Evento: XXIV Jornada de Pesquisa

ANÁLISE DO PROCESSAMENTO DE WORK UNITS PELO MOTOR DE EXECUÇÃO DA PLATAFORMA DE INTEGRAÇÃO GUARANÁ[1]
ANALYSE ABOUT PROCESSING OF WORK UNITS BY RUNTIME SYSTEM OF GUARANÁ INTEGRATION PLATFORM

Alessandra Lucero Silva[2], Fabricia Roos-Frantz[3], Rafael Z. Frantz[4]

[1] Pesquisa realizada no Programa de Pós-Graduação em Modelagem Matemática da Unijuí
[2] Aluna do Curso de Mestrado em Modelagem Matemática, Bolsista do CNPq - Brasil, alelucero182@gmail.com;
[3] Professora Doutora do Programa de Pós-Graduação em Modelagem Matemática, Orientador, fabriciar@gmail.com;
[4] Professor Doutor do Programa de Pós-Graduação em Modelagem Matemática, Coorientador, rzfrantz@gmail.com;

Resumo: Na comunidade de Engenharia de Software, existem estudos que avaliam a adequação dos testes estatísticos quando aplicados a conjuntos de dados experimentais com dadas características, tais como amostras não normalmente distribuídas, multimodais, com distribuição assimétrica, caudas pesadas e muitos outliers, frequentemente associadas a dados desta área. Aplicar testes estatísticos tradicionais a um conjunto de dados com estas características pode conduzir a erros, relacionados a aceitação e rejeição de hipóteses nulas, ou falhar em detectar importantes resultados. A área de Integração de Aplicações Empresariais oferece ferramentas para que as empresas possam integrar aplicações do seu ecossistema de software. A plataforma Guaraná é uma dessas ferramentas, a qual possui um motor que é responsável pela execução dos processos de integração. O presente trabalho trata da análise estatística descritiva de um conjunto de dados experimentais obtidos do processamento do motor de execução da plataforma Guaraná. Estes dados não satisfazem os pressupostos para análise paramétrica, pois possuem distribuição assimétrica, caudas pesadas, multimodais, com muitos outliers e são heterocedásticos. O objetivo deste artigo é sumarizar e descrever este conjunto de dados conforme a estatística descritiva. Para isto, primeiramente, verificou-se a distribuição do conjunto de dados utilizando Diagrama de caixa e Gráficos de densidade de kernel e após utilizou-se diferentes medidas de tendência central de forma a poder selecionar a medida mais adequada, considerando a distribuição dos dados analisada. Como resultado, observou-se que nos grupos amostrais, há a presença de outliers e que a medida de tendência central mais adequada é trimmed mean 20%. Concluiu-se que para este conjunto de dados os testes estatísticos da fase de Estatística Inferencial precisam ser robustos à não normalidade e heterocedasticidade.

Palavras-chave: Estatística Descritiva, Métodos Robustos, Densidade de Kernel, Engenharia de Software

Abstract: In the Software Engineering community, there are studies that assess the appropriateness of statistical testing when applied to experimental datasets with characteristics

such as non-normally distributed, multimodal, asymmetric distribution, heavy tails samples and many outliers, often associated with data of this area. Applying traditional statistical tests to a dataset with these characteristics can lead to errors related to accepting and rejecting null hypotheses, or failing to detect important results. Enterprise Application Integration provides tools for companies to integrate applications from their software ecosystem. The Guarana platform is one of these tools, which has a motor that is responsible for executing the integration processes. The present work deals with the descriptive statistical analysis of an experimental data set obtained from the Guaraná platform runtime system processing. These data do not satisfy the assumptions for parametric analysis because they have asymmetric distribution, heavy tails, multimodal, with many outliers and are heteroscedastic. The purpose of this paper is to summarize and describe this dataset according to descriptive statistics. For this, we firstly verified the distribution of the dataset using Boxplot and Kernel Density Graphs and then we used different measures of central tendency in order to select the most appropriate measure, considering the data distribution analyzed. As a result, it was observed that in the sample groups, there is the presence of outliers and that the most appropriate central tendency measure is trimmed mean 20%. It was concluded that for this data set the statistical tests of the Inferential Statistics phase need to be robust to non-normality and heteroscedasticity.

Keywords: Descriptive estatistic, Robust Methods, Kernel Density, Software Engineering

1 Introdução

Enterprise Application Integration – EAI – em português, Integração de Aplicações Empresariais é uma área da Engenharia de Software que se preocupa em oferecer melhores formas de realizar a integração de aplicações, ou seja, o compartilhamento de dados e funcionalidades entre quaisquer aplicações e banco de dados conectados em uma empresa, conforme Linthicum (2000).

Para realizar a integração das aplicações do ecossistema de software de uma empresa são utilizadas plataformas de integração (RITTER, MAY, RINDERLE-MA, 2017). Muitas dessas plataformas dão suporte à integração baseada na troca de mensagens entre aplicações. As mensagens carregam dados que precisam ser compartilhados entre as aplicações para que ocorra uma interação entre elas. O motor de execução das plataformas de integração é responsável por efetuar a troca de mensagens. Portanto, o bom funcionamento do processo de integração ao conectar aplicações depende do desempenho do motor de execução (FRANTZ, 2012).

Em Engenharia de Software, como as demais áreas, é necessário que seus processos de desenvolvimento e produtos sejam pensados e analisados constantemente. Para que isso ocorra pode-se realizar atividades experimentais que auxiliam na predição, avaliação e compreensão dos produtos de Engenharia de Software, conforme Basili, Selby e Hutchens (1986). Realizar experimentos é importante para geração de dados úteis para responder as questões de pesquisa, avaliar hipóteses ou criar novas, segundo Perry, Porter e Votta (2000).

Para uma adequada análise estatística dos dados experimentais, é indicado analisar o contexto do

experimento e o tipo de dado. No trabalho de Kitchenham et al. (2017), os autores discutem testes estatísticos mais adequados para conjuntos de dados de Engenharia de Software, capazes de lidar com características como não normalidade, presença de outliers (valores discrepantes) e distintas variâncias entre grupos. Eles propõem que os pesquisadores em Engenharia de Software usem métodos robustos, definidos como métodos “insensíveis a desvios de suposições relacionados a um modelo subjacente específico” (p. 581, tradução nossa).

Kitchenham et al. (2017) fazem um estudo sobre testes estatísticos robustos paramétricos e não paramétricos para lidar com amostras que não são normalmente distribuídas ou que entre grupos não possuem a mesma distribuição ou testes para lidar com diferenças de tendência central. Eles afirmam que são necessárias mais pesquisas referentes aos testes estatísticos robustos para engenharia de software, pois os conjuntos de dados provenientes de engenharia de software frequentemente são multimodais, com distribuição assimétrica, com caudas pesadas e com muitos outliers; aplicar testes estatísticos tradicionais a um conjunto de dados com as características citadas pode conduzir a erros, relacionados a aceitação e rejeição de hipóteses nulas, ou falhar em detectar importantes resultados.

O presente trabalho trata da análise estatística descritiva de um conjunto de dados experimentais obtidos do processamento do motor de execução da plataforma Guaraná. Estes dados não satisfazem os pressupostos para análise paramétrica, pois possuem distribuição assimétrica, caudas pesadas, multimodais, com muitos outliers e são heterocedásticos. O objetivo deste artigo é sumarizar e descrever este conjunto de dados conforme a estatística descritiva. Para isto, primeiramente, verificou-se a distribuição do conjunto de dados utilizando Diagrama de caixa e Gráficos de densidade de kernel e após utilizou-se diferentes medidas de tendência central de forma a poder selecionar a medida mais adequada, considerando a distribuição dos dados analisada.

O restante deste artigo está organizado em seções: na seção 2 são apresentados conceitos relacionados a estatística descritiva e motor de execução da plataforma de integração Guaraná, na seção 3 são apresentados o experimento, os procedimentos metodológicos e a forma como a análise e discussão dos dados é conduzida, na seção 4 são apresentados os resultasos e na seção 5 são tecidas as considerações finais e trabalhos futuros.

2 Referencial Teórico

Nesta seção, para um melhor entendimento da problemática abordada, alguns conceitos relacionados a estatística descritiva e ao motor de execução da plataforma de integração Guaraná são apresentados.

2. 1 Estatística Descritiva

As análises de dados em estatística podem ser realizadas em duas etapas: Estatística Descritiva e Estatística Inferencial. Estatística Descritiva é a etapa da estatística em que os dados a serem

analisados são organizados, descritos e sumarizados. As descrições podem ser gráficas, por distribuições de frequência ou por medidas. A Estatística Inferencial é a etapa em que são feitas as análises e interpretações, para tomada de decisão sobre os dados, com base em testes de hipóteses e estimação de parâmetros.

As análises, em ambas etapas, podem ser classificadas em univariada, quando as análises são encaminhadas de tal forma a verificar apenas um fator na amostra, bivariada quando as análises são realizadas para verificar como dois fatores impactam na amostra, e multivariada, quando se quer analisar como mais de dois fatores afetam a amostra. Consoante os objetivos da pesquisa apresentada neste artigo, a seguir são elucidadas as medidas e gráficos que foram utilizados para descrever a amostra da população.

2.2 Diagrama de Caixa e Densidade de Kernel

Nesta subseção são apresentadas duas ferramentas de visualização de dados, que permitem observar características como a distribuição de um conjunto de dados, a saber: diagrama de caixa (em inglês, boxplot) e densidade de Kernel.

Um diagrama de caixa é uma ferramenta de visualização de dados que fornece informações quanto à distribuição dos dados. Este diagrama é construído considerando quantis, o limite inferior e o limite superior. Os quantis são medidas de posição que particionam o conjunto de dados em partes iguais, por exemplo o quartil, particiona o conjunto em 4 partes iguais, percentil, particiona o conjunto em 100 partes iguais, e o decentil, particiona o conjunto em 10 partes iguais.

Para o diagrama de caixa são utilizados os quartis: o 1º quartil ($Q1$) coincide com o percentil 25, que significa que pelo menos 25% dos dados do conjunto são menores do que $Q1$ e que pelo menos 75% dos dados do conjunto são maiores do que $Q1$, o 2º quartil ($Q2$) coincide com o percentil 50 e também com a mediana do conjunto, que significa que pelo menos 50% dos valores são menores do que $Q2$ e 50% dos valores são maiores do que $Q2$, e o 3º quartil ($Q3$) coincide com o percentil 75, que significa que 75% dos valores do conjunto são menores do que $Q3$ e 25% dos valores são maiores do que $Q3$.

No diagrama de caixa, os limites inferior e superior são representados por hastes que partem da caixa: o limite inferior pela haste inferior e o limite superior pela haste superior. O limite inferior LI, o menor valor da amostra admitido para não ser um outlier é determinado como $LI = Q1 - 1,5 \cdot IIQ$, em que $IIQ=Q3-Q1$, que refere-se ao intervalo interquartil, uma medida de dispersão. O limite superior LS, o maior valor da amostra admitido para não ser um outlier é determinado como $LS = Q3 + 1,5 \cdot IIQ$. A constante 1,5 é escolhida para garantir que cerca de 86,64% dos dados serão capturados para mais e para menos, considerando a curva normal. O diagrama de caixa é muito útil para fornecer informações sobre valores discrepantes, os outliers, assim denominados os valores que são menores do que o valor do limite inferior e os valores que são maiores do que o limite superior.

Uma outra forma de visualização de dados é pela função de densidade de probabilidade. Conforme

Silverman (1986), uma função de densidade de probabilidade f descreve a distribuição natural de X. Esta relação é descrita pela integral abaixo.

$$P(a < X < b) = \int_a^b f(x)dx \quad \forall\, a < b$$

Estimação de densidade é uma estimativa da função de densidade de uma amostra de dados, realizada quando a distribuição do conjunto de dados é desconhecida A estimação de densidade pode ser paramétrica, em que é conhecido que a variável X pertence à uma família de distribuições paramétricas e então, são calculados e fornecidos alguns parâmetros que são suficientes para estimar a função f de distribuição de probabilidade; por exemplo a distribuição normal (ou gaussiana), em que são calculados a média μ e a variância σ^2 , substituídos na fórmula da distribuição normal, e então é conhecida a estimativa para a distribuição da variável X . Além disso, a estimação de densidade pode ser não-paramétrica, em que não é assumido o conhecimento de qual família de distribuições de densidade pertence a variável X. Dessa forma, a distribuição da variável X é calculada de forma mais livre do que pela abordagem paramétrica.

Estimações de densidade podem ser utilizados na exploração e apresentação dos dados para investigar as propriedades de um dado conjunto. Estimações de densidade são adequadas para indicar características como assimetria, caudas longas e multimodalidade no conjunto de dados. As estimações de densidade podem fornecer argumentos para conclusões ou apontar o tipo de análise que deve ser feita.

A estimativa da função de densidade pode ser feita com um estimador de kernel. A função de densidade de Kernel f , com um Kernel K é dada por:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h})$$

Em que h é o parâmetro suavizador, ou bandwidth (largura de banda). Por definição a função de Kernel K deve satisfazer a seguinte condição:

$$\int_{-\infty}^{+\infty} K(x)dx = 1$$

Conforme Silverman (1986), K geralmente é uma função de densidade de probabilidade simétrica, por exemplo o Kernel gaussiano, expresso como:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$$

Em seu trabalho, Silverman (1986) apresenta a função de densidade de Kernel para dois conjuntos

de dados distintos. Em cada um deles, é utilizado o Kernel gaussiano, e experimentado distintos valores para h afim de verificar como é afetada a função graficamente. Ao variar o parâmetro suavizador h, percebeu que se h é um valor muito pequeno, ficam mais evidentes as características locais do conjunto de dados, dificultando interpretações e inferências sobre as características mais globais e gerais do conjunto estudado; em contrapartida, para valores de h grandes, características podem ser mascaradas, por exemplo, se uma amostra é bimodal, para valor de h grande, esta características pode não ser evidenciada.

No diagrama de caixa e no gráfico de densidade de kernel é possível verificar a amplitude dos dados, a diferença entre os valores da média e mediana, a moda, a distribuição dos dados, a presença de outliers, o tamanho das caudas da distribuição e se as caudas da distribuição são pesadas. No entanto, apesar de que o diagrama de caixa é uma boa ferramenta para comparar distintos grupos amostrais, em Kitchenham et al (2017), os autores demonstram que o gráfico de densidade de Kernel é mais indicado para explorar a distribuição dos dados. Em uma de suas amostras, os autores verificaram que, após a transformação logarítmica, a amostra no diagrama de caixa indicava ter distribuição normal, no entanto, ao construir o gráfico de densidade de Kernel, a distribuição era diferente da distribuição normal. Assim, para potencializar a descrição dos dados é interessante analisar as informações do diagrama de caixa e do gráfico de densidade de Kernel juntos.

2.3 Medidas de Posição e Medidas de Dispersão

No que tange as medidas, estas podem ser classificadas em medidas de posição e medidas de dispersão. Enquanto as medidas de posição avaliam para qual valor os dados amostrais tendem, as medidas de dispersão avaliam quão dispersos os dados amostrais estão das medidas de tendência central. A média, a mediana e a moda são medidas de posição. As medidas de dispersão mais utilizadas são a amplitude, a variância, o desvio padrão e o coeficiente de variação.

A média, a mediana e a moda são medidas de posição, que servem para localizar a distribuição de frequência sobre uma variável em questão. A média e a mediana são também medidas de tendência central; são métricas que localizam o centro da distribuição de frequências. No entanto, a mediana é calculada considerando um conjunto ordenado e procura determinar um valor central que divide o conjunto em duas partes com igual número de elementos. Teoricamente, a mediana é uma medida de tendência central menos afetada pela presença de outliers, pois seu valor é calculado considerando, no máximo, os dois valores centrais de um conjunto ordenado (COSTA NETO, 2002). Se ordenado, os outliers presentes na amostra localizam-se nas posições iniciais e/ou finais.

Para avaliar o impacto do maior valor de um conjunto de dados de Engenharia de Software no cálculo de medidas de tendência central, Kitchenham et al (2017) calcularam distintas medidas de posição considerando o conjunto inteiro e o conjunto sem o maior valor e, foi medida, em porcentagem, quantas unidades a medida de posição foi deslocada, bem como o erro padrão de cada métrica, quando possível, para o conjunto inteiro e o conjunto sem o maior valor. O erro

padrão é uma medida para avaliar a confiabilidade das medidas de tendência central, calculado como $SE = \frac{sd}{\sqrt{n}}$, em que sd é o desvio padrão e né a quantidade de elementos do conjunto. Como resultado, identificaram que a mediana teve uma variação de 11%, maior do que o esperado.

Existem métodos para lidar com problemas ligados à contaminação de amostras por outliers, estes contribuem para condições de não normalidade em uma amostra e para que grupos amostrais não possuam a mesma variância. Um método comum é remover os outliers e então calcular a média dos valores remanescentes e utilizar a variância, no entanto, Wilcox and Keselman (2003) pontuam que este método baseado na média pode falhar ao detectar outliers e que ao simplesmente desconsiderar os valores extremos, os dados remanescentes não são tão independentes, o que invalida o cálculo do erro padrão.

O método trimmed mean é um método para cálculo de medida de tendência central baseado na remoção de $X\,\%$ de menores e de maiores valores de um conjunto de dados. Wilcox (2012, apud KITCHENHAM, 2017) comenta que a melhor escolha para a porcentagem de remoção é desconhecida, mas a porcentagem padrão é 20%. No entanto, é interessante escolher a porcentagem de remoção $X\,\%$ com base em uma investigação e exploração de como é a distribuição dos dados, por meio de diagrama de caixa e gráfico de densidade, para averiguar a presença e a quantidade de outliers. Também pode ser utilizado um método para detecção de outliers, proposto por Wilcox (2012, apud KITCHENHAM, 2017).

Para calcular a medida de tendência central de um conjunto de dados pelo método trimmed mean, primeiro deve-se realizar o trimming do conjunto de dados. O trimming é a etapa em que é retirada a porcentagem de menores e maiores valores, como segue:

Primeiramente, os dados do conjunto precisam ser ordenados em ordem crescente. Definida porcentagem de remoção de dados, são calculados os índices inferior e superior, cujos índices irão dividir o conjunto de dados em quantis. O índice i_{bottom} é o índice inferior, cujos dados dos índices menores do que i_{bottom} serão descartados e o índice i_{top} é o índice superior, cujos dados dos índices maiores serão descartados.

$$i_{bottom} = floor(0{,}0X \cdot N) + 1$$

$$i_{top} = N - i_{bottom} + 1$$

Em que, $0{,}0X = X/100$, N é a quantidade de dados do conjunto original e $floor$ trunca o valor para o primeiro inteiro inferior. Então, os dados remanescentes fazem parte do quantil $i_{bottom} \leq i_x \leq i_{top}$. Realizado o trimming, para o trimmed mean calcula-se a média aritmética dos dados remanescentes.

Para obter a variância do trimmed mean, é necessário, no conjunto original, realizar o método winsorizing, que consiste em repor os dados dos índices excluídos pelo dado correspondente à i_{bottom}, se os dados excluídos correspondem aos índices menores do que i_{bottom}, ou à i_{top}, se os

dados excluídos correspondem aos índices maiores do que i_{top} .

A variância do trimmed mean s_{tr}^2 é calculada por:

$$s_{tr}^2 = \frac{s_w^2}{N(1 - \frac{X}{100})}$$

Em que s_w^2 é a variância de winsorized mean, calculado na forma padrão. O erro padrão do trimmed mean é s_{tr}^2.

2.4 Motor de Execução da Plataforma de Integração Guaraná

Durante a execução de um processo de integração, quando uma tarefa recebe em suas entradas todas as mensagens necessárias para habilitar seu processamento, o motor gera uma work unit na fila de tarefas prontas para serem processadas. A work unit é uma tarefa associada as suas mensagens de entrada que será executada quando houver threads disponíveis para processá-la. Considerando a solução de integração do Café que é composta por 14 tarefas e 6 portas, uma mensagem é considerada processada se o motor de execução processou as 20 work units produzidas para esta mensagem. O procedimento ocorre da seguinte forma: uma mensagem entra no fluxo de integração através de uma porta, e para esta atividade acontecer, é produzida uma work unit referente a esta porta e a esta mensagem. Logo que produzida, esta work unit é alocada em uma fila de work units, se é a primeira na fila e há threads disponíveis ela é executada. Assim que executada é produzida uma nova work unit correspondente à próxima tarefa que precisa ser executada. Esta é enviada para a fila de work units, se é a primeira na fila e há threads disponíveis ela é executada; e assim, por diante. Salienta-se que neste modelo de execução, baseado em tarefas, uma mesma thread não necessariamente irá processar todas as work units produzidas para uma mesma mensagem.

3 Metodologia

Os dados utilizados neste artigo são observações sobre o processamento de work units pelo motor de execução da plataforma de integração Guaraná, versão 1.4. O experimento foi projetado por outros autores, pertencentes ao Grupo de Pesquisa em Computação Aplicada, e executado em uma máquina com 16 processadores Intel Xeon CPU E5-4610 V4, 1.8 GHz, 32 GB, operando sistema Windows Server 2016 Datacenter 64 bits.

No experimento para avaliar o processamento de work units pelo objeto de estudo motor de execução da plataforma foram testados o fator quantitativo taxa de entrada de mensagens (msg/), com 41 níveis, e o fator quantitativo threads, com 50 níveis, e 25 repetições. O experimento fatorial é em Delineamento Inteiramente Casualizado (DIC) 41x50x25.

Variáveis Independentes: As variáveis independentes referem-se a taxa de chegada de mensagens e o número de threads. A taxa de chegada de mensagens refere-se à carga de mensagens que são

enviadas ao motor de execução para serem processadas. As threads são os recursos computacionais disponíveis para que o motor execute uma solução de integração.

Variáveis Dependentes ou Variáveis-Resposta: As variáveis dependentes são o número de work units processadas (quantitativo discreto) e as mensagens processadas (quantitativo discreto). As work units são estruturas associadas a cada tarefa da solução de integração, para cada mensagem devem ser processadas 20 work units. As mensagens processadas são as mensagens que entraram no fluxo de integração por meio da taxa de chegada de mensagens e que foram completamente executadas.

Um parâmetro do experimento é o tempo de execução que refere-se ao tempo estipulado para que o motor execute o processo, que neste caso são 60 segundos. Neste aspecto, ao referir-se à uma mensagem completamente processada, significa que foi totalmente processada antes que o tempo de execução se esgotasse. Logo o tempo total de processamento de uma mensagem é de no máximo 60 segundos. Para cada uma das 25 repetições de cada combinação de taxa de chegada de mensagens e threads, o motor executou a solução durante 60 segundos.

A população é formada por 51250 observações para cada variável-resposta. Para este artigo, foi analisada uma amostra dessa população, formada por 20500 observações. Esta amostra refere-se às observações da variável-resposta processamento de work units, testadas com os 41 níveis do fator taxa de entrada de mensagens e 20 níveis do fator thread, e 25 repetições. O tipo de análise realizada será a univariada, serão descritos e sumarizados a amostra quanto ao fator thread.

A apresentação de resultados e discussão será apresentada em três etapas. Na subseção 4.1, etapa 1, são analisados os dados referentes às 25 repetições de cada combinação de níveis dos fatores analisados. A análise consiste em construir o diagrama de caixa e o gráfico de estimativa de densidade de Kernel, construídos com a linguagem R, utilizando RStudio (TEAM, 2015) e o pacote ggplot2 (WICKHAM, 2016). A análise é feita para verificar a distribuição das 25 repetições do processamento de work units de cada combinação de níveis e mostrar a presença de outliers, justificando a escolha da medida de tendência central trimmed mean 10%. Para facilitar a visualização detalhada da “dispersão” entre as observações para as diferentes quantidades de threads, estas foram divididas em dois conjuntos de dados. O conjunto 1 conterá os níveis do fator thread 2, 4, 6, 8, 10, 12, 14, 16, 18 e 20, que será referido como ‘conjunto 1’ e o conjunto 2 que conterá os níveis de thread 22, 24, 26, 28, 30, 32, 34, 36, 38 e 40, referido como ‘conjunto 2’.

Em cada uma das figuras, há 4 imagens diferentes, em que as imagens a,b referem-se aos dados sobre processamento de work units com uma determinada taxa de entrada de mensagens e o conjunto 1 e as imagens c,d referem-se aos dados sobre processamento de work units com uma determinada taxa de entrada de mensagens e o conjunto 2. As imagens a,c mostram os diagrama de caixa e as imagens b,d mostram os gráficos de densidade de kernel. Nos diagramas de caixa também são representados como os dados (pontos vermelhos) se distribuem sobre o gráfico.

Determinada esta medida, há a subseção 4.2, referente à etapa 2, em que serão analisados, por

meio do diagrama de caixa e do gráfico de densidade de Kernel, os dados sobre o processamento de work units em relação a cada nível do fator thread da amostra selecionada, para determinar qual medida será utilizada para representar o processamento de work units considerando os níveis do fator thread. As figuras mostradas na subseção 4.2 seguem o mesmo padrão da subseção 4.1.

Na subseção 4.3, etapa 3 de análise, é apresentada uma tabela para auxiliar na apresentação dos valores de 3 medidas de tendência central, média, mediana, trimmed mean 10% e trimmed mean 20%, e seus respectivos erros-padrão, da 2ª à 9ª coluna. Em cada linha, da 2ª à 21ª, são mostrados os valores das medidas para cada nível do fator thread.

4 Resultados e Discussão

Nesta seção são mostrados os diagramas de caixa e gráficos de densidade de Kernel considerando a amostra do processamento de work units pelo motor de execução da plataforma de integração Guaraná.

4.1 Descrição de dados amostrais em cada combinação de níveis de fatores thread e taxa de entrada de mensagens

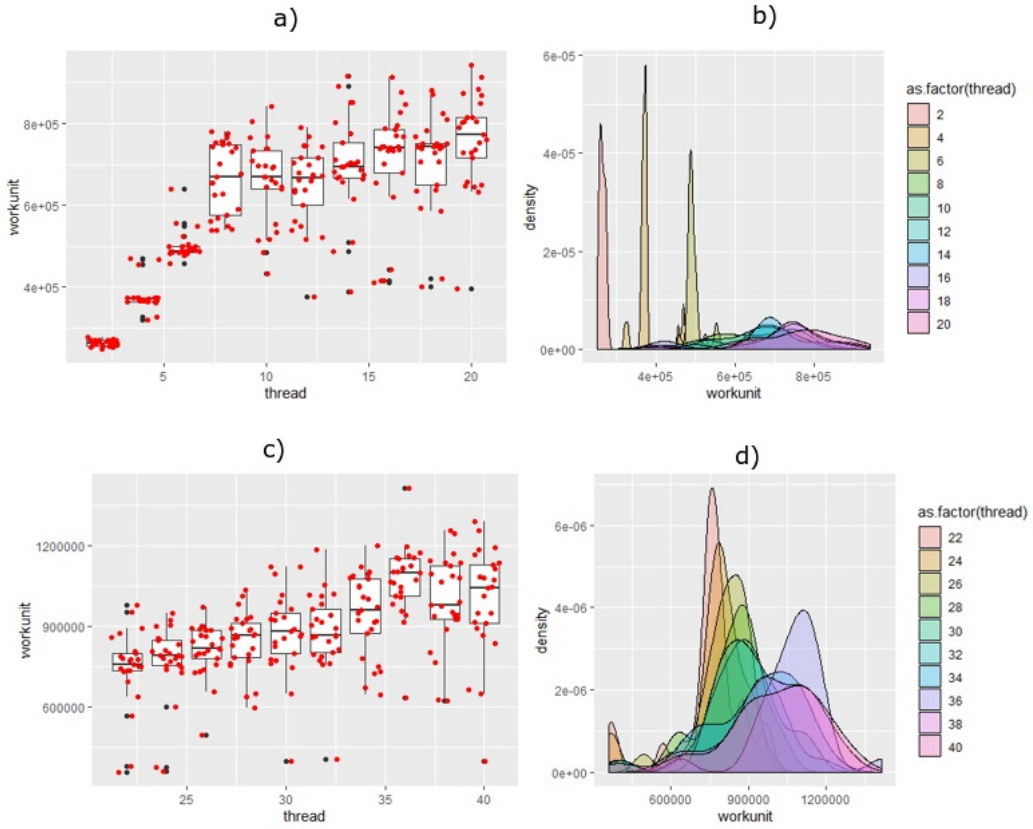
Nesta subseção são apresentados os diagramas de caixa e gráficos de densidade de kernel para os dados sobre o processamento de work units considerando dois casos. Foram selecionados dois casos para mostrar como foi realizada a análise que se estendeu aos demais níveis do fator taxa de entrada. O primeiro caso está relacionado ao processamento de work units com a taxa de 2000 msg/s e variando a quantidade de threads disponíveis no pool global. O segundo caso está relacionado ao processamento de work units com a taxa de 8000 msg/s e variando a quantidade de threads disponíveis no pool global.

Na figura 1 são analisados os dados sobre o processamento de work units com uma taxa de entrada de mensagens 2000 msg/s. Na figura 1.a pode-se observar que os dados sobre o processamento de work units com 2, 4 e 6 threads são menos distribuídos, logo são mais concentrados em um valor; com as demais quantidades de threads do conjunto 1, os dados sobre o processamento de work units são mais distribuídos. Nota-se também que há vários outliers, representados por pontos pretos. Ao comparar com o gráfico na figura 1.b percebe-se que o processamento de work units com 2, 4 e 6 threads realmente concentra-se em uma determinada quantidade. Além disso, pode-se perceber que para as demais quantidades de threads do conjunto 1, os dados são muito distribuídos e que a presença de outliers, valores discrepantes menores, contribui para que a distribuição de densidade dos dados analisados tenha caudas esquerdas pesadas. Portanto, a distribuição de densidade de cada um dos 10 conjuntos de dados não é gaussiana.

Quanto aos dados do processamento de work units com o conjunto 2 de threads, nota-se pela Figura 1.c, que os dados são distribuídos de forma equilibrada, mas há a presença de outliers, principalmente valores inferiores. Na Figura 1.d, percebe-se que as amostras possuem

distribuição com caudas pesadas à esquerda, devido à presença de outliers inferiores, além disso, é possível observar em alguns gráficos que a amostra possui duas modas (multimodal), por exemplo o processamento de work units com 38 threads disponíveis ao pool global.

Figura 1: Diagrama de caixa e gráficos de densidade de kernel de dados de processamento de work units com taxa de 2000 msg/s.



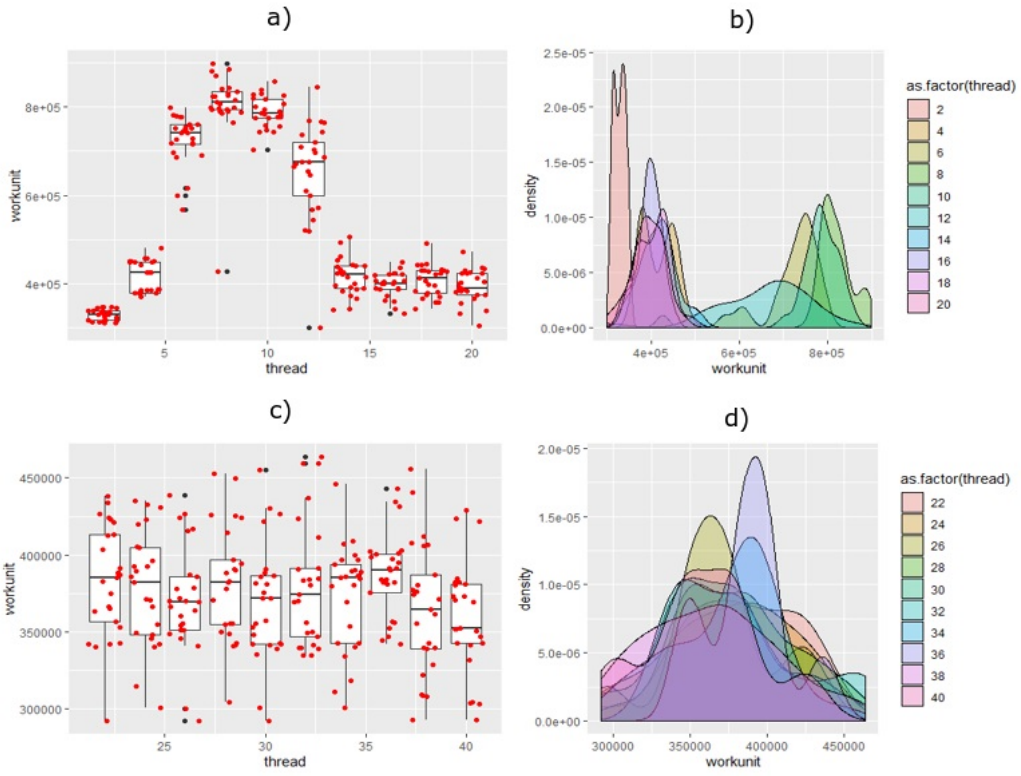
Fonte: Próprio autor

Na figura 2 são analisados os dados sobre o processamento de work units com a taxa de entrada de mensagens de 8000 msg/s e os dois conjuntos de threads. Na Figura 2.a, verifica-se que os conjuntos de dados representados são pouco distribuídos, entretanto também são detectados outliers, a maioria outliers inferiores. Na Figura 2.b, vê-se nos gráficos de densidade que a distribuição de cada amostra não é gaussiana, e revela como, pelo formato da curva como a presença de outliers contribui para caudas pesadas à esquerda. Comentando especificamente os dados sobre o processamento de work units com 4 threads alocadas no pool global, percebe-se pelo apresentado na Figura 2.a que é possível que o conjunto seja multimodal, neste caso com duas modas, e esta possibilidade é reforçada ao observar a distribuição dos dados na Figura 2.b.

Sobre os dados de processamento de work units com o conjunto 2, percebe-se que os dados de que cada conjunto são bastante distribuídos, possuem grande amplitude, e possuem outliers, principalmente outliers superiores.

Pelos gráficos de densidades, tanto as imagens b,d na figura 1 quanto as imagens b,d na figura 2, mostra que os conjuntos de dados entre si não possuem a mesma variância, ou seja, não são homocedásticos.

Figura 2: Diagramas de caixa e gráficos de densidade de kernel de dados brutos de processamento de work units com taxa de 8000 msg/s.



Fonte: Próprio autor

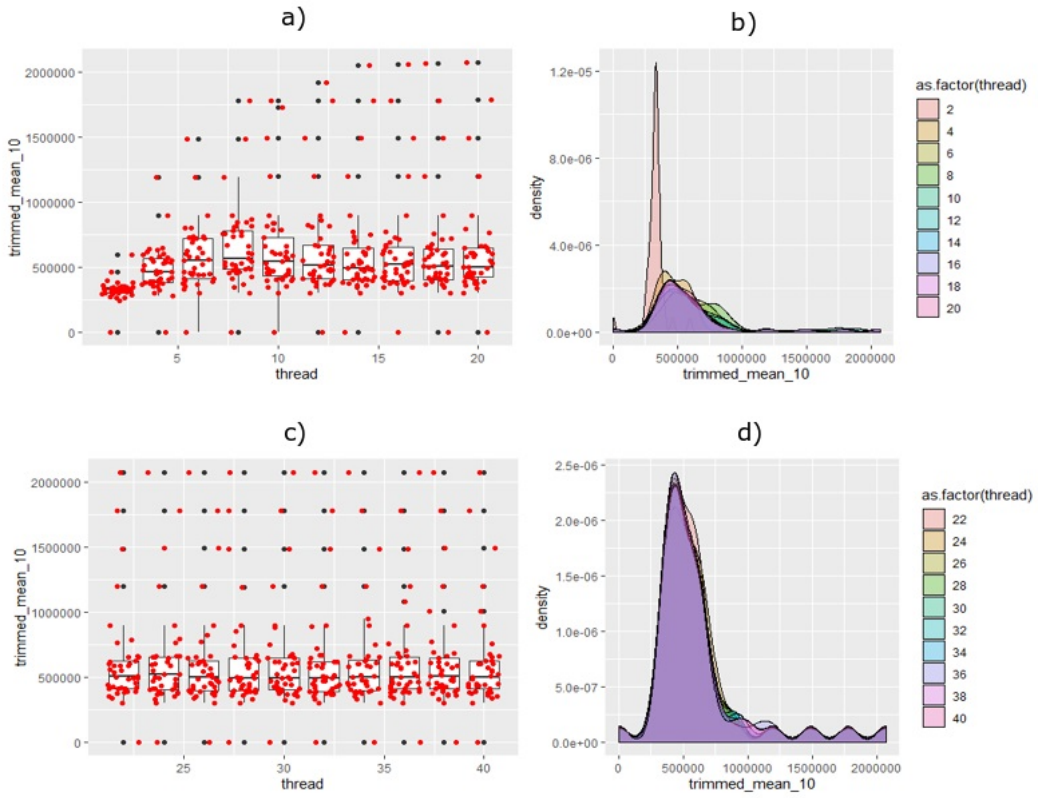
Considerando as características dos dados, discutidas acima, optou-se pela medida de posição para cada nível do fator thread será calculada denominada trimmed mean 10%, em outros termos, 10% dos menores valores e 10% dos maiores valores são retirados de cada conjunto de combinação de taxa de entrada de mensagem e thread, que contém 25 repetições. Assim, são retirados 4 dados (2 menores e 2 maiores). Foi escolhida a taxa de retirada de 10% pois verificou-se que na maioria dos diagramas de caixa estavam representados outliers, entretanto, a quantidade de outliers em cada conjunto não é superior a 2, que representa 10% de cada amostra

analisada.

4.2 Descrição de dados amostrais em relação a cada nível do fator thread

Após decidir por uma medida de tendência central que representa o processamento de work units em cada combinação de taxa de entrada de mensagens, é necessário averiguar as características do novo conjunto de dados. Assim, é analisado as trimmed means 10% de cada processamento de work units de acordo com as quantidades de threads analisadas. Dessa forma, cada nível do fator threads representa um conjunto com 41 dados, cada dado refere-se à trimmed mean 10% de cada taxa de entrada de mensagens. A análise é feita a partir dos diagramas de caixa e densidades de kernel para cada nível do fator thread, mostrados na figura 3.

Figura 3: Diagrama de caixa e gráficos de densidade de kernel de dados referentes à trimmed mean de cada nível do fator thread, considerando 41 taxas de entrada de mensagens em cada conjunto



Fonte: Próprio autor

Sobre o conjunto 1, na Figura 3.a, é possível perceber que os dados não são tão distribuídos, mas

há presença outliers inferiores e principalmente superiores. Na Figura 3.b, percebe-se que a distribuição dos dados é assimétrica e que possui uma cauda direita pesada, causada pela presença de outliers superiores. Além disso, as distribuições dos dados amostrais do processamento com os níveis de 2 à 12, possuem formato não semelhante entre si. Ainda nessa figura, é possível notar que à medida que os grupos se referem à quantidade maior de threads utilizada para processamento, tal como os níveis de 14 à 20, os dados amostrais tendem a ter uma mesma distribuição e que os valores com maior probabilidade de ocorrer em cada distribuição convergem para o mesmo, cerca de 500000 work units.

Na Figura 3.c, observa-se que os diagrama de caixa são semelhantes aos diagramas de caixa da Figura 3.a, inclusive quanto à presença de outliers. Na Figura 3.d, é possível visualizar que as distribuições possuem formato muito semelhante entre si, e que o valor com maior probabilidade de ocorrência converge para o mesmo.

4.3 Seleção de medida de tendência central

Na tabela 1 são mostradas 4 métricas de tendência central e seus respectivos erros padrão para cada nível do fator threads.

Tabela 1: Dados referentes à distintas medidas de tendência central e respectivos erros padrão para cada nível do fator thread estudado

Threads	$\bar{x}$	$SE\bar{x}$	*md*	*SEmd*	*tr10%*	*SEtr10%*	*tr20%*	*SEtr20%*
T2	327333	12005.4	330445	6132.62	327524	6299.77	328794	5092.57
T4	485752	28233.8	462095	31845.1	474054	19882.7	469829	22581.6
T6	574836	38806.5	550777	49533.8	551770	30163.6	547780	35240
T8	653815	47990.3	564420	52727.3	619154	32611.1	612366	38209.2
T10	634925	55760.7	543517	41720.5	570763	34973.1	554016	38886.4
T12	630430	58486.5	517527	41518.4	558888	35730.8	540113	40172.9
T14	611104	60238.6	495168	37789.1	531366	34286.1	514279	31442.6
T16	611580	60286.5	521060	39089.8	531997	33966.5	517642	31172.4
T18	610721	60225.4	506849	37147.4	530412	33534.9	515395	28739.1
T20	617174	60189.1	503360	32315.6	538344	33192.4	524285	27759.4
T22	615114	60148.8	504145	35584.8	536125	32922.9	521325	26283.7
T24	617018	60435.2	521324	46472.5	538499	34018.9	525945	29314.6
T26	609408	60632	496428	36847.6	528925	34338.9	512116	28288.1
T26	609906	60706.7	493845	38612.2	529550	34507.5	512402	28363.2
T30	611248	60830.9	490721	38330.2	531664	34671.9	513332	28755
T32	607318	60910.4	496064	37959.5	526358	34619.3	507996	29152
T34	612995	61004.1	500597	39100.2	533758	37124.4	512946	29218.4
T36	620324	61427.3	500581	37878.4	542534	43732.4	516902	30436
T38	617136	61161	504000	36689.6	538457	40234.8	516150	30235.4
T40	614183	61178.2	501423	33358.1	534999	40367.1	512508	29097.6

Fonte: Próprio autor

Considerando as características discutidas nas etapas 1 e 2 de análise dos dados do processamento de work units foi selecionada uma medida de tendência central para representar cada nível do fator thread. Como as amostras possuem outliers, não é possível utilizar como medida a média de cada conjunto, pois esta medida é altamente influenciada por valores extremos. Neste caso, a média é elevada pela presença de outliers superiores. A mediana é considerada robusta quanto à presença de outliers, no entanto, a linha central que marca o valor da mediana nos diagramas de caixa, é levemente descentralizada, indicando que os valores amostrais tendem a ser mais próximos da origem. Essa característica é corroborada pelas distribuições nas Figuras 3.b.d.

A medida trimmed mean 10% retira 10% dos menores valores e 10% dos maiores valores (total de 20% de dados excluídos) enquanto que a medida trimmed mean 20% retira 20% dos menores valores e 20% dos maiores valores (total de 20% de dados excluídos). Optou-se por escolher a medida trimmed mean 20% devido a quantidade de outliers superiores ser próxima a 20% da quantidade de elementos em cada grupo e devido a esta medida, em geral, possuir erro padrão menor do que os erros padrão das demais medidas.

5 Considerações Finais

Neste artigo foi realizada a descrição e análise univariada, correspondente à etapa de estatística descritiva, de uma amostra de Engenharia de Software, utilizando diagrama de caixa e gráfico de densidade de kernel para visualização da distribuição dos dados amostrais. Após observar a distribuição dos dados, suas propriedades e características fundamentaram a escolha de uma medida de tendência central para cada nível do fator thread.

Como resultado da análise, identificou-se que alguns grupos possuem distribuições distintas entre si, apesar desta quantidade ser relativamente pequena, há evidências que indicam heterocedasticidade entre os grupos. Além disso, é evidente que a distribuição de cada grupo não é gaussiana (normal). Portanto, na fase de Estatística Inferencial deverão ser adotados métodos que não exigem pressupostos paramétricos; testes estatísticos robustos à heterocedasticidade entre os grupos e a não normalidade das distribuições. Também foi observado que nos grupos amostrais há a presença de outliers e que a medida de tendência central mais adequada é trimmed mean 20%.

Como resultados futuros, testes serão realizados a fim de averiguar se há diferenças significativas entre o processamento de work units para cada nível do fator thread. Se confirmada a diferença significativa, serão aplicados testes post-hocs, para identificar quais níveis diferem entre si. Ademais, pretende-se estabelecer grupos de níveis do fator taxa de entrada de mensagens e, nestes grupos, analisar o processamento de work units.

Agradecimentos

O presente trabalho foi realizado com apoio do CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil.

Referências

BASILI, V. R.; SELBY, R. W.; HUTCHENS, D. H. **Experimentation in software engineering**. IEEE Transactions on software engineering, IEEE, n. 7, p. 733–743, 1986.

COSTA NETO, P. L. O. **Estatística**. 2ª edição. São Pulo: Editora Edgard Blücher LTDA, 2002.

FRANTZ, R. Z. **Enterprise application integration: an easy-to-maintain model-driven engineering approach**. Diss. Universidad de Sevilla, 2012.

KITCHENHAM, B. *et al*. **Robust statistical methods for empirical software engineering**. Empirical Software Engineering 22.2 (2017): 579-630.

LINTHICUM, D. S. **Enterprise application integration**. Addison-Wesley Professional, 2000.

MAIR, P.; WILCOX, R. **Robust statistical methods in r using the wrs2 package**. Harvard Univ (2016).

PERRY, D. E.; PORTER, A. A.; VOTTA, L. G. **Empirical studies of software engineering: a roadmap**. In: ACM. Proceedings of the conference on The future of Software engineering. [S.l.], 2000. p. 345–355.

RITTER, D.; MAY, N.; RINDERLE-MA, S. **Patterns for emerging application integration scenarios: A survey**. Information Systems 67 (2017): 36-57.

RSTUDIO TEAM (2015). **RStudio: Integrated Development for R. RStudio**, Inc., Boston, MA, URL http://www.rstudio.com/.

SILVERMAN, B. **Density Estimation for Statistics and Data Analysis**. Monographs on Statistics and Applied Probability, London, 1986.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York, 2016.

WILCOX, R. R.; KESELMAN, H. J. **Modern robust data analysis methods: measures of central tendency**. Psychological methods 8.3 (2003): 254.