Evento: XXVI Jornada de Pesquisa

MÉTODOS ESTATÍSTICOS NO CONTEXTO DE BIG DATA[1]

STATISTICAL METHODS IN THE BIG DATA

Carolina Hilda Schleger[2], Benjamim Zucolotto[3], Rafael Zancan Frantz[4]

[1] Projeto de Pesquisa realizado no programa de Pós – Graduação em Modelagem Matemática e Computacional.
[2] Mestranda do Programa de Pós – Graduação em Modelagem Matemática e Computacional da UNIJUI.
[3] Professor do Programa de Pós – Graduação em Modelagem Matemática e Computacional da UNIJUI.
[4] Professor do Programa de Pós – Graduação em Modelagem Matemática e Computacional da UNIJUI.

RESUMO

A geração e o armazenamento de dados nas diferentes áreas do conhecimento estão cada vez mais presentes na sociedade com o decorrer dos anos. A quantidade e a qualidade das informações podem ser refinadas para a tomada de decisão sobre determinado aspecto. Nestes, casos, os métodos estatísticos se apresentam como uma poderosa ferramenta para tratamento de dados, e principalmente no contexto de *Big Data* em que as informações possuem alto volume de dados, velocidade e variedade. Assim, o estudo da aplicabilidade do comportamento dos diferentes métodos estatísticos no contexto de *Big Data* se faz importante para uma análise estatística eficiente. Ao longo do texto é abordado sobre métodos paramétricos e não paramétricos robustos diante de dados, principalmente, como comportamento de distribuição não normal fazendo uso da representação pelo gráfico de caixa e de densidade de Kernel.

Palavras-chave: Métodos estatísticos. Big Data. Análise estatística.

ABSTRACT

The generation and storage of data in different areas of knowledge are increasingly present in society over the years. The quantity and quality of information can be refined for decision making on a particular aspect. In these cases, statistical methods are presented as a powerful tool for data treatment, and especially in the context of Big Data in which information has a high volume of data, speed and variety. Thus, the study of the applicability of the behavior of different statistical methods in the context of Big Data is important for an efficient statistical analysis. Throughout the text, robust parametric and non-parametric methods are discussed in relation to data, mainly as non-normal distribution behavior making use of the representation by the box plot and Kernel density.

Keywords: Statistical Methods. Big Data. Statistical analysis.

INTRODUÇÃO

O conhecimento traz em sua essência o mistério do desconhecido que move

fronteiras imagináveis pela sociedade. Durante o decorrer dos séculos muitos cientistas enfrentaram desafios para realizar e divulgar suas descobertas demonstrando suas concepções baseadas em experimentos rigorosos que por vezes contestavam verdades absolutas. Os diferentes ramos que caracterizam a sociedade atual geram e armazenam uma vasta quantidade de dados muitas vezes incalculáveis em sistemas computacionais. Estes dados apresentam informações relacionadas à, por exemplo, culturas sociais, financeiras, educacionais, judiciais e de saúde, podendo ser armazenados em arquivos de texto, planilhas, imagens, áudio e vídeos (GANDOMI; HAIDER, 2015).

Neste contexto, a área de *Big Data* busca compreender o comportamento destes dados verificando a existência de padrões através da análise exploratória de dados (AED). Os dados oriundos de *Big Data* trazem desafios para análises estatísticas pela dificuldade de serem processados e/ou analisados usando ferramentas tradicionais (ZIKOPOULOS; EATON; IBM, 2011). Essas dificuldades são decorrentes das complexidades dos dados sobre as diferentes fontes que o compõe, os erros de medição, valores discrepantes e valores ausentes. Essas características, como correlações espúrias, podem causar descobertas científicas falsas e inferências estatísticas erradas (FAN; HAN; LIU, 2014).

Os métodos estatísticos tradicionais têm sua concepção baseada na teoria do limite central sobre uma distribuição normal de amostras. Entretanto, a AED de Big Data apresenta, em sua maioria, distribuições não normais. Sobre este contexto, o teste de hipóteses estatística pode acarretar dois tipos de erros (ARCURI; BRIAND, 2014). Erro do tipo I ou falso positivo, ocorre quando é rejeitada a hipótese nula quando ela é verdadeira. O nível de probabilidade adotado é conhecido como $\propto$, definido entre os valores de 0.05 ou 0.01 (ZIMMERMAN, 2000). O erro do tipo II ou falso negativo, ocorre quando aceitamos a hipótese nula quando esta é realmente falsa. O nível de probabilidade, β, é aceitável quando apresenta valor igual ou menor que 0.2 (KITCHENHAM et al., 2017). O nível de probabilidade ou significância determina o poder estatístico de um método referente à probabilidade de rejeitar corretamente a hipótese nula. A confirmação da hipótese definida é apenas confirmada quando é comparado o teste estimado com o fator das condições reais.

Situações que apresentam dados com características distantes da normalidade teórica podem comprometer o desempenho das técnicas clássicas que são baseadas em suposições de

idealismo sobre a distribuição (HOAGLIN; MOSTELLER; TUKEY, 2000). Kitchenham et al. (2017) realizou uma pesquisa sobre o desempenho de métodos estatísticos paramétricos e não paramétricos em dados com diferentes características de distribuição com o intuito de verificar a robustez. Métodos robustos aumentam a confiabilidade do poder estatístico na AED. O estudo possibilitou visualizar que a determinação do método a ser utilizado deve se apoiar na análise da distribuição dos dados anteriormente por meio de gráficos. O Gráfico de Caixa e de Densidade de Kernel são os mais recomendados pelos estatísticos por proporcionarem a visualização do comportamento dos dados detalhadamente (KITCHENHAM, 2015; KITCHENHAM et al., 2017; SILVERMAN, 1986; WILCOX, 2017).

Métodos paramétricos tem a análise baseada em amostras com a distribuição dos dados normal ou quase normal como os testes T e ANOVA. Para amostras com dados com distribuições não normais é necessário a transformação dos dados para o emprego de métodos paramétricos. Nestes casos, aconselha-se a realização da análise por meio de métodos não paramétricos como o teste de Welch, Kruskal-Wallis e Wilcoxon Mann Whitney (WILCOX, 2017; KITCHENHAM et al., 2017). Diante deste contexto, o texto tem o intuito de apresentar as principais vantagens e desvantagens de métodos paramétricos e não paramétricos sobre os diferentes tipos de comportamentos dos dados derivados do contexto de *Big Data*.

BIG DATA E A RELAÇÃO COM OS MÉTODOS ESTATÍSTICOS

A civilização moderna está conectada com aparatos tecnológicos dos quais muitas profissões possuem dependência para manter suas atividades no mercado competitivo. O armazenamento e monitoramento de informações como dados ambientais, financeiros, médicos, educacionais, agrícolas, entre tantos outros, estão cada vez mais presentes e são cada vez mais urgentes nas diferentes áreas do conhecimento (FAN; HAN; LIU, 2014; GANDOMI; HAIDER, 2015). O grande volume de dados produzidos em tempo real ou não, e de forma contínua podem apresentar características heterogêneas e estruturas de formas distintas, por exemplo, tabelas, texto, imagem, áudio, vídeo. Os dados não estruturados e semi-estruturados constituem hoje em dia a maioria das informações geradas (CUKIER, 2010).

A estrutura de dados complexos está muito presente na neurociência, que busca compreender a organização hierárquica, complexa e funcional da rede de conectividade do cérebro, identificando e explorando como o cérebro muda com a presença de doenças como Alzheimer, esquizofrenia, transtorno de déficit de atenção e hiperatividade, depressão, ansiedade, entre outros. Uma técnica de neuroimagem não invasiva muito utilizada chamada Inventário Freiburg de *mindfulness*, FMI, é utilizada para determinar os correlatos neurais de processos mentais em humanos (WALACH et al., 2006). As imagens de FMI permitem explorar a associação entre a conectividade do cérebro e respostas potenciais perante as doenças ou estado psicológico. Os dados obtidos de diferentes sujeitos são massivos e dimensionalmente elevados. Periodicamente uma solução de inteligência artificial examina diferentes informações e produz novos dados de imagem examinando centenas de vezes o cérebro do sujeito, gerando uma imagem 3D com curso no tempo que contém mais de centenas de milhares de voxel. Entretanto, existem enormes desafios na análise dos dados gerados, pois podem conter erros oriundos desde ruídos devido ao limite tecnológico das técnicas de medidas das imagens, até possível movimentação da cabeça do sujeito na realização da imagem (FAN; HAN; LIU, 2014).

Os dados gerados atualmente pela civilização apresentam grandes volumes e alta velocidade de dados complexos e variáveis que requerem técnicas avançadas e tecnologias para permitir a captura, armazenamento, distribuição, gestão e análise das informações; esse tipo de dado é identificado com o termo *Big Data* (GANDOMI; HAIDER, 2015). *Big Data* é normalmente definido pelas características dos 5Vs: volume, velocidade, variedade, veracidade e valor. O volume consiste na enorme quantidade de dados, a velocidade refere-se à criação dos dados em tempo real, e a variedade trata das características dos dados serem estruturados, semiestruturados ou não estruturados (LANEY, 2001). Entretanto, muitos autores destacam que apenas estas características não são suficientes para definir um conjunto de dados como *Big Data*, e acrescentam também: exaustividade, refinado e exclusivamente indexical, relacionalidade, extensionalidade e escalabilidade, veracidade, valor e variabilidade (BOYD; CRAWFORD, 2012; MARZ; WARREN, 2013; GANDOMI; HAIDER, 2015; KITCHIN; MCARDLE, 2016).

As características que auxiliam a identificar se um conjunto de dados pode ser

intitulado de *Big Data* nem sempre aparecem concomitantemente. Podemos considerar um conjunto de dados sem apresentar as características de volume e variedade, mas é fundamental que estes tenham velocidade e exaustividade, pois estes dois não são critérios de qualificação (KITCHIN; MCARDLE, 2016). O volume não é considerado o mais importante pelo motivo de não haver um estabelecimento de limites numéricos para determinar o tamanho necessário que o conjunto de dados deve conter. Por não apresentar limites destaca-se a importância da exaustividade, pois demonstra que os dados estão sendo gerados continuamente repetidas vezes por um período. Entretanto, a presença da velocidade não deixa de ser considerada, pois a geração de dados deve ser em tempo real e de forma continua. Logo, a análise e divulgação dos dados deve ser no menor tempo possível para que não seja obsoleto em relação ao tempo presente em que as informações estão sendo criadas.

Big Data oportuniza descobrir padrões populacionais sutis e heterogeneidades que não são possíveis com dados de pequena escala. Amostras maiores possibilitam revelar padrões ocultos associados a pequenas subpopulações e pontos fracos em comum em toda a população (FAN; HAN; LIU, 2014). O processo geral de "extração de *insights*", destes dados, envolve dois processos: gerenciamento e análise. O processo de gerenciamento engloba processos e tecnologias de suporte para obter e armazenar dados e prepará-los para a análise. O processo de análise refere-se às técnicas utilizadas para a análise destes *insights* (LABRINIDIS; JAGADISH, 2012). Esta análise geralmente é realizada utilizando métodos estatísticos convencionais. Ao mesmo tempo que os dados proporcionam novas descobertas, também trazem desafios para a análise estatística. Porque a maioria das técnicas de alta dimensão tratam apenas problemas de acúmulo de ruído e correlações espúrias, mas não de endogeniedade acidental. Kitchenham et al. (2017) descreve que estes problemas são encontrados em conjuntos de dados não normais de Engenharia de Software e causam baixo desempenho de testes estatísticos clássicos. Logo, recomenda reanalisar as técnicas usadas para estas situações. O segundo desafio está na capacidade dos métodos desenvolvidos em escalonar a grande quantidade de dados computacionalmente de forma eficiente (GANDOMI; HAIDER, 2015). Contanto, percebe-se a proximidade entre as informações que definem o contexto de Big Data e as técnicas estatísticas.

MÉTODOS ESTATÍSTICOS PARAMÉTRICOS E NÃO PARAMÉTRICOS

A análise estatística é importante nas avaliações de resultados e para vários tipos de dados existem diferentes métodos de análise estatística, ou seja, "mais adequados". Dados com distribuição não normal são constantemente submetidos a métodos estatísticos robustos devido a confiabilidade dos resultados (KITCHENHAM, 2015). Métodos estatísticos que demonstram bom desempenho são aqueles que focam sua ação para o corpo central da distribuição dos dados destinando pouca importância aos dados das extremidades (MOSTELLER; TUKEY, 1977; WILCOX; KESELMAN, 2003; KITCHENHAM et al., 2017).

Para determinar qual método estatístico deve ser utilizando para que haja confiabilidade nos resultados, é preciso atentar aos comportamentos dos dados da pesquisa, principalmente quanto ao tipo de distribuição. A distribuição dos dados pode ser classificada em normal e não normal. A distribuição normal apresenta simetria dos dados em torno da média e um desvio padrão baixo. Uma distribuição não normal apresenta uma média descentralizada tendendo para alguma das extremidades, são assimétricas para a direita ou para a esquerda, possuem existência de dados bimodais, presença de *outliers* e um desvio padrão alto.

Na Figura 1 é possível visualizar as principais características de cada distribuição nos histogramas e nos gráficos de caixa. A Figura 1 (a) representa uma amostra com distribuição considerada normal. As Figuras 1 (b), (c) e (d) representam dados com distribuição não normal, mas a Figura 1 (b) apresenta dados bimodais, a Figura 1 (c) uma distribuição assimétrica à direita e a Figura 1 (d) uma distribuição dos dados de assimetria à esquerda com a presença de *outliers* no início.

A verificação do comportamento da distribuição dos dados de uma amostra de tamanho n pode ser realizada por meio de métodos gráficos, métodos numéricos e teste de normalidade. Os testes de normalidade geralmente utilizados são o de Shapiro-Wilk, Kolmogorov-Smirnow, Lilliefors e de Anderson-Darling. Um estudo de comparação entre estes, demonstrou que o teste Shapiro-Wilk é indicado como mais poderoso sobre uma amostra de tamanho maior que trinta (FARRELL; ROGERS-STEWART, 2006). Os métodos gráficos se destacam por sua facilidade de interpretação e análise, mas para que haja

evidências conclusivas e válidas é recomendado utilizar um dos métodos citados anteriormente concomitantemente (RAZALI; YAP, 2011).

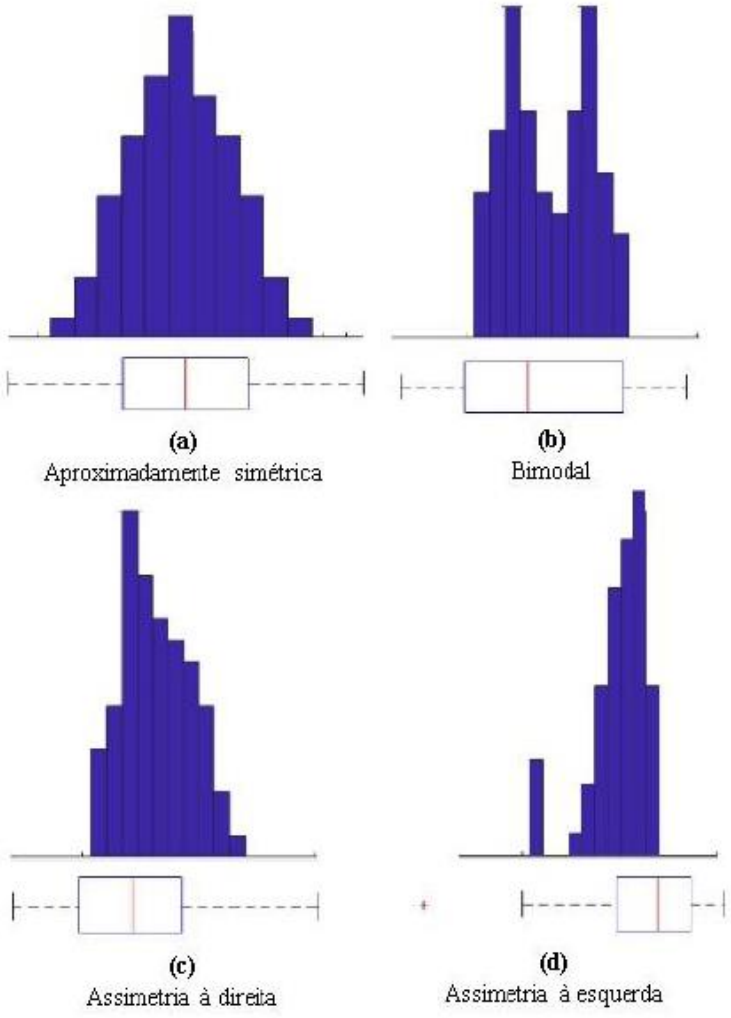

Figura 1 – Exemplos de distribuição de uma amostra de dados. Dados autor 2021..

A alta sensibilidade dos testes de normalidade levou Kitchenham et al. (2017) a recomendar a utilização conjunta do gráfico de caixa e do gráfico de densidade de Kernel por possibilitarem a visualização da distribuição dos dados. Os dois gráficos possibilitam ao pesquisador analisar a distribuição dos dados com clareza. O gráfico de densidade de Kernel possibilita a visualização mais detalhada da distribuição e, o gráfico de caixa se destaca sobre a análise da existência de *outliers*.

No gráfico de caixa é possível visualizar a organização dos dados em quartis, limite

inferior e superior, sua amplitude e os *outliers* existentes. Os quarto quartis são medidas de tendência central que dividem os dados em parcelas. Os limites inferiores e superiores são simbolizados pelas hastes localizados nos extremos da caixa. Os *outliers* são os dados localizados antes do limite inferior ou depois do limite superior, em outros termos, são dados que apresentam valores distintos do restante da amostra.

O gráfico de densidade de Kernel é uma abordagem de estimativa de uma função de probabilidade não paramétrica que não realiza suposições sobre a distribuição subjacente à amostra aleatória (HARPOLE et al., 2014). A densidade de Kernel é utilizada quando a distribuição do conjunto de dados não é conhecida pelo pesquisador, ao contrário de uma estimativa paramétrica em que se tem o conhecimento sobre a possível distribuição, por exemplo, uma distribuição normal.

Na Figura 2 é possível visualizar as informações que caracterizam o gráfico de caixa e de densidade de Kernel. Neles é representado o mesmo conjunto de dados de um experimento que demonstram certa distribuição. No gráfico de caixa é possível visualizar a mediana próxima da origem e a presença de *outliers* no limite superior do gráfico. Com gráfico de densidade de Kernel é possível analisar com mais detalhes a distribuição destes dados confirmando a presença de dados próximos da origem e indícios que a distribuição é bimodal.

Em experimentos com distribuição normal ou quase normal são geralmente utilizados métodos estatísticos paramétricos e os métodos não paramétricos sobre dados com distribuição não normal. É possível usar métodos paramétricos sobre dados que apresentam comportamento irregular realizando a transformação dos mesmos. As técnicas de transformação dos dados com o intuito de obter uma distribuição normal do conjunto de dados mais conhecidas são: logarítmica, raiz quadrada, arcoseno raíz quadrada, boxcox e recíproca. A transformação logarítmica é frequentemente mais utilizada por aplica a transformação em toda amostra e não apenas em alguns grupos (WILCOX, 2017). Apesar de ser considerada mais confiável esta ação, há casos que a transformação dificulta a realização de uma análise posterior e que os dados continuem apresentando a não normalidade em sua distribuição. Além disso, podem não proteger contra o baixo poder estatístico ao lidar com distribuições assimétricas ou para a direita ou para a esquerda, recomendando-se para estes

casos a utilização de métodos não paramétricos (WILCOX; KESELMAN, 2003).

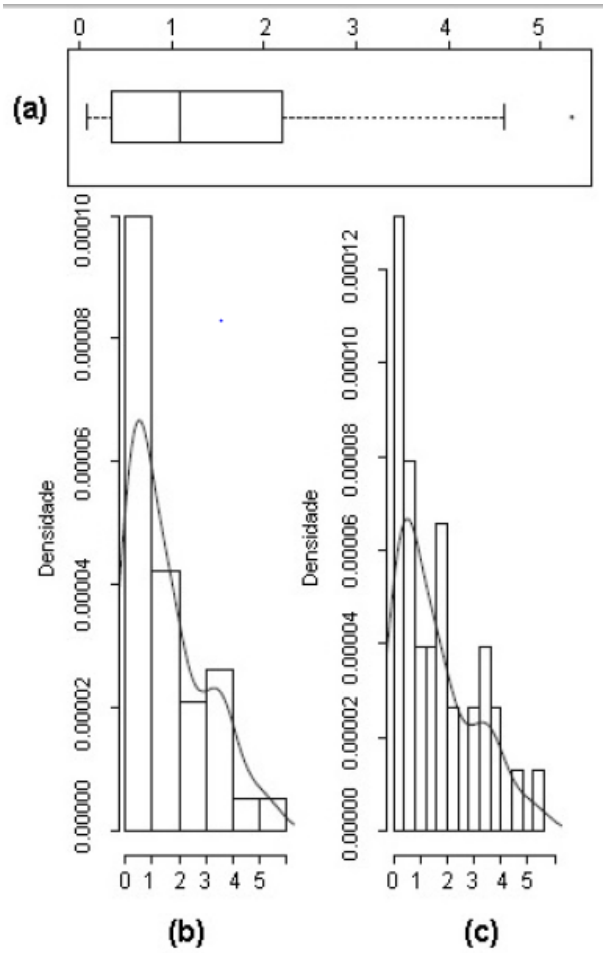

Figura 2 – Conjuto de dados representados em Gráfico de Caixa e dois Gráficos de Densidade de Kernel. Adaptado de Kitchenham et al. (2017).

Os métodos paramétricos, como o Teste T, Teste F, ANOVA, *Trimmed mean* ou média parada; possuem melhor desempenho quando os valores dos dados analisados são distribuídos em torno do valor central, com frequência e desvio padrão baixos, o que classificamos como dados normais. Em sua maioria, estes métodos têm sua definição apoiada no teorema do limite central perante pequenas amostras, como o Teste T (KITCHENHAM et al., 2017). O Teste T de Student compara a diferença entre as médias de até dois grupos de amostras. Kitchenham et al. (2017) relata sobre a importância de haver amostras com grandes

quantidades de dados para evitar problemas com dados não normais, pois amostras pequenas com estas características aumentam significativamente a problematicidade da análise do teste. Em comparação sobre a gravidade do problema entre amostras com combinações não normais e amostras com apenas uma propriedade não normal, a primeira possui maior gravidade. Isso ocorre por que pode haver um alto desvio padrão por causa dos *outliers* e a inflação da variância, acarretando no aumento da probabilidade do erro do tipo II. Um dos motivos do aumento da probabilidade do erro tipo II é a presença de contaminação nas amostras de dados, que acarretam uma variância maior do que a distribuição não contaminada. Isso causa um desvio padrão alto e a presença de *outliers* que ocultam possível inflação da variância (WILCOX; KESELMAN, 2003; KITCHENHAM et al., 2017). É possível o teste T ser robusto perante uma distribuição com características não normais se a variância for diferente em cada grupo. Logo, se os tamanhos dos grupos forem iguais, os dados em cada grupo são normalmente distribuídos e os tamanhos das amostras são maiores que quinze (RAMSEY, 1980).

A realização de transformação de dados não normal é frequentemente utilizada para podemos usar o teste paramétrico com o intuito de colocar limites de confiança na média da amostra e deter um alto poder estatístico. Entretanto, deve-se ter o cuidado para que este recurso não se distancie das características necessárias para o emprego do método. A distribuição lognormal, é um exemplo disto, independente de uma amostra com tamanho igual ou maior que 20. A probabilidade, sobre a definição de $\propto$ igual a 0.1; de a cauda inferior exibir um erro de tipo I é 0.11 em vez de 0.05 e; em relação a uma cauda superior a probabilidade de um erro tipo I é 0.02 em vez de 0.05. Então, a probabilidade real é 0.13 em vez de 0.1 (WILCOX; KESELMAN, 2003).

De forma geral o Teste T, o Teste F e os métodos de classificação não possuem resistência sobre dados distorcidos, não demonstrando confiabilidade por apresentarem problemas na variação das aproximações.

Outra métrica considerada robusta é a mediana que utiliza uma ou duas observações. Mas, não é recomendado por Kitchenham et al. (2017), justamente por utilizar apenas alguns dados que tornam as estimativas de erro padrão da mediana ineficientes, principalmente sobre valores de dados duplicados. Na mesma estratégia, em utilizar apenas uma porcentagem das

amostras dos dados, está à remoção de *outliers*. Os métodos que usam a média e a variância com os dados restantes depois da remoção de *outliers* apresentam dois problemas: podem falhar na detecção posterior de *outliers* e as observações restantes tornam-se dependentes, invalidado o cálculo do seu erro padrão (WILCOX; KESELMAN, 2003). Mesmo assim, é possível utilizar esta estratégia com segurança como o método de *Trimmed mean* (WILCOX, 2017).

O método *Trimmed mean* consiste na remoção de uma porcentagem de valores menores e maiores de um conjunto de dados. A porcentagem de 20% remoção é um padrão confiável que mantém um equilíbrio na confiabilidade estatística e a probabilidade de um erro do tipo I (KITCHENHAM, 2015; WILCOX, 2017).

Os testes não paramétricos por vezes não são utilizados para análise estatística por apresentarem um grau de liberdade menor que os testes paramétricos. Deve-se atentar que por mais que seja realizada uma técnica de transformação de dados com o intuito de obter uma distribuição normal, este continua apresentando irregularidades e, mesmo que seja possível sua transformação, seu resultado estatístico pode dificultar na interpretação posterior. Os métodos não paramétricos se tornam uma ótima alternativa nestes casos e mais confiáveis diante de características de distribuições não normais (KITCHENHAM et al., 2017). Os métodos que se destacam são o método de Cliff, Teste Yuen, Teste Welch e o Teste MWW (RAMSEY, 1980; BERGMANN; LUDBROOK; SPOOREN, 2000; ZIMMERMAN, 2000; ARCURI; BRIAND, 2014; WILCOX, 2017).

Os testes T e F são afetados por variâncias desiguais de grupos de tratamento, principalmente quando o tamanho das amostras é diferente (ZIMMERMAN, 2000). Uma alternativa para o teste T é o teste U de Mann-Whitney (MWW) e Kruskal-Wallis baseadas na suposição de não haver valores duplicados. Contudo recomenda-se utilizar o teste de Welch, pois apresenta robustez diante de variáveis desiguais (RAMSEY, 1980). O Teste de Welch é uma variação do Teste T definido por Welch (1938).

O teste de MWW é baseado na conversão de dois conjuntos de dados independentes de seu tamanho por meio da classificação destes dados independente a qual conjunto pertença (KITCHENHAM et al., 2017). O teste de MWW exposto a grupos de amostra com variações diferentes, indiferente destas serem iguais, aumenta significativamente a probabilidade de

erros do Tipo I ocultos, excedendo o nível de significância. Este fato pode acontecer principalmente em conjuntos de amostras que são alimentadas consecutivamente, acarretando o aumento do valor da média e a variância pelo motivo de seu cálculo depender significativamente do tamanho das amostras (KITCHENHAM et al., 2017). Desta forma, as equações para a média e variância das classificações deste método não convergem em um valor finito, tornando-o pouco confiável para amostras grandes.

O Teste Yuen é diferente do Teste Welch, pois é uma derivação do método de *trimmed mean* que é projetado para permitir variâncias desiguais (YUEN, 1974). Este teste tem destaque sobre experimentos com medidas repetidas, grupos múltiplos e experimentos fatoriais; permitindo a testagem de combinações lineares entre os valores médios (KITCHENHAM et al., 2017). Entretanto se à distribuição entre os dados diferem em assimetria, a diferença entre as médias pode ser maior que o teste demonstra, nestes casos, o Teste Welch é mais poderoso (WILCOX, 2017).

CONSIDERAÇÕES FINAIS

O método estatístico a ser utilizado para análise dos dados recomenda ser decido após à analise do comportamento dos dados quanta sua origem e sua distribuição estrutural. Pelo motivo de os métodos estatísticos apresentam diferentes comportamentos de confiança dependo das características dos dados, principalmente se estes apresentam uma não normalidade. Uma ferramenta que possibilita está visão ao pesquisador é o gráfico de densidade de Kernel e o gráfico de caixa em conjunto. O gráfico de densidade de Kernel proporciona a visão detalhada da distribuição dos dados da pesquisa não ocultando informações, entretanto, o gráfico de caixa possibilita visualizar a existência de *outliers* sem margem de duvida.

Em comparação ao substituto ao Teste T sobre amostras com distribuição não normal e apresentando uma variância desigual de grupos de tratamento com a existência de valores duplicados o teste Welch apresenta melhor desempenho.

Ao verificar grande influência dos outliers sobre a distribuição dos dados se recomenda o teste Trimmed mean ou seu substituto não paramétrico Teste Yuen que suporta variância desiguais. Entretanto deve-se ter o cuidado de visualizar quando a simetria das

amostras, caso houver diferença se recomenda o Teste Welch.

REFERÊNCIAS BIBLIOGRÁFICAS

ARCURI, A.; BRIAND, L. A hitchhiker's guide to statistical tests for assessing randomized algorithms in software engineering. *Software Testing, Verification and Reliability*, v. 24, n. 3, p. 219–250, 2014. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/stvr.1486>.

FAN, J.; HAN, F.; LIU, H. Challenges of big data analysis. *National Science Review*, Oxford University Press, v. 1, n. 2, p. 293–314, jun. 2014. ISSN 2053-714X. Publisher Copyright: © The Author(s) 2014. Published by Oxford University Press on behalf of China Science Publishing & Media Ltd. All rights reserved. Copyright: Copyright 2021 Elsevier B.V., All rights reserved.

GANDOMI, A.; HAIDER, M. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, v. 35, n. 2, p. 137–144, 2015. ISSN 0268-4012. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0268401214001066>.

HOAGLIN, D. C.; MOSTELLER, F.; TUKEY, J. W. *Understanding Robust and Exploratory Data Analysis*. 1. ed. [S.l.]: Wiley-Interscience, 2000. ISBN 0471384917.

KITCHENHAM, B. Robust statistical methods: Why, what and how: Keynote. In: *Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering*. New York, NY, USA: Association for Computing Machinery, 2015. (EASE'15). ISBN 9781450333504. Disponível em: <https://doi.org/10.1145/2745802.2747956>.

KITCHENHAM, B. et al. Robust statistical methods for empirical software engineering. *Empirical Software Engineering*, v. 22, n. 2, p. 579–630, Apr 2017. ISSN 1573-7616. Disponível em: <https://doi.org/10.1007/s10664-016-9437-5>.

KITCHENHAM, B. et al. Systematic literature reviews in software engineering – a systematic literature review. *Information and Software Technology*, v. 51, n. 1, p. 7–15, 2009. ISSN 0950-5849. Special Section - Most Cited Articles in 2002 and Regular Research Papers. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950584908001390>.

KITCHIN, R.; MCARDLE, G. What makes big data, big data? exploring the ontological characteristics of 26 datasets. *Big Data Society*, v. 3, n. 1, p. 2053951716631130, 2016. Disponível em: <https://app.dimensions.ai/details/publication/pub.1011507222andhttps: //doi.org/10.1177/2053951716631130>.

LABRINIDIS, A.; JAGADISH, H. V. Challenges and opportunities with big data. *Proc. VLDB Endow.*, VLDB Endowment, v. 5, n. 12, p. 2032–2033, ago. 2012. ISSN 2150-8097. Disponível em: <https://doi.org/10.14778/2367502.2367572>.

LANEY, D. *3D Data Management: Controlling Data Volume, Velocity, and Variety*. [S.l.], 2001. Disponível em: <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>.

MARZ, N.; WARREN, J. *Big Data: Principles and best practices of scalable real-time data systems*. [S.l.]: Manning, 2013.

MOSTELLER, F.; TUKEY, J. Data analysis and regression: a second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods, Reading, Mass.: Addison-Wesley*, -1, 01 1977.

RAMSEY, P. H. Exact type 1 error rates for robustness of student's t test with unequal variances. *Journal of Educational Statistics*, v. 5, n. 4, p. 337–349, 1980. Disponível em: <https://doi.org/10.3102/10769986005004337>.

RAZALI, N. M.; YAP, B. Power comparisons of shapiro-wilk, kolmogorov-smirnow, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, v. 2, n. 1, p. 21–33, 01 2011. ISSN 978-967-363-157-5.

SCOTT, D. W. v. *Multivariate density estimation: theory, practice, and visualization*. New York : Wiley, 1992. ISBN 0471547700. Disponível em: <http://lib.ugent.be/catalog/rug01:001051093>.

SILVERMAN, B. *Density Estimation for Statistics and Data Analysis*. Taylor & Francis, 1986. (Chapman & Hall/CRC Monographs on Statistics & Applied Probability). ISBN *Referências* 229780412246203. Disponível em: <https://books.google.co.ao/books?id=e-xsrjsL7WkC>.

WALACH, H. et al. Measuring mindfulness—the freiburg mindfulness inventory (fmi). *Personality and Individual Differences*, v. 40, n. 8, p. 1543–1555, 2006. ISSN 0191-8869.

Disponível em: <https://www.sciencedirect.com/science/article/pii/S0191886906000262>.

WELCH, B. L. The significance of the difference between two means when the population variances are unequal. *Biometrika*, [Oxford University Press, Biometrika Trust], v. 29, n. 3/4, p. 350–362, 1938. ISSN 00063444. Disponível em: <http://www.jstor.org/stable/2332010>.

WILCOX, R. Kernel density estimators: An approach to understanding how groups differ. *Understanding Statistics*, v. 3, p. 333–348, 2004.

Front matter. In: WILCOX, R. (Ed.). *Introduction to Robust Estimation and Hypothesis Testing (Fourth Edition)*. Fourth edition. Academic Press, 2017, (Statistical Modeling and Decision Science). p. i–iii. ISBN 978-0-12-804733-0. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128047330000147>.

WILCOX, R. R. Nonparametric estimation. In: . *Handbook of Computational Econometrics*. John Wiley Sons, Ltd, 2009. cap. 5, p. 153–182. ISBN 9780470748916. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470748916.ch5>.

WILCOX, R. R.; KESELMAN, H. J. Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, US, v. 8, n. 3, p. 254–274, 2003. Disponível em: <https://doi.org/10.1037/1082-989X.8.3.254>.

YUEN, K. K. The two-sample trimmed t for unequal population variances. *Biometrika*, v. 61, n. 1, p. 165–170, 04 1974. ISSN 0006-3444. Disponível em: <https://doi.org/10.1093/biomet/61.1.165>.

ZIKOPOULOS, P.; EATON, C.; IBM. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. 1st. ed. [S.l.]: McGraw-Hill Osborne Media, 2011. ISBN 0071790535.

ZIMMERMAN, D. W. Statistical significance levels of nonparametric tests biased by heterogeneous variances of treatment groups. *The Journal of General Psychology*, Routledge, v. 127, n. 4, p. 354–364, 2000. PMID: 11109998. Disponível em: <https://doi.org/10.1080/00221300009598589>.