados que seguem a distribuição binomial, isto é, valores expressos em porcentagem (% de plantas que florescem, % de plantas acamadas) é a angular. Na distribuição binomial, a variância está estreitamente relacionada com a média. Isto significa que, na distribuição binomial, a média é igual a Np, onde N é o número de provas e p, a probabilidade de ocorrência de determinado evento e a variância é Npq, onde $q = 1 - p$ é a probabilidade de não ocorrência do referido evento. Portanto, se forem tiradas amostras de várias distribuições binomiais, os tratamentos e as variâncias residuais não serão independentes. Muitos dados experimentais expressos em porcentagem podem seguramente ser submetidos a uma análise de variância sem transformação, visto que, para certos valores de p (proporção média), a distribuição binomial se aproxima muito da normal. Assim, para porcentagens de contagens de 100 ou mais indivíduos que fiquem entre 20% e 80%, pode-se esperar que os testes sejam válidos na análise de variância. Mas, para porcentagens baseadas em menos do que 20% ou maiores do que 80%, é usualmente necessária uma transformação antes de se realizar a

análise de variância. A transformação apropriada é $arcsen\sqrt{X_{\%}/100}$, isto é, o ângulo cujo seno é a raiz quadrada da porcentagem/100 (SILVA, 2013).

Com relação à transformação angular, os dados percentuais podem ser classificados em três grupos (CLARK; LEONARD, 1939 citados por SILVA, 2013) os quais são os dados contínuos expressos em percentagens, pela divisão de cada valor por um valor arbitrário constante; dados contínuos expressos em percentagens para indicar concentrações; e dados discretos, baseados em um número determinado de ensaios ou casos. Quando os dados contínuos são expressos em percentagens pela divisão de cada valor por um valor arbitrário constante, o que se está fazendo, na realidade, é uma transformação da unidade de mensuração. Percentagens desse tipo serão tratadas estatisticamente como se os dados estivessem em sua forma original. Dados de rendimento podem ser expressos em porcentagem da testemunha ao invés de em rendimentos reais em kg ha^{-1}, por exemplo. Nesses casos, a análise de variância pode ser feita sem necessidade de transformação. Os dados contínuos são muitas vezes expressos em percentagens para mostrarem concentrações, porque uma comparação de concentrações é o principal objetivo do estudo. Estes tipos de percentagens como pureza de sementes, dada pela massa de sementes puras/massa total de sementes; folhosidade dada pela massa foliar/massa total da planta; teor de proteína no grão, dado pela massa de proteína/massa do grão; teor de açúcar na raiz, expresso pela massa de açúcar/massa de raiz são muito comuns. Tais concentrações, como regra, não estarão sujeitas a qualquer transformação para equalizar a variância. Todavia, a técnica para a análise de cada problema, que fornecerá dados percentuais, deve ser considerada cuidadosamente na decisão se uma transformação será ou não empregada para remover um dado tipo de heterogeneidade. Quando se dispõe de dados discretos, baseados em um número determinado de ensaios ou casos, a transformação angular será aplicada. Como exemplos desse tipo de dados podem ser citados; as percentagens de germinação, expressas pelo número de sementes germinando/número total de sementes e as percentagens de plantas doentes, dada pelo número de plantas doentes/número total de plantas.

Conforme Silva (2013), a transformação raiz quadrada pode ser aplicada numa situação onde a análise preliminar de certos dados pode revela que as amplitudes de variação dos tratamentos e as médias desses tratamentos não são independentes, mas que as variâncias dos tratamentos são proporcionais às respectivas médias. É usual transformar tais dados calculando a raiz quadrada dos valores observados e analisar essas raízes quadradas. Essa transformação é apropriada para dados descritos pela distribuição de Poisson, assim chamada em homenagem ao matemático Siméon Denis Poisson, que a descreveu, que acontece quando se trata do número de ocorrências de certo evento, com um número muito grande de observações e a probabilidade de que ele ocorra em uma observação é pequena. Por exemplo, o número anual de suicídios em uma população humana, o número de parafusos defeituosos produzidos por uma indústria etc. Na experimentação com plantas, a distribuição de Poisson ocorre na contagem de árvores doentes em pequenas áreas e no número de determinado tipo de frutos encontrados em uma árvore.

Quando essa contagem atinge 50 ou mais, para cada unidade considerada, não há necessidade de qualquer modificação no método normal de análise. Entretanto, contagens inferiores a 50 podem indicar uma distribuição de Poisson sendo, por isso, aconselhável extrair a raiz quadrada da contagem. Quando ocorrem valores iguais a zero, ou quando a maioria das contagens é menor do que 20 (ou 10), a transformação da raiz quadrada tende a corrigir excessivamente os dados e, nesses casos, é costume adicionar ½ ou 3/8 a cada valor, antes de se extrair a raiz quadrada. Então, se X obedece à distribuição de Poisson, $V(X) = f(m) = m$, e a transformação apropriada é $Y = \sqrt{X}$.

Segundo ainda Silva (2013), a transformação logarítmica pode ser útil numa situação onde uma análise preliminar dos dados de um experimento pode revelar que há uma relação definida entre desvios padrões dos tratamentos e as médias desses efeitos.

Quando os desvios padrões dos tratamentos são diretamente proporcionais às respectivas médias, ou seja, $s/m = k$, como são frequentemente os casos da determinação da contagem do número de árvores por hectare, a medição dos comprimentos totais das raízes por plântula, a transformação logarítmica decimal (em alguns casos pode ser vantajoso utilizar os logaritmos naturais) é a adequada. Completa-se então a análise à maneira usual.

Como o logaritmo de zero não tem significado, os valores zero não podem ser modificados diretamente por esta transformação. Quando ocorrem valores zero nos dados a analisar, é costume adicionar 1 a todos os valores observados e transformar esses valores em logaritmo $(X + 1)$. A adição de 1 a cada valor não afeta grandemente os resultados e torna possível completar a análise. Quando aparecem valores negativos da variável transformada, pode-se adicionar um valor constante a cada observação antes da transformação de modo a tornar positivos todos os dados transformados. Portanto, se o desvio padrão é aproximadamente proporcional à média, então $Y = log\ X$.

Conforme Silva (2013), os testes para determinar se a transformação é apropriada na análise dos dados de um experimento, pode haver dúvidas sobre a necessidade de transformação e, em caso afirmativo, sobre qual a transformação que deverá utilizar-se. Convém lembrar que pequenos desvios dos valores observados para a normalidade não têm relativamente, qualquer importância. Mesmo desvios bastante grandes têm pequenos efeitos. As decisões de que devem ser aplicadas transformações podem ser tomadas pela experiência do seu uso, frequentemente, o pesquisador analisa as mesmas características em vários experimentos.

Pode-se ganhar bastante experiência analisando os dados de determinada característica com e sem transformação dos dados, observando-se os casos em que a transformação modificou os testes de significância. Há poucas dificuldades em determinar se a transformação angular é apropriada, visto que ela é aplicada em circunstâncias bem caracterizadas, isto é, para percentagens resultantes de observações de menos de 100 indivíduos ou percentagens que incluem valores inferiores a 20% ou superiores a 80%.

Para transformações logarítmicas e da raiz quadrada, a escolha correta pode ser rapidamente determinada, colocando-se, em papel logarítmico, as variâncias dos tratamentos individuais em função das médias correspondentes.

Pode-se notar, então, qualquer tendência que esteja presente e, também, reduzir-se a transformação requerida. Se os pontos grafados se encontram sobre uma linha reta que passa pela origem do gráfico com o declive 1, deverá ser usada a transformação da raiz quadrada. Se a declividade da reta através dos pontos marcados tem um valor 2, deve ser usada a transformação logarítmica. Isso pode ser constatado da seguinte maneira (SILVA, 2013):

No caso da distribuição de Poisson, tem-se:

$$V(X) = K_1 m$$

$$log\ V(X) = log\ m + log\ K_1$$

ou, generalizando-se,

$$log\ V(X) = log\ m_i + log\ K_i$$

Essa é a equação de uma reta em que o valor de $b = 1$, isto é, $tg\ \alpha = 1 =$ declividade. Para o caso da distribuição logarítmica, o que se tem é:

$$s_i/m_i = K_i$$

$$s_i^2/m_i^2 = K_i^2$$

$$V(X_i)/m_i^2 = K_i^2$$

$$logV(X_i) = 2logK_i + logm_i$$

$$tg\ \alpha = 2 = declividade$$

Para Silva (2013), as características com dados transformados, podem ser visualizadas na Tabela 4 onde são apresentados alguns tipos de transformação de dados, de experimentos com milho, obtidos em trabalhos publicados em 250 artigos de algumas revistas brasileiras. Constata-se que as transformações ocorreram mais com dados de características de patógenos ou pragas do que com dados de características da referida cultura.

Convém mencionar que, no levantamento realizado, foram encontrados estudos em que a transformação de dados não foi efetuada nem mesmo em dados que geralmente não seguem a distribuição normal, como, por exemplo, os de percentagens de germinação de sementes.

Tabela 4. Tipo de transformação de dados de característica avaliadas em experimentos com milho em levantamento realizado em 250 artigos de algumas revistas brasileiras. Mossoró-RN, 2023.

Características	Transformação	Observações	Autores
Notas de severidade de doenças foliares	$(x)^{0,5}$	-	Nihei & Ferreira, 2012
Número de larvas da praga sobrevivente após 48 horas	$(x+1)^{0,5}$	-	Mendes et al., 2011
Número de larvas da praga sobrevivente após 48 horas	$(x+1)^{0,5}$	-	Mendes et al., 2011
Número de larvas da praga sobrevivente na fase pré-imaginal	$(x)^{0,5}$	-	Mendes et al., 2011
Percentagens de germinação e debulha	$arcsen(x/100)^{0,5}$	-	Stanisavljevic et al., 2010
Percentagem de grãos atacados	$arcsen(x/100)^{0,5}$	-	Stanisavljevic et al., 2010
Percentagem de insetos mortos	$arcsen(x/100)^{0,5}$	-	Stanisavljevic et al., 2010
Percentagens de mortalidade de insetos e de plantas com sintomas foliares	$arcsen(x/100)^{0,5}$	No caso de x = 0 e x = n utilizaram-se arc sen $(1/4n)^{0,5}$ e arc sen $(4n\text{-}1/4n)^{0,5}$, respectivamente	Oliveira et al., 2007
Nota de dano da praga e número de espigas	$(x+1)^{0,5}$	Escala de 0 a 5	Figueiredo et al., 2006
Percentagens de mortalidade e de emergência de insetos	$arcsen(x/100)^{0,5}$	-	Silva et al., 2005
Número médio de ninfas	$log\ (x\ +\ 5)$	-	Gonçalvez & Sousa e Silva, 2003
Número de entrenós do milho	$(x+0{,}5)^{0,5}$	-	Silva, 1963
Rendimento de fubá, germen, volume específico comparativo e escore total comparativo	$arcsen(x/100)^{0,5}$	-	Lima et al., 1988
Nota de incidência da ferrugem	$(x)^{0,5}$	Escala de 1 a 9	Lima et al., 1996
Número de insetos	$(x+0{,}5)^{0,5}$	-	Ceccon et al., 2004

Tabela 5. Tipo de transformação de dados de característica avaliadas em experimentos com milho em levantamento realizado em 250 artigos de algumas revistas brasileiras. Mossoró-RN, 2023 (continuação).

Número de lagartas, número de posturas e sobrevivência (%) das lagartas	$(x)^{0,5}$ ou $log\ (x\ +\ 5)$	O tipo de transformação depende da natureza dos dados	Viana & Potenza, 2000
Índice de massa de ovos (IMO) de nematoide	$(x + 0{,}5)^{0,5}$	-	Sawazaki et al., 1998
Ovos por grama de raiz, fator de reprodução (FR) = população final/população inicial de nematoide	$log\ (x + 0{,}5)^{0,5}$	-	Sawazaki et al., 1998

Conforme Silva (2013), na apresentação dos resultados da análise estatística de dados que foram transformados implica alguns problemas que não são encontrados na análise de dados não transformados. Todos os testes de significância devem ser realizados com os dados transformados. É, portanto, necessário indicar as médias e os erros padrões dos dados transformados quando se apresentam os resultados da análise. É também conveniente, por razões práticas, relacionar as médias transformadas aos dados originais. O método usual para fazer isso consiste em desconverter as médias dos dados transformados, elevar ao quadrado as médias dos dados transformados pela transformação da raiz quadrada e encontrar o antilogaritmo das médias dos dados transformados através da transformação logarítmica. Infelizmente, as médias obtidas por desconversão não são, em geral, as mesmas ou idênticas às medidas obtidas por análise dos dados sem transformação por médias simples dos valores originais. A razão da diferença entre as duas séries de médias é que as médias calculadas a partir dos dados transformados são menos afetadas pelos valores extremos do que as médias obtidas diretamente. Para a transformação angular, a diferença é geralmente pequena e pode ser desprezada. Com a transformação logarítmica (com ou sem adição de 1), deve juntar-se a cada média transformada 1,15 vezes a variância das observações transformadas, antes da desconversão para as unidades originais das observações. Para a transformação da raiz quadrada (com ou sem $\frac{1}{2}$ ou $\frac{3}{8}$), deve-se adicionar às médias derivadas a variância das observações transformadas. Talvez segundo o autor seja conveniente apresentar duas colunas: uma com os dados transformados, juntamente com o teste de médias, e outra coluna com os dados não transformados.

Já a transformação de Yeo-Johnson criada por Yeo e Johnson (2000) se justifica, pois, segundo estes autores a contribuição de Box-Cox (1964) foi o maior passo na determinação de uma maneira objetiva de se efetuar transformação de dados, entretanto, como a transformação de Box - Cox é válida apenas para valores positivos de X, havia espaço para algum tipo de melhoria. Embora seja possível

efetuar uma troca de parâmetros, em caso de valores negativos para utilização da transformação de Box-Cox, existe o inconveniente de tal ação afetar a teoria que suporta a definição do intervalo de confiança de λ. Yeo e Johnson (2000) propuseram uma nova família de transformação de dados, válida tanto para valores positivos como para valores negativos da variável X. Sua fórmula, definida como uma função $\Psi: R\ X\ R \to R$, é apresentada a seguir:

$$\Psi^{(\lambda,x)} = \begin{cases} \left\{\frac{(X+1)^{\lambda}-1}{\lambda}\right\} & (X \geq 0, \lambda \neq 0) \\ log(X+1) & (X \geq 0, \lambda = 0) \\ \frac{-\{(-X+1)^{2-\lambda}-1\}}{2-\lambda} & (X < 0, \lambda \neq 2) \\ -log(-X+1) & (X < 0, \lambda = 2) \end{cases}.$$

Fica claro que outros tipos de situações podem ocorrer e assim poder escolher outros tipos de transformações. De uma maneira geral, no caso de uma relação linear do coeficiente b entre o logaritmo das variâncias e o logaritmo das médias, a transformação a ser utilizada é, desde que b seja diferente de 2, a seguinte: $Y = X^{1-\frac{b}{2}}$. Nestas condições, tem-se, com efeito, o resultado dado por:

$$log\, \sigma_X^2 = b\, log\ \ m_X + log\, k\ ,$$

onde

$$\sigma_X^2 = k m_X^b \text{e, } \sigma_Y^2 \cong k\ m_X^b \left(\frac{dX^{1-\frac{b}{2}}}{dX}\right)_{m_X}^2 = k\ \left(1-\frac{b}{2}\right)^2 = k'$$

Muitos exemplos destas transformações são mostrados particularmente por Taylor (1961). Métodos mais rigorosos foram igualmente propostos por Box e Cox (1964) e por Kruskal (1965), para realizar de maneira empírica a escolha de uma transformação ótima. Na prática corrente, estes métodos são, contudo, muitos trabalhosos e não parece necessário levar de forma demasiada e longa a preocupação em utilizar uma transformação ideal. Mas independentemente das dificuldades de cálculo, o principal inconveniente das mudanças de variáveis reside, com efeito, em complicar a interpretação dos resultados. Esta dificuldade é de importância relativamente secundária para as transformações raiz quadrada e logarítmica, por exemplo, mas pode ser muito mais considerável em transformações mais complexas. Sendo assim parece razoável, limitarmo-nos, tanto quanto possível, as transformações mais simples.

Conforme Murteira (1993), a motivação para o uso de transformações de um conjunto de observações para facilitar a análise exploratória de dados, são de diversas ordens, tais como: facilitar a interpretações naturais, simetrizar a coleção de dados, estabilizar a dispersão de várias coleções de dados, linearizar a relação entre duas variáveis e simplificar a estrutura de uma tabela de dupla entrada, segundo o autor ao medir temperaturas em graus Celsius (C), mas se esta vier dada em graus Fahrenheit (F), a

primeira coisa que deve ser feita é passar para graus Celsius através da fórmula ou transformação, $C = \frac{5}{9}(F - 32)$.

A transformação acima consiste numa simples mudança de escala e de origem, $y = \frac{(x-A)}{B}, B > 0$ e pertence à família das transformações lineares. As transformações lineares são úteis, principalmente para o cálculo, mas não alteram a forma da coleção de valores.

Conforme ainda aquele autor, noutras situações é natural medir o tempo necessário para se dar a ocorrência de um fenômeno, por exemplo, o psicólogo que estuda o comportamento de ratos e mede o tempo que levam para percorrer um labirinto se deparam com uma dificuldade, como aquela de haver casos em que o tempo parece não ter limite. É então mais natural trabalhar com o inverso dos tempos, $y = \frac{1}{x}, x = tempo$; onde y assume praticamente o valor zero para os que nunca mais acabam e cresce com a rapidez dos que chegam ao fim do percurso.

Segundo ainda Murteira (1993), o cálculo de logaritmos neperianos ou decimais, tem a vantagem de revelar com clareza o padrão de crescimento dos dados, isto se dá, porque se os pontos transformados $y = ln\, x$ se dispõem linearmente e assim pode-se avançar-se com uma lei de crescimento exponencial. Ele ainda afirma que a simetrização das distribuições de valores numéricos tem uma razão simples, as quais são: As coleções de dados simétricos são susceptíveis de uma análise mais simples pelo equilíbrio que se verifica entre os desvios positivos e negativos em relação ao centro. As medidas de localização são mais fáceis de serem compreendidas quando as distribuições de valores são simétricas, pois é sabido que para tais coleções a média e a mediana coincidem com o centro da distribuição de valores, bem como a moda.

De acordo com Murteira (1993), as transformações mais empregadas na análise exploratória e na descrição de dados de observação, são as transformações potência, e elas se caracterizam por ser funções elementares, pois os valores transformados calculam-se facilmente, são estritamente crescentes, contínuas e regulares, com derivadas de todas as ordens. Em consequência da monotonia, permitem preservar a ordem dos dados. Isto é, a mediana transformada é a mediana da coleção transformada, os quartis transformados são os quartis da coleção transformada, etc. Em suma, as médias resumo transformadas são com pequenas diferenças resultante das interpolações, as medias resumo da coleção transformada. A forma geral das transformações potência é a seguinte, $\begin{cases} t_P(x) = a\ x^P + b, & P \neq 0 \\ t_P(x) = c\ ln\ \ x + d, & P = 0 \end{cases}$, onde ln designa o logarítimo neperiano, a, b, c, d e P são números reais com a restrição $P > 0 \Rightarrow a > 0\ e\ P < 0 \Rightarrow a < 0$. A escolha de P decorre dos objetivos da análise dos dados. Bussab (1986), afirma que em estudo de regressão quando ocorre ausência de um modelo teórico relacionando as duas variáveis, a investigação do diagrama de dispersão pode sugerir uma relação adequada para os dados. Sabe-se também, que uma dada relação gráfica pode ser aproximada por diferentes funções matemáticas, e que algumas delas

possuem propriedades mais convenientes do que outras. O autor afirma ainda que o interesse é em funções que através de transformações de variáveis reduzem-se a modelos lineares do tipo $y_i = \alpha + \beta x_i$. O conhecimento da forma de diversas famílias de curvas ajuda o pesquisador a decidir por um dado modelo. Para ajudar o investigador na escolha de uma dada função linearízável, será apresentado a seguir um conjunto de formas mais usuais de relações entre variáveis (Tabela 6). O autor afirma ainda que os modelos obtidos pela análise gráfica exigem cuidados especiais tanto na sua utilização para previsão como para interpretação. Não deve haver extrapolações fora do intervalo analisado, e na interpretação das estimativas dos parâmetros. Na maioria dos casos há interesse na relação do modelo original e não no modelo transformado (linearizado).

Tabela 6. Tipos de modelos de regressão e transformações. Mossoró – RN, 2023.

Intervalos das estimativas dos parâmetros	**Tipo de Modelo**	**Função**	**Transformações**	**Forma Linear**	**Observações**
α, β e x positivos, e $\alpha, x > 0 \ e \ \beta < 0$	Função Potência	$y = \alpha x^{\beta}$	$y' = log\, y\,,\ x' = log\, x$	$y' = log\, \alpha + \beta\ x'$	y > 0, x > 0
$\beta > 0 \ e \ \beta < 0$	Função Exponencial	$y = \alpha e^{\beta x}$	$y' = ln\, y$	$y' = \alpha + \beta\ x'$	y > 0
$\beta > 0$	Função Logarítmica	$y = \alpha + \beta\, log\, x$ $y = \alpha + \beta\, ln\, x$	$x' = log\, x$	$y = \alpha + \beta\ x'$	x > 0
$\beta < 0$; $\alpha, \beta > 0 \ e \ x > \frac{\beta}{\alpha}$ e $\alpha > 0 \ e \ x > \frac{\beta}{\alpha}, \beta < 0$	Função Hiperbólica	$y = \frac{x}{\alpha x - \beta}$	$y' = \frac{1}{y},\ x' = \frac{1}{x}$	$y' = \alpha - \beta\ x'$	$y \neq 0,\ x \neq 0$
----------	Função Logística	$y = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$	$y' = ln\left(\frac{y}{1-y}\right)$	$y' = \alpha + \beta\ x$	$0 < y < 1$

A função ou modelo de regressão $g(x)$ pode expressar-se na forma $g(x) = \alpha c_1(x) + \beta c_2(x)$, isto, é, como um modelo linear de dois parâmetros, do qual a regressão linear é um caso particular $(c_1(x) = x \text{ e } c_2(x) = 1)$. O modelo matemático $g(x) = \alpha c_1(x) + \beta c_2(x)$ é, do ponto de vista do método dos mínimos quadrados, considerado um modelo linear do sistema real dado os parâmetros α e β aparecerem linearmente combinados, embora as funções $c_1(x) = x$ e $c_2(x)$ possam ser não lineares. Nesta situação, o sistema normal é, tal como no caso da reta, um sistema linear 2×2 em α e β.

De fato, quando uma função $g(x)$ na forma $g(x) = \alpha c_1(x) + \beta c_2(x)$ se substitui na equação

$$E(g) = \sum_{i=1}^{m} [x_i - g(x_i)]^2,$$

as equações normais

$$\frac{\partial}{\partial \alpha} E(g) = 0 \text{ e } \frac{\partial}{\partial \beta} E(g) = 0$$

convertem-se no sistema linear. A hipérbole com $g(x) = \frac{\alpha}{x} + \beta$ é um modelo linear de dois parâmetros da forma $g(x) = \alpha c_1(x) + \beta c_2(x)$, com $c_1 = \frac{1}{x}$ e $c_2 = 1$. No caso de um modelo de tipo exponencial, como é, por exemplo, o modelo típico da curva de esgotamento das reservas subterrâneas, caracterizado por $g(x) = \alpha\ e^{\beta x}$ o sistema

$$\frac{\partial}{\partial \alpha} E(g) = 0 \text{ e } \frac{\partial}{\partial \beta} E(g) = 0$$

não será linear. A função $g(x) = \alpha\ e^{\beta x}$não é um modelo linear (não é soma ponderada de duas funções que seja possível expressar como $g(x) = \alpha c_1(x) + \beta c_2(x)$. A substituição de $g(x)$ na Equação

$$E(g) = \sum_{i=1}^{m} [x_i - g(x_i)]^2$$

conduz ao problema de minimizar a função

$$E(\alpha, \beta) = \sum_{i=1}^{m} \left[y_i - \alpha\ e^{\beta x_i}\right]^2$$

que, por sua vez, conduz a um sistema 2×2 não linear nas variáveis α e β. que não tem, regra geral, solução analítica.

Na linearização observa-se o seguinte: Na discussão anterior permitiu verificar as dificuldades adicionais na resolução do problema quando o modelo $g(x)$ é não linear. Com o objetivo de contornar estas dificuldades adota-se, por vezes, uma técnica de linearização do problema. É importante observar que os parâmetros assim obtidos (linearização) não são ótimos, de acordo com o critério dos mínimos quadrados. Isto porque se ajusta o problema linearizado e não o original. No entanto, em termos práticos, as duas soluções são, de modo geral, muito próximas. Na linearização de um modelo não linear tem-se que.

Por exemplo, linearize $y = \alpha\ x^{\beta}$. Para esta função em particular, a linearização consiste em logaritmizar ambos os membros de $y = \alpha\ x^{\beta}$ e observar que:

$$y = \alpha x^{\beta} \Leftrightarrow \underbrace{ln\, y}_{Y} = \beta \underbrace{ln\, x}_{X} + \underbrace{ln\, \alpha}_{b},$$

com:

$$Y = ln\, y\,, X = ln\, x\,, \alpha = \beta \text{ e } b = ln\, \alpha.$$

A Tabela abaixo (Tabela 7) ilustra outros exemplos de linearização de funções (MARTINS. et al., 2010).

Tabela 7. Exemplos de linearização de alguns modelos não lineares, Mossoró, RN, 2023.

i	$y = g_i(x)$	Forma linear $Y = aX + b$	$X =$	$Y =$	$a =$	$b =$
1	$y = \alpha\ e^{\beta x}$	$ln\, y = \beta x + ln\, \alpha$	x	$ln\, y$	β	$ln\, \alpha$
2	$y = \frac{\alpha}{\beta + x}$	$y = \frac{1}{\beta} xy + \frac{\alpha}{\beta}$	xy	y	$\frac{-1}{\beta}$	$\frac{\alpha}{\beta}$
3	$y = \frac{\alpha x}{\beta + x}$	$y = -\beta \frac{y}{x} + \alpha$	$\frac{y}{x}$	y	$-\beta$	α

Nota: Um eventual ponto (0,0) deve ser eliminado do conjunto de dados antes de utilizar as funções $g_1(x)$ e $g_3(x)$(ver colunas $X =$ e $Y =$).

Segundo Cadima (2015), o pesquisador deve tomar cuidados com as transformações, isto é, duas prevenções gerais, em relação à utilização de transformações das variáveis:

(i) há que ter cuidado em relação à possibilidade de uma dada transformação poder resolver um problema (como o das variâncias heterogêneas), mas simultaneamente criar outro (como por exemplo, destruindo uma normalidade admissível dos erros aleatórios);

(ii) convém ter alguma discrição na utilização de transformações: a grande variedade de possíveis transformações faz com que seja possível encontrar uma transformação que, de forma espúria, resolva os problemas associados a um conjunto de dados específico, mas sem que isso reflita uma solução geral e robusta para os problemas associados ao fenômeno sob estudo.

Em algumas transformações de variáveis as vezes, é possível contornar violações às hipóteses de normalidade dos erros aleatórios ou homogeneidade de variâncias através de transformações de variáveis. Por exemplo:

a) Se $Var(\varepsilon i) \propto E[Yi\,]$ então $Y \rightarrow \sqrt{Y}$.

b) Se $Var(\varepsilon i) \propto (E[Yi])^2$ então $Y \rightarrow ln\, Y$.

c) Se $Var(\varepsilon i) \propto (E[Yi])^4$ então $Y \rightarrow \frac{1}{Y}$, são propostas usuais para estabilizar as variâncias.

Os exemplos acima são casos particulares da família Box-Cox de transformações:

$$\text{Y} \rightarrow \frac{Y^{\lambda}-1}{\lambda}, \text{ para } \lambda \neq 0 \text{ e } ln\,(Y), \text{ para } \lambda = 0.$$

Com relação as transformações linearizantes que são aquelas que visam linearizar uma relação original não linear entre x e y, deve-se tomar precauções sobre tais transformações. Neste caso os estimadores que minimizam a soma de quadrados dos resíduos nas relações linearizadas não são os que produzem as soluções ótimas de um problema de minimização de somas de quadrados de resíduos na relação não-linear original. As transformações não levam em conta os erros aleatórios.

As hipóteses de erros aleatórios aditivos, normais, de variância homogênea, média zero e independentes terão de ser válidas para as relações lineares entre as variáveis transformadas.

Segundo Pedrosa e Gama (2004), no caso de uma relação probabilística linear, o modelo de regressão é $Y_i = E(Y_i|x_i) + E_i = \beta_0 + \beta_1 x_i + E_i$. A linearidade de $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$relativamente aos parâmetros $\beta_0 \ e \ \beta_1$ e relativamente a x_i é evidente. Já por exemplo, o modelo $Y_i = E(Y_i|x_i) + E_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$, embora seja linear relativamente aos parâmetros $\beta_0 \ \beta_1 \ e \ \beta_2$, não é linear relativamente a x_i, e o modelo $E(Y|x_i) = \beta_0 e^{\beta_1 x_1}$é linear relativamente a β_0, mas não é linear nem relativamente linear a β_1 nem a x_i. Segundo os autores um modelo de regressão é linear quando é linear nos parâmetros e que a teoria de regressão linear se pode aplicar sempre que o modelo é linear nos parâmetros.

Há, no entanto, situações em que o modelo, embora não seja linear nos parâmetros, pode ser transformado facilmente num modelo linear, efetuando transformações de variáveis. Um exemplo típico é $Y_i = \beta_0 x_i^{\beta_1} E_i$ em que E_i (erro de observação i) é tal que $ln\, E_i \sim N(0, \sigma^2)$. Este modelo não é linear no parâmetro β_1, mas, calculando logaritmos em ambos os membros, obtém-se $ln\, Y_i = ln\, \beta_0 + \beta_1\, ln\, x + ln\, E_i$. Fazendo $Y_i^* = ln\, Y_i, x_i^* = ln\, x_i, \beta_0^* = ln\, \beta_0$e $E_i^* = ln\, E_i$, o modelo transforma-se em $Y_i^* = \beta_0^* + \beta_1 x_i^* + E_i^*$que é linear nos parâmetros β_0^* e β_1.

Noutras situações é conveniente o uso de outras transformações. Listamos algumas delas, na Tabela 8 a seguir.

Tabela 8. Algumas transformações para usar o modelo de regressão linear. Mossoró, RN, 2023.

Relação funcional	Transformação	Modelo de aplicação
$y = \beta_0 + \beta_1 \frac{1}{x}$, β_1>0 e β_1<0	$x^* = \frac{1}{x}$	$y = \beta_0 + \beta_1 x^*$
$y = \beta_0 e^{\beta_1 x}$, β_1>0 e β_1<0	$y^* = ln\, y$ $\beta_0^* = ln\, \beta_0$	$y^* = \beta_0^* + \beta_1 x$
$y = \beta_0 x^{\beta_1}$, β_1>1, β_1<0 e $0 < \beta_1 < 1$	$x^* = ln\, x$ $y^* = ln\, y$ $\beta_0^* = ln\, \beta_0$	$y^* = \beta_0^* + \beta_1 x^*$

Regressão linear é um modelo do tipo $y = \alpha + \beta(x - \bar{x})$, sendo esta reta chamada reta de regressão e β de coeficiente de regressão. O método de regressão linear simples pode ser aplicado a outros casos, além da relação linear entre duas variáveis, isto porque mediante uma adequada transformação de variáveis, as funções podem ser linearizadas. Vejam alguns exemplos de modelos de regressão na tabela 9 a seguir (BONINI; BONINI, 1972).

Tabela 9. Exemplos de funções de regressão linearizadas, mediante transformações de variáveis. Mossoró, RN, 2023.

Função real	Transformação	Regressão linear
$y = \frac{1}{\alpha + \beta x}$	$u = \frac{1}{y}$	$u = \alpha + \beta x$
$y = \alpha + \frac{\beta}{x}$	$u = \frac{1}{x}$	$y = \alpha + \beta u$
$y = Ax^{\beta}$	$u = \log x$ $v = \log y$ $\alpha = \log A$	$V = \alpha + \beta u$
$y = Ae^{\beta x}$	$u = \ln y$ $\alpha = \ln A$	$u = \alpha + \beta x$
$y = a10^{\beta t}$	$u = \log y$ $\alpha = \log a$	$u = \alpha + \beta t$

Naghettini e Pinto (2007) descrevem em sua obra sobre hidrologia estatística que na regressão não linear existem algumas funções linearizáveis que podem ser linearizadas mediante o uso de transformações adequadas permitindo a aplicação da regressão linear simples. Um exemplo pode ser a função potencial a seguir.

Realizando a anamorfose logarítmica dessa função, obtém-se:

$$ln\, y = ln(ax^b), y = ax^b, ln\, y = ln\, a + ln\, x^b, \text{ e}$$

$$ln\, y = ln\, a + b\, ln\, x.$$

Alterando as variáveis de forma que

$$z = ln\, y, k = ln\, a \text{ e } v = ln\, x,$$

a equação $ln\, y = ln\, a + b\, ln\, x$ se transforma na equação da reta: $z = k + bv$. Trabalhando com as variáveis transformadas $z = ln\, y$, e $v = ln\, x$, é possível estimar os parâmetros k e b com as equações, respectivamente,

$$a = \frac{\sum_{i=1}^{n} y_i}{n} - b\frac{\sum_{i=1}^{n} x_i}{n} = \bar{y} - b\bar{x}$$

e

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i x_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

Calculando o antilogaritmo de k estima-se o parâmetro a da equação $y = ax^b$. De forma análoga, a função $y = ab^x$ pode ser resolvida utilizando as variáveis x e a transformada $ln\, y$. Existem muitas outras funções linearizáveis, como por exemplo, $y = (a + bx)^{-2}$, que estão listadas na Tabela 9.

Porém, como o processo de linearização pode envolver a transformação da variável dependente Y, em alguns casos as hipóteses da regressão podem não ser atendidas, após a modificação, prejudicando a aplicação dos testes estatísticos descritos anteriormente.

Os autores apresentam um exemplo com os valores médios de vazões máximas anuais e as respectivas áreas de drenagem de 22 estações fluviométricas que compõem uma região homogênea de um estudo de regionalização de vazões máximas da bacia do alto São Francisco. Analisando o diagrama de dispersão, percebe-se que a relação entre as variáveis área de drenagem e média da vazão máxima anual pode ser expressa por uma função potencial como a equação $y = ax^b$, ou seja, $Q = kA^b$. Os parâmetros k e b podem ser estimados por meio da regressão linear simples, após a linearização da equação $Q = kA^b$.

A linearização é realizada por anamorfose logarítmica como apresentado a seguir: $ln\, Q = ln\, k + b\, ln\, A$. Assim, para concretização da regressão linear simples é necessário calcular os logaritmos da área de drenagem e das médias das vazões máximas anuais.

Por outro lado, no caso da regressão linear múltipla estuda-se o comportamento de uma variável dependente Y em função de duas ou mais variáveis independentes X_i. Se a variável Y variar linearmente com as variáveis X_i, pode-se adotar um modelo geral com a seguinte forma: $Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P$ onde Y é a variável dependente ou prevista; $X_1, X_2, \cdots, X_P$ são as variáveis independentes ou explicativas e $\beta_1, \beta_2, \cdots, \beta_P$ são os coeficientes de regressão.

As transformações de um modelo de regressão múltipla, deve ser realizado, pois em alguns casos, a violação do pressuposto de homocedasticidade dos resíduos pode ser superada, por meio dessas transformações tanto da variável dependente, como das variáveis explicativas ou de ambas. Além disso, a transformação de variáveis pode permitir a linearização de uma relação não linear.

De uma forma geral, a modificação das variáveis para alcançar os critérios de homocedasticidade não é uma tarefa fácil. As transformações mais utilizadas são a de raiz quadrada, a logarítmica e a recíproca, conforme apresentado a seguir:

$$Y = \beta_0 + \beta_1\sqrt{X_1} + \beta_2\sqrt{X_2} + \cdots + \varepsilon$$

$$Y = \beta_0 + \beta_1\, ln\, X_1 \; + \beta_2\, ln\, X_2 + \cdots + \varepsilon$$

$$Y = \beta_0 + \beta_1 \frac{1}{X_1} + \beta_2 \frac{1}{X_2} + \cdots + \varepsilon$$

As transformações de modelos não lineares podem ser obtidas por meio de anamorfose logarítmica, tal como exemplificado a seguir. Modelo multiplicativo do tipo:

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \varepsilon$$

Após a transformação obtêm-se:

$$ln\, Y \; = ln\, \beta\_0 + \beta\_1 \;\; ln\, X\_1 + \beta\_2 \;\; ln\, X\_2 + ln\, \varepsilon$$

No caso de um modelo exponencial

$$Y = e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2)} \varepsilon$$

A transformação logarítmica resulta em:

$$ln\; Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ln\varepsilon$$

Os autores apresentam um exemplo referente a um estudo de regionalização de vazões mínimas com 7 dias de duração na bacia do rio Paraopeba, no qual foi aplicado o método index-flood, definiu-se uma região homogênea com 15 estações fluviométricas. Definiram-se as seguintes variáveis: vazões mínimas (Q), área de drenagem (A), declividade (I) e densidade de drenagem (DD).

O modelo de regressão adotado será do tipo multiplicativo como apresentado a seguir: $Q = \beta_0 A^{\beta_1} X_2^{\beta_2} X_3^{\beta_3}$. Após a transformação logarítmica obtêm-se: $ln\; Q = ln\beta_0 + \beta_1 lnA + \beta_2 lnX_2 + \beta_3 lnX_3$. Assim, para calcular ou estimar os parâmetros da equação $ln\; Q = ln\beta_0 + \beta_1\; lnA + \beta_2\; lnX_2 + \beta_3\; lnX_3$ necessário calcular os logaritmos das variáveis independentes e dependentes.

Tabela 10. Transformações para linearização de diferentes tipos de funções, Mossoró, RN, 2023.

	Tipo de Função	Coordenadas		Equação na Forma Linear
		Abscis	Ordenada	
1	$y = a + bx$	x	y	$[y] = a + b[x]$
2	$y = be^{ax}$	x	log y	$[\log y] = \log b + (a \log c)[x]$
3	$y = ax^{b}$	log x	log y	$[\log y] = \log a + b[\log x]$
4	$y = a_0 + a_1 x + ax^2$	$x - x_0$	$\frac{y - y_0}{x - x_0}$	$\left[\frac{y - y_0}{x - x_0}\right] = a_1 + 2a_1 x_0 + a_2[(x - x_0)]$
5	$y = a + \frac{b}{x}$	$\frac{1}{x}$	y	$[y] = a + b\left[\frac{1}{x}\right]$
6	$y = \frac{x}{(a + bx)}$	x	$\frac{x}{y}$	$\left[\frac{x}{y}\right] = a + b[x]$
7	$y = \frac{a}{(b + cx)}$	x	$\frac{1}{y}$	$\left[\frac{1}{y}\right] = \left(\frac{b}{a}\right) + \left(\frac{c}{a}\right)[x]$
8	$y = c + be^{ax}$	x	$\log \frac{\Delta y}{\Delta x}$	$\left[\log \frac{dy}{dx}\right] = \log(ab) + (a \log e)[x]$
9	$y = c + ax^{b}$	log x	$\log \frac{\Delta y}{\Delta x}$	$\left[\log \frac{dy}{dx} = \log(ab) + (b - 1)[\log x]\right]$
10	$y = c + \frac{b}{x - a}$	$x - x_0$	$\frac{x - x_0}{y - y_0}$	$\left[\frac{x - x_0}{y - y_0}\right] = \frac{a - x}{c - y_0} + \frac{1}{c - y_0}[x - x_0]$
11	$y = c + \frac{x}{a + bx}$	x	$\frac{x - x_0}{y - y_0}$	$\left[\frac{x - x_0}{y - y_0}\right] = (a + bx_0) + \frac{b(a + bx_0)}{a}[x]$
12	$y = d + cx + be^{ax}$	x	$\log \frac{\Delta^2 y}{\Delta x^2}$	$[y - be^{x}] = d + c[x]$ $\left[\log \frac{d^2 y}{dx^2} = \log(a^2 b) + (a \log e)[x]\right]$
13	$y = dc^{x}b^{m}$, onde $m = a^{x}$	x	$\log \frac{\Delta^2(\log y)}{\Delta x^2}$	$\left[\frac{\log d^2(\log y)}{dx^2}\right] = \log\left[\frac{(\log b)(\log a)^2}{(\log e)^2}\right] + (\log a)[x]$
14	$y = de^{cx} + be^{ax}$	$\frac{y_{k+1}}{y_k}$	$\frac{y_{k+2}}{y_k}$	$[ye^{-cx}] = d + b[e^{(a-c)x}]$ $[\log y - a^{x} \log b] = \log d + (\log e)[x]$

				$\left[\frac{y_{k+2}}{y_k}\right] = -e^{(a+c)\Delta x} + \left(e^{a\Delta x} + e^{c\Delta x}\right)\left[\frac{y_{k+1}}{y_k}\right]$

Continuação da Tabela 10

	Tipo de Função	**Coordenadas**		**Equação na Forma Linear**
		Absciss	**Ordenada**	
15	$y = e^{ax}(d\cos bx + csenbx)$	$\frac{y_{k+1}}{y_k}$	$\frac{y_{k+2}}{y_k}$	$\left[\frac{y_{k+2}}{y_k}\right] = -e^{2a\Delta x} + \left(2e^{a\Delta x}\cos b\Delta x\right)\left[\frac{y_{k+1}}{y_k}\right]$ $\left[\frac{yc^{-ax}}{\cos bx}\right] = d + c[\tan bx]$

Observação: Nas equações 14 e 15, y_k, y_{k+1}, y_{k+2} são valores consecutivos para um incremento Δx. Fonte: Yevjevich (1964), pág. 8-49. Citado por Naghettini e Pinto (2007).

Conforme Gujarati (2000), nos estudos envolvendo modelos de regressão em econometria, o pesquisador pode examinar diversas hipóteses sobre o padrão de heterocedasticidade, como descrito a seguir. Se a variância do erro u_i^2 é proporcional ao quadrado das variáveis explicativas X_i^2, isto é, $E(u_i^2) = \sigma^2 X_i^2$, pode-se transformar o modelo original da seguinte maneira dividir o modelo original $Y_i = \beta_1 + \beta_2 X_i + u_i$, por X_i, e assim fica $\frac{Y_i}{X_{i_i}} = \frac{\beta_1}{X_i} + \beta_2 + \frac{u_i}{X_i} = \beta_1 \frac{1}{X_i} + \beta_2 + \upsilon_i$, em que υ_i é o termo de perturbação (erro) transformado, igual a $\frac{u_i}{X_i}$. É fácil verificar agora que $E(\upsilon_i^2) = E\left(\frac{u_i}{X_i}\right)^2 = \frac{1}{X_i^2} E(u_i^2) = \sigma^2$, usando $E(u_i^2) = \sigma^2 X_i^2$.consequentemente, a variância de υ_i é agora homocedástica, e pode-se passar a aplicar o método dos mínimos quadrados ordinários à equação transformada $\frac{Y_i}{X_i} =_i \beta_1 \frac{1}{X_i} + \beta_2 + \upsilon_i$, regredindo, $\frac{Y_i}{X_i}$ sobre $\frac{1}{X_i}$. Segundo ainda o autor vale a pena observar que na regressão transformada $\frac{Y_i}{X_i} =_i \beta_1 \frac{1}{X_i} + \beta_2 + \upsilon_i$, o termo do intercepto β_2 é o coeficiente de inclinação na equação original e o coeficiente de inclinação β_1 é o termo de intercepto no modelo original. Por isso, para retornar ao modelo original, tem-se que multiplicar a $\frac{Y_i}{X_i} =_i \beta_1 \frac{1}{X_i} + \beta_2 + \upsilon_i$ estimada por X_i. Uma hipótese 2 é aquela onde a variância do erro é proporcional a X_i, ou seja $E(u_i^2) = \sigma^2 X_i$ e assim a transformação da raiz quadrada é mais recomendada. Caso se acredite que a variância de u_i, em vez de ser proporcional a X_i elevado ao quadrado, seja proporcional ao próprio X_i, então o modelo original pode ser transformado da seguinte maneira: $\frac{Y_i}{\sqrt{X_i}} =_i \frac{\beta_1}{\sqrt{X_i}} + \beta_2\sqrt{X_i} + \frac{u_i}{\sqrt{X_i}} = \beta_1 \frac{1}{\sqrt{X_i}} + \beta_2\sqrt{X_i} + \upsilon_i$, em que $\upsilon_i = \frac{u_i}{\sqrt{X_i}}$e com $X_i > 0$. Dada a hipótese 2, pode-se facilmente verificar que $E(\upsilon_i^2) = \sigma^2$, uma situação de homocedasticidade. Portanto, pode-se passar a aplicar o método dos mínimos quadrados ordinários a $\frac{Y_i}{\sqrt{X_i}} = \beta_1 \frac{1}{\sqrt{X_i}} + \beta_2\sqrt{X_i} + \upsilon_i$,

regredindo $\frac{Y_i}{\sqrt{X_i}}$ sobre $\frac{1}{\sqrt{X_i}}$ e $\sqrt{X_i}$. Note uma importante característica do modelo transformado, ele não tem o termo de intercepto. Por isso, tem-se de usar o modelo de regressão pela origem para estimarβ_1 e β_2. Após o pesquisador rodar o modelo $\frac{Y_i}{\sqrt{X_i}} = \beta_1 \frac{1}{\sqrt{X_i}} + \beta_2\sqrt{X_i} + \upsilon_i$, ele pode retornar ao modelo original simplesmente multiplicando $\frac{Y_i}{\sqrt{X_i}} = \beta_1 \frac{1}{\sqrt{X_i}} + \beta_2\sqrt{X_i} + \upsilon_i$ por $\sqrt{X_i}$. Ainda segundo o autor uma terceira hipótese é que a variância do erro é proporcional ao quadrado do valor médio de Y, ou seja, $E(u_i^2) = \sigma^2[E(Y_i)]^2$, sendo que esta equação postula que a variância de u_i é proporcional ao quadrado do valor esperado de Y. Então, $E(Y_i) = \beta_1 + \beta_2 X_i$, portanto, se a equação original for transformada como segue tem-se que: $\frac{Y_i}{E(Y_i)} = \frac{\beta_1}{E(Y_i)} + \beta_2 \frac{X_i}{E(Y_i)} + \frac{u_i}{E(Y_i)} = \beta_1 \left(\frac{1}{E(Y_i)}\right) + \beta_2 \frac{X_i}{E(Y_i)} + \upsilon_i$., em que $\upsilon_i = \frac{u_i}{E(Y_i)}$, pode se ver que $E(\upsilon_i^2) = \sigma^2$, ou seja, as perturbações ou erros υ_i são homocedásticas. Consequentemente, é a regressão $\frac{Y_i}{E(Y_i)} = \beta_1 \left(\frac{1}{E(Y_i)}\right) + \beta_2 \frac{X_i}{E(Y_i)} + \upsilon_i$ que satisfará a hipótese de homocedasticidade do modelo clássico de regressão linear. A transformação $\frac{Y_i}{E(Y_i)} = \beta_1 \left(\frac{1}{E(Y_i)}\right) + \beta_2 \frac{X_i}{E(Y_i)} + \upsilon_i$, porém não é operacional, pois $E(Y_i)$ depende de β_1 e β_2, que não são conhecidos. Naturalmente, é conhecido o modelo $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$, que é um estimador de $E(Y_i)$. Por isso pode-se proceder em duas etapas. Primeiro, roda-se a regressão usual de mínimos quadrados ordinários, desconsiderando o problema da heterocedasticidade, e obtém-se $\hat{Y}_i$. Em seguida, usando os $\hat{Y}_i$ estimados, transforma-se o modelo assim como segue $\frac{Y_i}{\hat{Y}_i} = \beta_1 \left(\frac{1}{\hat{Y}_i}\right) + \beta_2 \left(\frac{X_i}{\hat{Y}_i}\right) + \upsilon_i$, em que $\upsilon_i = \left(\frac{u_i}{\hat{Y}_i}\right)$. No passo 2, roda-se a regressão $\frac{Y_i}{\hat{Y}_i} = \beta_1 \left(\frac{1}{\hat{Y}_i}\right) + \beta_2 \left(\frac{X_i}{\hat{Y}_i}\right) + \upsilon_i$. Embora $\hat{Y}_i$ não sejam exatamente $E(Y_i)$, eles são estimadores consistentes, ou seja, conforme o tamanho da amostra aumenta indefinidamente, eles convergem para a verdadeira $E(Y_i)$. Por isso, a transformação $\frac{Y_i}{\hat{Y}_i} = \beta_1 \left(\frac{1}{\hat{Y}_i}\right) + \beta_2 \left(\frac{X_i}{\hat{Y}_i}\right) + \upsilon_i$ cumprirá satisfatoriamente seu papel na prática se a amostra for razoavelmente grande. Ainda conforme o autor, pode existir uma quarta hipótese onde se diz que uma transformação em log do tipo $ln\, Y_i = \beta_1 + \beta_2\, ln\, X_i + u_i$, muitas vezes reduz a heterocedasticidade quando comparada com a regressão $Y_i = \beta_1 + \beta_2 X_i + u_i$. Este resultado ocorre porque a transformação em log comprime as escalas nas quais as variáveis são medidas reduzindo assim uma diferença de dez vezes entre dois valores para uma diferença de duas vezes. Assim segundo o autor, o número 80 é 10 vezes o número 8, mas $ln\, 80$ que é igual a 4,3280 é cerca de o dobro do $ln\, 8$ que é igual a 2,0794. Uma vantagem adicional da transformação em log é que o coeficiente de inclinação β_2 mede a elasticidade de Y com relação a X, ou seja, a variação percentual em Y para uma variação percentual em X. Por exemplo, se Y for consumo e X for renda, β_2

em $ln\, Y_i = \beta_1 + \beta_2\, ln\, X_i + u_i$ medirá a elasticidade-renda, enquanto no modelo original β_2 mede apenas a taxa de variação do consumo médio para uma variação de uma unidade na renda.

Para concluir a discussão das medidas corretivas, deve-se enfatizar que todas as transformações examinadas anteriormente são ad hoc, pois se está apenas especulando sobre a natureza de σ_i^2. Sendo assim o autor afirma que de todas as transformações discutidas anteriormente a que irá funcionar depende da natureza do problema e da gravidade da heterocedasticidade (GUJARATI, 2000).

Gujarati (2000) descreve que as transformações que foram examinadas anteriormente apresentam alguns problemas adicionais que não podem ser esquecidos e que são relatados a seguir.

1) quando se avança para além do modelo de duas variáveis, pode-se não saber a princípio qual das vaiáveis X deve ser escolhida para transformar os dados. (por questão prática, porém, pode-se representar graficamente contra cada variável e decidir qual variável X pode ser usada para transformar os dados);

2) a transformação em log discutida na hipótese 4 não é aplicável se alguns dos valores Y e X for zero ou negativo (pode-se às vezes usar $ln(Y_i + K)$ ou $ln(X_i + K)$, em que K é um número positivo escolhido de tal maneira que todos os valores de Y e X se tornem positivos).

3) quando há o problema da correlação espúria. Este termo, devido ao estatístico inglês Karl Pearson, se refere a situação em que se verifica que a correlação está presente entre as razões das variáveis mesmo que as variáveis originais não tenham correlação ou sejam aleatórias, por exemplo, se X_1, X_2 e X_3 forem mutuamente não correlacionadas $r_{12} = r_{13} = r_{23} = 0$ e verifica-se que razões (valores das) $\frac{X_1}{X_3}$ e $\frac{X_2}{X_3}$ são correlacionadas, então há uma correlação espúria. Mais genericamente, uma correlação pode ser descrita como espúria se ela for induzida pelo método de como os dados foram manipulados e não estiver presente no material original. Assim no modelo $Y_i = \beta_1 + \beta_2 X_i + u_i$, Y e X podem não ter correlação, mas no modelo transformado $\frac{Y_i}{X_i} = \beta_1 \left(\frac{1}{X_i}\right) + \beta_2 \left(\frac{Y_i}{X_i}\right)$e $\frac{1}{X_i}$ muitas vezes mostram ter correlação. 4) quando σ_i^2não forem diretamente conhecidos e forem estimados a partir de uma ou mais transformações que foram apresentadas e discutidas anteriormente, todos os procedimentos de teste com o uso dos testes t, F, etc. são, rigorosamente falando, válidos somente em grandes amostras. Por isso o pesquisador, precisa tomar cuidado ao interpretar os resultados baseados nas diferentes transformações em amostras pequenas ou finitas.

Conforme Kmenta (2000), as transformações de varáveis em análise de regressão em séries temporais, sobre consumo, renda e riqueza, pode ser realizada em razão da alta multicolinearidade entre renda e riqueza em tais dados é que, com o tempo, ambas as variáveis tendem a se mover na mesma direção, e um meio de minimizar esta dependência é proceder como segue. Se a relação $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$ for válida no instante t, deve ser válida também no instante $t - 1$, porque de

qualquer forma a origem de tempo é arbitrária, sendo assim tem-se que $Y_{t-1} = \beta_1 + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + u_{t-1}$. Se subtrair $Y_{t-1} = \beta_1 + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + u_{t-1}$ de $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$, obtém-se $Y_t - Y_{t-1} = \beta_2(X_{2t} - X_{2,t-1}) + \beta_3(X_{3t} - X_{3,t-1}) + v_t$, em que $v_t = u_t - u_{t-1}$. A equação $Y_t - Y_{t-1} = \beta_2(X_{2t} - X_{2,t-1}) + \beta_3(X_{3t} - X_{3,t-1}) + v_t$é conhecida como forma de primeira diferença, porque o modelo é rodado no pacote estatístico com a regressão não sobre as variáveis originais, mas sobre as diferenças dos sucessivos valores das variáveis. O modelo de regressão de primeira diferença muitas vezes reduz a severidade da multicolinearidade, pois embora os níveis de X_2 e X_3 possam ser altamente correlacionados, não há nenhuma razão, a princípio, para se acreditar que suas diferenças também sejam altamente correlacionadas. A transformação para a primeira diferença, porém, cria alguns problemas adicionais. O termo de erro v_t que aparece em $Y_t - Y_{t-1} = \beta_2(X_{2t} - X_{2,t-1}) + \beta_3(X_{3t} - X_{3,t-1}) + v_t$ pode não satisfazer uma das hipóteses do modelo de clássico de regressão linear, que é aquele em que os erros não possuem correlação serial. Se o termo de erro original u_t for seriamente independente ou não correlacionado, o termo de erro v_t obtido anteriormente terá na maioria dos casos correlação serial. Além disso, há uma perda de uma observação em virtude do procedimento de diferenciação, por isso os graus de liberdade se reduzem em um. Em uma amostra pequena, este pode ser um fator que pelo menos deve ser levado em conta. Além do mais, o método da primeira diferença pode não ser apropriado em dados de corte em que não haja nenhuma ordenação lógica das observações.

Gujarati (2000), afirma que no ajuste do modelo de regressão logístico dado pelo modelo de equação $L_i = ln\left(\frac{P_i}{1-P_i}\right) = \beta_1 + \beta_2 + u_i$ a partir de dados agrupados ou replicados, e se o tamanho da amostra n razoavelmente grande, e também se cada observação em uma determinada classe de valores X_i se distribuir independentemente como uma variável binomial, então, ou seja, os erros u_i, dado por $u_i \cap N\left[0, \frac{1}{N_i P_i(1-P_i)}\right]$ segue a distribuição normal com média zero e variância igual a $\frac{1}{[N_i P_i(1-P_i)]}$. Portanto, o termo do erro u_i no modelo $logit$ é heterocedástico. Sendo assim, em vez de usar o método dos mínimos quadrados ordinários para estimar os parâmetros do modelo, tem-se que usar o método dos mínimos quadrados ponderados, e para fins empíricos substitui-se o P_i desconhecido pelo $\hat{P}$, e usa-se a seguinte estimador $\hat{\sigma}^2 = \frac{1}{N_i P_i(1-P_i)}$da variância dos erros como estimador de σ^2. Conforme ainda o autor, os passos na estimação do modelo de regressão logística são os seguintes: 1) para cada nível do valor X, calcula-se a probabilidade estimada de possuir um bem ou de ocorrer o evento como $\hat{P}_i = \frac{n_i}{N_i}$, 20), e para cada X_i obtém-se o $logit$ como $\hat{L}_i = ln\frac{\hat{P}_i}{[(1-\hat{P}_i)]}$; 3) para resolver o problema da heterocedasticidade, transforma-se $L_i = ln\left(\frac{P_i}{1-P_i}\right) = \beta_1 + \beta_2 + u_i$, como segue: $\sqrt{w_i}L_i = \beta_1\sqrt{w_i} + \beta_2\sqrt{w_i}X_i + \sqrt{w_i}u_i$, que passa a ser escrita como: $L_i^* = \beta_1\sqrt{w_i} + \beta_2\sqrt{w_i}X_i^* + v_i$, em que os pesos $w_i = N_i\hat{P}_i(1-\hat{P}_i)$; $L_i^* =$

L_i, transformado ou ponderado; $X_i^* = X_i$ transformado ou ponderado; e v_i é o termo de erro transformado. É fácil verificar que o termo de erro transformado v_i é homocedástico, tendo em mente que a variância de erro original é $\sigma_u^2 = \frac{1}{N_i P_i(1-P_i)}$; 4) Estima-se $\sqrt{w_i}L_i = \beta_1\sqrt{w_i} + \beta_2\sqrt{w_i}X_i + \sqrt{w_i}u_i$ pelo método dos mínimos quadrados ordinários, lembrando que método dos mínimos quadrados ponderados é o método dos mínimos quadrados ordinários sobre os dados transformados. Vale a pena lembrar e reparar que em $\sqrt{w_i}L_i = \beta_1\sqrt{w_i} + \beta_2\sqrt{w_i}X_i + \sqrt{w_i}u_i$ não há um termo de intercepto introduzido explicitamente, por isso quando o pesquisador for rodar os dados no pacote estatístico no computador deve ele usar a rotina da regressão passando pela origem para estimar $\sqrt{w_i}L_i = \beta_1\sqrt{w_i} + \beta_2\sqrt{w_i}X_i + \sqrt{w_i}u_i$; 5) o autor sugere estabelecer ou construir intervalos de confiança e aplicar testes de hipóteses na estrutura usual dos mínimos quadrados ordinários, mas tendo em mente que, rigorosamente falando, todas as conclusões serão válidas se a amostra for razoavelmente grande. Portanto em amostras pequenas, os resultados estimados devem ser interpretados com cautela.

Fonseca et al. (1982), descreve o cálculo do coeficiente de correlação linear simples de Pearson (r) para dados agrupados em intervalos de classes, pois segundo os autores, muitas os valores das variáveis X e Y são dispostos em uma tabela ou distribuição de frequência bidimensional. Neste caso, ao invés de calcular o coeficiente de correlação a partir de valores individuais das variáveis, utiliza-se uma tabela onde tais valores encontram-se agrupados em classes, e para efeito de cálculo toma-se os pontos médios das classes como os valores das variáveis que irão compor a fórmula, por isso deve-se ter os seguintes símbolos: X_i são os pontos médios das classes em que os valores da variável X foram agrupados; Y_i os pontos médios das classes em que os valores da variável Y foram agrupados; f_{x_i} é a frequência simples absoluta da i-ésima classe de valores de X ou frequência marginal de X; f_{y_i} é a frequência simples absoluta da i-ésima classe de valores de Y ou frequência marginal de Y; $f_{x_i y_i}$= frequência simples absoluta conjunta da i-ésima classe dos valores de X e j-ésima classe dos valores de Y ou frequência correspondente aos pares de pontos médios. Segundo os autores os resultados são obtidos diretamente da fórmula original do coeficiente de correlação, isto, é:

$$r = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} X_i Y_j - \frac{\left(\sum_{i=1}^{n} X_i \sum_{j=1}^{n} Y_j\right)}{n}}{\sqrt{\left[\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right]\left[\sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}\right]}}$$

e assim tem-se,

$$r = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} X_i Y_j f_{x_i y_j} - \left(\sum_{i=1}^{n} X_i f_{x_i}\right)\left(\sum_{j=1}^{n} Y_j f_{y_j}\right)}{\sqrt{\left[n\sum_{i=1}^{n} X_i^2 f_{x_i} - \left(\sum_{i=1}^{n} X_i f_{x_i}\right)^2\right]\left[n\sum_{i=1}^{n} Y_j^2 - \left(\sum_{i=1}^{n} Y_j f_{y_{ij}}\right)^2\right]}}$$

e de acordo com a propriedade do coeficiente de correlação, podem-se usar variáveis transformadas para facilitar os cálculos. A fórmula poderá ser modificada da seguinte maneira.

$$U = \frac{X - X_0}{c} \text{ e } U = \frac{Y - Y_0}{d}$$

Os valores

$$r = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} X_i Y_j f_{x_i y_j} - \left(\sum_{i=1}^{n} X_i f_{x_i}\right)\left(\sum_{j=1}^{n} Y_j f_{y_j}\right)}{\sqrt{\left[n\sum_{i=1}^{n} X_i^2 f_{x_i} - \left(\sum_{i=1}^{n} X_i f_{x_i}\right)^2\right]\left[n\sum_{i=1}^{n} Y_j^2 - \left(\sum_{i=1}^{n} Y_j f_{y_{ij}}\right)^2\right]}}$$

são constantes arbitrárias ou médias provisórias, respectivamente, de X e de Y. Costuma-se atribuir para X_0 e Y_0 os valores dos pontos médios das classes de maior frequência. As constantes arbitrárias c e d são normalmente escolhidas como os valores correspondentes às respectivas amplitudes dos intervalos de classe. Lembrando que $f_{x_i,y_i} = f_{u,v}$, $f_{x_i} = f_u$ e $f_{y_i} = f_v$, e assim deve-se ter o seguinte coeficiente.

$$r_{(X,Y)} = r_{(U,V)} = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} UV f_{u,v} - (\sum_{i=1}^{n} U f_u)(\sum_{j=1}^{n} V f_v)}{\sqrt{[n\sum_{i=1}^{n} U^2 f_u - (\sum_{i=1}^{n} U f_u)^2]\ [\ n\sum_{i=1}^{n} V^2 f_v - (\sum_{i=1}^{n} V f_v)^2]}}$$

Segundo Fonseca et al. (1982), o coeficiente de correlação linear simples de Pearson (r) será definido como a razão entre a covariância e a raiz quadrada do produto das variações de X e de Y, simbolicamente ele é dado por.

$$r = \frac{\sum_{i=1}^{n}(X - \bar{X})\,(Y - \bar{Y})}{\sqrt{[\sum_{i=1}^{n}(X - \bar{X})^2][\sum_{i=1}^{n}(Y - \bar{Y})^2]}} = \frac{cov(X,Y)}{S_x S_y}$$

e dividindo-se o numerador e o denominador por n que o tamanho da amostra de pares de valores, o coeficiente r será definido como a razão da covariância e o produto dos desvios padrão de X e de Y.

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(X - \bar{X})\,(Y - \bar{Y})}{\frac{1}{n}\sqrt{[\sum_{i=1}^{n}(X - \bar{X})^2][\sum_{i=1}^{n}(Y - \bar{Y})^2]}} = \frac{cov(X,Y)}{S_x S_y}$$

Segundo ainda estes autores outras expressões podem ser encontradas para calcular o coeficiente de correlação linear simples de Pearson (r), usando variáveis transformadas como $x = X - \overline{X}$ e $y = Y - \overline{Y}$

$$r = \frac{covariância(X,Y)}{\sqrt{\left[\frac{\sum_{i=1}^{n}(X-\bar{X})^2}{n}\right]\left[\frac{\sum_{i=1}^{n}(Y-\bar{Y})^2}{n}\right]}} = \frac{cov(X,Y)}{S_x S_y}, \text{ ou ainda, } r = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j}{\sqrt{[\sum_{i=1}^{n} x_i^2\ ]\ [\ \sum_{i=1}^{n} y_i^2\ ]}}$$

Já na análise de regressão linear simples, os estimadores dos parâmetros $\hat{a}\ e\ \hat{b}$ da equação $\hat{Y}_i = \hat{a}\ +\ \hat{b}X_i + e_i$ utilizando as variáveis transformadas como $x = X - \overline{X}$e $y = Y\ - -\overline{Y}$, são dados por

$$\hat{a}\ =\ \overline{Y}\ ,\quad \hat{b}\ =\ \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j}{\sum_{i=1}^{n} x_i^2}$$

Segundo Fonseca et al. (1982), para contornar o problema de inferência estatística quando a hipótese de nulidade afirma que o verdadeiro coeficiente de correlação linear simples de Pearson "$\boldsymbol{\rho}$" da população assume valores que não são nulos podendo ser qualquer valor diferente de zero, e para que seja possível construir intervalos de confiança e aplicar testes de hipóteses, é necessário uma transformação do coeficiente de correlação amostral r, usando para isso uma variável transformada "$\boldsymbol{Z_r}$", Fisher (1915) demonstrou que essa transformação do coeficiente de correlação $\boldsymbol{r}$ em $\boldsymbol{Z_r}$ produziria uma variável normalmente distribuída com média zero. Sendo assim esse autor derivou uma aproximação normal de "$\boldsymbol{r}$" em "$\boldsymbol{Z_r}$", cuja distribuição amostral possui distribuição teórica de probabilidade normal ou gaussiana aproximada ou assintótica cuja média e desvio padrão são mostrados a seguir.

Sendo assim, $\boldsymbol{Z_r = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right]}$ possui distribuição normal assintótica com média: $\boldsymbol{\mu_{Z_r} = \frac{1}{2}\left[\ln\left(\frac{1+\rho}{1-\rho}\right) + \frac{\rho}{n-1}\right]}$, variância dada por: $\boldsymbol{\sigma^2_{Z_r} \cong \frac{1}{n-3}}$, onde o desvio padrão de $\boldsymbol{Z_r}$ é dado por $\boldsymbol{S_{Z_r} = \frac{1}{\sqrt{n-3}}}$, onde $\boldsymbol{n}$ é a dimensão da amostra.

Então a variável pivotal é dada por: $\boldsymbol{Z = \frac{Z_r - Z_\rho}{\frac{1}{\sqrt{n-3}}} \cap N(0,1)}$.

A variância de "$\boldsymbol{Z}$" não envolve o verdadeiro valor do parâmetro e foi essa propriedade de estabilização da variância que conduziu Fisher a propor a transformação da tangente hiperbólica inversa dada por $\boldsymbol{Z_r = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right]}$. Essa transformação não é válida para $\boldsymbol{\rho\ =\ -1}$ ou $\boldsymbol{\rho\ =\ 1}$ e para $\boldsymbol{\rho\ =\ 0}$, a distribuição exata de "$\boldsymbol{t}$" deve ser preferida. Dessa forma, a estatística $\boldsymbol{Z = Z_C = \frac{(Z_r - \mu_Z)}{\sigma_Z}}$ deve ser usada para realizar testes de hipóteses e obter intervalos de confiança para o coeficiente populacional $\boldsymbol{\rho}$.

Existem tabelas prontas para $\boldsymbol{Z_r}$ em função de vários valores de $\boldsymbol{r}$.

Portanto o intervalo de confiança para $\boldsymbol{Z_r}$ com $\boldsymbol{100(1 - \alpha)\%}$ de confiança é dado por:

$$\boldsymbol{P\left[Z_r - Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}} \le Z_R \le Z_r + Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}}\right] = 1 - \alpha}$$

como $\boldsymbol{Z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)}$, e assim o intervalo fica como mostrado a seguir.

$$\boldsymbol{P\left[\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}} \le Z_R \le \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) + Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}}\right] = 1 - \alpha}$$

Agora, portanto para transformar o intervalo obtido num intervalo de confiança para o coeficiente de correlação populacional "$\boldsymbol{\rho}$" utiliza-se novamente a tabela que relaciona $\boldsymbol{Z_r}$ com $\boldsymbol{r}$, ou então calculando os limites do intervalo através da expressão $\boldsymbol{r = \frac{e^{2Z_r} - 1}{e^{2Z_r} + 1}}$. Vale salientar que o intervalo

de confiança para $\boldsymbol{Z_r}$ está centrado em torno de $\boldsymbol{Z_r}$, enquanto que o intervalo de confiança resultante para o coeficiente "$\boldsymbol{\rho}$" não está centrado em torno de $\boldsymbol{r}$.

Portanto o intervalo de confiança para o coeficiente "$\boldsymbol{\rho}$" é o seguinte:

$$\boldsymbol{P\left[\frac{e^{2Z_{r_{inf}}}}{e^{2Z_{r_{inf}}}} \leq \rho \leq \frac{e^{2Z_{r_{sup}}}}{e^{2Z_{r_{sup}}}}\right] = 1-\alpha}$$

onde

$$\boldsymbol{Z_{r_{inf}} = Z_r - Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}}}$$

$$\boldsymbol{Z_{r_{sup}} = Z_r + Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}}}$$

$$Z_r = \frac{1}{2}\, ln\left(\frac{1+r}{1-r}\right)$$

Sendo assim temos que:

$$\boldsymbol{P\left[\frac{e^{2\left(\frac{1}{2} ln\left(\frac{1+r}{1-r}\right) - Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}}\right)}}{e^{2\left(\frac{1}{2} ln\left(\frac{1+r}{1-r}\right) - Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}}\right)}} \leq \rho \leq \frac{e^{2\left(\frac{1}{2} ln\left(\frac{1+r}{1-r}\right) + Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}}\right)}}{e^{2\left(\frac{1}{2} ln\left(\frac{1+r}{1-r}\right) + Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}}\right)}}\right] = 1-\alpha}$$

É possível também obter os valores limites para um intervalo que contenha o verdadeiro valor do coeficiente de correlação populacional ϱ utilizando-se gráficos, que se referem, respectivamente, a coeficientes de confiança de 95 % e 99% de probabilidade. A utilização desses gráficos é feita da seguinte maneira: para um coeficiente de confiança $100(1-\alpha)$ % fixado, toma-se o valor observado de $r = r_0$ na amostra de tamanho n em estudo e localiza-se este valor no eixo das abscissas; a parti dele levanta-se uma paralela ao eixo das ordenadas até encontrar as duas curvas referentes ao n em questão. De cada um dos pontos de encontro traça-se uma reta paralela ao eixo das abscissas até encontrar o eixo dos valores de ρ, obtendo-se assim os limites inferior e superior do intervalo de confiança para ρ.

Frequentemente pode ser de interesse do pesquisador, obter várias correlações e desejar-se saber se uma delas é significativamente mais alta que as demais. Vale a pena mostrar o caso de duas amostras independentes, cujos coeficientes de correlação se deseja comparar. Se duas correlações se basearem em amostras independentes, pode –se escrever o seguinte.

$$\begin{cases} \boldsymbol{H_0: \rho_1 = \rho_2} \\ \boldsymbol{H_1: \rho_1 \neq \rho_2} \end{cases} \text{ ou } \begin{cases} \boldsymbol{H_0: \rho_1 - \rho_2 = 0} \\ \boldsymbol{H_1: \rho_1 - \rho_2 \neq 0} \end{cases}$$

$$\boldsymbol{\sigma_{Z_1 - Z_2} = \sqrt{\sigma_{Z_1}^2 + \sigma_{Z_2}^2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

e assim tem-se,

$$\boldsymbol{E[Z_1 - Z_2] = E[Z_1] - E[Z_2] = \mu_{Z_1} - \mu_{Z_2} = 1,1513\, log\left[\frac{1 + (\rho_1 - \rho_2)}{1 - (\rho_1 - \rho_2)}\right] = 0}$$

$$\boldsymbol{z_c = \frac{(Z_1 - Z_2) - (\mu_{Z_1} - \mu_{Z_2})}{\sigma_{Z_1 - Z_2}} = \frac{(Z_1 - Z_2) - 0}{\sigma_{Z_1 - Z_2}}}$$

Bussab e Morettin (2002), mostram como comparar coeficientes de correlação de duas populações através de mudanças de variáveis. Isto é, suponha que ρ_1 e ρ_2 sejam os coeficientes de correlação de duas populações, dos quais retiramos duas amostras independentes, de tamanho n e m, respectivamente. Deste modo, as variáveis aleatórias $Z_1 = \frac{1}{2} ln \frac{1+r_1}{1-r_1}$ e $Z_2 = \frac{1}{2} ln \frac{1+r_2}{1-r_2}$, são independentes e terão, respectivamente, as distribuições $Z_1 \sim N\left(\mu_1; \frac{1}{n-3}\right)$ e $Z_2 \sim N\left(\mu_2; \frac{1}{m-3}\right)$ com $\mu_1 = \frac{1}{2} ln \frac{1+\rho_1}{1-\rho_1}$ e $\mu_2 = \frac{1}{2} ln \frac{1+\rho_2}{1-\rho_2}$. Segue-se que a variável aleatória $D = Z_1 - Z_2$ terá distribuição normal, com média $\mu_D = \mu_1 - \mu_2 = \frac{1}{2} ln \left(\frac{1+\rho_1}{1-\rho_1}, \frac{1-\rho_2}{1+\rho_2}\right)$ e variância $\sigma_D^2 = \frac{1}{(n-3)} + \frac{1}{(m-3)}$. Quando, $\rho_1 = \rho_2$ temos que $\mu_D = 0$. Este resultado permite testar se dois coeficientes de correlação são iguais ou não.

Segundo Fonseca et al. (1982), o método dos mínimos quadrados empregado na estimação dos parâmetros da equação de regressão, pode ser usado para avaliar a equação de uma reta ou curva de tendência em análise de séries temporais, e neste caso, a variável dependente é o Y e o tempo t, a variável independente. Efetuando a mudança de variável $(t \rightarrow X)$, onde $X_i = t_i - t_0$, sendo t_0 o termo médio da série, e assim tem-se a simplificação de cálculos. Ou seja, para estimarmos os parâmetros $\alpha \; e \; \beta$ da curva de tendência linear $\hat{Y}_i = \hat{a} \; + \; \hat{b} X_i$, deve-se efetuar o seguinte cálculo.

$$\hat{a} \; = \frac{\sum_{i=1}^{n} Y_i}{n} \; = \overline{Y} \; , \quad \hat{b} \; = \; \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} X_i Y_j}{\sum_{i=1}^{n} X_i^2}$$

SOUZA (1998), Afirma que antes de ajustar modelos de regressão o pesquisador deve fazer a chamada análise residual, cuja ideia é a de utilizar os resíduos da regressão no estabelecimento de técnicas visuais e formais para detectar desvios das hipóteses clássicas do modelo linear. Segundo este autor quando os resíduos ε_i são independentes e possuem distribuição normal com média zero e variância não homogênea, ou seja, $N(0, w_i \sigma^2)$ e $w_i = f(x_i' \beta)$ para alguma função $f(y)$, a forma corriqueira de corrigir o problema é mediante uma transformação não-linear da variável dependente. Quando o processo é bem sucedido, induz não somente homocedasticiade como também um melhor ajuste da distribuição normal aos resíduos. Para ilustrar intuitivamente o processo, seja $\varphi(y)$ a transformação de interesse. O desenvolvimento de Taylor de $f(y)$ numa vizinhança de $\mu_i = E(y_i)$ fornece a aproximação linear $\varphi(y) = \varphi(\mu_i) + \varphi'(\mu_i)(y_i - \mu_i)$. Portanto, com um apelo ao Teorema Central de Limite se necessário, aproximadamente, $\varphi(y_i) \sim N(\varphi(\mu_i), \sigma^2 f(\mu_i)[\varphi'(\mu_i)]^2)$. Se puder se escolher $\varphi(y)$ de

modo a satisfazer $\varphi'(y) = \frac{1}{\sqrt{f(y)}}$, a variável aleatória $\varphi(y_i)$ terá variância constante. Exemplos simples desta configuração se obtêm quando $f(y) = y$ e $f(y) = y^2$. Nestes casos tem-se $\varphi(y) = 2\sqrt{y}$e $\varphi(y) = ln(y)$ respectivamente. A transformação da raiz quadrada, por exemplo, está indicada quando $y_i$tem distribuição de Poisson. Em geral se $f(y) = y^{\alpha}$com $\alpha \neq 2$então $\varphi(y) = \frac{y^{1-\frac{\alpha}{2}}}{1-\frac{\alpha}{2}}$. Se y se distribui como uma variável binomial, então $f(y) = y(1-y)$ e $\varphi(y) = 2arcsen\left(\sqrt{y}\right)$ estabiliza a variância e normaliza y. Quando do uso desta transformação em proporções y_i, estas devem ser computadas com base no mesmo número de observações. Caso contrário, será necessário o uso de mínimos quadrados ponderados, pois as observações serão heterocedásticas. É claro que se espera também que $\varphi(\mu_i) = x_i'\beta$. O problema, numa dada aplicação, é escolher a transformação adequada. Candidatos típicos em potencial são $\varphi(y) = \sqrt{y}$ e $\varphi(y) = ln(y)$. Um método mais formal para a escolha de λ na família

$$\varphi_\lambda(y) = \begin{cases} -y^\lambda & \lambda < 0 \quad y > 0 \\ y^\lambda & \lambda > 0 \quad y > 0 \\ ln(y) & \lambda = 0 \quad y > 0 \end{cases}$$

foi sugerido por Box e Cox (1964) citados por Souza (1998) com uma motivação diferente. Na realidade Box e Cox (1964) citado por Souza (1998), consideram a família modificada

$$\varphi_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \quad y > 0 \\ \\ ln(y) & \lambda = 0 \quad y > 0 \end{cases}$$

A hipótese fundamental feita por Box e Cox (1964), citados por Souza (1998), é a de que para algum λ as observações transformadas $\varphi_\lambda(y)$ podem ser tratadas como normalmente e independentemente distribuídas com variância constante e esperança matemática definida pelo modelo linear $E\left(\varphi(y)\right) = X\beta$. Box e Cox (1964), citados por Souza (1998) afirmam que, como os resultados da análise de variância não são afetados por transformações lineares, a análise induzida pela Família dada por,

$$\varphi_\lambda(y) = \begin{cases} -y^\lambda & \lambda < 0 \quad y > 0 \\ y^\lambda & \lambda > 0 \quad y > 0 \\ ln(y) & \lambda = 0 \quad y > 0 \end{cases}$$

É equivalente à análise induzida pela Família dada por

$$\varphi_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \quad y > 0 \\ \\ ln(y) & \lambda = 0 \quad y > 0 \end{cases}$$

Deve-se observar, contudo, que esta afirmação só é verdadeira para modelos com intercepto. Em modelos sem intercepto a Família dada por

$$\varphi_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \quad y > 0 \\ ln(y) & \lambda = 0 \quad y > 0 \end{cases}$$

conduz a procedimentos que não são invariantes por transformações de escala, o que não ocorre com a Família dada por

$$\varphi_\lambda(y) = \begin{cases} -y^\lambda & \lambda < 0 \quad y > 0 \\ y^\lambda & \lambda > 0 \quad y > 0 \\ ln(y) & \lambda = 0 \quad y > 0 \end{cases}$$

De acordo ainda com Souza (1998), com base na especificação normal da variável transformada o método de Box e Cox se resume em estimar λ pela técnica de máxima verossimilhança. A abordagem tem sido objeto de crítica principalmente na literatura econométrica. A dificuldade conceitual é que exceto no caso $\lambda = 0$ não se pode ter $\varphi_\lambda(y)$ normalmente distribuído, pois $\varphi_\lambda(y) > \frac{-1}{\lambda}$se $\lambda > 0$ e $\varphi_\lambda(y) < \frac{-1}{\lambda}$ se $\lambda < 0$com $y > 0$. Outro problema é o de tratar o valor estimado de λ como fixo, isto é, não estocástico, nas análises subsequentes. Apesar disto, o método tem bastante apelo como ferramenta exploratória. O procedimento segue. Inicialmente escolhe-se valores para λ num determinado intervalo. Draper e Smith (1981) citados por Souza (1998), sugerem em princípio os intervalos (-1,1) ou (-2,2). Para cada λ calcula-se então as quantidades $w_i = \frac{\varphi_\lambda(y_i)}{[\dot{y}]^{(\lambda-1)}}$onde $\dot{y}$ é a média geométrica das observações y_i. Seja $SSE(\lambda)$ a soma de quadrados dos resíduos da regressão dos w_i em X. O estimador de máxima verossimilhança de λ é o valor $\hat{\lambda}$ que minimiza $SSE(\lambda)$ ou equivalentemente que maximiza $L(\lambda) = -\frac{1}{2} n\, ln\left(\frac{SSE(\lambda)}{n}\right)$. Um intervalo de confiança para λ ao nível de $100(1-\alpha)$ % é definido pela região $L(\hat{\lambda}) - L(\lambda) \leq \frac{1}{2}\chi_1^2(1-\alpha)$ onde $\chi_1^2(1-\alpha)$ é o quantil de ordem $1-\alpha$ da distribuição quadrado com um grau de liberdade. O método expedito para determinar esta região bem como o ponto $\hat{\lambda}$ é pelo gráfico de $L(\lambda)$ contra λ. O valor $\hat{\lambda}$ é encontrado por inspeção visual. A reta $y = L(\hat{\lambda}) - \frac{1}{2}\chi_1^2(1-\alpha)$ determina dois pontos da curva $L(\lambda)$ que correspondem aos extremos do intervalo para λ.

De acordo com Ogliari (2015), Algumas medidas para contornar problemas no ajuste em modelo de regressão devem ser utilizados conforme seja detectado os seguintes problemas: o modelo de regressão linear simples não é adequado, então deve-se usar um modelo apropriado ou o uso de transformações de dados; já se ocorrer não linearidade do modelo de regressão, então o pesquisador deve mudar o modelo, como por exemplo as seguintes equações.

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2$$

$$E(Y) = \beta_0 \beta_1^{\,X} \quad \text{(Exponenci al)}$$

$$E(Y) = \frac{\beta_0}{1+\beta_1 \exp(\beta_2 X_i)} \quad \text{(logístico)}$$

Ou então usar transformação de variáveis. Segundo ainda este autor se ocorrer variâncias heterogêneas, então se deve adotar o método de mínimos quadrados ponderados para estimar os parâmetros, ou ainda usar transformação de dados. Se os erros forem correlacionados então usar modelos que levam em consideração a dependência entre os erros, como por exemplo, modelos de séries temporais, modelar a matriz de covariâncias e ainda usar transformação, como a seguinte $(Y_t^{'} = Y_t - \rho Y_{t-1})$. Já a falta de normalidade dos erros geralmente vem junto com falta de homogeneidade de variâncias. Frequentemente, a mesma transformação estabiliza a variância e aproxima para normalidade, portanto, primeiro usar uma transformação para estabilizar a variância. Na omissão de variável preditora importante, deve ser feito uma mudança no modelo, isto, é usar um modelo de Regressão linear múltipla. Se ocorrer presença de dados aberrantes ou Outliers, o pesquisador deve usar procedimentos de estimação robustos, como o método dos mínimos quadrados reponderados iterativamente, pois os métodos de mínimos quadrados e máxima verossimilhança produzem estimativas distorcidas.

Segundo ainda Ogliari (2015), a aplicação de transformação da variável Y ou da variável preditora X, ou de ambas, frequentemente é suficiente para tornar o modelo de regressão linear simples apropriado para os dados transformados. No uso de transformações para não linearidade do modelo, o autor considera algumas transformações quando a distribuição dos erros é aproximadamente normal e com variância constante. Segundo ele deve-se realizar uma transformação apenas na variável X. Ele mostra ainda que os padrões de relação entre X e Y, são aqueles mostrados na Figura abaixo (Figura 2).

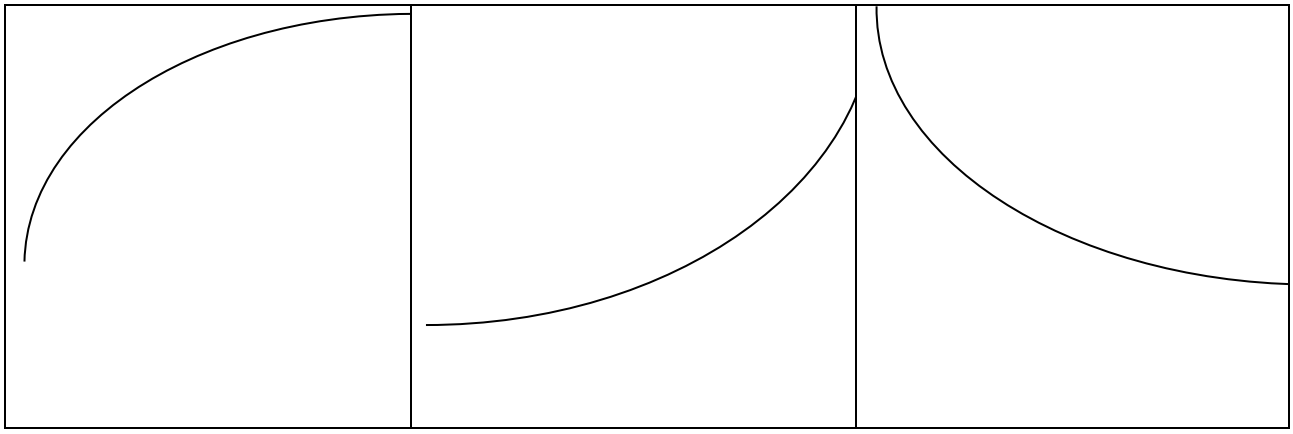


Figura 3. Padrões de relacionamento entre as variáveis X e Y no estudo de regressão.Mossoró, RN, 2023.

As transformações sugeridas por ele são as seguintes:

$X' = log_{10} X$ $\quad$ $X' = X^2$ $\quad$ $X^{'} = 1/X$

$X' = \sqrt{X}$ $\quad$ $X' = exp(X)$ $\quad$ $X^{'} = exp(-X)$

As transformações para não normalidade e heterocedasticidade, devem ser usadas, pois variâncias heterogêneas e não normalidade dos erros frequentemente aparecem juntas. Necessita-se fazer uma transformação em *Y*, pois a forma e a dispersão em *Y* precisam ser modificadas. A transformação em *Y* pode também eliminar o problema de não linearidade do modelo. Outras vezes uma transformação também em X é necessária para manter ou obter uma relação linear. A Figura 3 a seguir ilustra algumas formas de relacionamento onde a assimetria e as variâncias aumentam com a resposta média $E(Y)$.

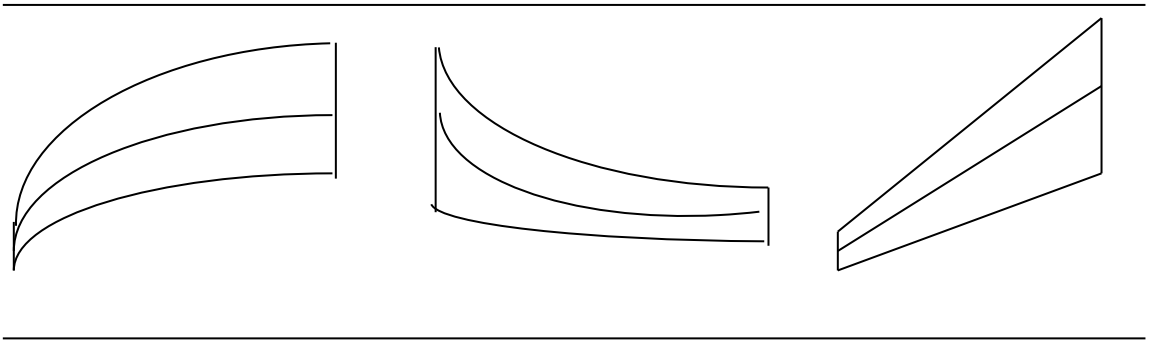


Figura 4. Formas de relacionamento entre as variáveis X e Y no estudo de regressão, onde a assimetria e as variâncias aumentam com a resposta média E(Y). 2023.

As transformações sobre a variável respostas Y são as seguintes:

$$Y' = \sqrt{Y}, Y' = log_{10} Y \quad e \quad Y' = 1/Y$$

Mas segundo o autor uma transformação em X pode ser útil ou necessário. No entanto deve ser realizada uma análise de resíduos (OGLIARI, 2015).

Conforme ainda Ogliari (2015), o uso da chamada transformação Box-Cox, automaticamente identifica uma transformação a partir de uma família de transformações potência de Y. A família de transformações potência é dada por: $Y' = Y^{\lambda}$

Onde λ é um parâmetro a ser determinado a partir dos dados da amostra. Esta família inclui, por exemplo,

$$\lambda = 2 \rightarrow \mathrm{Y}' = Y^2 \quad \lambda = 0{,}5 \rightarrow \mathrm{Y}' = \sqrt{Y}, \quad \lambda = -0{,}5 \rightarrow \mathrm{Y}' = \frac{1}{\sqrt{Y}}$$

$$\lambda = 0 \rightarrow \mathrm{Y}' = log_e Y \text{ (por definição)} \qquad \lambda = \text{-}1{,}0 \rightarrow \mathrm{Y}' = \frac{1}{Y}$$

O modelo de regressão com erros normais com a variável resposta pertencente a família de transformação potência fica: $Y_i^{\lambda} = \beta_0 + \beta_1 X_i + \varepsilon_i$

O procedimento Box-Cox usa o método de máxima verossimilhança para estimar λ, β_0, β_1 e σ^2. A função de verossimilhança é dada por:

$$L(\beta_0, \beta_1, \sigma^2, \lambda) = \frac{1}{(2\pi\sigma^2)^{n/2}} exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(Y_i^{\lambda} - \beta_0 - \beta_1 X_i)^2\right]$$

Desta forma, o procedimento de Box-Cox encontra a estimativa de máxima verossimilhança de λ para usar na transformação potência.

O procedimento simples para obter uma estimativa de λ , pode ser feito usando a análise de regressão padrão do modelo de regressão linear simples. Deve ser feita uma busca numérica, como menor soma de quadrados dos erros (SQE) para uma faixa de valores de lambda, por exemplo: $\lambda = -2$ $\lambda =$ -1 $\lambda =$ -0,5 $\lambda = 0$ $\lambda = 0{,}5$ $\lambda = 1$ $\lambda = 2$. Para cada valor de λ, as observações Y_i^{λ} são padronizadas do seguinte modo:

$$W_i = \begin{cases} K_1(Y_i^{\lambda} - 1) & \lambda \neq 0 \\ K_2(log_e\, Y_i) & \lambda = 0 \end{cases}, \quad K_2 = (\prod_{i=1}^{n} Y_i)^{1/n} \quad e \quad K_1 = \frac{1}{\lambda K_2^{\lambda-1}}$$

Faz-se a regressão das observações W_i sobre X e obtêm-se as somas de quadrados dos erros (SQE). Pode-se mostrar que a estimativa de máxima verossimilhança de λ é o valor de λ para a qual a SQE é mínima.

As pressuposições básicas para que a análise de variância seja válida são as seguintes: Aditividade: os efeitos dos termos do modelo matemático devem ser aditivos. Independência: Já os resíduos, erros ou desvios $e_{i,j}$, devidos aos efeitos de fatores não controlados, devem ser independentes. Isso implica que os efeitos de tratamentos sejam independentes, que não haja correlação entre eles. Isso pode não ocorrer quando os tratamentos são níveis ou doses crescentes de produtos ou materiais, tais como, adubos, inseticidas, fungicidas, herbicidas, droga, lâminas de água, períodos, densidades de plantio, etc., onde a aplicação da técnica de análise de variância (ANOVA) deve ser realizada ajustando-se modelos de regressão. Homogeneidade (ou homocedasticidade) de variâncias: os resíduos, erros ou desvios $e_{i,j}$, devidos aos efeitos de fatores não controlados, devem possuir uma variância comum σ^2. Isso mostra que a variabilidade das repetições de um tratamento deve ser semelhante à dos outros tratamentos, isto é, os tratamentos devem possuir variâncias homogêneas. Normalidade: os resíduos, erros ou desvios $e_{i,j}$, devidos aos efeitos de fatores não controlados, devem possuir uma distribuição teórica ou especial de probabilidade do tipo normal ou Gaussiana. Isso implica que os dados experimentais devem aderir a uma distribuição de probabilidade normal. Convém ressaltar que a não verificação de pelo menos uma destas pressuposições afeta tanto o nível de significância ou valor p, como a sensibilidade do teste F de Fisher aplicado na análise de variância (ANOVA). Assim, qualquer ausência de ao menos uma dessas pressuposições, deve ser corrigida antes da aplicação da técnica da análise de variância (ANOVA), mediante o uso de transformação dos dados ou de técnicas de regressão dos modelos lineares generalizados, ou ainda eliminando tratamentos discrepantes e se não for possível subdividi-los em grupos e testá-los separadamente através de resíduos apropriado a cada grupo.

Os procedimentos utilizados para corrigir a ausência de homocedasticidade, de normalidade e de aditividade são frequentemente possíveis por meio do uso de transformação dos dados originais para

outras formas mais convenientes. Se existir heterogeneidade e se for do tipo regular, a solução é buscar uma transformação adequada. Por outro lado, se ocorrer heterogeneidade irregular deve-se eliminar tratamentos discrepantes. Se não for possível, subdividi-los em grupos e testá-los separadamente através de resíduos apropriado a cada grupo.

Sendo assim, medidas para corrigir a heterogeneidade da variância, podem e devem ser implementadas para validação das análises estatísticas posteriores, e como o sintoma mais comum dos dados experimentais que viola uma ou mais das pressuposições da análise da variância (ANOVA) é a heterogeneidade de variância, tem-se que duas são as medidas usadas para corrigir a heterogeneidade da variância, as quais são descritas mais adiante. Primeira é aquela que trata do método de transformação de dados para variâncias que estão funcionalmente relacionadas com a média. No entanto um diagnóstico correto do tipo específico de heterogeneidade de variância presente nos dados deve ser feito antes que uma medida corretiva apropriada possa ser selecionada, e assim os procedimentos são os seguintes: i) Para cada tratamento, determine a variância e a média sobre as repetições (a amplitude pode ser usada no lugar da variância); ii) Construa ou trace um diagrama de dispersão entre o valor médio e a variância (ou a amplitude). O número de pontos no diagrama é igual ao número de tratamentos e iii) Visualmente deve-se examinar o diagrama de dispersão para identificar se existe algum padrão de relacionamento, entre a média e a variância.

O outro mecanismo é aquele que trata do método de subdividir os erros das variâncias que não são funcionalmente relacionadas a média. A heterogeneidade de variância na qual nenhum relacionamento entre variância e a média existe, é quase sempre devido a presença de um ou mais tratamentos cujos erros associados são diferentes daqueles dos outros tratamentos. Estes erros excepcionalmente altos e baixos são geralmente devido a duas principais causas: i) Eles envolvem tratamentos que, por sua natureza, exibe variâncias grandes ou pequenas; ii) Eles envolvem erros grosseiros; isto é, alguns valores grandes ou pequenos que podem ter sido registrado erroneamente em algumas parcelas resultando em variâncias do erro grandes dos tratamentos envolvidos. A partição do erro é um procedimento comumente usados para manusear dados que tem heterogeneidade de variância que não são funcionalmente relacionadas a média. A partição dos erros não deve ser usada, contudo, quando a heterogeneidade é devida a erros grosseiros. Em outras palavras, a partição do erro deve ser aplicada apenas após a presença de erros grosseiros ter sido completamente examinada e eliminada.

Agora o procedimento para detectar erros grosseiros e aplicar o método de partição do erro, pode ser feito das seguintes maneiras: i) identificar tratamentos que tenham diferenças extremamente grandes entre observações de diferentes repetições; ii) para cada um destes tratamentos, identificar a parcela especifica cujo valor é grande e diferente do resto (isto é, parcelas com valores excepcionalmente grandes e pequenos). Para essa parcela em questão, examinar os registros ou diário dos dados para ver se algumas observações ou anotações foram observadas pelo pesquisador para explicar o

valor extremo. No croqui de campo, marcar as parcelas que tem valores extremos por colocar um sinal + na parcela com valor excepcionalmente alto e um sinal – nas parcelas com valor excepcionalmente baixo. Examine a proximidade das parcelas com sinais + e – para localizar possíveis causas que estejam relacionadas a localização da parcela na área experimental. Para as parcelas cujas causas dos valores extremos foram identificadas como erros grosseiros, recuperar os valores corretos se possível. Se a recuperação não é possível, os dados suspeitos devem ser rejeitados e os dados declarados perdidos. Para parcelas cujas causas de valores extremos não podem ser determinadas, os dados suspeitos devem ser retidos.

Quando o pesquisador detectar a heterogeneidade da variância, a transformação de dados apropriada para ser usada depende do tipo especifico de relacionamento entre a variância e a média. Os tipos de transformações são as seguintes: i) transformação raiz quadrada $(\sqrt{y})$, a qual é apropriada para estabilizar as variâncias das observações da distribuição de Poisson, frequentemente usada para dados de contagens, onde a variância é igual a média. É apropriada também para dados consistindo de números inteiros pequenos, dados de contagem provenientes de, por exemplo: número de ervas daninha por parcela, números de colônias de bactérias por lâmina, número de insetos capturados em armadilha luminosa, número de insetos por folha, número de bactérias em uma placa de Petri, número de sementes germinadas por parcela, número de plantas infestadas, número de parasitas por animal, etc. Esta transformação também é apropriada para dados provenientes de uma escala de notas que também devem ser transformados, como por exemplo, apropriada ocasionalmente para dados de percentagem oriundos de dados de contagem. Se as percentagens estiverem entre 0 e 30% ou 70 e 100%, mas não em ambas, a transformação raiz quadrada é recomendada. As percentagens entre 70% e 100%, devem ser preferencialmente subtraídas de 100, antes de se fazer a transformação. A transformação raiz quadrada é, ainda, indicada no caso de percentagens, fora dos limites acima considerados, quando as observações estão claramente numa escala contínua. A transformação raiz quadrada $Y = \sqrt{y}$ terá uma variância constante $\sigma_y^2 = 0{,}25$ para todos os valores μ_y. Se a média é pequena, $\mu_y < 3$, então a transformação $Y = \sqrt{(y + 3/8)}$ é superior a $\sqrt{y}$ para estabilizar as variâncias. Quando ocorrem zeros ou valores abaixo de 10, usar $\sqrt{y + 0{,}5}$ ou $\sqrt{y + 1{,}0}$, em lugar de $\sqrt{y}$; ii) a transformação logarítmica $log\ y$ ou $ln\ y$, é apropriada a dados onde existe uma proporcionalidade entre as médias e os desvios padrões (ou amplitudes) dos diversos tratamentos, ou seja, todas as amostras apresentam o mesmo coeficiente de variação. Apropriada também quando os efeitos principais são multiplicativos em vez de aditivos. Nessa situação, tal transformação, além de estabilizar as variâncias produz aditividade nos efeitos do modelo matemático e tende a normalizar a distribuição dos erros. Indicada para números inteiros com grande amplitude de variação (por exemplo, em estudo com número de microrganismos, cuja variação está entre 50 e 200.000), como por exemplo, o número de insetos por parcela, número de ovos por planta ou por

unidade de área, número de bactérias, esporos, de grão de pólen, dados de adição de vitaminas em animais. Quando a amostragem possui dados iguais a zero ou muito próximos de zero, usa-se $log\ (y + 1)$. A base 10, por conveniência, no emprego da transformação logarítmica é a mais usada, no entanto, pode-se empregar qualquer base; iii) transformação angular ou arcoseno $(p' = arc\ sen\ \sqrt{p})$, apropriada a dados em que a média é proporcional a variância, ou seja, oriundos de uma distribuição binominal, como aqueles expressos em proporções $p = \left(\frac{y}{n}\right)$ou percentagens $P = 100\left(\frac{y}{n}\right)$. Existem tabelas apropriadas para essa transformação, nas quais entramos diretamente com a percentagem (P) e obtemos $arcsen\sqrt{\frac{p}{100}}$. A transformação angular é usada para homogeneizar a variância residual dos dados de percentagem ou normalizar a distribuição binomial, especialmente quando as percentagens observadas estiverem todas entre 0 e 30% ou entre 70 e 100%. Se as percentagens dos dados estiverem entre 30 e 70%, torna-se desnecessária a transformação, e pode-se analisar diretamente os dados originais. Se os dados extrapolam esta amplitude, usa-se então a transformação. A transformação também é desnecessária quando as porcentagens são resultantes da divisão de valores observados nas parcelas por um valor constante (valor representativo), como a média do tratamento testemunha ou quando são representativas de concentração, como teor de N na folha, pureza da semente, teor de sacarose da cana-de-açúcar, teor de proteína do trigo, etc. A transformação é necessária em dados de percentagem provenientes de dados discretos num total de casos, como, por exemplo, percentagem de germinação (número de sementes germinadas/número total de sementes), percentagem de plantas doentes (número de plantas doentes/número de plantas consideradas), etc. A transformação angular não é boa quando $p = \frac{0}{n} = 0$ ou $p = \frac{n}{n} = 1$. A transformação é melhorada por substituir $\frac{0}{n}$ por $\frac{1}{4n}$ e $\frac{n}{n}$ por $1 - \frac{1}{4n}$, antes de transformar os dados, onde n é o número total de unidades sob observação.

Anscombe (1948), propôs uma transformação ainda melhor: $p' = \sqrt{n + 0{,}5}\ arcsen\sqrt{\frac{y + \frac{3}{8}}{n + ¾}}$. Não é apropriada a dados de percentagem que não são originados de dados de contagem. Por exemplo, percentagem de proteína em arroz, índice de infecção, percentagem de lucro, etc.

A análise de variância e outros métodos associados à distribuição normal são aplicados sobre os dados transformados. Quando a análise é completada, a média aritmética das contagens transformadas pode ser transformada de volta para a escala original tornando-se uma média derivada, isto é, para uma transformação raiz quadrada $(\sqrt{y})$, a média das contagens transformadas deve ser elevada ao quadrado. Para uma transformação $log\ (y + 1)$, deve-se obter o $antilog$ da média transformada e subtrair 1. De modo geral, as médias derivadas são menores que as médias das contagens na escala original. Portanto, pequenos ajustes devem ser feitos sobre as médias derivadas.

Elliott (1979) recomenda as seguintes correções ao aplicar transformação de dados: i) na transformação raiz quadrada, a média dos dados transformados em raiz quadrada deve ser elevada ao quadrado, e a seguir somada com a variância dos dados transformados $[Y = \bar{y}^2 + V(\bar{y})]$ ii) já na transformação logarítmica, deve-se adicionar à média da variável transformada 1,15 vezes no valor da variância da variável transformada $Y = antilog[\bar{y} + 1{,}15V(\bar{y})]$, e a seguir obter o antilog da média derivada. O valor final é geralmente um bom estimador da média obtida diretamente das contagens; iii) para a transformação arco seno os valores transformados podem ser transformados na escala original, usando-se a seguinte expressão: $p = (senp')^2$ e multiplicando por 100 para expressá-lo em percentagem (P).

Vale salientar que existe uma relação extremamente importante entre a distribuição normal, gaussiana, em forma de sino, de chapéu de Napoleão ou ainda campanular e a transformação de dados. Por exemplo, se existem chances iguais de um evento ocorrer na distribuição binomial ($p = q = 0{,}5$) e n aproxima-se de infinito, então a série de probabilidades dada por $(p + q)^n$ aproxima-se de uma curva simétrica em forma de sino. Esta é a curva da distribuição normal, que é a distribuição associada com variáveis contínuas, tais como, medições, pesagens, etc. A distribuição normal raramente é aceitável para estudar contagens, mas ela é importante por causa do grande número de métodos estatísticos inferenciais e de ajuste de curvas de regressão associados a ela, tais como: teste t, análise de variância, análise de regressão, coeficiente de correlação, etc. O uso destes métodos envolve as condições: i) os erros ou resíduos devem possuir distribuição Normal com média zero e variância comum σ^2 ou seja, eij $\sim N(0, \sigma^2)$; ii) os dados devem seguir uma distribuição normal. iii) a variância da amostra deve ser independente da média e constante nas várias amostras; iv) os componentes da variância devem ser aditivos; v) os erros devido aos efeitos dos fatores não controlados ou acaso devem ser independentes. A distribuição binomial positiva é aproximadamente normal se o número de unidades amostrais é grande ($n \geq 30$) e a variância da amostra não é menor que 3 (a variância é $S^2 = npq$). Portanto, a aproximação normal pode ser usada quando $0{,}4 \leq p \leq 0{,}6$ para $10 \leq n \leq 30$, ou quando $0{,}1 \leq p \leq 0{,}9$ para $n \geq 30$, e não pode ser usada quando $n < 10$.

A distribuição de Poisson é muito assimétrica para baixos valores do parâmetro λ (estimado por $\hat{m} = s^2$), mas aproxima-se da normalidade quando λ cresce, e é aproximadamente normal quando λ é maior do que 10. A distribuição binomial negativa é assimétrica para uma grande faixa de variação da média quando k é pequeno (isto $k < 3$), mas aproxima-se da normalidade quando k aumenta e a média é razoavelmente grande ($\mu = 10$ ou mais). Quanto k tende para ∞, a distribuição binomial negativa é idêntica à distribuição de Poisson. Como a média e a variância tendem a crescer juntas em todas as três distribuições, a segunda condição de independência entre a média e a variância nunca é satisfeita. Portanto, alguns métodos, incluindo o teste t, análise de variância, análise de regressão, etc. não podem

ser aplicados sem o risco de erros consideráveis. Esta dificuldade pode ser superada, trocando cada contagem por uma função matemática adequada das contagens. As contagens são então transformadas, e a transformação correta geralmente normaliza a distribuição de frequência das contagens, elimina a dependência entre a média e a variância e assegura que os componentes da variância sejam aditivos para a aplicação da análise de variância (ANOVA).

É importante destacar que a escolha da transformação correta depende da distribuição de frequências das contagens originais, como mostra a Tabela 11 a seguir:

Tabela 11. Tipos de transformações utilizadas em função da natureza da distribuição de frequências dos dados. Mossoró, RN, 2023.

Distribuição de Frequências dos Dados	**Relação Entre Média e Variância** $\hat{m}$ e s^2	**Condições Especiais**	**Transformação Recomendada**
Poisson	$s^2 = \hat{m}$	Nenhum valor < 10	$\sqrt{y}$
Poisson	$s^2 = \hat{m}$	Alguns valores < 10	$\sqrt{y + 0{,}5}$
Binomial	$s^2 < \hat{m}$ $s^2 = \hat{m}(1 - \hat{m})/n$	Proporções binomiais	$\text{arcseno}\sqrt{\text{proporção}}$
Binomial Negativa	$s^2 > \hat{m}$	$2 < k < 5$	$log\ (y + k/2)$
	$s^2 > \hat{m}$	$k > 5$	$\text{arcseno hiperb}\sqrt{\frac{y + 0{,}375}{k - 0{,}750}}$
Empírica	$s^2 = C^2\hat{m}$		$\sqrt{y}$
Desconhecida Empírica	$s^2 > \hat{m}$ $s^2 = C^2\hat{m}^2$	Nenhum zero	$log\ (y)$
Desconhecida	$s^2 > \hat{m}$	Alguns zeros	$log\ (y + 1)$

Um método bastante simples de se encontrar a transformação mais adequada é através da lei de potência de Taylor, a qual afirma que a variância (σ^2) de uma população é proporcional a uma potência fracionária da média aritmética (μ), ou seja $\sigma^2 = a\mu^b$ e portanto, $log\sigma^2 = loga + blog\mu$ onde, a e b são parâmetros populacionais. O parâmetro "a" depende principalmente do tamanho das unidades amostrais. Já o parâmetro "b" é um índice de dispersão e varia continuamente de zero (0) no caso da distribuição regular ($\sigma^2 < \mu$), a infinito (∞) no caso de distribuições altamente contagiosas ($\sigma^2 > \mu$). No

caso da distribuição aleatória, temos $a = 1$ e $b = 1$. Obtida uma estimativa de b, pode-se encontrar facilmente a transformação mais adequada para que os métodos associados com a distribuição normal sejam utilizados. A transformação apropriada pode ser obtida por: $y = y^p$ onde, $P = 1 - \frac{b}{2}$.

Os procedimentos usados para se encontrar a transformação apropriada, podem ser implementados pelas seguintes ações: i) obter as médias e variâncias para cada amostra; ii) obter o logaritmo da média e logaritmo da variância para cada amostra; iii) estimar os parâmetros a e b através de uma regressão linear de $y = log(s^2)$ sobre $x = log(\hat{m})$, isto é: log $s^2 = log\ a + blog\ (\hat{m})$, onde: "$b$" é o coeficiente de inclinação (angular) ou de regressão mostrando o quanto varia em média o logaritmo da variância em função da variação de uma unidade no logaritmo da média. Com esta técnica, os dados originais são convertidos em nova escala resultando em um novo conjunto que é esperado satisfazer a condição de homogeneidade de variância. Devido uma escala comum de transformação ser aplicada a todas as observações, os valores comparativos entre os tratamentos não são alterados e comparações entre eles permanecem válidos.

Pedrosa e Gama (2004) descrevem como é que através da aplicação do teorema do limite central são geradas observações ou realizações de variáveis aleatórias com distribuição de probabilidade normais. Segundo este teorema, a soma das n variáveis aleatórias independentes e identicamente distribuídas, $U_1, U_2, \cdots, U_n \sim U(0,1)$, onde todas têm média $\mu_U = 0{,}5$ e variância $\sigma_U^2 = \frac{1}{12}$, tem uma distribuição aproximadamente normal, com média 0,5 n e variância $\frac{n}{12}$.

Ou seja, a variável aleatória

$$X = \frac{\sum_{i=1}^{n} U_i - 0{,}5n}{\sqrt{\frac{n}{12}}}$$

é aproximadamente normal, com média $\mu_X = 0$ e variância $\sigma_x^2 = 1$. Esta aproximação será tanto melhor quanto maior for n. Contudo, a maioria da literatura de simulação sugere usar o valor $n = 12$, que tem a vantagem de simplificar o cálculo computacional, evitando uma raiz quadrada e uma divisão, resultando em $X = \sum_{i=1}^{n} U_i - 6$. Este algoritmo tem a grande desvantagem de não ser exato, isto é, não produz observações com distribuição exatamente normal, apesar de ser extremamente simples e eficiente. Conforme ainda os autores um dos métodos mais antigos e populares para gerar realizações de variáveis aleatórias normais estandardizadas foram desenvolvidos por Box e Müller (1958). Embora não seja tão eficiente como alguns métodos mais recentes, é fácil de usar e aplicar. Este método, que será apresentado em seguida sem justificativa, precisa de dois números aleatórios independentes, U_1 e U_2, e transforma-os em duas observações normais estandardizadas usando as transformações: $X_1 = \sqrt{-2\ ln\ U_1}\ cos(2\pi U_2)$, $X_2 = \sqrt{-2\ ln\ U_1}\ sen(2\pi U_2)$. Assim usamos dois números aleatórios para gerar duas observações

normais estandardizadas. A geração de observações de $Y \sim N(\mu, \sigma^2)$ pode ser conseguida a partir das realizações de $X \sim N(0,1)$, fazendo simplesmente $Y = \mu + \sigma X$.

Considere duas variáveis X e Y, com função densidade conjunta $f(x, y)$ e suponha que se queira obter a densidade das variáveis Z e W, tais que $Z = h_1(X, Y)$ e $W = h_2(X, Y)$. Suponha que se possa expressar x e y em função de z e w, isto é, $x = g_1(z, w)$ e $y = g_2(z, w)$. Supondo que as derivadas parciais de x e y, em relação a z e w, existam e sejam contínuas, pode-se obter a densidade conjunta de Z e W através de, $g(z, w) = f(g_1(z, w), g_2(z, w)|J)$, onde J é o jacobiano da transformação que leva (x, y) em (z, w), dado por

$$J = \begin{vmatrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial z} & \frac{\partial y}{\partial w} \end{vmatrix}$$

No caso unidimensional, $Y = h(X)$, J é simplesmente $\frac{\partial x}{\partial y}$, com $x = h^{-1}(y)$ (BUSSAB; MORETTIN, 2002).

Segundo Bussab e Morettin (2002), pode-se provar que após o uso de transformações nas variáveis, as somas algébricas e somas de quadrados e produtos, podem tornar os cálculos menos trabalhoso. Veja seguir como essas somas de valores são equivalentes, o que torna a obtenção de seus resultados uma tarefa mais rápida, fácil e sem perda de precisão, as quais são:

i) $\sum_{i=1}^{4}(x_i - \bar{x}) = 0$;

ii) $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$;

iii) $\sum_{i=1}^{k} n_i(x_i - \bar{x})^2 = \sum_{i=1}^{k} n_i x_i^2 - n\bar{x}^2$;

iv) $\sum_{i=1}^{k} f_i(x_i - \bar{x})^2 = \sum_{i=1}^{k} f_i x_i^2 - \bar{x}^2$.

No estudo de correlação e regressão linear simples tem-se que:

$$SQX = S_{XX} = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n} = \sum_{i=1}^{n} X_i^2 - n\left[\bar{X}\right]^2,$$

$$SQY = S_{YY} = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} = \sum_{i=1}^{n} Y_i^2 - \left[\bar{Y}\right]^2$$

$$SPXY = S_{XY} = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y} = \sum_{i=1}^{n}\sum_{j=1}^{n} X_i Y_j - \frac{\left(\sum_{i=1}^{n} X_i \sum_{j=1}^{n} Y_j\right)}{n}$$

Sendo que, $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, $\overline{Y} = \frac{\sum_{j=1}^{n} Y_i}{n}$, onde o modelo ajustado é dado por:

$$\hat{Y}_i = \hat{a} + \hat{b}X_i, \hat{a} = \overline{Y} - -\hat{b}\overline{X},$$

$$\hat{b} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} X_i Y_j - \frac{\left(\sum_{i=1}^{n} X_i \sum_{j=1}^{n} Y_j\right)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}} = \frac{(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{j=1}^{n} x_i y_j - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - \bar{x}^2}$$

Para a análise de correlação linear simples, tem-se que:

$$r = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} X_i Y_j - \frac{\left(\sum_{i=1}^{n} X_i \sum_{j=1}^{n} Y_j\right)}{n}}{\sqrt{\left[\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right]\left[\sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}\right]}} = \frac{SPXY}{\sqrt{SQX.SQY}}$$

Bussab e Morettin (2002) descrevem a importante transformação de Box– Müller, da seguinte maneira: Considere as variáveis aleatórias X e Y, independentes e ambas tendo distribuição normal padrão com média zero (0) e variância igual a um (1), ou seja, $N(0,1)$. É sabido que: $R^2 = X^2 + Y^2$E $tg\theta = \frac{Y}{X}$. a densidade conjunta de x e y é,

$$f(x,y) = \frac{1}{2\pi} e^{\frac{-(x^2+y^2)}{2}}$$

Considere a transformação de variáveis $r = x^2 + y^2$, e $\theta = arctg\left(\frac{y}{x}\right)$. A densidade conjunta de r e θ é obtida usando o resultado $g(z,w) = f(g_1(z,w), g_2(z,w)|J)$. Tem-se que $x = \sqrt{r}\cos\theta$, $y = \sqrt{r}sen\theta$ e o jacobiano da transformação é $|J| = \frac{1}{2}$. Segue-se que a densidade de r e θ é $f(r,\theta) = \frac{1}{2\pi} e^{-r^2}\frac{1}{2}$, $0 < r < \infty$, $0 < \theta < 2\pi$. Dessa relação pode-se concluir que $R = r^2$ e θ são independentes, com $R^2 \sim Exp(2)$, $\theta \sim U(0,\ 2\pi)$. Portanto, pode-se escrever que,

$$X = R\cos\theta = \sqrt{-2\log U_1}\cos(2\pi U_2) \text{ e}$$

$$y = R\ sen\theta = \sqrt{-2\log U_1}\, sen(2\pi U_2).$$

Aqui usou-se o fato de que, se $R^2 \sim Exp(2)$, gerado um número aleatório (NA) NAU_1, vem que $-2\log U_1 \sim Exp(2)$ e se $\theta \sim U(0,\ 2\pi)$, então gerado um NAU_2, vem que $2\pi U_2 \sim U(0,\ 2\pi)$. O método de Box–Müller gera valores de duas normais padrões independentes Z_1 e Z_2. Logo, se o pesquisador quiser gerar valores da distribuição conjunta de X e Y, independentes e normais, com $X \sim N(\mu_x, \sigma_x^2)$ e $Y \sim N(\mu_y, \sigma_y^2)$ basta considerar $X = \mu_x + \sigma_x Z_1$ e $Y = \mu_y + \sigma_y Z_2$. Já na simulação de uma distribuição gama, os autores afirmam que se a variável aleatória x possui distribuição gama isto é, $X \sim Gama\ (r,\beta)$, com r inteiro, então $X = Y_1 + Y_2 + \cdots + Y_r$, onde cada $Y_i \sim Exp\ (\beta)$ e as variáveis aleatórias Y_i independentes. Logo, para gerar um valor de uma distribuição $Gama\ (r,\beta)$, com $r > 0$, inteiro, basta gerar r valores de uma distribuição exponencial de parâmetro β e depois soma-los.

Bussab e Morettin (2002), também mostram o processo de simulação de uma distribuição exponencial onde eles afirmam que se a variável aleatória T tiver densidade dada por

$$f(t) = \frac{1}{\beta} e^{-\frac{t}{\beta}}, \quad t > 0,$$

a sua função densidade de probabilidade acumulada é dada por $F(t) = 1 - e^{-\frac{t}{\beta}}$, logo tem-se que resolver a equação $F(x) = u$, para gerar t, ou seja, $x = F^{-1}(u)$, para isso usando-se um gerador de número aleatório (NA), produz-se um $NA\ u$, marca-se esses valor no eixo das ordenadas de $F(x)$, por meio da função inversa de $F(x)$ obtém-se o valor x da variável aleatória (v.a.) X no eixo das abscissas.Tomando logaritmo na base e, tem-se que $1 - u = e^{-\frac{t}{\beta}}$, e assim $log(1-u) = -\frac{t}{\beta}$ obtém-se, $t = -\beta\ log(1-u)$. Logo, gerado um Na, um valor da distribuição $Exp(\beta)$ é dado por $-\beta\ log(1-u)$. Pode-se reduzir um pouco os cálculos se for usado o seguinte fato: se $U \sim u(0,1)$, então $1 - U \sim u(0,1)$. Resulta que se pode gerar valores de uma exponencial por meio de $t = -\beta\ log(u)$. Conforme ainda os mesmos autores na simulação de distribuição normal há vários métodos para gerar variáveis aleatórias normais, mas uma observação importante é que basta gerar uma variável aleatória normal padrão, pois qualquer outra pode ser obtida desta. De fato, gerado um valor z_1 da variável aleatória $Z \sim N(0,1)$, para gerar um valor de uma variável aleatória $X \sim N(\mu, \sigma^2)$ e basta usar a transformação $z = \frac{(x-\mu)}{\sigma}$ para obter $x_1 = \mu + \sigma z_1$. Este método, embora simples, não é prático, sob o ponto de vista computacional. Há outros métodos mais eficientes. Alguns são variantes do método de Box-Muller (1958). Neste método são geradas duas variáveis aleatórias Z_1 e Z_2, independentes, e com distribuição normal de média zero e variância um, ou seja, $N(0,1)$, por meio das transformações

$$Z_1 = \sqrt{-2\ log\ U_1}\ cos(2\pi U_2) \text{ e}$$

$$Z_2 = \sqrt{-2\ log\ U_1}\ sen(2\pi U_2).$$

Onde U_1 e U_2 são variáveis aleatórias com distribuição uniforme em $[0,1]$. Portanto, basta gerar dois números aleatórios $NA\ u_1$ e u_2 e depois gerar z_1 e z_2 usando $Z_1 = \sqrt{-2\ log\ U_1}\ cos(2\pi U_2)$ e $Z_2 = \sqrt{-2\ log\ U_1}\ sen(2\pi U_2)$.Já a simulação de uma distribuição normal bidimensional, pode-se considerar o seguinte: o método de Box-Muller gera valores de duas normais padrões independentes, Z_1 e Z_2.Logo, se quiser gerar valores da distribuição conjunta de X e Y, independentes e normais, com $X \sim N(\mu_x, \sigma_x^2)$ e $Y \sim N(\mu_y, \sigma_y^2)$, basta considerar $X = \mu_x + \sigma_x z_1$, $Y = \mu_y + \sigma_y z_2$.Na simulação de uma distribuição qui- quadrado, pode-se usar o seguinte: como sabe-se que, se $Z \sim N(0,1)$ e $Y = Z^2$, então $Y \sim \chi^2(1)$. Por outro lado, uma variável aleatória W com distribuição Qui-quadrado com n gruas de liberdade, ou seja, $\chi^2(n)$ pode ser escrita como $W = Z_1^2 + Z_2^2 + \cdots + Z_n^2$, onde as variáveis aleatórias $Z_1^2, \cdots, Z_n^2$ são normais padrões, independentes. Portanto, para simular um valor de uma variável aleatória com

distribuição Qui-quadrado, com n gruas de liberdade, basta gerar n valores de uma variável aleatória normal padrão com média zero e variância igual a um, isto é, $N(0,1)$ e considerar a soma de seus quadrados.

Uma abordagem sobre transformação Gaussiana dos dados experimentais, em estudos geoestatísticos é feita por Soares (2000), onde ele afirma que a grande vantagem desta abordagem incide na sua simplicidade de implementação, pois a função densidade de probabilidade (fdp) de qualquer ponto x_0 em uma área ou espaço A fica definida com a estimação (krigagem simples) de $Y(x_0)$. No entanto, há duas grandes dificuldades que se colocam na aplicação desta abordagem: i) em primeiro lugar, raramente se encontra nas ciências da terra e ambiente um conjunto de dados experimentais que tenha uma distribuição de probabilidade marginal Gaussiana; ii) mesmo que os valores, realização de um conjunto de variáveis aleatórias, tenham uma distribuição marginal de probabilidades Gaussiana, é necessário aferir se a hipótese de multigaussianidade pode ser considerada apropriada às distribuições conjuntas das diferentes variáveis. O primeiro ponto é relativamente fácil de ser ultrapassado através de uma transformação Gaussiana dos experimentais $Z(x_\alpha) = \Phi\big(Z(x_\alpha)\big), \quad \alpha = 1, \cdots, N$, em que $Y(x)$ segue uma lei Gaussiana de média nula e variância unitária. Esta transformação Φ pode ser obtida por uma aproximação polinomial, polinômios de Hermite (MATHERON, G., 1974; MUGE, F. 1982 citados por SOARES, 2000), ou por uma simples transformação gráfica que é o método mais simples e apropriado para esta aplicação. Dadas as funções de distribuição das variáveis $Z(x)$ e $Y(x)$,$F(z) = prob\{Z(x) < z\}$ e $G(y) = prob\{Y(x) < y\}$, então o valor z, que corresponde ao valor gaussiano y, satisfaz $F(z) = G(y)$. Generalizando, $Y(x_\alpha) = \Phi\big(Z(x_\alpha)\big) = G^{-1}\left(F\big(Z(x_\alpha)\big)\right), \alpha = 1, \cdots, N$. Em termos concretos, a transformação Gaussiana pode ser resumida nos seguintes passos: i) ordenação crescente dos N valores originais $Z(x_i)$: $Z_1(x_i) \leq Z_2\big(x_j\big) \leq \cdots \leq Z_N(x_k)$, correspondendo a cada valor desta série ordenada uma frequência acumulada. Por exemplo, ao valor de ordem $K, Z_k(x_i)$, corresponde $P_k = \frac{(k-0{,}5)}{N}$. O primeiro e o último valores de uma série de 100 têm as seguintes frequências: $P_1 = \frac{0{,}5}{100} = 0{,}005$ e $P_{100} = \frac{99{,}5}{100} = 0{,}995$; ii) o correspondente valor Gaussiano de cada um dos valores da série $Z_k(x_i)$ será: $Y(x_i) = G\big[F\big(Z_k(x_i)\big)\big] = G^{-1}(P_k)$, sendo k a ordem de $Z_k(x_i)$ na série ordenada. Com esta transformação dos valores experimentais $Z(x)$, com uma distribuição de probabilidades $F(z)$ em valores Gaussianos $Y(x)$ e admitindo que estes seguem uma lei conjunta multiGaussiana, então todo o este formalismo multigaussiano pode ser aplicado.

Concretamente, num primeiro passo, são transformados os valores experimentais $Y(x_\alpha) = \Phi\big(Z(x_\alpha)\big), \quad \alpha = 1, \cdots, N,$. Depois de calculados os variogramas dos valores transformados $Y(x_\alpha)$, para qualquer ponto no espaço é calculada a função densidade de probabilidade (fdp) local de acordo

com $G(x_{0;y}) = G\left[\frac{y-[Y(x_0)]^*}{\sigma_E^*(x_0)}\right]$ que é a função de distribuição de probabilidade no ponto x_0 a qual fica definida com base nos dois parâmetros estimados por krigagem simples, a média e a variância, $G(x_0, y) = prob\{Y(x_0) < y\}$.

Todo este processo é baseado no conceito de formalismo multiGaussiano o qual é um modelo hipotético de caracterização de funções de distribuição de probabilidades de dados experimentais (amostra) o qual admite que essa mesma distribuição pode ser caracterizada pela conjunção de múltiplas distribuições gaussianas. É vulgarmente utilizada na caracterização de fenômenos das ciência da terra e é inclusive adotada no procedimento de simulação de variáveis espaciais simulação sequencial gaussiana no ramo da geoestatística. A vantagem desta assunção é a simplicidade de implementação. Se para um dado sub-conjunto do conjunto total dos dados experimentais calcularmos a média e variância podemos, efetivamente, caracterizar uma distribuição gaussiana. Se admitirmos o conjunto de variáveis aleatórias $\{Y(x), x \in A\}$ como seguindo uma lei conjunta multigaussiana, qualquer par, $Y(x_1)$ e $Y(x_2)$, localizado respectivamente em x_1 e x_2, consiste numa distribuição biGaussiana que pode ser caracterizada pela função de covariância $C_Y(x_1, x_2)$. Deste modo a função de distribuição de probabilidades local (ou condicional - referindo-se a um sub-conjunto do conjunto $Y(x)$ é também gaussiana determinada pela média e variância dessa mesma distribuição local. Assumindo que pretendemos saber a distribuição local num qualquer local no espaço, x_0, com média e variância (WIKIPÉDIA, 2015):

$$E\{Y(x_0)|(Y(x_1), \cdots, Y(x_N))\} \text{ e,}$$
$$var\{Y(x_0)|(Y(x_1), \cdots, Y(x_N))\}$$

Conforme ainda Soares (2000), os valores da função de distribuição de probabilidades $F(z')$, para qualquer valor de corte z', são obtidos pela transformação inversa φ. Primeiramente, é calculado o valor de y'correspondente a z':$y = \varphi(z) = G^{-1}[F(z)]$. Seguidamente, $F(x_0, z')$é obtido a partir de $G(x_0, y')$, estimada em $G(x_{0;y}) = G\left[\frac{y-[Y(x_0)]^*}{\sigma_E^*(x_0)}\right]$, $F(x_0, z') = G(x_0, y')$. Se os valores de corte z' não coincidirem com os valores experimentais $Z(x_\alpha)$, uma vez que $F(z, x)$ é monotonicamente crescente, aquela transformação inversa pode ser calculada por uma interpolação linear ou de potência (GOOVAERTS, 1997, citado por SOARES, 2000).

A segunda grande dificuldade de aplicação desta metodologia reside exatamente na hipótese de multiGaussianidade da variável transformada. No entanto, em termos práticos, é suficiente assegurar a biGaussianidade da lei de distribuição entre os diferentes pares de variáveis. Um dos métodos para verificar se os valores transformados $Y(x_i)$ seguem uma lei de distribuição biGaussiana consiste em construir a bidistribuição entre os pares de variáveis $Y(x_i)$ e $Y(x_i + h)$para cada vetor **h** e compará-la

com a expressão teórica da lei biGaussiana (ABRAMOWITZ; STEGUN, 1972 E VENTSEL, 1973, citados por SOARES, 2000).

Dentre os vários métodos para gerar valores de uma variável aleatória Z_1 a partir da sua função de distribuição $F_{Z_1}(z)$, existem o método de transformação inversa, método de aceitação/rejeição e método de composição (LAW; KELTON, 1991, citados por SOARES, 2000), Soares (2000) mostra com algum pormenor o método de transformação inversa que é dos mais simples e utilizados segundo o autor. Na apresentação dos modelos de simulação sequencial o autor assume simplificação na notação de $F_{Z_1}(z)$ escrevendo $F(Z_1)$. Uma vez que $F(Z_1)$ é uma função não decrescente, a função inversa $F^{-1}(y)$ pode ser definida para qualquer valor y compreendido entre 0 e 1. Considere uma outra variável U com uma distribuição uniforme entre 0 e 1: $U \in [0,1]$. Para se simular uma valor $z$da variável Z_1, basta gerar um valor u da distribuição uniforme U e calcular o valor $z = F^{-1}(u)$.é fácil provar que se U é uniformemente distribuída no intervalo $[0,1]$ então $Z_1 = F^{-1}(U)$ tem uma lei de distribuição $F(Z_1)$,

$$P(Z_1 \leq z_1) = P[F^{-1}(U) \leq z_1] = P[U \leq F(z_1)] = F(Z_1)$$

No caso de Z_1 ser uma variável discreta, o método de transformação inversa é o mesmo, mas aplicado a um histograma cumulativo com um número finito de classes. Conforme ainda o mesmo autor na simulação de um conjunto de variáveis aleatórias, o princípio da simulação sequencial é extremamente simples e baseia-se na aplicação da relação de Bayes em passos sequenciais sucessivos. Admita que se queira simular somente dois valores, z_1e z_2, a partir da função $F(Z_1, Z_2)$. Pela relação de Bayes, $F(Z_1, Z_2) = F(Z_2|Z_1)F(Z_1)$, a simulação daqueles valores pode ser feita em dois passos sucessivos: primeiro simula-se uma valor z_1 a partir da distribuição $F(Z_1)$, por exemplo, através do método da transformação inversa, e seguidamente simula-se o valor z_2 a partir da distribuição condicional $F(Z_2|Z_1 = z_1)$, obtendo-se assim o par de valores z_1 e $z_2$das variáveis Z_1e Z_2. A relação de Bayes pode ser generalizada para um conjunto de variáveis:

$$F(Z_1, Z_2, Z_3, \cdots, Z_N) = F(Z_1)F(Z_2|Z_1)F(Z_3|Z_1, Z_2) \cdots F(Z_N|Z_1, Z_2, \cdots, Z_{N-1})$$

Assim o conjunto de valores $Z_1, \cdots, Z_N$ com lei de distribuição conjunta $F(Z_1, Z_2, Z_3, \cdots, Z_N)$ pode ser obtido da simulação sequencial das diferentes distribuições condicionais.

Os principais objetivos de uma transformação nos estudos de regressão são linearizar a relação entre as variáveis explicativas e a variável resposta e obter uma variável Y transformada com distribuição normal e variância constante. Inicialmente é interessante pensar que, para obter a linearidade, deve-se transformar apenas as variáveis explicativas, porque uma transformação em Y modifica sua normalidade e variância. Em um modelo de regressão simples, o qual tem apenas uma variável explicativa, a necessidade de transformar a covariável é facilmente percebida quando o diagrama de dispersão entre as duas variáveis apresenta um comportamento não linear. Se neste caso for ajustado um modelo de regressão linear, o gráfico dos resíduos padronizados versus os valores ajustados apresenta um

comportamento sistemático que indica a função adequada para transformar a variável explicativa. Para se obter uma variável Y com a variância mais homogênea é utilizada a transformação logarítmica. Quando a variável resposta é positiva, esta transformação também modifica o seu intervalo de variação para $(-\infty,+\infty)$, o qual é mais coerente com a definição da distribuição normal. Além disso, os valores ajustados transformados para a escala original sempre serão positivos. A escolha da transformação mais adequada a partir da análise gráfica é muito subjetiva, por esta razão foram desenvolvidos métodos formais de seleção. Serão descritos a seguir os métodos de Tukey e Box-Cox.

Na aplicação do teste de Tukey além de verificar se os dados necessitam de uma transformação também se determina, se for o caso, o expoente da transformação potência que deve ser utilizada. Os procedimentos do teste são: i) Ajusta-se o modelo aos dados originais; ii) Inclui-se no modelo mais uma covariável Z com seus elementos definidos por $z_i = \hat{y}_i^{\,2}$ e testa-se a significância do parâmetro $\hat{\beta}_z$ desta nova covariável a um dado nível α; iii) Se o parâmetro for significante ao nível α, conclui-se que a não aditividade é significativa e alguma transformação potência Y_i^{λ} é necessária aos dados observados. Caso contrário, deve-se permanecer com o modelo ajustado aos dados originais; iv) Calcula-se o expoente da função potência que é dado por $\lambda = 1 - 2\hat{\beta}_z \overline{Y}$. Se o seu valor for próximo de zero, pode-se aceitar a transformação logarítmica.

Já Box-Cox (1964) propuseram um método para a família de transformações potência que fornece: i) estrutura linear simples; ii) constância da variância do erro; iii) independência entre as observações e iv) normalidade. A transformação potência é modificada para que a variável transformada seja contínua em $\lambda = 0$. A expressão obtida é a seguinte,

$$Y_i(\lambda) = \begin{cases} \dfrac{Y_i^{\lambda} - 1}{\lambda}, & \text{para } \lambda \neq 0 \\ \log Y_i, & \text{para } \lambda = 0. \end{cases}$$

Então, $Y(\lambda) = \left(Y_1(\lambda),\ldots, Y_n(\lambda)\right)'$ é um vetor de dimensão $nx1$, podendo-se ajustar o modelo $Y(\lambda) = X\beta + \varepsilon$ aos dados transformados. O método de máxima verossimilhança de estimação de λ é constituído das três etapas abaixo: i) arbitra-se valores para λ. Os valores de λ são escolhidos num determinado intervalo. Inicialmente, o intervalo pode ser $\lambda = \{-3,\ -2,\ -1,\ 0,\ 1,\ 2,\ 3\}$; ii) Calcula-se, para cada valor de λ, o máximo da log-verossimilhança dado por,

$$\text{lmax}(\lambda) = -\frac{1}{2} n \log(\hat{\sigma}^{\,2}(\lambda)) + \log(J(\lambda, Y)), \text{ onde,}$$

$$\hat{\sigma}^2(\lambda) = Y'(\lambda)(I - X(X'X)^{-1}X')Y(\lambda)/n,$$

$$J(\lambda, Y) = \prod_{i=1}^{n} \frac{\partial Y_i(\lambda)}{\partial Y_i}$$

$$J(\lambda, Y) \quad = \prod_{i=1}^{n} Y_i^{\lambda-1}, \quad \forall\lambda\,;$$

iii) Depois de calcular $\mathrm{lmax}(\lambda)$ para os valores do intervalo, verifica-se se o gráfico de $\mathrm{lmax}(\lambda)$ versus λ contém o ponto de máximo da curva. Se isto ocorrer, o procedimento está terminado e o valor de λ correspondente ao ponto de máximo é o estimador de máxima verossimilhança de λ. Caso contrário, é necessário ampliar o intervalo de variação dos valores para λ. Finalmente, o intervalo de $100(1 - \backslash alpha)\%$ de confiança para λ é dado por $\{\lambda: l\,max(\hat{\lambda}) - l\,max(\lambda) \leq 1/2\,{\chi_1}^2(1-\alpha)\}$, onde $l\,max(\hat{\lambda})$ é a ordenada correspondente ao ponto de máximo da curva $\mathrm{lmax}(\lambda)$ versus λ.

No ajustamento dos modelos de regressão linear para modelar o relacionamento entre a variável resposta Y e as variáveis explicativas $\mathrm{X}_1,...,\mathrm{X}_p$ é muito frequente à inadequação das hipóteses de linearidade da relação e constância da variância das componentes de Y. No parágrafo anterior foi estudada a transformação de Box-Cox que tinha o objetivo de resolver estes dois problemas simultaneamente. Nelder e Wedderburn (1972) apresentaram um exemplo, com dados de tuberculose, onde não é possível encontrar um valor para λ, a constante da transformação, que produza linearidade e constância da variância ao mesmo tempo. Eles verificaram, inclusive, que, enquanto a transformação raiz quadrada produzia normalidade do erro, a transformação logarítmica era necessária para obter aditividade dos efeitos sistemáticos. Neste mesmo trabalho, Nelder e Wedderburn também desenvolveram uma classe de modelos bem mais ampla do que a classe dos modelos clássicos de regressão onde as suposições básicas, tais como, variância constante e linearidade, não são mais exigidas. A idéia básica destes modelos é transformar os valores ajustados e não os dados para obter linearidade bem como especificar uma função de variância. Tais modelos foram denominados de modelos lineares generalizados.

Em um estudo de caso as análises foram realizadas utilizando o pacote estatístico (software) R, Version 3.1.2 (2014).

CAPÍTULO 3

Os dados do exemplo a seguir (Tabela 12), foram obtidos em um ensaio conduzido na França e se referem a comparação do número de ervas daninhas, utilizando uma transformação raiz quadrada, num experimento destinado a comparar nove tratamentos herbicidas aplicados a uma cultura de alho francês (Allium porrum), onde contaram-se o número de ervas daninhas anuais presentes em cada parcela ou unidade experimental (DAGNELIE, 1973).

Tabela 12. Comparação do número de ervas daninhas: valores iniciais e valores transformados em raiz quadrada$\left[\left(\sqrt{X+0{,}5}\right).10\right]$, somas e somas de quadrados dos desvios. Mossoró – RN, 2023.

Tratamentos	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**
Valores iniciais: $X_{i,j}$	40	8	5	13	6	3	26	6	2
	48	7	11	9	16	6	20	20	3
	16	18	25	36	10	20	6	7	12
	18	26	7	37	0	18	21	10	0
Somas	122	59	48	95	32	47	73	43	17
Somas dos quadrados dos desvios	763	243	244	659	136	217	221	123	85
Valores transformados: $Y_{i,j}$	64	29	23	37	25	19	51	25	16
	70	27	34	31	41	25	45	45	19
	41	43	50	60	32	45	25	27	35
	43	51	27	61	7	43	46	32	7
Somas	218	150	134	189	105	132	167	129	77
Somas de quadrados dos desvios	645	395	425	721	623	504	395	243	409

As variâncias parecem ser sensivelmente proporcionais aos valores das médias. Justifica-se, portanto, a aplicação a estes dados de uma transformação do tipo raiz quadrada ($Y = \sqrt{X}$) .

O Cálculo das somas e das somas dos quadrados dos desvios dos valores transformados permite verificar que não há uma relação importante entre médias e variâncias, reduzindo-se a razão entre os valores extremos das somas dos quadrados dos desvios de 9 para 3.

A Tabela 13 fornece igualmente os resultados da análise da variância, tanto para os valores iniciais como para os valores transformados. Como se verifica muitas vezes, a mudança de variável não provoca alterações importante no valor do teste F (F observado ou F calculado).

Tabela 13. Comparação do número de plantas daninhas: análise de variância dos valores iniciais e dos valores transformados em raiz quadrada. Mossoró – RN, 2023.

Fonte de variação	Graus de liberdade	Somas de quadrados dos desvios	Quadrados médios	Teste F
Tratamentos	8	2118	264,8	2,46*
Blocos	3	103	34,3	------
Interação (Resíduo)	24	2587	107,8	------
Totais	35	4808	------	------
Tratamentos	8	3635	454,4	2,68*
Blocos	3	288	96,0	------
Interação (Resíduo)	24	4071	169,6	------
Totais	35	7994	------	------

A mudança de variáveis pode, contudo, apresentar uma grande importância no que se refere, por exemplo, às comparações particulares de médias duas a duas e no cálculo dos limites de confiança das médias ou das suas diferenças. Com efeito, o quadrado médio residual, ou, neste caso, o quadrado médio da interação, é uma média das variâncias relativas aos diferentes tratamentos, e esta média pode ser demasiado elevado nos tratamentos de fraca dispersão e demasiado reduzida nos tratamentos de grande dispersão.

As transformações de dados são necessárias, para garantir ao pesquisador realizar inferências estatísticas seguras, diminuindo custos associado com alta precisão e elevada confiabilidade nos resultados experimentais ou em levantamentos.

CONSIDERAÇÕES FINAIS SOBRE TRANSFORMAÇÕES DE DADOS.

As técnicas da análise de variância e de regressão são amplamente utilizadas em várias áreas da pesquisa científica. Entretanto, para aplicação dessas técnicas, faz-se necessário atender a certas pressuposições do modelo. Contudo, na prática, nem sempre tais pressuposições são satisfeitas.

Quando se trabalha com dados que não atendem às pressuposições do modelo para aplicação dessas técnicas, dois procedimentos devem ser seguidos: primeiro, devem-se buscar novos métodos de análise que deem melhor ajuste; segundo, devem-se submeter os dados a uma transformação que atenda às pressuposições do modelo. Transformar os dados significa trocar a escala da variável original por uma outra escala.

A idéia básica é que, se para a variável original as pressuposições não são atendidas, pode existir uma transformação, tal que, na nova escala elas sejam razoavelmente atendidas.

As transformações, além de corrigir os desvios das pressuposições, podem também, ter outras aplicações, como linearizar um modelo de regressão ou eliminar a interação de um modelo de análise de variância, tornando-o aditivo, porém, em muitas situações práticas, a transformação necessária para estabilizar a variância pode não ser a mesma para obter uma resposta linear (Draper & Smith, 1981). As transformações mais comumente usadas são discutidas em: Box & Cox (1964); Steel & Torrie (1980); Snedecor & Cochran (1989).

MODELO LINEAR GERAL E SUAS PRESSUPOSIÇÕES

As técnicas de análise de variância e de regressão são tratadas através do modelo linear geral (Searle, 1987), dado por:

$$y = X\beta + \varepsilon,$$

em que: y é o vetor $(n \times 1)$ das observações; X é a matriz conhecida $(n \times p)$ de incidência dos parâmetros, obtida de acordo com o delineamento experimental e o modelo usado; β é o vetor $(p \times 1)$ dos parâmetros, desconhecido; ε é o vetor $(n \times 1)$ de erros aleatórios; n o número de observações e p o número de parâmetros do modelo.

Para o desenvolvimento teórico dessas técnicas são feitas certas pressuposições. Segundo Searle (1987) as pressuposições básicas que se deve admitir para a validade da análise de variância e de regressão, são as seguintes:

i) Aditividade: os efeitos dos fatores que ocorrem no modelo devem ser aditivos;

ii) Independência: os erros aleatórios devem ser independentes;

iii) Homocedasticidade ou homogeneidade de variâncias: os erros aleatórios devem possuir uma variância comum σ^2. Isso significa que a variabilidade das repetições de um tratamento deve ser semelhante à dos outros tratamentos, isto é, os tratamentos devem possuir variâncias homogêneas;

iv) Normalidade: os erros aleatórios devem possuir uma distribuição normal de probabilidade.

Essas pressuposições podem ser resumidas na forma: $\varepsilon \sim i.i.d.N\ 0\ ;\ I\sigma^2$,

e como consequência $y \sim N\ X\beta\ ;\ I\sigma^2$.

As suposições de homogeneidade de variâncias, normalidade e aditividade, nem sempre podem ser controladas pelo pesquisador. Já a suposição de independência dos erros, em geral, pode ser assegurada através de um esquema de aleatorização apropriado, na fase de planejamento do experimento.

Os testes apropriados para se verificar cada suposição do modelo linear são os seguintes, dentre outros:

i) Teste para não-aditividade:
   - teste "F" da análise de variância;
   - teste de Tukey para não-aditividade.

ii) Teste para independência dos erros:
   - teste de aleatoriedade;
   - teste de correlação serial de Durbin-Watson.

iii) Teste de heterocedasticidade:
   - teste de Cohran;
   - teste de Hartley;
   - teste de Bartlett.

iv) Teste de normalidade:
   - teste de χ^2 de Pearson;
   - teste de Kolmogorov-Smirnov;
   - teste de Lilliefors;
   - teste de Shapiro-Wilk.

ESCOLHA DA TRANSFORMAÇÃO

O estudo das relações entre médias e variâncias de tratamentos pode sugerir uma possível transformação, de modo que se tenha variância constante e variâncias independentes das médias (Steel & Torrie, 1980).

Assim, seja Y uma variável aleatória com média, $E(Y) = \mu$, e variância $V(Y)$. Admite-se que a variância seja uma função da média, dada por:

$$V(Y) = f(\mu)$$

O que se pretende é determinar uma função de Y, $g(Y)$, que transforme Y numa nova variável Z, com variância independente da média.

Tem-se, então que:

$$Z = g(Y)$$

Para se determinar uma transformação adequada, usa-se expansão em série de Taylor de 1ª ordem em $Z = g(Y)$ em torno de μ.

Desse modo, tem-se que:

$$Z \approx g(\mu) + g'(\mu)(Y - \mu)$$

cuja esperança é dada por:

$$E(Z) \approx g(\mu)$$

em que $E(Y - \mu) = 0$.

Consequentemente,

$$Var(Z) = E[Z - E(Z)]2$$
$$= E[g(\mu) + g'(\mu)(Y - \mu) - g(\mu)]2$$
$$= E[g'(\mu)(Y - \mu)]2$$
$$= [g'(\mu)]2\, E[Y - \mu]2$$
$$= [g'(\mu)]2\, V[Y].$$

Tem-se que:

$$V(Y) = f(\mu)$$

Logo,

$$Var(Z) = [g'(\mu)]2\, f(\mu) = C$$

em que C é uma constante, independente de μ.

Desse modo,

$$[g'(\mu)]^2 = \frac{C}{f(\mu)}$$

E assim,

$$g'(\mu) = \sqrt{\frac{C}{f(\mu)}}.$$

Portanto,

$$g(\mu) = \int \sqrt{\tfrac{C}{f(\mu)}}\, d\mu.$$

Substituindo μ por y, obtém-se:

$$g(y) = \sqrt{Var(Z)} \int \tfrac{1}{\sqrt{f(y)}}\, dy. \qquad (1)$$

Para o caso em que a variável aleatória Y segue uma distribuição binomial, tem-se que:

$$E(Y) = \pi = \mu$$
$$\mathrm{V}(Y) = \frac{\pi(1-\pi)}{N} = \frac{\mu(1-\mu)}{N}$$

A variância de Y é, portanto, uma função da média μ.

Conforme (1), tem-se:

$$g(y) = \sqrt{C} \int \frac{1}{\sqrt{\frac{y(1-y)}{N}}}\, dy$$
$$= \sqrt{C} \int \frac{1}{\frac{\sqrt{y(1-y)}}{\sqrt{N}}}\, dy$$

$$= \sqrt{C} \int \frac{\sqrt{N}}{\sqrt{y(1-y)}} dy$$

$$= C\sqrt{N} \int \frac{1}{\sqrt{y(1-y)}} dy$$

$$= C\sqrt{N} \int \frac{1}{\sqrt{y-y^2}} dy.$$

Fazendo $y = t^2$, obtém-se:

$$dy = 2t\, dt$$

Assim,

$$g(y) = C\sqrt{N} \int \frac{1}{\sqrt{t^2 - t^4}} 2tdt$$

$$= C\sqrt{N} \int \frac{2t}{\sqrt{t^2(1-t^2)}} dt$$

$$= C\sqrt{N}2 \int \frac{t}{t\sqrt{1-t^2}} dt$$

$$= C\sqrt{N}2 \int \frac{1}{\sqrt{1-t^2}} dt$$

$$= C\sqrt{N}2\text{arco seno } t + C1.$$

Tem-se que $t = \sqrt{y}$, e ignorando-se as constantes, a transformação adequada fica:

$$g(y) = \text{arco seno}\sqrt{y},$$

em que y são proporções.

A variância de Z é dada por:

$$Var[Z] = [g'(y)]2\, V[Y]$$

$$= \left[\frac{1}{\sqrt{1-y}} \frac{1}{2\sqrt{y}}\right]^2 \frac{\mu(1-\mu)}{N}$$

$$= \frac{1}{(1-y)} \frac{1}{4y} \frac{\mu(1-\mu)}{N}$$

$$= \frac{1}{(1-\mu)} \frac{1}{4\mu} \frac{\mu(1-\mu)}{N}.$$

Logo,

$$Var[Z] = \frac{1}{4N}$$

Portanto, observa-se pela expressão acima que a variância independe da média μ. Uma análise descritiva e exploratória dos dados, pode ajudar na escolha da transformação, mas, não existem regras gerais que garantam o sucesso na escolha da transformação.

As transformações mais utilizadas (Steel & Torrie, 1980) são:

i) Transformação Logarítmica:

$$log\ (y\ +\ K) \text{ ou } ln(y\ +\ K), K \geq 0.$$

Recomendada quando é constatada uma certa proporcionalidade entre a média e a variância.

É muito usada em estudos biológicos, em estudos de efeito de drogas e também utilizada para se obterem relações lineares.

ii) Transformação Raiz Quadrada:

$$\sqrt{y + K}, K \geq 0.$$

Frequentemente utilizada em dados provenientes de contagens, que geralmente seguem uma distribuição de Poisson, na qual a média é igual à variância.

iii) Transformação Recíproca:

$$\frac{1}{y_i} \text{ se } y_i \neq 0, i\ =\ 1, 2, \dots, n, \text{ ou } \frac{1}{(y_i + K)} \text{ se } y_i\ =\ 0.$$

Essa transformação é muito usada em análise de sobrevivência de plantas e animais, em estudos farmacológicos e em estudos de densidades de plantas por parcela.

iv) Transformação Angular:

$$\text{arco seno}\sqrt{y/n},$$

sendo y/n proporções.

Usada em dados expressos em porcentagens, que geralmente seguem uma distribuição binomial.

TRANSFORMAÇÃO DE BOX-COX

Box & Cox (1964) propuseram uma família de transformações dada por:

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{\lambda} - 1}{\lambda}, \lambda \neq 0 \\ \log Y, \lambda = 0 \end{cases}, \qquad (2)$$

em que Y é a variável resposta e λ o parâmetro da transformação.

A idéia básica é considerar λ como um parâmetro adicional e desconhecido do modelo, e estimá-lo pelos métodos padrões da inferência estatística.

A suposição feita por Box & Cox (1964) é que para algum λ, as observações transformadas por (2) podem ser tratadas como sendo normalmente distribuídas e independentes, com variância constante σ^2 e com esperança definida pelo modelo linear, dada por:

$$E\left[Y^{\lambda}\right] = X\beta$$

Assim,

$$Y^{\lambda} \sim N\left(X\beta;\ I\ \sigma^2\ \right).$$

Estimação do parâmetro λ

Seja $Y = (y_1, y_2, \dots, y_n)'$ o vetor de observações da variável resposta. Suponha que para algum λ desconhecido, o vetor de observações transformados $Y^{(\lambda)} = \left(y_1^{(\lambda)}, y_2^{(\lambda)}, \dots, y_n^{(\lambda)}\right)'$ satisfaça às pressuposições do modelo linear.

Assim, assumindo que, para uma escolha adequada de λ, o modelo linear é tal que:

$$Y^{\lambda} = X\beta + \varepsilon,$$

em que $\varepsilon \sim N(0;\ I\sigma^2)$.

Considerando o método da máxima verossimilhança para estimação do parâmetro λ, a função de verossimilhança de Y^{λ} é dada por:

$$L(\beta, \sigma^2, \lambda; Y) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{\frac{-(Y^{(\lambda)} - X\beta)'(Y^{(\lambda)} - X\beta)}{2\sigma^2}\right\} J(\lambda; Y),$$

em que $J(\lambda; Y)$ é o jacobiano da transformação definido por:

$$J(\lambda; Y) = \prod_{i=1}^{n} \frac{dy_i^{(\lambda)}}{dy_i} = \prod_{i=1}^{n} y_i^{(\lambda-1)}.$$

O logaritmo da função de verossimilhança fica:

$$\ell(\beta, \sigma^2, \lambda; Y) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\left[\left(Y^{(\lambda)} - X\beta\right)'\left(Y^{(\lambda)} - X\beta\right)\right] + (\lambda - 1)\sum_{i=1}^{n}\ln(y_i) \quad (3)$$

Os estimadores de máxima verossimilhança dos parâmetros envolvidos em $\ell(\beta, \sigma^2, \lambda; Y)$ são obtidos em duas etapas:

i) Para λ fixo, estimam-se β e σ^2.

Derivando-se (3) em relação a β e a σ^2, obtém-se:

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{\sigma^2} X'\left[Y^{(\lambda)} - X\beta\right];$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\left[Y^{(\lambda)} - X\beta\right]'\left[Y^{(\lambda)} - X\beta\right].$$

Igualando-se a zero ambas as derivadas, tem-se que os estimadores são dados por:

$$\hat{\beta}(\lambda) = (X' \rightleftarrows X)^{-1} X' Y^{(\lambda)};$$

$$\hat{\sigma}^2(\lambda) = \frac{1}{n}\left[Y^{(\lambda)} - X\hat{\beta}\right]'\left[Y^{(\lambda)} - X\hat{\beta}\right] = \frac{S(\lambda;Y)}{n},$$

em que $S(\lambda; Y)$ é a soma de quadrados do resíduo referente a variável transformada.

Para λ fixo, o máximo de $\ell(\beta, \sigma^2, \lambda; Y)$ é dado por:

$$\ell_{\max}(\lambda) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}^2(\lambda)) + (\lambda - 1)\sum_{i=1}^{n}\ln(y_i).$$

ii) A segunda etapa consiste em se determinar o estimador de máxima verossimilhança de λ.

Uma forma simples, porém, equivalente, para $\ell_{\max}(\lambda)$ é obtida, trabalhando-se com a transformação normalizada, dada por:

$$Z^{(\lambda)} = \frac{Y^{(\lambda)}}{J(\lambda;Y)^{\frac{1}{n}}}$$

Como $J(\lambda;Y) = \prod_{i=1}^{n} y_i^{(\lambda-1)}$, tem-se que $J(\lambda;Y)^{\frac{1}{n}} = \dot{Y}^{\lambda-1}$, sendo $\dot{Y}$ a média geométrica das observações da variável Y.

Assim,

$$Z^{(\lambda)} = \frac{Y^{(\lambda)}}{\dot{Y}^{\lambda-1}} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda \dot{Y}^{\lambda-1}}, \lambda \neq 0 \\ \dot{Y}\log(Y), \lambda = 0 \end{cases}$$

em que $\dot{Y} = \sqrt{\prod_{i=1}^{n} y_i}$.

Na transformação normalizada, $J(\lambda;Y) = 1$ e, portanto, o logaritmo da função de verossimilhança parcialmente maximizada é dado por:

$$\ell_{\max}(\lambda) = -\frac{n}{2}\log(\hat{\sigma}^2(Z)) = -\frac{n}{2}\log\left\{\frac{S(\lambda;Z)}{n}\right\}, \qquad (4)$$

em que $S(\lambda;Z)$ é a soma de quadrados do resíduo referente a variável normalizada.

Assim, a estimativa de máxima verossimilhança, $\hat{\lambda}$, é o valor de λ que maximiza $\ell_{\max}(\lambda)$, dado em (4) ou, equivalentemente, minimiza $S(\lambda;Z)$.

Box & Cox (1964) sugerem uma solução numérica mais simples possível. Para vários valores de λ num intervalo conveniente, determina-se $\ell_{\max}(\lambda)$ em função de λ e então determina-se graficamente o valor de λ que minimiza essa função. Esse, então, será o estimador de máxima verossimilhança $\hat{\lambda}$, ficando assim determinada a transformação de BOX-COX.

Método prático para estimar λ

Na aplicação do procedimento de BOX-COX, Draper & Smith (1981) sugerem que se tomem de 11 a 21 valores de λ, no intervalo de [-2; 2]. Nas proximidades do ponto de máximo, pode-se tomar valores adicionais de λ para tornar a determinação de $\hat{\lambda}$ mais precisa. Entretanto, na prática, nem sempre se utiliza o valor de $\hat{\lambda}$ obtido, mas sim, um valor mais próximo da sequência . . . , -2, $-\frac{3}{2}$, -1, $-\frac{1}{2}$, 0, $\frac{1}{2}$, 1, $\frac{3}{2}$, 2, . . .

Esse procedimento facilita a interpretação da variável transformada. Essa substituição é bastante razoável quando o valor adotado estiver contido no intervalo de confiança para λ.

A construção do intervalo de confiança para λ, baseia-se no fato de que $-2log\,\Lambda$, tem distribuição aproximadamente χ^2, em que Λ é a razão de verossimilhanças (Mood *et al.*, 1974). Assim, um intervalo de confiança, com um coeficiente de confiança de $100(1-\alpha)\%$ para λ, é dado por:

$$2\left[\ell_{\max}(\hat{\lambda}) - \ell_{\max}(\lambda)\right] \leq \chi^2_{1,\alpha},$$

em que χ^2_1 é o quantil de ordem 1 - α da distribuição χ^2 com um grau de liberdade.

Uma outra alternativa de transformação é através do uso da teoria de modelos lineares generalizados, que também envolve uma transformação, não dos dados, mas da média, e assim essa transformação é conhecida como função de ligação.

Uma função usada na análise estatística com os modelos lineares generalizados incluem uma função de ligação que associa os valores esperados da resposta aos preditores lineares no modelo. Uma função de ligação transforma as probabilidades dos níveis de uma variável de resposta categórica em uma escala contínua que é ilimitada. Depois de concluída a transformação, a relação entre os preditores e a resposta pode ser modelada com regressão linear. Por exemplo, uma variável de resposta binária pode ter dois valores exclusivos. A conversão desses valores em probabilidades faz com que a variável de resposta varie de 0 a 1. Quando você aplica uma função de ligação adequada às probabilidades, os números resultantes variam de $-\infty$ até $+\infty$. A Forma geral da função de ligação é dada pela equação a seguir: $g\,(\mu_i) = X_i'\beta$. Os modelos lineares generalizados **(**GLMs**)** é uma generalização dos modelos lineares ordinários. Os modelos lineares generalizados são usados quando os resíduos ou erros do modelo matemático apresentam distribuição diferente da normal, simétrica ou gaussiana. A natureza da variável resposta é um indicador do tipo de distribuição dos resíduos que será obtida nos modelos matemáticos. Por exemplos, variáveis de contagem são inteiras e apresentam os valores limitados no zero. Esse tipo de variável, em geral, tem uma distribuição de erros assimétrica para valores baixos e uma variância que aumenta com a média dos valores preditos, violando duas premissas dos modelos lineares. Os casos mais comuns de modelos lineares generalizados são de variáveis resposta de contagem, proporção e binária, muito comum nos estudos em ciências agrárias, biológicas, em ecologia e evolução. Deve-se adotar os modelos lineares generalizados principalmente quando a variável resposta é expressa em: contagens simples; contagem expressa em proporções; número de sucesso e tentativa; variáveis binárias, como por exemplo números de animais sadios e doentes, número de insetos morto e vivo, contagem de pulgões por folha de uma espécie vegetal e tempo para o evento ocorrer aplicados em modelos de sobrevivência.

Uma das formas de se compreender como é a estrutura dos modelos matemáticos representativos dos modelos lineares generalizados é separar o modelo em dois componentes: a relação determinística entre as variáveis (resposta e preditora) e o componente aleatório dos resíduos (distribuição dos erros). Em um modelo linear ordinário a relação entre as variáveis é uma proporção constante, o que define uma relação funcional de uma reta. Quando se tem uma contagem, essa relação pode ter uma estrutura

funcional de um modelo exponencial. Para esses casos, os modelos lineares generalizados utilizam uma função de ligação log para linearizar a relação determinística entre as variáveis. Portanto, a estrutura determinística dos modelos lineares generalizados é definida por um preditor linear, associada à função de ligação. O componente aleatório dos resíduos, no caso de uma variável de contagem, segue, em geral, uma distribuição teórica de probabilidade de Poisson. A variável aleatória discreta ou descontínua que possui distribuição teórica de probabilidade de Poisson é uma variável aleatória definida por apenas um parâmetro (λ), equivalente à média, chamada de lambda, que é o número médio de eventos que ocorrem em um intervalo contido considerado de tempo, distância, área ou volume, ou seja, é uma constante positiva dada, conhecida como taxa de Poisson. A distribuição de probabilidade de Poisson tem uma propriedade única que é aquela onde seu desvio padrão é igual à sua média. Portanto, se a média aumenta, o desvio acompanha esse aumento e a distribuição passa a ter a maior variabilidade. O preditor linear está associado à estrutura determinística do modelo e está relacionado à linearização da relação, aqui definido como η: $\eta = \alpha + \beta x$. A função de ligação é o que relaciona o preditor linear com a esperança do modelo: $\eta = g^{-1}[E(y)]$. Ou seja, nos modelos lineares generalizados não é a variável resposta que tem uma relação linear com a preditora, e sim o preditor linear que tem uma relação linear com as preditoras.

REFERÊNCIAS

ABRAMOVITZ, M.; STEGUN, I., Handbook of mathematical functions with formulas, graphs and mathematical tables, 9ª Ed., Dover, New York, U.SA. 1972. 1060 p.

AITKEN, A. C. Studies in practical mathematics: the evaluation of the latent roots and latent vectors of a matrix. Proceedings of the Royal Society of Edinburgh, v. 57, p. 269-304, 1937.

ALBUQUERQUE, J. J. L. Estatística experimental. Universidade Federal do Ceará (UFC). Centro de Ciências. Departamento de Estatística e Matemática Aplicada. Fortaleza, CE. 1986.

ALMEIDA, V. A. F. Métodos quantitativos para ciência da computação experimental. Departamento da ciência da computação. Universidade Federal de Minas Gerais, DCC – UFMG. 2005. 34 p. (notas de aula).

BALESTRASSI, P. P.; PAIVA, A. P. Estatística aplicada. UNIFEI - Universidade Federal de Itajubá.

Instituto de Engenharia de Produção & Gestão. Itajubá, MG, 2007.263 p. (apostila).

ANDRADE, D. F.; OGLIARI, P. J. Estatística para as ciências agrárias e biológicas: com noções de experimentação. Editora da Universidade Federal de Santa Catarina - EDUFSC. 3ª edição, 2013 475 p.

ANSCOMBE, F.J. The transformation of Poisson, binomial, and negative binomial data. Biometrika Vol. 35, nº ¾. Dec.. Rothamsted, Harpenden. Inglaterra.1948. p. 246 – 254.

AYRES, M., AYRES JR., M., AYRES, D.L., SANTOS, A.S. BioEstat. Versão 5.0, Sociedade Civil Mamirauá, MCT – CNPq, Belém, Pará, Brasil. 2007.

BANZATTO, D.A.; KRONKA, S. N. Experimentação agrícola. Fundação de Estudos e Pesquisas em Agronomia, Medicina Veterinária e Zootecnia – FUNEP. Faculdade de Ciências Agrárias e veterinárias – FCAV/UNESP. Campus de Jaboticabal. 2006, SP. 237 p.

BARBIN, D. Planejamento e análise estatística de experimentos agronômicos. Editora Midas Ltda. Arapongas, PR. 2003. 208 p.

BARTLETT, M.S. The use of transformations. Biometrics, n. 3, p. 39 -52. USA. 1947.

BERRY, D.A. Logarithmic transformations in ANOVA. Biometrics, Washington, v.43, p.439-456, 1987.

BLISS, C. I. The transformation of percentages for use in the analysis of variance. Ohio Journal of Science: Volume 38, Issue 1. n. 1. U.S.A. January, 1938. p .9-12.

BLISS, C.I. The method of probits. Science. U.S.A.12 January. v. 79, n. 2037. 1934: p.38-39.

BONINI, E. E.; BONINI, S. E. Estatística teoria e exercícios. L. P. M. São Paulo. 1972. 443 p.

BOX, G. E. P.; COX, D. R. An analysis of transformations. Journal of the Royal Statistics Society. Serie B (Statistical Methodological), New York. v.26, n.2, p.211-243, discussion p.244-252. 1964.

BOX, G .E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. 1954. Annals of Mathematical Statistics, 25: 290-302.

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

BOX, G. E. P.; MÜLLER, M. E. A note on the generation of random normal deviates. Annals of Statistics, 29, pp 610-611, 1958.

BRIEGER, F.G. Limites unilaterais e bilaterais na análise estatística. Campinas, SP. Bragantia, 6: 479-545.

BRITO, M. Metodologia da investigação agrária. Escola Superior Agrária – Instituto Politécnico de Viana do Castelo. Portugal. 70 p. Acesso em 27/11/2014.<www.ci.esapl.pt/mbrito/metodologia%20da%20investigação%20agrária.pdf.>

BUSSAB, W. O. Análise de variância e de regressão. Atual Editora LTDA. São Paulo, SP. 1986. 151 p. (Coleção Métodos Quantitativos).

BUSSAB, W. O. ; MORETTIN, P. A. Estatística Básica. São Paulo: SARAIVA, 5.ed. 2002, 526p.

BUSTOS, O. Procedimentos robustos. 4º Simpósio Nacional de Probabilidade e Estatística. 163p. 1980.

BUSTOS, O. Outliers e robustez. Revista Brasileira de Estatística. 49:7-30. 1988.

CADIMA, J. < http://www.isa.utl.pt/dm/estdel/estdel/aulasRLS.pdf.> Acesso em 16/06/2015.

CALADO, V.; MONTGOMERY, D. Planejamento de experimentos usando o Statistica. Rio de Janeiro: E – Papers Serviços Editoriais, 2003. 260 p.

CAMPOS, K. A.; PAIXÃO, C. A.; MORAIS, A.R. Transformação de dados como alternativa a análise de variância univariada. Sigmae, Alfenas, v.2, n.3, p. 57-64. 2013.

CECCON, G.; RAGA, A.; DUARTE, AP.; SILOTO, R.C. Efeito de inseticidas na semeadura sobre pragas iniciais e produtividade de milho safrinha em plantio direto. Bragantia, v.63, n. 2, p. 227-237, fev. 2004.

CLARK, A.G.; LEONARD, WH. The analysis of variance with special reference a data expressed as percentages. Journal of the American Society of Agronomy, v. 31, n. 1, p. 55-66. 1939.

COCHRAN, W., G. Some difficulties in the statistical analysis of replicated experiments. Emp. Journal Exp. Agric., v.6, p.157-175, 1938.

COCHRAN, W., G.; COX., G., M. Experimental designs. 2nd ed. London, John Wiley. 1957. 611 p.

COCHRAN, W., G.; COX., G., M. Diseños experimentales. México, Editorial Trillas.1971. 664 p.

COCHRAN, W.G. Some consequences when the assumptions forthe analysis of variance are not satisfied. Biometria, 1947.

CONAGIN A; NAGAI V; IGUE T. Efeito da falta de normalidade em testes de homogeneidade das variâncias. 1993. Bragantia. 57: 203-214.

COUTO; M.R.M.; LÚCIO, A.D.; LOPES, S.J. CARPES, R.H. Transformações de dados em experimentos com abobrinha italiana em ambiente protegido. Ciência Rural, Santa Maria, v.39, n.6, p.1701-1707, set, 2009.

DAGNELIE, P. Estatística: Teoria e métodos: 2º Volume. Publicações Europa – América LDA. Biblioteca Universitária, Lisboa, Portugal, 1973. 536 p.

DE MUNTER, P. Sur les transformations des variables aléatoires, Bullettin Society Belge of Statistics, n. 97, p. 111, 1958.

DEMÉTRIO, C.G.B. Transformação de dados. Efeitos sobre a análise da variância. Piracicaba, SP, ESALQ, 1978. 113 P. (Dissertação de mestrado).

DIAS; L. A. S.; BARROS, W. S. Biometria experimental. Universidade Federal de Viçosa - UFV. Viçosa, MG. 2009. 408 p.

DRAPER, N. R.; SMITH, H. Applied regression analysis. 2ª Edição. John Wiley & Sons, Inc. New York, U.S.A. 1981. 736 p.

DRAPER, N. R.; SMITH, H. Applied regression analysis. 3ª Edição. John Wiley & Sons, Inc. New York, U.S.A. 1998. 736 p.

ELLISON, A. M.; GOTELLI, N. J. Princípios de Estatística em Ecologia. Artmed Editora, São Paulo, SP. 2010. 532 p.

ELLIOTT, J. M. Some methods for the statistical analysis of samples of benthic invertebrates. Freshwater Biological Association. 25, 1979.156p.

FAZOLIN, M.; COSTA, C.R.; DAMACENO, J.E.O.; ABULQUERQUE, E.S.; CAVALCANTE, A.S.S.; ESTRELA, J.L.V. Fumigação de milho para o controle do gorgulho utilizando caule de *Tanaecium nocturnum* (Bignoniaceae). Pesquisa Agropecuária Brasileira, v. 45, n. 1, p. 1-6, jan. 2010.

FERNANDEZ, G. C. J. Residual analysis and data ytansformations: important tools in statistical analysis. Hortscience, v. 27, n. 4, p. 297-300, 1992.

FERREIRA, P., V Estatística Experimental Aplicada à Agronomia. Editora da Universidade Federal de Alagoas: EDUFAL. Maceió-AL., 3ª Ed., 2000. 419 p.

FIGUEIREDO, M.L.C.; MARTINS-DIAS, A.M.P.; CRUZ, I. Relação entre a lagarta-do-cartucho e seus agentes de controle biológico natural na produção de milho. Pesquisa Agropecuária Brasileira, v.41, n. 12, p. 1693-1698, dez. 2006.

FINNEY, D.J. 1960. An introduction to the theory of experimental design. Univ. Chicago, 1960. 222p.

FISHER, R. A. Frequency distributions of the values of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika, London, v. 10, n. 4, p. 507 – 521, 1915.

FISHER, R. A. The use of multiple measurements in taxonomic problems. Annals of Eugenics, London, v. 7, n. 2, p. 179-188, 1936. http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x

FISHER, R.A.; YATES, F. Statistical tables for biological, agricultural and medical research. London: Hafner Pub. Co., 1963. 146 P.

FONSECA, J. S.; MARTINS G. A.; TOLEDO, G. L. Estatística aplicada. Editora Atlas. 2ª Edição. São Paulo, SP. 1982.272 p.

GABRIEL K.R. Analysis of variance of proportions with unequal frequencies. Journal American Statistical Association. USA. n. 58, p. 1133 – 1157., 1963.

GOMEZ, K. A.; GOMEZ, A. A. Statistical Procedures for Agricultural Research. 2 nd Edition. New York, USA. John Wiley & Sons, 1984 - 704 p.

GONÇALVES, P.A.S.; SOUSA E SILVA, C.R. Efeito de espécies vegetais em bordadura em cebola sobre a densidade populacional de tripés e sirfídeos predadores. Horticultura Brasileira, v. 21, n. 4, p. 731-733, out./dez. 2003.

GOOVAERTS, P. Geostatistics for natural resources characterization, Oxford University Press. 1977.483 p.

GRIFFITHS, D.A. Interval estimation for the three-parameter lognormal distribution via the likelihood function. Journal of the Royal Statistical Society. Series C (Applied Statistics), New York, v.29, n.1, p.58-68, 1980.

GRIMM, H. transformations von Zufallsvariablen. **Biometrische Zeitschrift**, n. 2, p. 164 – 182, 1960.

GUJARATI, D. N. Econometria básica. Terceira edição. Makron Books. São Paulo, SP. 2000. 882 p.

HEALY, M.J.R.; TAYLOR, L.R. Tables for power – law transformations, Biometrika, n. 49, p . 557 – 559, London. 1962.

HEATH, O.V.S. A estatística na pesquisa científica. São Paulo: EPU: Editora da Universidade de São Paulo, 1981.95 p.

HILL, B.M. The three - parameter lognormal distribution and Bayesian analysis of a point-source epidemic. Journal of the American Statistical Association, Alexandria. v.58, n.301, p.72-84, 1963.

HOAGLIN, D. C. F. MOSTELLER; J. W. TUKEY. Análise exploratória de dados. Técnicas robustas. Edições Salamandra. Lisboa.1992. 446p

KLECZKOWSKI, A. The transformation of local lesion counts for statistical analysis. Annals of Applied Biology, n. 36, p. 139 – 152, 1949.

KRUSKAL J B. analysis of factorial experiments by estimating monotone transformations of the data, Journal of the Royal Statistical Society , Série B. n. 27, p . 251 – 263. London, 1965.

LAW, A.; KELTON, D. Simulation modeling and analysis, 2ª ed. New York: McGraw-Hill, 1991.672 p.

LI, J.C.R. Statistical inference, vol.1 Edwards, ann Arbor, London, 658 p. 1964a.

LI, J.C.R. Statistical inference, vol.2 Edwards, ann Arbor, London, 575 p. 1964b.

LIMA, M.; PATERNIANI, M.E.A.G.Z.; DUDIENAS, C.; SIQUEIRA, W.J.; SAWAZAKI, E.; SORDI, G. Avaliação da resistência à ferrugem tropical em linhagens de milho. Bragantia, v. 55, n. 2, p. 269-273, fev. 1996.

LIMA, M.; VITTI, P.; GALLO, P.B. Populações de milho: características agronômicas e tecnológicas. Bragantia, v. 47, n. 1, p. 55-62, jan. 1988.

LIMA, P.C.; R. R. LIMA. Estatística Experimental: Guia de estudos. CEAD - Universidade Federal de Lavras, Lavras, MG. 2014. 186 p.

LÚCIO, A.D.; COUTO, M.R.M.; LOPES, S.J.; STORCK, L. Transformação box - cox em experimentos com pimentão em ambiente protegido. Horticultura Brasileira, v. 29, n. 1, jan.- mar. 2011.

LÚCIO, A.D.; SCHWERTNER, D.; HAESBAERT, F.M.; SANTOS, D.; BRUNES, R.R.; RIBEIRO, A.L.P.; LOPES, S.J. Violação dos pressupostos do modelo matemático e transformação de dados. Horticultura Brasileira, v. 3, jul – set. 2012.

MAcARTHUR, R. H.; E. O. Wilson, E. O. The theory of island biogeography. Princeton University Press, Princeton, N J. 1967. 203 p.

MARTINS, C. M. T.; MENDES, M. G. T.; ABREU, J. M.; ALMEIDA, J. P. L.; LIMA, J. P.; LIMA, I. P. Hidrologia urbana: conceitos básicos. Série cursos técnicos 1. Universidade de Coimbra, Organização das Nações Unidas Para a Educação, a Ciência e a Cultura. Coimbra, Portugal. 2010. 210 p.

MATOS, M. A. Manual operacional para a regressão linear. Faculdade de Engenharia da Universidade do Porto – FEUP, PORTO, POTUGAL, 1995. 26 p. (apostila).

MATHERON, G. Les fonctions de transfer dês petits panneaux, Note Geoestatistique. Nº 127, Les Cahiers Du Centre de Morphologie Mathèmatique de Fontainebleau. 1974.

MENDES, S.M.; BOREGAS, K.G.B.; LOPES, M.E.; WAQUIL, M.S.; WAQUIL, J.M. Respostas da lagarta-do-cartucho a milho geneticamente modificado expressando a toxina Cry 1A(b). Pesquisa Agropecuária Brasileira, v. 46 n. 3, p. 239-244, mar. 2011.

MISCHAN, M. M.; PINHO, S. Z. Experimentação agronômica: dados não – balanceados. Universidade Estadual Paulista – UNESP. Fundação do Instituto de Biociências – FUNDIBIO. 1ª Edição. Botucatu, SP. 1996. 472 p.

MUGE, F. As funções de recuperação globais como instrumento de planejamento mineiro, Tese de Doutoramento, Instituo Superior Técnico – Universidade Técnica de Lisboa (IST-UTL), Lisboa, Portugal.1982. 485 p.

MURTEIRA, B.J.F. Análise exploratória de dados - estatística descritiva. Editora McGraw – Hill de Portugal Ltda. Lisboa, 1993. 341 p.

NAGHETTINI, M.; PINTO, E.J.A. Hidrologia estatística. CPRM: Serviço geológico do brasil. Belo Horizonte, MG. 2007. 561 p.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. Journal of the Royal Statistical Society. Serie A, London Vol. 135, nº. 3, p. 370-384, 1972.

NIHEI, T.H.; FERREIRA, J.M. Análise dialética de linhagens de milho com ênfase na resistência a doenças foliares. Pesquisa Agropecuária Brasileira, v. 47, n.3, p. 369-377, mar. 2012.

NUNES, R.P. Métodos para a pesquisa agronômica. Departamento de Fitotecnia. Centro de Ciências Agrárias. Universidade Federal do Ceará. Fortaleza, CE. 1998. 564 p.

OGLIARI, P. J.< www.inf.ufsc.br/~ogliari/.../Diagnosticonaanalisederegressao.ppt.> Universidade Fe deral de Santa Catarina UFSC. Florianópolis, SC. 2015. Acesso em 04/08/2015.

OLIVEIRA, C. M.; OLIVEIRA, E.; CANUTO, M.; CRUZ, I. Controle químico da cigarrinha-do-milho e incidência dos enfezamentos causados por molicutes. Pesquisa Agropecuária Brasileira, v. 42, n.3, p. 297-303, mar. 2007.

PADOVANI, C. R. P.; ARAGON, F. F. Programa computacional para método de discriminante de Fisher. Energia na Agricultura. Botucatu, v. 20, n. 1, p 1-10. 2005.

PAGANO, M.; GAUVREAU, K. Princípios de bioestatística. Thomson Learning, São Paulo, SP. 2006.524 p.

PEARSON, E.S.; HARTLEY, H.O. Biometrika tables for statisticians. Vol. 1., Cambridge, University Press, London. 1970, 270p.

PEDROSA, A. C.; GAMA, S. M. A. Introdução Computacional à Probabilidade e Estatística. Porto Editora. Porto - Portugal.2004, 607 p.

PÉREZ, F.; SILVA, G.; TAPIA, M.; HEPP, R. Variación anual de las propiedades insecticidas de Peumus boldus sobre Sitophilus zeamais. Pesquisa Agropecuária Brasileira, v. 42, n. 5, p.633-639, maio 2007.

PIMENTEL GOMES, F. Curso de Estatística Experimental. 12ª Edição. Livraria Nobel, Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiróz". Piracicaba, SP. 1987. 480 p.

PIMENTEL-GOMES, F. Curso de estatística experimental. 15ª Edição. Piracicaba: FEALQ, 2000. 451 p.

PIMENTEL-GOMES, F. Curso de estatística experimental. 15ª Edição. Piracicaba: FEALQ, 2009. 451 p.

PRESTON, F. W. The canonical distribution of commonness and rarity: Part I. Ecology. v. 43, n. 2. 1962, p. 185–215.

RADTKE, J.J. Desvendando a estatística com o R commander. UTFPR - Universidade Tecnológica Federal do Paraná. 2015.50 p. (apostila).

R VERSION 3.1.2. Viena, Áustria: Foundation for Statistical Computing, 2014. (Software).

RESENDE, M.D.V. Genética biométrica e estatística no melhoramento de plantas perenes. Embrapa. Informação Tecnológica. Brasília - DF. 2002. 975 p.

RESENDE, M.D.V. Matemática e estatística na análise de experimentos e no melhoramento genético. Editora da Embrapa. Colombo, PR: Embrapa Florestas, 2007. 562 p.

RIBEIRO JÚNIOR, J. I. Análise estatística no SAEG. Universidade Federal de Viçosa – UFV. Centro de Ciências Exatas e Tecnológicas, Viçosa, MG. Departamento de Informática. 2001. 301 p.

SAMPAIO, I.B.M. Estatística aplicada à experimentação animal. 2ª Edição. FEPMVZ - Editora Fundação de Ensino e Pesquisa em Medicina Veterinária e Zootecnia (FEPMVZ). Escola de Veterinária da Universidade Federal de Minas Gerais (UFMG). Belo Horizonte, MG, 2002. 265 p.

SAWAZAKI, E.; LORDELLO, A.I.L.; LORDELLO, R.R.A. Herança da resistência de milho a Meloidogyne javanica. Bragantia, v. 57, n. 2, p. 259-265, fev. 1998.

SILVA, G.; ORREGO, O.; HEPP, R.; RAPIA, M. Búsqueda da plantas com propiedades insecticidas para el control de Sitophilus zeamais em maíz almacenado. Pesquisa Agropecuária Brasileira, v. 40, n.1, p. 11-17, jan. 2005.

SILVA, I.P.; SILVA, J.A.A. Métodos estatísticos aplicados à pesquisa científica: Uma abordagem para profissionais da pesquisa agropecuária. Universidade federal Rural de Pernambuco. UFRPE. Recife, PE. 1999. 305 p.

SILVA, P. S. L. Métodos para pesquisas com plantas. Mossoró: Editora da Universidade federal do Semi-Árido – Edufersa. 2013. 264 p.

SILVA, W.J. Variabilidade do número de internódios e altura da espiga em linhagens e híbridos de milho. Bragantia, v. 22, n.8, p. 81-89, set. 1963.

SIQUEIRA, A.L. Uso de transformação em análise de variância e análise de regressão. Dissertação (Mestrado em Matematica e Estatistica)- Universidade de São Paulo, SP. 1983. 146p.

SNEDECOR, G.W.; W.G. COCHRAN. Statistical methods, 6ª Edição. The Iowa State University Press, 1967. Ames, Iowa. U.S.A. 593 p.

SOARES, A. Geoestatística para as ciências da terra e do ambiente. Coleção Ensino da Ciência e da Tecnologia. IST Press. Instituto Superior Técnico, Lisboa, Portugal. 2000. 220 p.

SOKAL, R.R.; ROHLF, F. J. Biometry: the principles of statistics in biological research. New York, 3 rd ed. Freeman & Company,1995, 887 p.

SOUZA, G. S. Introdução aos modelos de regressão linear e não linear. Empresa Brasileira de Pesquisa Agropecuária. Brasília: Embrapa – SPI/ Embrapa – SEA, 1998. 505p.

STANISAVLJEVIĆ, R.; DJOKIĆ, D.; MILENKOVIĆ, J.; DJUKANOVIĆ, L.; STEVOVIĆ, V.; DODIG, D. Disiccation, postharvest maturity and seed aging of tall oat-grass. Pesquisa Agropecuária Brasileira, v. 45, n.11, p. 1297-1302, nov. 2010.

STEEL, R.G.D.; TORRIE, J.H. Principles and procedures of statistics. New York: Mc – Graw Hill, 1960. 490 p.

STEEL, R.G.D.; TORRIE, J.H. Principles and procedures of statistics: a biometrical approach. New York: Mc – Graw Hill, 1981. 633 p.

STEEL, R.G.D.; TORRIE, J.H.; DICKEY, D. A. Principles and procedures of statistics: a biometrical approach. 3º ed. New York: Mc – Graw Hill, 1997. 633 p.

TAYLOR, L.R. Aggregation, variance and the mean. Nature, n. 189, p. 732 – 735, USA. 1961.

THӦNI, H. Transformations of variables used in the analysis of experimental and observational data. A review. Iowa State University, Statistical laboratory, USA. Technology. Rep 7, 61 p. 1967.

THÖNI, H., Transformations of variables used in the analysis of experimental and observacional data. A Review. Iowa State University. Ames, Iowa, U.S.A.1978.

THOMPSON, R. The estimation of heritability with unbalanced data.Biometrics, v. 33, p. 485-504, 1977.

TUKEY, J. W. One degree of freedon for non-additivity. Biometrics. 5:232-242.1949.

TUKEY, J.W. On the comparative anatomy of transformations, Annals of Mathematical Statistics, n . 28, p. 602 – 632, USA. 1957.

VENTSEL, H. Théorie des probabilités, Ed. MIR, Moscou, 1973.563 p.

VIANA, P.A.; POTENZA, M.R. Avaliação de antibiose e não-preferência em cultivares de milho selecionados com resistência à lagarta-do-cartucho. Bragantia, v. 59, n. 1, p. 27-33, jan. 2000.

WEISBERG, S. Applied Linear Regression. New York, John Wiley and Sons, 1980. 283 p.

WIKIPÉDIA: <https://pt.wikipedia.org/wiki/Formalismo_multigaussiano>. Acesso em 06/08/2015.

YAMAMURA K. Transformation using (x+0.5) to stabilize the variance of populations. Journal Researches on Population Ecology. 1999. 42: 229-234.

YEO, I.K., JOHNSON, R.A. A New Family of Power Transformation to Improve Normality or Symmetry. Biometrika, 87, 954-959, 2000.

YEVJEVICH, V. M. Section 8-II Statistical and probability analysis of hydrological data. Part II Regression and correlation analysis. In: CHOW, V. T. Handbook of applied hydrology. Ed. McGraw-Hill. 1964. 1495 p.

ZAR, J. H. Biostatistical analysis.Prentice Hall, Upper Saddle River, NJ,3rd ed, 1996. 718p.

ZAR, J.H. Biostatistical analysis. Edition: 5th. New York: Prentice Hall, 2010. 960 p.

ZIMMERMANN, F.J. Estatística aplicada à pesquisa agrícola. Embrapa Arroz e Feijão. Santo Antônio de Goiás. 2004. 402 p.

APÊNDICE

ALFABETO GREGO

NOME DA LETRA	SÍMBOLOS	
	MAIÚSCULAS	MINÚSCULAS
Alfa	Α	α
Beta	Β	β
Gama	Γ	γ
Delta	Δ	δ
Épsilon	Ε	ε
Zeta	Ζ	ζ
Eta	Η	η
Téta	Θ	θ
Iota	Ι	ι
Capa	Κ	κ
Lambda	Λ	λ
Um(mi)	Μ	μ
Nu(ni)	Ν	ν
Csi	Ξ	ξ
Omicron	Ο	ο
Pi	Π	π
Ró	Ρ	ρ
Sigma	Σ	σ
Tau	Τ	τ
Úpsilon(ipsilon)	Υ	υ
Fi	Φ	φ
Chi(qui)	Χ	χ
Psi	Ψ	ψ

ÍNDICE REMISSIVO

A

aleatória, 16, 24, 26, 44, 58, 87, 97, 98, 100, 101, 103, 110, 111, 116
amostragem, 14, 50, 51, 94
análise de variância, 10, 13, 14, 30, 33, 39, 42, 55, 60, 62, 88, 92, 96
análise exploratória, 11, 12, 68, 69
arco seno, 16, 33, 35, 36, 46, 48, 57, 96, 112, 113
assimetria, 14, 19, 28, 30, 33, 38, 91

B

Bayes, 104
binomial negativa, 51, 96
Box-Cox, 18, 20, 28, 43, 53, 67, 73, 91, 105, 106, 113

C

classes, 83, 84, 104
condicional, 6, 103, 104
contagem, 5, 12, 15, 21, 23, 25, 28, 33, 34, 38, 39, 47, 56, 63, 64, 94, 95, 96, 116
Cúbica, 13
curtose, 19

D

delineamento, 10, 11, 14, 33, 109
desvio, 23, 32, 38, 40, 41, 43, 47, 49, 54, 56, 58, 64, 85, 117
desvio padrão, 23, 38, 40, 41, 43, 47, 50, 54, 56, 58, 64, 85, 117
distribuição, 5, 6, 10, 11, 12, 14, 15, 17, 18, 21, 25, 26, 27, 30, 33, 36, 38, 39, 40, 42, 45, 47, 48, 49, 50, 51, 52, 53, 55, 56, 57, 58, 59, 60, 62, 63, 65, 69, 82, 83, 85, 87, 89, 90, 92, 94, 95, 96, 97, 98, 100, 101, 102, 103, 104, 109, 111, 113, 115, 116

E

erros, 4, 5, 10, 11, 14, 16, 18, 30, 32, 33, 40, 42, 44, 45, 47, 51, 57, 58, 59, 60, 61, 62, 67, 73, 74, 80, 82, 89, 90, 91, 92, 93, 94, 96, 109, 116
escala, 10, 13, 15, 22, 24, 29, 32, 33, 35, 36, 42, 45, 47, 56, 57, 58, 59, 62, 69, 89, 94, 95, 96, 98, 105, 108, 116
estandardização, 43
eventos, 15, 117
exponencial, 6, 27, 40, 41, 42, 69, 72, 77, 100, 101, 116

F

Fisher, 5, 30, 31, 48, 49, 56, 59, 85, 92, 123
função de ligação, 6, 116

H

Hartley, 34, 47, 110
heterocedasticidade, 10, 19, 32, 42, 47, 51, 79, 81, 82, 90, 110
hiperbólico, 51
homogeneidade, 5, 14, 18, 32, 34, 39, 42, 45, 51, 60, 62, 73, 90, 98, 109, 119

I

independentes, 10, 11, 33, 39, 42, 45, 55, 59, 62, 63, 74, 76, 77, 86, 87, 92, 96, 98, 100, 101, 109, 110, 113
intervalo, 16, 24, 33, 41, 43, 49, 56, 60, 68, 70, 85, 86, 89, 104, 105, 106, 115, 117

J

jacobiano, 99, 100, 114

L

linear, 4, 5, 6, 11, 12, 13, 18, 22, 23, 24, 25, 27, 29, 30, 31, 32, 33, 36, 38, 40, 41, 42, 51, 68, 72, 73, 74, 75, 76, 80, 82, 83, 84, 85, 87, 88, 89, 90, 91, 98, 99, 103, 104, 105, 106, 109, 113, 114, 116, 122, 124
logit, 52

M

média, 10, 11, 12, 14, 16, 20, 21, 25, 27, 32, 34, 38, 39, 40, 41, 42, 43, 45, 47, 49, 50, 51, 53, 55, 56, 57, 58, 59, 60, 62, 64, 67, 69, 74, 76, 82, 85, 87, 89, 91, 93, 94, 95, 96, 97, 98, 100, 101, 102, 103, 108, 110, 111, 112, 113, 114, 116, 117
mínimos quadrados, 32, 44, 54, 72, 79, 82, 87, 88, 89
modelo matemático, 10, 15, 20, 47, 59, 72, 92, 94, 116, 122
monotonicamente, 103

N

não linear, 4, 5, 6, 36, 72, 124

não paramétrica, 5, 18, 36

P

parcela, 23, 35, 39, 59, 60, 93, 94, 107, 113
Poisson, 6, 12, 14, 15, 25, 28, 33, 34, 38, 39, 45, 54, 55, 56, 57, 60, 63, 65, 88, 94, 96, 97, 113, 116, 118
porcentagem, 16, 23, 35, 39, 48, 49, 56, 57, 60, 62, 63
preditor, 6, 116
pressuposto, 76
probabilidade, 4, 10, 11, 18, 26, 29, 32, 34, 36, 38, 47, 49, 52, 53, 62, 63, 82, 85, 86, 92, 98, 101, 102, 109, 116

Q

Qui-quadrado, 34, 38, 101

R

raiz quadrada, 5, 12, 15, 18, 25, 28, 30, 33, 34, 36, 38, 39, 41, 45, 49, 50, 51, 54, 56, 57, 58, 63, 65, 67, 68, 76, 79, 84, 88, 94, 95, 98, 106, 107, 108
recíproca, 13, 28, 33, 53, 76
regressão, 4, 5, 7, 10, 13, 18, 23, 27, 36, 38, 40, 41, 42, 44, 51, 52, 54, 60, 69, 71, 72, 74, 75, 76, 77, 79, 81, 82, 84, 87, 89, 90, 91, 92, 96, 98, 99, 104, 106, 108, 109, 116, 119, 122, 124, 131
resíduos, 13, 16, 17, 20, 23, 27, 31, 41, 42, 44, 74, 76, 87, 89, 91, 92, 96, 104, 116

T

Taylor, 15, 68, 87, 97, 110
Teorema, 87
tratamentos, 7, 10, 11, 12, 14, 15, 20, 21, 22, 25, 27, 31, 34, 36, 38, 40, 45, 47, 55, 56, 57, 59, 60, 61, 62, 63, 64, 65, 92, 93, 94, 98, 107, 108, 109, 110

U

unidade experimental, 16, 107

V

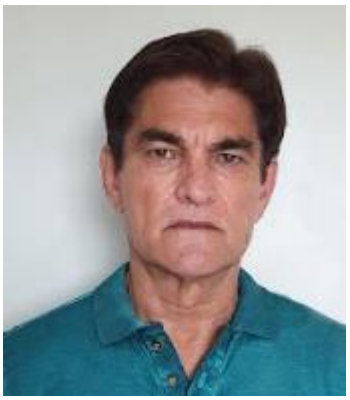
variância, 4, 5, 7, 10, 11, 12, 13, 14, 15, 16, 18, 21, 22, 23, 25, 26, 27, 29, 30, 31, 32, 33, 36, 38, 39, 40, 41, 42, 44, 45, 47, 49, 50, 51, 53, 54, 55, 56, 57, 58, 59, 60, 62, 63, 67, 74, 79, 82, 85, 87, 88, 90, 92, 93, 94, 95, 96, 97, 98, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 116, 119, 120, 124

SOBRE OS AUTORES

Janilson Pinheiro de Assis, Natural de João Câmara, RN
Engenheiro Agrônomo graduado em Engenharia Agronômica (1987) na ESCOLA SUPERIOR DE AGRICULTURA DE MOSSORÓ - ESAM. Concluiu o Curso de Pós-graduação-Mestrado (1990) em Engenharia Agronômica (Fitotecnia-Estatística experimental) na Universidade Federal do Ceará (UFC). Realizou o curso de Pós-graduação Doutorado (2014) em Produção Vegetal - Fitotecnia na Universidade de São Paulo (USP), com pesquisa direcionada para a área de estatística e modelagem aplicada à agricultura. Orientou e orienta estudantes bolsistas de monitoria, de residência acadêmica e iniciação científica. Atualmente, é Professor Titular da Universidade Federal Rural do Semi-Árido (UFERSA), Foi Professor da Escola Superior de Agricultura de Mossoró (ESAM) onde leciona a disciplina de Estatística há 33 anos, possui seis livros publicados, 11 apostilas e mais de 25 artigos completos publicados/aceitos em revistas nacionais e internacionais além de 20 resumos simples/expandido. É revisor de dez revistas nacionais e internacionais. Contato: (85) 99826636.

Isaac Reinaldo Pinheiro de Lima
Natural de Natal, RN, É estudante de graduação do curso de Bacharelado em Tecnologia da Informação da Universidade Federal do Rio Grande do Norte-UFRN (2019- 2023). Ingressou através do Processo de Seleção ENEM-Exame Nacional de Ensino Médio e do programa de seleção Unificado-SISU do Ministério da Educação. Já foi selecionado para ingresso (onde cursou parcialmente) curso de Graduação de Sistemas de Informação na Universidade de São Paulo-USP, EACH. Atualmente, é Bolsista do programa PSH – Painel de Segurança Hídrica na Universidade Federal do Rio Grande do Norte, Além de Estagiário de tecnologia na Empresa VTEX, possui um livro publicado. Já Concluiu cursos de Robótica e programação Computacional, possui um artigo científico publicado em revista internacional com Qualis da Capes B1. Contato: (84) 98810 9278.

Joelma de Assis França, Natural de João Câmara, RN
Professora do ensino fundamental do estado do Rio Grande do Norte, e Da Prefeitura da Cidade de João Câmara , RN, Com atuação durante mais de vinte anos, É Graduada em Licenciatura em Ciências Biológicas Pela Universidade federal do Rio Grande do Norte-UFRN, EM Natal, RN, Atualmente é Mestranda do Programa de Pós Graduação CIÊNCIA DA EDUCAÇÃO na Universidade Federal do Rio Grande do Norte (UFRN) em Natal , RN, com Longa experiência na área de Docência, E Administrativa como Diretora de Escola, Possui um Livro Publicado Durante o Ano de 2023 Sobre Transformação de Dados Estatísticos. Contato: (84) 988539347.



Roberto Pequeno de Sousa

Engenheiro Agrícola, graduado em Engenharia Agrícola (1981) na Universidade Federal da Paraíba (UFPB). Mestre (1985) em Engenharia Civil (Recursos Hídricos - Irrigação) na Universidade Federal da Paraíba (UFPB). Doutor (2013) em Agronomia - Fitotecnia na Universidade Federal Rural do Semi-Árido (UFERSA). Atualmente, é Professor Titular da Universidade Federal Rural do Semi-Árido (UFERSA), leciona a disciplina de Estatística Experimental, possui sete livros publicados, 62 artigos completos publicados/aceitos em revistas nacionais e internacionais, 42 resumos simples/expandido. É revisor de cinco revistas nacionais e internacionais. Contato: (84)99411-5032.

Paulo César Ferreira Linhares

Engenheiro Agrônomo, graduado em Engenharia Agronômica (2002) na Escola Superior de Agricultura de Mossoró (ESAM). Mestre em Fitotecnia (2007) e Doutorado em Fitotecnia (2009) pela Universidade Federal Rural do Semi-Árido (UFERSA). Atualmente é Pesquisador na área de Produção Orgânica de Hortaliças da Universidade Federal Rural do Semi-Árido (UFERSA), possui quatro livros publicado, 112 artigos publicados em revistas nacionais e internacionais. 100 resumos simples/expandido. 32 orientações de trabalho de conclusão do curso de Agronomia. 22 orientações de Dissertação de Mestrado. 02 coorientações de Doutorado. 07 participações em bancas de dissertação de mestrado. 03 participações em tese de Doutorado. 25 participações em trabalhos de conclusão do curso de Agronomia. Pioneiro na região semiárida na utilização da jitirana como adubo verde na produção de hortaliças. Líder do grupo de pesquisa jitirana. Contato: paulolinhares@ufersa.edu.br

Robson Pequeno de Sousa

Doutor em Engenharia Elétrica pela Universidade Federal da Paraíba (2000), mestrado em Engenharia Elétrica pela universidade Federal de Pernambuco (1991) e graduação em Bacharelado em Matemática pela Universidade Federal da Paraíba (1985). Professor Doutor Associado D do Departamento de Computação da Universidade Estadual da Paraíba e membro efetivo do Programa de Pós-Graduação em Ciência e Tecnologia em Saúde – PPGCTS-UEPB. Coordena o Laboratório de Análises de Imagens e Sinais –LAIS do Núcleo de Tecnologias Estratégicas em Saúde – NUTES. Contatos: Fone: 083993129256.

Telde Natel Custódio

Natural de Lavras/MG. Possui graduação em Engenharia Agrícola pela Escola Superior de Agricultura de Lavras, atual Universidade Federal de Lavras (1988), mestrado em Agronomia com área de concentração em Estatística e Experimentação Agropecuária pela Universidade Federal de Lavras (1999), doutorado em Agronomia com área de concentração em Estatística e Experimentação Agronômica pela Escola Superior de Agricultura "Luiz de Queiroz" da Universidade de São Paulo (2004). Atualmente é professor na área de Estatística pela Universidade Federal de São João Del-Rei (UFSJ). Tem experiência docente na área de Probabilidade e Estatística. Atua principalmente nos seguintes temas: Estatística Experimental Superfície de resposta, Análise de regressão, Modelos lineares generalizados, Modelos lineares generalizados mistos, Meta-análise. Contato: natel@ufsj.edu.br. (35)99979-0218.

Walter Martins Rodrigues

Possui graduação em Licenciatura em Matemática pela USP (1987), onde também cursou o mestrado concluído em 2020 e doutorado em Matemática finalizado em 2005, na área de Representação de álgebras. Atualmente é professor Titular da Universidade Federal Rural do Semi-Árido (UFERSA). Exerceu a atividade de coordenador Pedagógico e pró-reitor adjunto de Graduação no período de agosto de 2013 a fevereiro de 2015. Coordenou o mestrado Profissional em Matemática da UFERSA de 2020 até março de 2023. Trabalha com pesquisa voltada para modelagem Matemática aplicada, Álgebra e Modelagem Estatística. Contato: Fone: 084988933633.

Pantanal Editora
Rua Abaete, 83, Sala B, Centro. CEP: 78690-000
Nova Xavantina – Mato Grosso – Brasil
Telefone (66) 99682-4165 (Whatsapp)
https://www.editorapantanal.com.br
contato@editorapantanal.com.br