CIÊNCIA e NATURA

ISSN 2179-460X
ACESSO ABERTO

Ci. e Nat., Santa Maria, v. 45, e9, 2023 • https://doi.org/10.5902/2179460X67509
Submissão: 02/09/2021 • Aprovação: 14/12/2022 • Publicação: 23/06/2023

Estatística

Avaliação da normalidade, validade dos testes de médias e opções não-paramétricas: contribuições para um debate necessário

Evaluation of normality, validity of mean tests and non-parametric options: contributions to a necessary debate

André Mundstock Xavier de Carvalho[I], **Éder Matsuo**[I], **Marcelo da Silva Maia**[I]

[I] Universidade Federal de Viçosa, MG, Brasil

RESUMO

A experimentação é uma importante base metodológica para as inovações no setor agrícola. Apesar disso, vários aspectos podem ainda ser aperfeiçoados nas análises estatísticas clássicas utilizadas nas pesquisas agrícolas. O objetivo desta revisão foi discutir alguns elementos conceituais e resultados de pesquisas sobre a validade de testes estatísticos usualmente aplicados na experimentação e apresentar algumas recomendações que podem melhorar a qualidade das análises comumente empregadas no âmbito dos modelos fixos. São apresentados elementos úteis para a discussão sobre os testes de médias, sobre a avaliação da condição de normalidade e sobre opções não-paramétricas de análise. O entendimento das hipóteses estatísticas e dos subtipos de erro tipo I, por exemplo, pode auxiliar numa melhor interpretação de resultados e na escolha do teste de médias. Algumas dúvidas sobre a avaliação do requisito de normalidade dos resíduos, aqui exploradas, também podem auxiliar pesquisadores num melhor uso das ferramentas estatísticas paramétricas. Por fim, apresenta-se um fluxograma de decisão geral e uma breve discussão exemplificada sobre algumas opções de análises não-paramétricas, com ênfase nas diferenças entre os métodos clássicos e os métodos baseados em modelos generalizados.

Palavras-chave: Pressuposições da ANOVA; Testes de comparação múltipla; GLMM

ABSTRACT

Experimentation is an important methodological basis for innovations in the agricultural sector. Nevertheless, several aspects can be improved in the classical statistical analysis used in agricultural research. The objective of this review was to discuss a few conceptual elements and research results about the validity of statistical tests usually applied in experimentation and present some recommendations that can improve the quality of the analyzes commonly used in the scope of fixed models. Useful elements for the discussion of tests of means, assessment of the condition of normality, and non-parametric analysis options are presented. Understanding the statistical hypotheses and Type I error subtypes, for example, can help in better result interpretation and choice of means test. Some doubts about the evaluation of the normality requirement of the residues explored here can also help researchers better use parametric statistical tools. Finally, we present a general decision flowchart and a brief exemplified discussion of some non-parametric analysis options with emphasis on the differences between classical methods and methods based on generalized models.

Keywords: ANOVA assumptions; Multiple comparison tests; GLzM

1 INTRODUÇÃO

Os métodos de análises estatísticas são um dos pilares para uma interpretação mais segura dos dados das pesquisas científicas e, consequentemente, para a inovação agrícola. A ciência estatística avança permanentemente e centenas de procedimentos já foram desenvolvidos ou aprimorados para atender as especificidades de cada área do conhecimento, ainda que muitos deles possuam aplicação ampla. Em cada área do conhecimento, no entanto, há um conjunto de métodos e procedimentos mais usualmente empregados (GOTELLI; ELLISON, 2011). É importante que os profissionais de cada área possuam um adequado entendimento sobre estes procedimentos para poder melhor participar das pesquisas e das discussões inerentes à ampliação dos conhecimentos científicos na sua área (LÚCIO; SARI, 2017). A expansão da cultura científica tanto entre os grupos profissionais especializados quanto na sociedade em geral é sensível ao entendimento destes procedimentos estatísticos usuais. Esta perspectiva vai ao encontro da necessidade de aprimoramento de métodos estatísticos relativamente simples para que possam ser rapidamente incorporados ao domínio dos pesquisadores não estatísticos (LITTLE, 2013).

O dinamismo da ciência estatística tem evidenciado que alguns procedimentos usualmente aplicados deveriam ter suas recomendações revisadas (CARVALHO *et al.*, 2021). No entanto, novas recomendações podem gerar desconforto quanto à validade de milhares de trabalhos previamente publicados apoiados em procedimentos cuja indicação foi revista. Além disso, a construção de novos consensos pode ser lenta devido a fatores como resistência, desatualização e conflitos de interesses. Se por um lado é natural que a ciência estatística avance e forneça novos métodos e entendimentos, por outro, deve-se compreender que, como uma ciência exata, os conhecimentos anteriormente gerados por ela são válidos, ao menos parcialmente, já que foram validados empiricamente. A física "moderna" exemplifica esta questão, sendo que o "novo", representado simplificadamente pela física quântica e pela relatividade, não invalidou o "velho", representado pela física Newtoniana. Estão apenas em domínios de validade distintos. Igualmente, na estatística os métodos clássicos de análise tendem a conviver com os métodos mais recentes. Não significa, no entanto, que todas as clássicas recomendações de métodos de análise permanecem inalteradas, assim como não significa que todas as técnicas mais recentes sempre superam os métodos estatísticos clássicos. Afinal, o que se busca com as análises estatísticas? Ferramentas objetivas para auxiliar na interpretação dos dados ou algo mais? Só há uma opção válida para a análise dos dados? O que seria mais essencial nas ferramentas estatísticas: métodos válidos ou modelos precisos?

Dessa forma, o objetivo desta revisão é discutir alguns aspectos conceituais e resultados de pesquisas sobre a validade de testes estatísticos usualmente aplicados na experimentação agrícola e apresentar algumas recomendações. Especificamente, são apresentados alguns elementos úteis para a discussão sobre os testes de médias, o requisito de normalidade e opções para análises não-paramétricas. Para tal, este artigo está estruturado em quatro temas principais organizados numa sequência lógica para facilitar o entendimento geral.

2 O DILEMA DOS ERROS TIPO I E II NA INFERÊNCIA ESTATÍSTICA

Na inferência estatística há dois tipos de erros principais, os erros tipo I (α) e tipo II (β). O erro tipo I corresponde ao resultado "falso positivo" (aceitar erroneamente H_1)

e o tipo II ao "falso negativo" (aceitar erroneamente H_0). O erro tipo II também pode ser entendido como "falta de poder", uma vez que erro β (%) = 100 – poder (%). Na maior parte dos casos, entende-se que o erro do tipo I é um erro mais grave, pois a aceitação de H_1 irá implicar em aceitar algo novo, aceitar que a mudança seria válida. Enquanto que o erro tipo II irá implicar em aceitar que o novo não difere (afinal H_0 é sempre baseada em um sinal de igualdade), mantendo as coisas como estão. Estabelecendo um paralelo com a área jurídica, o erro α seria equivalente a condenar um inocente e o erro β seria equivalente a absolver um culpado, que por falta de provas suficientes teve que ser absolvido (LOUREIRO; GAMEIRO, 2011). Por este motivo, os testes foram concebidos para terem erro α sob controle.

Para a maior parte das condições experimentais, onde não há condições do *n* ser exageradamente grande, os erros α e β são aproximadamente complementares. Significa que se tentarmos reduzir o erro α quase sempre iremos aumentar o erro β e vice-versa. E como os testes foram concebidos para controlar o erro α, infelizmente, o erro β será variável, podendo ser muito grande quando o coeficiente de variação (CV) é alto ou quando o *n* é pequeno ou quando as diferenças reais entre os tratamentos são de pequena magnitude em relação à magnitude do erro experimental. Nas condições avaliadas por Conagin e Pimentel-Gomes (2004), por exemplo, o teste de Tukey apresentou taxas de erro tipo II entre 27,2 e 97,7% para diferenças reais entre médias de 30 e 10% (r = 4), mesmo com valores de CV de apenas 10%. Nas situações simuladas por Borges & Ferreira (2003) o erro tipo II do teste Tukey, mesmo com diferenças reais correspondentes a 4 desvios padrão, foi superior a 40%. Em outras palavras, ao menos na perspectiva dos métodos Fisherianos clássicos (PATRIOTA, 2014), quando um teste não rejeita H_0 não há segurança em se afirmar que H_0 é verdadeira, pois quase sempre o erro β é relativamente grande.

Este é um dilema frequentemente negligenciado nas pesquisas agrícolas e impõe uma enorme restrição para concluir, com segurança, que os "tratamentos não diferem entre si". A variabilidade do erro tipo II evidencia que não-rejeitar H_0 é um resultado mais próximo de “inconclusivo” e que os testes, na verdade, só conseguem provar H_1. Afinal, mesmo um erro β de 10 ou 20% é grande o suficiente para questionarmos a não rejeição de H_0. Ou seja, na estatística também é válido lembrar que a ausência de evidências para um fenômeno não é sinônimo de evidência de ausência deste fenômeno. A própria maneira como comumente os resultados de um teste de médias são descritos já evidencia este descuido: “...médias seguidas por uma mesma letra não diferem entre si”. Não há provas de que as médias “não diferem entre si” pois o procedimento não testa se H_0 é verdadeira. Há apenas uma “falta de evidência suficiente para se afirmar que diferem”, já que o procedimento testa apenas se H_1 é verdadeira (ALVAREZ; ALVAREZ, 2013).

Compreender este dilema pode ajudar na redação das conclusões. Se o pesquisador almeja ser mais convincente terá que apoiar suas conclusões sobre as diferenças significativas encontradas e não sobre as não-diferenças. Além disso, esse dilema ajuda a compreender que exigir valores de CV abaixo de “x” ou “y” não aumenta a confiabilidade das conclusões quando estas estão baseadas em diferenças significativas. Exigir um CV baixo ou um índice de variação baixo ou um *n* alto somente faz sentido se a conclusão estiver apoiada em diferenças não-significativas. Evidentemente que experimentos bem conduzidos sob condições bem controladas tendem a ter CVs mais baixos que experimentos equivalentes malconduzidos ou com menos efeitos fixos. Mas o CV pode também estar em função de outros fatores como, por exemplo, a natureza dos preditores, a natureza das variáveis respostas, a presença de uma maior carga de efeitos aleatórios, a morfologia do organismo teste, o tamanho da unidade experimental, a interação entre os tratamentos e efeitos ambientais (KRAMER *et al.*, 2019; CARGNELUTTI FILHO *et al.*, 2020). Se mesmo com um CV alto as diferenças puderem ser evidenciadas com um teste confiável, num modelo estatístico apropriado e com seus requisitos cumpridos, a probabilidade de este resultado ser um falso positivo é a mesma que num experimento de CV baixo. Conceito equivalente aplica-se à exigência de um número mínimo de graus de liberdade (GL) para o resíduo na ANOVA. Evidentemente que quanto maior o GL para o resíduo maior tende a ser o poder dos testes estatísticos. No entanto, quando as conclusões são apoiadas nas diferenças encontradas, esta exigência não é importante (PIMENTEL-GOMES, 2009; DUTCOSKY, 2013).

3 AFINAL, EXISTE UM TESTE DE MÉDIAS MELHOR QUE OUTRO?

O conceito de erro tipo I, apesar de simples, pode ter desdobramentos mais complexos em função da forma como ele pode ser estimado. É bem aceito que o erro tipo I pode ser contabilizado ao menos de três maneiras: por comparação, por família e por experimento (GARCIA-MARQUES; AZEVEDO, 1995; KESELMAN, 2015). Considere, por exemplo, o teste Tukey aplicado em 100 experimentos com cinco tratamentos. Considere ainda que em cada experimento foram avaliadas 8 variáveis respostas. Neste caso, em cada variável resposta o teste será aplicado 10 vezes ($C_{5,2}$ = 10). Em cada experimento, o teste será aplicado, portanto, 80 vezes. A frequência de falsos positivos considerando o total de comparações duas-a-duas realizadas é a taxa de erro tipo I por comparação. A frequência de ao menos um falso positivo em cada variável resposta é a taxa de erro tipo I por família de comparações (ou *family wise error rate* – FWER). E a frequência de ao menos um falso positivo em um experimento (considerando as múltiplas inferências em todas as variáveis respostas) é a taxa de erro tipo I por experimento. Alguns autores consideram que este subtipo de erro tipo I deveria ser nomeado como *maximum family wise error rate* (MFWER) (BIRD; HADZI-PAVLOVIC, 2014). Em alguns casos, quando os estudos por simulação consideram apenas uma variável resposta por experimento, o erro tipo I por família coincide com o erro tipo I por experimento.

Em geral, a propriedade mais importante de um teste de médias é o seu controle do erro tipo I familiar (FWER), ou seja, a sua capacidade de não ultrapassar o valor nominal estipulado (geralmente 5%) nas mais diversas situações em que ele possa ser aplicado. A segunda propriedade mais importante de um teste de médias é o seu poder, ou capacidade de detectar diferenças reais, mesmo quando elas forem de pequena magnitude. Não cabe ao teste de médias decidir sobre a importância ou sobre a magnitude de uma diferença, função destinada às medidas de *effect size* (LOUREIRO; GAMEIRO, 2011) ou à uma análise econômica. Ao teste de médias cabe apenas apontar se a diferença é real.

Todos os testes de médias foram concebidos para controlar adequadamente as taxas de erro tipo I por comparação. No entanto, o controle das taxas de erro tipo I por família somente pôde ser verificado empiricamente após a consolidação dos métodos de Monte Carlo nas décadas de 1970 e 1980. Esta estratégia evidenciou que alguns testes usuais, como o teste Duncan e o teste LSD (teste *t*), não controlam adequadamente as taxas de erro tipo I por família. Nas condições avaliadas por Perecin e Barbosa (1988), por exemplo, as taxas de erro tipo I por família do teste LSD em experimentos com 5 tratamentos atingiu 25%. Significa dizer que em 25% das variáveis respostas analisadas havia ao menos um falso positivo entre as 10 comparações realizadas. Para 40 tratamentos, a taxa aumentou para 99,6% de falsos positivos por família. Ou seja, em quase todas as variáveis respostas ocorreu pelo menos um falso positivo entre as comparações realizadas em cada variável. É simples compreender que uma frequência tão alta de falsos positivos não permitirá conclusões seguras.

Dito isso, torna-se relativamente simples qualificar os testes de médias mais conhecidos. Os testes Tukey, Student-Newman-Keuls (SNK), Bonferroni, Holm e Dunnett possuem evidências empíricas de adequado controle do erro tipo I por família (PERECIN; BARBOSA, 1988; BORGES; FERREIRA, 2003; CONAGIN *et al*., 2008; GIRARDI *et al*., 2009; SOUSA *et al.*, 2012). No entanto, os resultados empíricos de controle de erro tipo I por família, sob nulidade parcial, são menos consistentes que sob nulidade total. Apesar disso, não há evidências consistentes de que os testes acima listados possuam estimativas de erro real muito acima de 10% quando sob alfa nominal de 5% e número não muito elevado de tratamentos (BORGES; FERREIRA, 2003; GONÇALVES *et al.*, 2015), algo que geralmente é bem tolerado. Nulidade parcial é a situação em que apenas parte dos tratamentos são simulados para terem diferenças reais entre si, situação relevante pela sua semelhança aos experimentos reais. O controle das taxas de erro tipo I por família, sob nulidade parcial, será mais fácil, como esperado, em situações com um menor número de médias a serem comparadas (geralmente $k < 10$).

O teste de agrupamento de Scott-Knott ainda desperta dúvidas sobre sua validade, pois acumularam-se evidências de que suas taxas de erro tipo I por família, sob nulidade parcial, são maiores que os demais testes citados acima. Frequentemente atingem 20% sob α nominal de 5% (Di RIENZO *et al.*, 2002; BORGES; FERREIRA, 2003; CONRADO *et al.*, 2017), uma frequência de erros realmente preocupante. Já os testes LSD (t para comparações múltiplas) e Duncan possuem grande volume de evidências de que não controlam adequadamente os falsos positivos por família (PERECIN; BARBOSA, 1988; CONAGIN *et al.*, 2008; GIRARDI *et al.*, 2009; SOUSA *et al.*, 2012), motivo pelo qual deveriam estar em desuso.

Quanto ao poder dos testes mais conhecidos, pode-se ordená-los em: Holm (para poucas comparações planejadas) > Dunnett > Scott-Knott > Bonferroni modificado por Conagin ~ SNK > Tukey. Esta distinção ajuda a compreender porque as comparações planejadas deveriam ser a primeira opção quando atendem aos objetivos da pesquisa. Também ajuda a compreender que geralmente há opções melhores que o teste Tukey, pois quase sempre se deseja testes poderosos. O teste de Bonferroni modificado por Conagin é o menos conhecido da lista e é interessante quando se precisa testar alguns poucos contrastes. Ele pode ser entendido como um procedimento de abordagem Bayesiana, pois utiliza o valor de F para tratamentos como uma informação prévia para definir um valor variável de diferença mínima significativa (CONAGIN, 2001).

Apesar do controle do erro tipo I por família de muitos testes univariados, nenhum deles controla as taxas de erro tipo I por experimento (mais precisamente o erro tipo I por inferência múltipla ou MFWER). E quanto mais variáveis respostas um experimento tem, maior a chance de ocorrer um falso positivo em alguma delas (GARCIA-MARQUES; AZEVEDO, 1995; KRAMER *et al.*, 2019). As taxas de erro tipo I por experimento para p variáveis independentes podem ser estimadas por $\alpha_{total}=1-(1-\alpha)^p$ (MANLY, 1995). Para p=6, por exemplo, o erro tipo I por experimento pode atingir, teoricamente, até 26,5%, sendo um pouco inferior para variáveis correlacionadas entre si. Portanto, este é um problema muito frequente, mas que pode ser parcialmente minimizado com algumas estratégias relativamente simples. Considerando que dificilmente um falso positivo ocorrerá, por acaso, no mesmo tratamento em duas variáveis respostas, uma opção é sempre avaliar duas ou mais variáveis altamente correlacionadas entre si. Se uma diferença significativa aparecer em apenas uma das variáveis atribui-se menor confiança nesta diferença, uma vez que pode ser um caso de erro tipo I por experimento. Esta estratégia pode implicar em aumento de custos, já que aumenta o número de variáveis a serem mensuradas, e também pode implicar em perda de poder, uma vez que diferenças reais isoladas podem ser desacreditadas.

Outra estratégia utilizada para aumentar a confiança nas diferenças encontradas e assim reduzir o risco de erro tipo I por experimento é exigir que os experimentos sejam integralmente repetidos. Geralmente também se justifica esta repetição, em experimentos de campo, para verificar quão extrapoláveis os resultados seriam em condições edafoclimáticas distintas, entre outros motivos (KRAMER *et al.*, 2019). Embora seja útil para superar limitações dos experimentos de modelos fixos, é uma estratégia cara e traz como consequências o desperdício de um grande volume de bons experimentos que, por razões diversas, não puderam ser repetidos. Além disso, exigindo-se condições edafoclimáticas distintas os efeitos fixos alteram-se e, consequentemente, o desempenho dos tratamentos também pode ser alterado. Embora esta estratégia possa enriquecer as discussões acerca dos resultados, se as alterações forem grandes não se poderá avaliar se algumas diferenças isoladas poderiam estar associadas a erros tipo I por experimento.

Por fim, uma terceira estratégia para minimizar as taxas de erro tipo I por experimento é utilizar procedimentos multivariados como a MANOVA (MANLY, 1995; COUTO *et al.*, 2020), índices baseados em Análise de Componentes Principais (ZHIYUAN *et al.*, 2011; PRIMPAS *et al.*, 2010) ou índices de seleção como o índice de Mulamba-

Mock ou o índice *Desirability* (CANDIOTI *et al.*, 2014). Embora bem conhecidos, estas opções são raramente utilizadas para esta finalidade, em parte porque o poder destes procedimentos é limitado. Com um poder limitado, estes procedimentos mais frequentemente invalidam uma diferença real isolada encontrada nos testes univariados do que conseguem validá-la. Apesar disso, provavelmente são as ferramentas mais adequadas e de baixo custo para reduzir as taxas de erro tipo I por experimento.

4 REQUISITOS DA ANOVA COM ÊNFASE NA NORMALIDADE

A análise de variância (ANOVA) é um dos procedimentos mais usuais na pesquisa agrícola. No entanto, a frequência de pesquisas que ignora os requisitos dos diferentes modelos clássicos de ANOVA é grande (TAVARES *et al.*, 2016; POSSATTO JÚNIOR *et al.*, 2019). Primeiramente deve-se recordar que a ANOVA tradicional, mesmo sendo conhecidamente robusta à violação de normalidade (SCHMIDER *et al.*, 2010), apoia-se na premissa de que a média caracteriza adequadamente a amostra. Esta premissa é válida apenas para as distribuições simétricas. Portanto, métodos não-paramétricos, geralmente baseados em postos e medianas, podem permitir inferências mais precisas quando há evidências claras de que a distribuição é assimétrica. Além disso, os resultados de Borges e Ferreira (2003) e alguns outros autores (SAWILOWSKY; BLAIR, 1992; CARVALHO, 2023) evidenciam que, para que as conclusões dos testes posteriores à ANOVA sejam válidas, os requisitos devem ser verificados e não simplesmente assumidos, pois a depender do tipo de distribuição não-normal, as taxas de erro tipo I podem ser inflacionadas. Apesar disso, alguns requisitos não são tão simples de serem avaliados. A questão da independência dos resíduos, por exemplo, segue sendo de avaliação complexa, pois os testes de não-independência (como o Durbin-Watson) não conseguem detectar todos os diferentes padrões de não-independência possíveis. Por este motivo, é comum se assumir a existência deste requisito pela simples inspeção do croqui experimental ou pela verificação da adequabilidade do modelo estatístico escolhido.

A questão da aditividade dos efeitos de tratamentos e blocos no modelo estatístico tem sido frequentemente negligenciada. Geralmente os experimentos em blocos são planejados para terem uma repetição por bloco, não sendo possível decompor a possível existência de interação entre blocos e tratamentos na própria ANOVA. Se blocos e tratamentos interagem (efeitos multiplicativos) não é possível distinguir corretamente seus efeitos, gerando estimativas errôneas para os componentes da ANOVA ou estimativas absurdas para dados perdidos (CARVALHO *et al.*, 2023a). Portanto, um teste de não-aditividade é imprescindível nos delineamentos em blocos (DBCs) com apenas uma repetição por bloco, sendo uma ferramenta simples que confere confiabilidade à ANOVA (NUNES, 1998). A questão da homocedasticidade ainda alimenta dúvidas quanto ao teste mais adequado para cada situação, especialmente pela grande permissividade do teste de Levene modificado por Brown-Forsythe, cujo uso é crescente (HINES; O'HARA-HINES, 2000; TAVARES *et al.*, 2016).

Por fim, a questão da normalidade é a que apresenta dúvidas mais frequentes, razão pela qual será aqui enfatizada. Dentre estas dúvidas, quatro delas serão abordadas neste artigo. Primeiramente deve-se recordar que a normalidade é um padrão de distribuição dos resíduos e não dos dados brutos. Significa dizer que os resíduos precisam ser calculados conforme o modelo estatístico previamente definido no planejamento experimental. No modelo inteiramente casualizado (DIC) simples ($y_{ij} = \mu + ti + e_{ij}$) ou no modelo em DIC sob esquema fatorial ($y_{ijk} = \mu + a_i + b_j + a_ib_j + e_{ijk}$) o cálculo dos resíduos é fácil e coincide com a simples subtração entre o valor da observação e o valor da média do tratamento em questão. Para modelos mais complexos, como DBC, parcelas subdivididas, faixas, etc., o cálculo dos resíduos torna-se mais complexo e alguns softwares não o fazem de maneira simples ou automática.

Em segundo lugar deve-se recordar que os testes de normalidade são, na realidade, testes de não-normalidade, pois H_1 = distribuição ≠ da normal. Ou seja, eles testam a hipótese dos resíduos em questão não se ajustarem à distribuição normal. O erro α, neste caso, corresponde à probabilidade de concluir que os resíduos não são normais quando, na realidade, eles são normais ou aproximadamente normais. Note que, diferente do teste F ou dos testes de médias, neste caso cometer erro α não é algo tão grave. Afinal, é mais grave analisar dados não-normais com testes paramétricos do que analisar dados normais com testes não-paramétricos. Note também que nos testes de normalidade, quando $p > 0,05$, a conclusão de que os dados possuem erros com distribuição normal está apoiada em não-rejeição de H_0. Como vimos anteriormente, a não-rejeição de H_0 pode estar carregada de erro tipo II. Uma não-rejeição de H_0 não pode ser interpretada seguramente como evidência de que H_0 esteja correta. Apesar disso, os testes de normalidade permitem uma avaliação suficiente de fortes violações da normalidade, o que é corroborado pelo adequado controle do erro tipo I real do teste F quando precedido por bom teste de normalidade. Mesmo assim, para aumentar a confiança nas conclusões de um teste de normalidade deve-se assegurar que as condições permitem um bom poder do teste. Segundo Torman et al. (2012), os testes de normalidade (exceto o Kolmogorov-Smirnov, que deveria estar em desuso) somente possuem poder elevado quando o n total é superior a ~30. Com $n < 30$, portanto, seria mais prudente tolerar um pouco mais de erro α para conseguirmos mais poder para o teste (afinal, quando se tolera um erro α maior o erro β diminui). Por este motivo, uma opção para evitar que dados não-normais sejam erroneamente analisados como normais é considerar um *p-valor* de 0,25 como valor crítico. Assim, um $p > 0,25$ indicaria forte evidência de normalidade, pois este α nominal aumentaria bastante o poder do teste em detectar a não-normalidade. A significância de 0,25 é aqui sugerida por analogia à recomendação de Perecin e Cargnelutti Filho (2008) para uma inferência mais segura sobre a não-significância da interação em experimentos fatoriais. Neste mesmo raciocínio, quando o n total é menor que 15 unidades experimentais devemos ter em mente que nenhum teste de normalidade é poderoso o suficiente, situação em que devemos sempre desconsiderar a realização de testes paramétricos.

Em terceiro lugar pode-se observar que, nas condições experimentais com repetições verdadeiras, não é usual verificar a normalidade dos resíduos das regressões que modelam as diferenças entre as médias dos tratamentos (BANZATTO; KRONKA, 2006; PIMENTEL-GOMES, 2009; STORCK *et al*., 2016). Embora este fato possa ser encarado como mau uso das ferramentas estatísticas paramétricas, há algumas razões práticas a serem consideradas. Como é bem conhecido, os modelos de ANOVA podem ser entendidos como modelos lineares de regressão, especificamente de regressão linear múltipla (MONTGOMERY, 2017). As ANOVAs de um experimento, portanto, mesmo nas situações com preditores apenas categóricos, envolvem sempre uma "regressão" pré-determinada, cujos resíduos precisam atender certos requisitos.

No entanto, simplificadamente pode-se entender que em um experimento com níveis quantitativos há duas regressões em questão: a regressão correspondente ao modelo estatístico previamente determinado pelo delineamento experimental e a regressão correspondente à modelagem da relação contínua entre as médias da variável resposta (Y) e os níveis quantitativos da variável preditora (X, não aleatório). Evidentemente, também é possível considerá-los em um único modelo. Nos estudos observacionais ou nos estudos por correlação, por outro lado, frequentemente não há um modelo estatístico pré-definido e, por isso, o modelo linear multivariado ajustado para os dados é o próprio modelo estatístico (GOTELLI; ELLISON, 2011), não sendo, neste caso, pré-determinado.

Nas condições experimentais, as regressões que modelam o comportamento das médias dos tratamentos podem ser estatisticamente testadas na ANOVA da regressão. Esta ANOVA permite obter uma segunda estimativa para a variância residual, o que permite usar a própria razão F para testar se os desvios da regressão são maiores que o desvio padrão experimental (CECON *et al.*, 2012). Não é estritamente necessário, portanto, construir intervalos de confiança para os regressores, tampouco testá-los pelo teste *t* paramétrico, uma vez que seria redundante em relação à ANOVA da regressão (NUNES, 1998; DANCEY, 2017). Note que para esta segunda regressão, já foi verificado se as médias são provenientes de dados com resíduos normais e variâncias homogêneas. E como há uma estimativa segura para a magnitude do desvio padrão experimental, há como inferir os intervalos mais prováveis nos quais as verdadeiras médias se encontram.

Portanto, se o modelo testado é o mais parcimonioso possível, teve seus parâmetros corretamente estimados para minimizar os desvios, possui um bom ajuste (R^2 ajustado elevado e menor critério de informação de Akaike corrigido - AICc), explica uma fração estatisticamente significativa da soma de quadrados de tratamentos e possui uma falta de ajuste não-significativa, dificilmente este modelo não será válido para a comparação empírica das médias dos tratamentos (PIEPHO e EDMONDSON, 2018). A avaliação dos resíduos da regressão, nesses casos, será de importância secundária. Além disso, dificilmente um modelo selecionado dessa forma terá desvios de regressão com grande assimetria quando o método utilizado (mínimos quadrados, Gauss-Newton ou outro) tenha convergido adequadamente, já que os desvios da regressão tendem a ser semelhantes aos desvios do modelo estatístico (SNEDECOR; COCHRAN, 1989). Deve-se ter em mente também que num fatorial duplo com 5 níveis qualitativos e 4 níveis quantitativos com 3 repetições, por exemplo, o *n* para os resíduos preditos para cada um dos modelos de regressão é de apenas 12, um valor insuficiente para uma avaliação segura da condição de normalidade.

Por fim, em quarto lugar, deve-se recordar que o padrão de distribuição dos resíduos do modelo estatístico para dados discretos não segue, em teoria, a distribuição normal. No entanto, este ponto é um dilema, pois se por um lado a distribuição Gaussiana é uma distribuição contínua, por outro o conceito de variável contínua é relativo. Afinal, o caráter contínuo inexiste nos dados reais devido à restrição de sensibilidade dos instrumentos (GOTELLI; ELLISON, 2011). Os instrumentos não permitem obter variáveis com infinitas possibilidades de números reais. Quando se mensura a massa (que é uma variável contínua) de pequenos insetos com uma balança de sensibilidade de 5 g, por exemplo, obtém-se dados com características de discretos. Numa análise de pH do solo, por exemplo, não há infinitas possibilidades de números reais para cada observação. Há apenas ~400 possibilidades, já que os valores possíveis estão quase sempre apenas entre 4,00 e 8,00 e os equipamentos para esta função tem sensibilidade de apenas 0,01. Por outro lado, quando se avalia o número de sementes de sorgo (uma variável discreta) em uma unidade experimental com 20 plantas úteis, obtém-se contagens tão volumosas que estes dados poderão ter características muitos semelhantes a dados contínuos.

A relatividade do conceito de contínuo não implica, obviamente, que todos os dados de contagem possam ser tratados, sem maiores cuidados, com procedimentos paramétricos de análise. Sugere apenas que a decisão talvez possa ser mais pautada no que os dados sugerem, e menos no que eles são em teoria. Evidentemente há riscos nessa abordagem, principalmente quando os dados têm pouco a dizer por estarem em pequeno número. Por este motivo, alguns economistas consideram que dados de contagem possam ser tratados como contínuos quando se tratar de contagens volumosas (MANN, 2015). É difícil definir um volume crítico, mas, na prática de análise de dados agronômicos, é razoável não considerar contagens menores que 30 possibilidades como minimamente volumosas. E mesmo em contagens volumosas, considerar como crítico para o teste de normalidade um *p-valor* > 0,25 é um procedimento um pouco mais cauteloso, que evita que dados com alguma suspeita de não-normalidade sejam analisados erroneamente com procedimentos paramétricos.

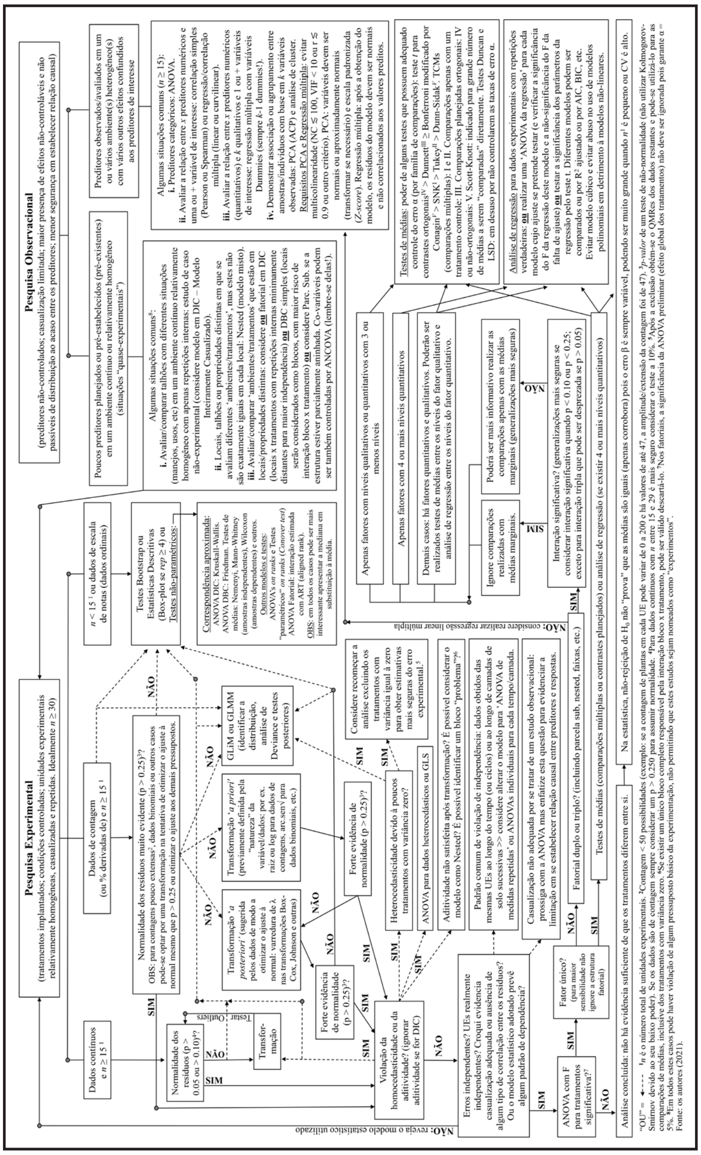
Para auxiliar no entendimento e adoção das sugestões aqui apresentadas, e outras adicionais, elaborou-se um fluxograma explicativo para os principais procedimentos da estatística experimental (para modelos fixos) e observacional quantitativa (Figura 1). Nele, as sugestões para tomada de decisão sobre quais procedimentos utilizar consideram, sempre que possível, mais de uma opção válida de análise. No caso de variáveis discretas, por exemplo, o fluxograma permite chegar à recomendação de uso de modelos lineares generalizados mistos (GLMM), testes não paramétricos, transformações a posteriori para otimizar o ajuste à normal ou mesmo uma clássica ANOVA paramétrica quando a contagem for volumosa e os dados apresentarem uma distribuição de erros próxima à normal (teste de normalidade com $p > 0,25$). Se, por um lado, o fluxograma da Figura 1 acrescenta um pouco de complexidade em relação às recomendações usualmente encontradas nos livros textos, por outro, resulta em algum avanço na qualidade e confiabilidade dos resultados.

5 ALGUMAS OPÇÕES DE ANÁLISE PARA DADOS COM RESÍDUOS NÃO-NORMAIS

Aparentemente a não verificação da normalidade em muitos trabalhos também está relacionada à dificuldade em encontrar uma opção adequada de análise quando a não-normalidade é detectada. Em alguns casos, as clássicas transformações de escala não permitem um bom ajuste à normal e os métodos não-paramétricos apresentam-se, para boa parte dos pesquisadores, como um grande conjunto de procedimentos com os quais eles não estão familiarizados. Vale lembrar também que alguns procedimentos paramétricos comuns na experimentação agrícola não possuem um equivalente não-paramétrico ou este não é acessível, ou amplamente divulgado.

Em sua quase totalidade, os métodos não-paramétricos clássicos, como Kruskal-Wallis, Friedman, Mann-Whitney, Spearman, etc., tem como primeiro passo a transformação para uma escala de postos (*rank transformation*). Esta é uma etapa chave, mas nesta escala apenas a ordem dos dados é preservada. As distâncias ou discrepâncias entre os valores são perdidas e isso geralmente implica em redução de poder destes procedimentos em comparação aos seus equivalentes procedimentos paramétricos. E esta é a razão principal pela qual quase sempre é optado primeiramente pelos métodos paramétricos.

Figura 1– Fluxograma para seleção dos principais métodos estatísticos empregados na pesquisa experimental e observacional agrícola



Fonte: Autores (2023)

De certa forma, graças às contribuições de Fisher, os modelos estatísticos mais usuais na experimentação agrícola estão todos sumarizados em modelos de ANOVAs paramétricas, com semelhanças diversas nas estimativas dos seus componentes. Algo semelhante não foi desenvolvido para unificar ou simplificar o grande e complexo conjunto de procedimentos não-paramétricos. Por este motivo, e baseando-se em evidências empíricas de estudos por simulação, o estatístico William Jay Conover propôs uma abordagem unificadora (CONOVER e IMAN, 1981; CONOVER, 1999). Ele observou que a aplicação dos procedimentos paramétricos clássicos em dados previamente submetidos à transformação rank (RT) permitia uma análise não-paramétrica válida. Aliado ao aprimoramento da técnica, como a *rank transformation type 2* (RT-2) e a aligned *rank transformation* (ART) (SALTER; FAWCET, 1993; WOBBROCK *et al.*, 2011), este novo conjunto de procedimentos permitiu também uma opção relativamente simples de análises não-paramétricas para desenhos em blocos, experimentos fatoriais, parcelas subdivididas e medidas repetidas. Gradativamente a resistência a estes procedimentos, também conhecidos como Conover tests, vem diminuindo (ZIMMERMANN, 2004; CONOVER, 2012; ZIMMERMAN, 2012; MONTGOMERY, 2017).

Por fim, aproximadamente a partir do ano 2000 uma nova abordagem para dados não-normais foi consolidada. Os modelos lineares generalizados (GLzM ou GLiM) e os modelos lineares generalizados mistos (GLMM) estenderam a teoria do modelo linear da ANOVA e abriram opções poderosas para análise de dados não-normais (STROUP, 2015). Simplificadamente, numa análise através de GLMM, os dados sugerem à qual distribuição de resíduos eles melhor se ajustam e a análise é realizada considerando esta distribuição, ou a que melhor se ajustar entre as que estiverem disponíveis no pacote ou software utilizado. Apesar da sofisticação, tal como nos métodos clássicos, os procedimentos posteriores baseados em GLMM igualmente não conseguem provar H0 em função da sempre variável taxa de erro tipo II. Igualmente, eles também não se apresentam como solução para o problema das elevadas taxas de erro tipo I por experimento (MFWER).

Os resultados das análises apresentadas nas Tabelas 1, 2 e 3 exemplificam que nem sempre a abordagem via GLMM resultará em testes mais poderosos que outras opções válidas de análises, ainda que o GLMM possua um domínio de validade menos restrito. Na Tabela 1, a análise através da ANOVA paramétrica usual poderia representar em risco aumentado de erro tipo I, uma vez que seus requisitos não podem ser negligenciados. Os dados em questão, apresentados por Stroup (2015), nem sempre satisfazem adequadamente os requisitos de normalidade (como era esperado por serem dados de contagem/proporções) ou de aditividade (ainda que a existência de apenas dois tratamentos dificulte uma melhor avaliação deste requisito). Os “dados 1”, por exemplo, apresentaram não-aditividade significativa pelo teste de aditividade de Tukey. Tal fato evidencia que os dados podem ter “violações” mais complexas do que apenas não-normalidade. Além disso, as transformações comumente recomendadas para dados desta natureza (raiz, log ou arcoseno da raiz) nem sempre atenderam adequadamente os requisitos simultaneamente (a saber, normalidade, homocedasticidade e aditividade). Estas situações, portanto, evidenciam a necessidade de se considerar a abordagem por transformações definidas *“a posteriori”*, abordagem em que a transformação ideal será escolhida de modo a otimizar o ajuste à normal e também atender aos demais pressupostos.

Vale lembrar que mesmo conhecendo a origem dos dados nem sempre é possível determinar qual será a transformação apropriada, devendo-se tentar outros tipos de transformação (NUNES, 1998). Este entendimento corrobora com a ideia de que, tal como no GLMM, as transformações clássicas também podem ser sugeridas pelos próprios dados e não necessariamente definidas *a priori* pela natureza teórica destes dados.

A análise dos “dados 1” disponíveis em Stroup (2015) por diferentes ferramentas não-paramétricas mostra que, neste caso, apenas a abordagem por GLMM permitiu evidenciar uma diferença possivelmente real (Tabela 1). No caso dos “dados 2”, no entanto, quatro abordagens não-paramétricas permitiram evidenciar a diferença. E, por fim, no caso dos “dados 3” apenas a abordagem por GLMM não teve poder suficiente para evidenciar as diferenças entre os tratamentos (Tabela 1). Vale lembrar que procedimentos como r*ank transformation* (RT), teste de Friedman e *bootstrap* Tukey possuem ampla evidência de adequado controle de suas taxas empíricas de erro tipo I por família (ZIMMERMAN, 2012; FERREIRA, 2014; CARVALHO *et al.*, 2023b). Estes dados, portanto, evidenciam que nem sempre a abordagem por GLMM será a única ou a melhor opção para dados não-normais. Identificar o tipo exato de distribuição dos erros será sempre desafiador para dados com n pequeno, ainda mais diante do problema da continuidade entre as dezenas de distribuições possivelmente existentes na natureza.

Tabela 1. Estimativa do erro tipo I nominal (*p-valor*) de diferentes procedimentos de análise estatística para três conjuntos de dados de contagem/proporções (dois tratamentos em oito blocos, disponíveis em Stroup (2015)

Dados 1	**ANOVA**	**RT1**	**RT-22**	**Friedman3**	**Bootstrap4**	**Raiz5**	**GLMM BN**
p-valor	**0,098**	**0,094**	**0,170**	**0,157**	**>0,05**	**0,081**	**0,032**
Dados 2	ANOVA	RT[1]	RT-2[2]	Friedman[3]	Bootstrap[4]	arcsen(y)^{0,5}	GLMM Binom
p-valor	0,113	0,014	0,020	0,034	>0,05	0,060	0,036
Dados 3	ANOVA	RT[1]	Johnson[6]	Friedman[3]	Bootstrap[4]	arcsen(y)^{0,5}	GLMM Beta
p-valor	0,016	0,004	0,004	0,005	<0,05	0,028	0,092

ANOVA corresponde ao teste F paramétrico inadequadamente aplicado negligenciando a verificação da condição de normalidade. [1]Transformação rank seguida de ANOVA. [2]*Rank in blocks* previamente à ANOVA. [3]Teste não-paramétrico de Friedman. [4]Bootstrap Tukey a 5% após 2000 reamostragens. [5]Transformações clássicas de raiz quadrada e arcoseno da raiz de y ou de y/100. Transformação simplificada de Johnson-Sb: $\ln[(y+1)/(\lambda-y)]-\ln(1/\lambda)$, com λ= 1.01. GLMM conforme ajuste às distribuições: binomial negativa (BN), binomial e beta (cálculos de Stroup, 2015)

Fonte: colunas [1, 2, 3, 4 e 5] cálculos dos autores e demais informações Stroup (2015)

A análise dos dados de número de pulgões por planta, apresentados por Carvalho (2019), por diferentes métodos não-paramétricos também evidenciou que podem existir diferenças entre eles (Tabela 2). Se for considerado que a distribuição ZIBN esteja mais próxima da natureza real dos dados, as conclusões obtidas são pouco distintas daquelas que seriam obtidas com a abordagem via transformação *rank* e teste SNK *on ranks*. Se a distribuição Poisson for, de fato, a mais adequada aos dados, a abordagem via GLMM apresenta-se como mais poderosa que as demais neste caso (Tabela 2). É importante, no entanto, não interpretar diferenças de poder entre os procedimentos como evidência de incoerência grave ou contradição grave, afinal o poder quase sempre é limitado nos procedimentos estatísticos. Como visto anteriormente, seria um erro conceitual afirmar que quando um teste "acusa" uma diferença significativa e outro teste "não-acusa" estamos diante de uma incompatibilidade ou contradição.

Afinal, o teste que não acusou a diferença não provou H_0. Se ambos controlam adequadamente o erro tipo I por comparação e por família, ambos são testes válidos, ainda que com diferenças de poder.

Tabela 2 – Número médio (após transformação) de pulgões por planta (dados de Carvalho, 2019) após a aplicação de diferentes produtos (Tratamentos 1 a 4) ou sob nenhuma aplicação (Tratamento 5) submetido à quatro procedimentos de análise estatística

	Medidas	Métodos Classicos				GLMM ou GLiM			
	Originais	RT (HSD)[1]		RT (SNK)[2]		Poisson[3]		ZIBN[4]	
Trat. 1	16,5	26,5	ab	26,5	b	23,4	c	42,4	b
Trat. 2	3,0	16,2	b	16,2	c	2,5	d	4,4	c
Trat. 3	3,5	15,5	b	15,5	c	2,4	d	4,6	c
Trat. 4	48,5	30,8	a	30,8	ab	32,7	b	67,6	b
Trat. 5	592,5	38,6	a	38,6	a	380,3	a	557,0	a

[1]Dados de postos (*rank transformation*) submetidos à ANOVA e ao teste Tukey a 5%. [2]Dados de postos submetidos à ANOVA e ao teste de SNK a 5%. [3]GLMM considerando distribuição Poisson com função de ligação log, ANODEV e teste Tukey a 5% (cálculos apresentados por Carvalho, 2019). [4]GLMM considerando distribuição Binomial negativa com inflação em zeros (ZIBN) com função de ligação log, ANODEV e teste Tukey a 5% (cálculos apresentados por Carvalho, 2019). Não há evidência suficiente de que médias seguidas por uma mesma letra, na coluna, diferem estatisticamente entre si pelos diferentes procedimentos.
Fonte: [1, 2] Cálculos dos autores e demais informações Carvalho (2019)

A análise dos dados de percentagem de plantas controladas por doses de um herbicida apresentados por Carvalho (2019) também evidenciou que nem sempre a abordagem via GLMM será mais poderosa que outras opções não-paramétricas ou via transformação (Tabela 3). Neste caso, a transformação também foi sugerida pelos dados, definida, portanto, *a posteriori* através da varredura do λ que permitiu a melhor aproximação à distribuição normal. Igual procedimento de varredura também pode ser realizado na conhecida transformação de Box-Cox (BOX; COX, 1964).

Tabela 3 – Diferentes opções de análise estatística de dados de percentagem de plantas controladas por doses de um herbicida (fatorial 3x7 em DBC com quatro repetições, dados de Carvalho, 2019)

Opções de análise:		**Doses (mg L^{-1})**						
		0	**5**	**10**	**15**	**20**	**25**	**30**
Escala original[1]	Espécie 1	0,0 a	10,0 c	10,6 b	24,4 c	43,1 c	71,9 b	99,4 a
p-valor JB:	Espécie 2	0,0 a	31,3 b	50,6 a	70,0 a	99,4 a	98,1 a	99,4 a
0,112	Espécie 3	0,0 a	37,5 a	49,4 a	63,1 b	75,0 b	99,4 a	100,0 a
RT[2]	Espécie 1	6,5 a	16,4 b	16,6 b	23,0 b	34,3 c	53,6 b	73,0 a
(transf. não-	Espécie 2	6,5 a	26,6 a	41,1 a	52,4 a	73,0 a	69,3 a	70,8 a
paramétrica)	Espécie 3	6,5 a	30,5 a	39,5 a	47,8 a	56,3 b	73,0 a	76,0 a
Johnson Sb[3]	Espécie 1	0,0 a	2,5 b	2,5 b	3,5 b	4,3 c	5,5 b	7,6 a
p-valor JB:	Espécie 2	0,0 a	3,8 a	4,6 a	5,4 a	7,6 a	7,4 a	7,6 a
0,891	Espécie 3	0,0 a	4,1 a	4,6 a	5,1 a	5,6 b	7,6 a	7,7 a
GLMM[4]	Espécie 1	0,0 a	10,0 b	10,6 b	24,4 b	43,1 c	71,9 b	99,4 a
Binomial	Espécie 2	0,0 a	31,3 a	50,6 a	70,0 a	99,4 a	98,1 a	99,4 a
(logit)	Espécie 3	0,0 a	37,5 a	49,4 a	63,1 a	75,0 b	99,4 a	100,0 a

[1]Dados na escala original apresentavam suspeita de não-normalidade (*p-valor* do teste de normalidade de Jarque-Bera < 0,250). [2]Dados de postos (*rank transformation*) submetidos à ANOVA e ao teste de Tukey a 5%. [3]Transformação simplificada de Johnson-Sb: $\ln[(y+1)/(\lambda-y)]-\ln(1/\lambda)$, com $\lambda = 105$. [4]GLMM considerando distribuição Binomial com função de ligação logit, ANODEV e teste de Tukey a 5% (cálculos apresentados por Carvalho, 2019). Comparações entre as doses deverão ser realizadas por análise de regressão (não apresentadas). Não há evidência suficiente de que médias seguidas por uma mesma letra (na coluna, para cada opção de análise) diferem entre si pelo teste de Tukey a 5% de probabilidade.

Fonte: [1, 2, 3] Cálculos dos autores e demais informações Carvalho (2019)

6 CONSIDERAÇÕES FINAIS

Se por um lado, métodos mais complexos ou mais recentes podem não estar sendo utilizados por repúdio, desatualização e insegurança de pesquisadores fora do âmbito da estatística, por outro é possível que tais métodos sejam, em alguns casos, supervalorizados para promover a atuação de alguns profissionais hiperespecializados ou supervalorizados por não se compreender o domínio de validade de cada método. O mais importante, portanto, é esclarecer as situações em que, de fato, justifica-se a adoção de métodos analíticos mais complexos em detrimento às opções simples válidas.

A ciência estatística forneceu uma grande diversidade de métodos que, por vezes, permitem diversas formas de se analisar um mesmo conjunto de dados. Ocasionalmente alguns métodos são descartados, principalmente quando se acumulam evidências de descontrole das taxas de erro tipo I por família. Evidentemente, em alguns casos, as diferenças de poder entre dois métodos de análise podem ser grandes. Em outros casos, no entanto, as diferenças de poder podem ser pequenas, sendo natural optar-se pelo método válido mais simples ou mais abrangente, uma vez que este exigirá menor tempo para que cada pesquisador domine a técnica. Não se deve confundir a qualidade de uma pesquisa com a sofisticação dos métodos. Em um contexto em que a pesquisa agrícola ainda enfrenta graves problemas básicos na qualidade das análises, parece sensato investir tempo e esforço na ampliação do domínio sobre as técnicas clássicas básicas. Evidentemente, tal sugestão não significa desencorajar a adoção de técnicas mais recentes e complexas quando estas são necessárias, nem desencorajar a ampliação de grupos de pesquisa multidisciplinares.

Portanto, é preciso manter-se atento às situações em que a parcimônia metodológica e científica é comprometida desnecessariamente, por vezes limitando a compreensão dos dados ou até restringindo um processo mais amplo de revisão por pares dos artigos científicos. Por fim, vale considerar que a formação ampla e multidisciplinar do pesquisador é cada vez mais importante para enfrentar adequadamente os desafios e demandas das complexidades do rural brasileiro (GONÇALVES JÚNIOR; CORRÊA, 2017). E, nesse contexto, a objetividade e parcimônia científica podem auxiliar o pesquisador a encontrar tempo para acessar melhor outras áreas de interface, necessárias para uma definição mais acertada de seus objetivos e metas de pesquisa.

AGRADECIMENTOS

Aos pesquisadores Marlon Corrêa Pereira e Fabrícia Queiroz Mendes pelas sugestões. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

REFERÊNCIAS

ALVAREZ, V. H.; ALVAREZ, G. A. M. Reflexões sobre a utilização de estatística para pesquisa em ciência do solo. **Boletim Informativo da Sociedade Brasileira de Ciência do Solo**, v. 38, n. 1, 28-35, 2013.

BANZATTO, D. A.; KRONKA, S. N. **Experimentação agrícola**. 4. ed. Jaboticabal: Funep, 2006.

BIRD, K.D.; HADZI-PAVLOVIC, D. Controlling the maximum familywise Type I error rate in analyses of multivariate experiments. **Psychological Methods**, v. 19, n. 2, 265–280, 2014.

BORGES, L. C.; FERREIRA, D. F. Power and type I error rates of Scott-Knott, Tukey and Student-Newman-Keuls's tests under residual normal and non-normal distributions. **Revista Matemática e Estatística**, v. 21, n. 2, 67-83, 2003.

BOX, G. E. P.; COX, D. R. An Analysis of Transformations. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 26, n. 2, 211-252, 1964.

CANDIOTI, L. V.; ZAN, M. M.; CAMARA, M. S.; GOICOECHEA, H. C. Experimental design and multiple response optimization - using the desirability function in analytical methods development. **Talanta**, v. 124, n. 2, 123-138, 2014.

CARGNELUTTI FILHO, A.; SOUZA, J. M.; PEZZINI, R. V.; NEU, I. M. M.; SILVEIRA, D. L.; PROCEDI, A. Optimal plot size for experiments with black oats and the common vetch. **Ciência Rural**, v. 50, e20190123, 2020.

CARVALHO, A.M.X. **Estatística Experimental e Observacional: uma nova abordagem sobre os métodos clássicos**. Rio Paranaíba: UFV-CRP, *in press*, 2023.

CARVALHO, A.M.X.; MENDES, F.Q.; BORGES, P.H.C.; KRAMER, M. A brief review of the classic methods of experimental statistics. **Acta Scientiarum - Agronomy**, v. 45, e56882, 2023a.

CARVALHO, F. J. **Modelos lineares generalizados na agronomia: análise de dados binomiais e de contagem, zeros inflacionados e enfoque bayesiano**. Tese (Doutorado em Fitotecnia). Universidade Federal de Uberlândia, Uberlândia, 2019.

CARVALHO, A.M.X.; SOUZA, M.R.; MARQUES, T.B.; SOUZA, D.L.; SOUZA, E.F.M. Familywise type I error of ANOVA and ANOVA on ranks in factorial experiments. **Ciência Rural**, v. 53, n. 7, e20220146, 2023b.

CECON, P. R.; SILVA, A. R.; NASCIMENTO, M.; FERREIRA, A. **Métodos estatísticos**. Viçosa: Editora da UFV, 2012.

CONAGIN, A. Tables for the calculation of the probability to be used in the modified bonferroni's test. **Brazilian Journal of Agriculture**, v. 76, 71-83, 2001.

CONAGIN, A.; PIMENTEL-GOMES, F. Escolha adequada dos testes estatísticos para comparações múltiplas. **Brazilian Journal of Agriculture**, v. 79, 288-295, 2004.

CONAGIN, A.; BARBIN, D.; DEMÉTRIO, C. G. B. Modifications for the Tukey test procedure and evaluation of the power and efficiency of multiple comparison procedures. **Scientia Agricola**, v. 65, n. 4, 428-432, 2008.

CONOVER, W. J. **Practical Nonparametric Statistics**. 3.. ed. John Wiley & Sons, 1999.

CONOVER, W. J. The rank transformation - an easy and intuitive way to connect many nonparametric methods to their parametric counterparts for seamless teaching introductory statistics courses. **WIREs Computational Statistics**, v. 4, 432-438, 2012.

CONOVER, W. J.; IMAN, R.L. Rank transformation as a bridge betwen parametric and nonparametric statistics. **The American Statistician**, v. 35, n. 3, 124-134, 1981.

CONRADO, T. V.; FERREIRA, D. F.; SCAPIM, C. A.; MALUF, W. R. Adjusting the Scott-Knott cluster analyses for unbalanced designs. **Crop Breeding and Applied Biotechnology**, v. 17, 1-9, 2017.

COUTO, M. R. M.; JACOBI, L. F.; MACHADO, G. S. Analise de variância multivariada aplicada a dados com medidas repetidas. **Ciência & Natura**, Santa Maria, v. 42, e01, 2020.

DANCEY, C. P.; REIDY, J. G.; ROWE, R. **Estatística sem matemática para as ciências da saúde**. Porto Alegre: Penso, 2017.

DI RIENZO, J. A.; GUZMÁN, A. W.; CASANOVES, F. A multiple-comparisons method based on the distribution of the root node distance of a binary tree. **Journal of Agricultural, Biological, and Environmental Statistics**, v. 7, 129-142, 2002.

DUTCOSKY, S. D. **Análise sensorial de alimentos**. 4. ed. Curitiba: Champagnat – PUCPress; 2013.

FERREIRA, D. F. SISVAR: A Guide for Its Bootstrap Procedures in Multiple Comparisons. **Ciência & Agrotecnologia**, v. 38, n. 2, 109-112, 2014.

GARCIA-MARQUES, T.; AZEVEDO, M. A inferência estatística múltipla e o problema da inflação do nível de alfa: a ANOVA como exemplo. **Psicologia**, v. 10, n. 1, 195-220, 1995.

GIRARDI, L. H.; CARGNELUTTI FILHO, A.; STORCK, L. Erro tipo I e poder de cinco testes de comparação múltipla de médias. **Revista Brasileira de Biometria**, v. 27, n. 1, 23-36, 2009.

GONÇALVES JÚNIOR, F. A.; CORRÊA, T. C. Reflexões sobre a hiperspecialização científica e suas conseqüências para a geografia. **Geografia, Ensino & Pesquisa**, v. 21, n. 3, 87-96, 2017.

GONÇALVES, B. O.; RAMOS, P. S.; AVELAR, F. G. Teste de Student-Newman-Keuls Bootstrap: proposta, avaliação e aplicação em dados de produtividade de graviola. **Revista Brasileira de Biometria**, v. 33, 445-470, 2015.

GOTELLI, N. J.; ELLISON, A. M. **Princípios de estatística em ecologia**. Porto Alegre: Artmed, 2011.

HINES, W. G. S.; O'HARA-HINES, R. Increased power with modified forms of the Levene (Med) test for heterogeneity of variance. **Biometrics**, v. 56, n. 2, 451-454, 2000.

KESELMAN, H. J. Per Family or Familywise Type I Error Control: "Eether, Eyether, Neether, Nyther, Let's Call the Whole Thing Off!". **Journal of Modern Applied Statistical Methods**, v. 14, n. 1, 24-37, 2015.

KRAMER, M. H.; PAPAROZZI, E. T.; STROUP, W. W. Best Practices for Presenting Statistical Information in a Research Article. **HortScience**, v. 54, n. 9, 1605-1609, 2019.

LITTLE, R. J. In praise of simplicity not mathematistry! Ten simple powerful ideas for the statistical scientist**. Journal of the American Statistical Association**, v. 108, n. 502, 359-369, 2013.

LOUREIRO, L. M. J.; GAMEIRO, M. G. H. Interpretação crítica dos resultados estatísticos: para lá da significância estatística. **Revista de Enfermagem Referência**, v. 3, n. 3, 151-162, 2011.

LÚCIO, A. D.; SARI, B. G. Planning and implementing experiments and analyzing experimental data in vegetable crops: problems and solutions. **Horticultura Brasileira**, v. 35, 316-327, 2017.

MANLY, B. F. J. **Multivariate statistical methods – a primer**. 2. ed. London: Chapman & Hall, 1995.

MANN, P. S. **Introdução à estatística**. 8ª Ed. Rio de Janeiro: LTC, 2015.

MONTGOMERY, D. C. **Design and Analysis of Experiments**. 9. ed. Wiley, Danvers, 2017.

NUNES, R. P. **Métodos para a pesquisa agronômica**. Fortaleza: Ed. UFC, 1998.

PATRIOTA, A. G. Uma medida de evidência alternativa para testar hipóteses gerais. **Ciência & Natura**, Santa Maria, v. 36, 14-22, 2014.

PERECIN, D.; BARBOSA, J. C. Uma avaliação de seis procedimentos para comparações múltiplas. **Revista de Matemática e Estatística**, v. 6, n. 1, 95-104, 1988.

PERECIN, D.; CARGNELUTTI FILHO, A. Efeitos por comparações e por experimento em interações de experimentos fatoriais. **Ciência e Agrotecnologia**, v. 32, n. 1, p. 68-72, 2008.

PIEPHO, H. P.; EDMONDSON, R. N. A tutorial on the statistical analysis of factorial experiments with qualitative and quantitative treatment factor levels. **Journal of Agronomy and Crop Science**, v. 204, n. 5, 429-455, 2018.

PIMENTEL-GOMES, F. **Curso de estatística experimental**. Piracicaba: FEALQ, 2009.

POSSATTO JUNIOR, O.; BERTAGNA, F. A. B.; PETERLINI, E.; BALERONI, A. G.; ROSSI, R. M.; ZENI NETO, H. Survey of statistical methods applied in articles published in Acta Scientiarum. Agronomy from 1998 to 2016. **Acta Scientiarum - Agronomy**, v. 41, e42641, 2019.

PRIMPAS, I.; TSIRTSIS, G.; KARYDIS, M.; KOKKORIS, G. D. Principal component analysis: Development of a multivariate index for assessing eutrophication according to the European water framework directive. **Ecological Indicators**, v. 10, 178-183, 2010.

SALTER, K. C.; FAWCET, R. F. The art test of interaction: a robust and powerful rank test of interaction in factorial models. **Communications in Statistics - Simulation and Computation**, v. 22, 137-153, 1993.

SAWILOWSKY, S. S.; BLAIR, R. C. A more realistic look at the robustness and Type II error properties of the t test to departures from population normality. **Psychological Bulletin**, v. 111, n. 2, 352–360, 1992.

SCHMIDER, E.; ZIEGLER, M.; DANAY, E.; BEYER, L.; BÜHNER, M. Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. **Methodology: European Journal of Research Methods for the Behavioral and Social Sciences**, v. 6, n. 4, 147–151, 2010.

SNEDECOR, G. W.; COCHRAN, W. G. **Statistical Methods**. 8. ed. Iowa: ISU Press. 1989.

SOUSA, C. A. D.; LIRA JUNIOR, M. A.; & FERREIRA, R. L. C. Avaliação de testes estatísticos de comparações múltiplas de médias. **Revista Ceres**, v. 59, n. 3, 350-354, 2012.

STORCK, L.; GARCIA, D. C.; LOPES, S. J.; ESTEFANEL, V. **Experimentação Vegetal**. 3. Ed. Santa Maria: Ed UFSM, 2016.

STROUP, W. W. Rethinking the Analysis of Non-Normal Data in Plant and Soil Science. **Agronomy Journal**, v. 107, n. 2, 811-827, 2015.

TAVARES, L. F.; CARVALHO, A. M. X.; MACHADO, L. G. An evaluation of the use of statistical procedures in soil science. **Revista Brasileira de Ciência do Solo**, v. 40: e0150246, 2016.

TORMAN, V. B. L.; COSTER, R.; RIBOLDI, J. Normalidade de variáveis: métodos de verificação e comparação de alguns testes não-paramétricos por simulação. **Revista HCPA**, v. 32, n. 2, 227-235. 2012.

WOBBROCK, J. O.; FINDLATER, L.; GERGLE, D.; HIGGINS, J. J. The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. In J. J. HIGGINS (Ed) **Proceedings of the ACM Conference on Human Factors in Computing Systems.** Vancouver, ACM Press New York, 2011, p. 143-146.

ZHIYUAN, W.; WANG, D.; ZHOU, H.; QI, Z. Assessment of soil heavy metal pollution with principal component analysis and Geoaccumulation Index. **Procedia Environmental Sciences**, v. 10, 1946–1952, 2011.

ZIMMERMANN, F. J. P. **Estatística Aplicada à Pesquisa Agrícola**. Santo Antônio de Goiás: Embrapa Arroz e Feijão, 2004.

ZIMMERMAN, D. W. A note on consistency of non - parametric rank tests and related rank transformations. **British Journal of Mathematical and Statistical Psychology**, v. 65, n. 1, 122-144, 2012.

Contribuição de autoria:

1 – André Mundstock Xavier de Carvalho

Doutor em Agronomia (Solos e Nutrição de Plantas) pela Universidade Federal de Viçosa

https://orcid.org/0000-0002-2806-6058 • andre.carvalho@ufv.br

Contribuição: primeira redação, investigação, análise formal

2 – Éder Matsuo

Doutor em Genética e Melhoramento pela Universidade Federal de Viçosa

https://orcid.org/0000-0002-2643-9367 • edermatsuo@ufv.br

Contribuição: investigação, revisão e edição.

3 – Marcelo da Silva Maia

Graduado em Agronomia pela Universidade Federal de Viçosa

https://orcid.org/0000-0003-0802-4675 • marcelo.maia@ufv.br

Contribuição: revisão e edição.

Como citar este artigo

CARVALHO, A. M. X.; MATSUO, E.; MAIA, M. S. Avaliação da normalidade, validade dos testes de médias e opções não-paramétricas: contribuições para um debate necessário. **Ciência e Natura**, Santa Maria, v. 45, e9, 2023. DOI 10.5902/217946067509. Disponível em: https://doi.org/10.5902/2179460X67509