ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO
MESTRADO EM CIÊNCIAS DA COMPUTAÇÃO

MARIANA NOLDE PACHECO DETONI

MAPEAMENTO E APLICAÇÃO DE TESTES ESTATÍSTICOS EM ENGENHARIA DE SOFTWARE

Porto Alegre
2020

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MAPEAMENTO E APLICAÇÃO DE TESTES ESTATÍSTICOS EM ENGENHARIA DE SOFTWARE

MARIANA NOLDE PACHECO DETONI

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Dr. Afonso Henrique Corrêa de Sales

Porto Alegre
2020

Ficha Catalográfica

D482m Detoni, Mariana Nolde Pacheco

Mapeamento e Aplicação de Testes Estatísticos em Engenharia de Software / Mariana Nolde Pacheco Detoni . – 2020.

110 p.

Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientador: Prof. Dr. Afonso Henrique Corrêa de Sales.

1. Estatística. 2. Teste Estatístico. 3. Engenharia de Software. 4. Pesquisa Quantitativa. I. Sales, Afonso Henrique Corrêa de. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

Mariana Nolde Pacheco Detoni

Mapeamento e Aplicação de Testes Estatísticos em Engenharia de Software

Tese/Dissertação apresentada como requisito parcial para obtenção do grau de Doutor/Mestre em Ciência da Computação do Programa de Pós-Graduação em Ciencia da Computação, Escola Politécnica da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovado em 20 de Março de 2020.

BANCA EXAMINADORA:

Prof. Dra. Tayana Uchôa Conte (UFAM)

Profa. Dra. Sabrina dos Santos Marczak (PUCRS)

Prof. Dr. Afonso Henrique Corrêa de Sales (PPGCC/PUCRS - Orientador)

DEDICATÓRIA

Dedico este trabalho aos meus pais e meu esposo, que não mediram esforços para me apoiar e me ajudar nessa etapa da minha vida.

“A flor que desabrocha na adversidade é a mais rara e mais bela de todas.”
(Mulan)

AGRADECIMENTOS

Para realizar esse trabalho, contei com o apoio de muitas pessoas e instituições aos quais estou profundamente grata. Mesmo considerando esta a parte mais difícil de escrever (porque a importância e significância das pessoas em nossas vidas não se avalia através de testes estatísticos...) e correndo o risco de injustamente não mencionar alguém quero deixar expresso os meus agradecimentos:

Aos pesquisadores que participaram do estudo, por aceitarem contribuir e por acreditarem no trabalho que foi desenvolvido.

Aos docentes e aos funcionários do Programa de Pós-graduação em Ciências da Computação da PUCRS, pelas orientações, apoio e conhecimento que compartilharam comigo.

À professora doutora Sabrina Marczak, que desde a apresentação do PEP me auxiliou muito com dicas e sugestões para o desenvolvimento desse estudo.

À professora doutora Tayana Conte, que veio de longe para auxiliar e contribuir na apresentação da presente dissertação.

Aos colegas do Programa de Pós-graduação em Ciências da Computação da PUCRS pelo apoio e auxilio desde o início do curso (nas disciplinas do mestrado) e com apoio e dicas para a escrita da dissertação. Para não correr o risco de esquecer alguém não vou citar nomes, mas desde já os meus agradecimentos.

Aos meus colegas de trabalho (antigos e atuais), que me auxiliaram nos momentos de ausência e me incentivaram desde o início do curso até as etapas finais.

Aos meus amigos que compreenderam a ausência neste período de dedicação aos estudos e pesquisa.

Aos meus pais Clarice e André pelo incentivo, apoio e compreensão. Desde sempre vocês estiveram do meu lado e dedicaram tudo que foi possível para me proporcionar grandes oportunidades. Vocês são a base de tudo.

Ao meu amado esposo pelo amor, incentivo, inspiração, compreensão e paciência em todos os momentos. Te agradeço não somente pelo amor e pelo apoio, como também por ter sido o grande incentivador para que eu me arriscasse e fosse atrás dos meus sonhos.

Por fim, ao meu orientador, Doutor Afonso Sales, por me incentivar, me auxiliar e principalmente por acreditar em mim. Obrigada pela grande oportunidade em me guiar durante todo este processo.

MAPEAMENTO E APLICAÇÃO DE TESTES ESTATÍSTICOS EM ENGENHARIA DE SOFTWARE

RESUMO

A Engenharia de Software costuma desenvolver e indicar o uso de diferentes ferramentas para construção de uma solução computacional. Além disso, um dos objetivos da área é a busca pela melhoria dos processos. Através da compreensão e da modificação adequada sobre os processos de engenharia de software existentes, pode ser possível reduzir os custos, reduzir o tempo de desenvolvimento e melhorar a qualidade dos produtos de software. Uma das formas de acompanhamento e avaliação dos processos da Engenharia de Software é através de avaliações quantitativas. Nas avaliações quantitativas, uma das formas mais usuais de analisar informações é através de técnicas estatísticas. Alguns estudos relataram desafios na utilização de técnicas de análise em Engenharia de Software, que abrange o pouco conhecimento sobre o tema, além de dificuldades na coleta e análise dos resultados obtidos. Com isso, esse estudo busca auxiliar os pesquisadores e profissionais da área de Engenharia de Software a conhecer mais os conceitos de análise de dados estatísticos e conseguir identificar possíveis tipos de técnicas estatísticas para realizar a análise dos seus estudos quantitativos. Para isso, foram construídos diferentes fluxogramas para escolha do testes estatístico na área de Engenharia de Software. Os fluxogramas foram construídos e avaliados com pesquisadores da área, visando uma aproximação e avaliação das principais necessidades.

Palavras-Chave: Estatística, Teste Estatístico, Engenharia de Software, Pesquisa Quantitativa.

MAPPING AND APPLICATION OF STATISTICAL TESTS IN SOFTWARE ENGINEERING

ABSTRACT

Software Engineering usually develops and indicates the use of different tools to build a computational solution. In addition, one of the goals of the area is improvement for process. Through the understanding and appropriate modification of existing Software Engineering processes, it may be possible to reduce costs, reduce development times and improve the quality of software products. One way of monitoring and evaluating Software Engineering processes is through quantitative assessments. In quantitative assessments, one of the most common ways of analyzing information is through statistical techniques. Some studies have reported challenges in using analysis techniques in Software Engineering, which encompasses little knowledge on the subject, in addition to difficulties in collecting and analyzing the results obtained. With this, this study seeks to help researchers and professionals in the Software Engineering area to learn more about the concepts of statistical data analysis and to be able to identify possible types of statistical techniques to carry out the analysis of their quantitative studies through the use flowcharts of statistical tests in the area of Software Engineering. The flowcharts were built and evaluated with researchers in the area, with the aim of approximating and evaluating the main needs.

Keywords: Statistics, Statistical Test, Software Engineering, Quantitative Research.

LISTA DE FIGURAS

SUMÁRIO

1. INTRODUÇÃO

Os aspectos de produção de um software (que variam desde as etapas iniciais de especificação até a manutenção de um sistema após a implementação) são de responsabilidade da área de Engenharia de Software. Um dos objetivos da área é a busca pela melhoria dos processos. Através da compreensão e da modificação adequada sobre os processos existentes de engenharia de software, pode ser possível reduzir os custos, reduzir o tempo de desenvolvimento e melhorar a qualidade dos produtos de software [129].

A Engenharia de Software costuma desenvolver soluções para diferentes áreas técnicas de conhecimento e em empresas de diferentes portes (pequenas, médias ou grandes empresas). Para isso, a área indica a possibilidade de uso de diferentes ferramentas para construção de um software e diversas técnicas para que essa construção ocorra da maneira mais padronizada e documentada possível [32].

Para melhoria dos processos em engenharia de software, um dos aspectos fundamentais é a correta avaliação dos sistemas computacionais projetados ou desenvolvidos. Existem diversas formas de analisar os diversos processos da Engenharia de Software, que variam desde avaliações qualitativas (onde há a comparação com o senso-comum ou com um referencial de base), até avaliações quantitativas (baseadas na formulação de valores específicos, sem considerar os méritos de geração desses valores) [32].

Uma das formas de analisar e avaliar os processos de gerenciamento, desenvolvimento, implementação, manutenção e avaliação de sistemas computacionais da Engenharia de Software é através de análises quantitativas, incluindo o uso de experimentos. O foco na análise de experimentos pode estar relacionado a fatores de melhoria de processos, que obtém avaliações mais precisas, rápidas e confiáveis dos dados ou processos estudados [32].

Uma das maneiras mais precisas de manejo e tratamento dos dados de análises quantitativas de dados é através de diferentes procedimentos estatísticos. As análises estatísticas auxiliam na coleta, classificação, sumarização, organização, análise e interpretação dos dados de experimentos [90].

Existem diferentes técnicas estatísticas para a manipulação e tratamento dos dados quantitativos. A escolha da técnica estatística a ser utilizada depende da natureza inicial dos dados. Dados categóricos ou binários exigem um tipo de análise, enquanto dados numéricos exigem outro tipo de abordagem. Além disso, como tratam-se de técnicas matemáticas, são necessárias algumas avaliações das características de natureza das informações, tais como distribuição de probabilidade dos dados, variabilidade entre as mensurações e outras medições. A correta utilização dessas técnicas produz resultados confiáveis e com possíveis métricas de erro mensuráveis, ajudando a garantir a melhor tomada de decisão em diferentes processos da Engenharia de Software [134].

Com isso, essa pesquisa visa a construção de um fluxo de testes estatísticos a serem aplicados na Engenharia de Software.

Além da apresentação do fluxograma, o presente estudo visa apresentar alguns tipos de análises quantitativas frequentemente utilizadas na Engenharia de Software e auxiliar na correta utilização e interpretação dos resultados dos testes estatísticos utilizados.

A apresentação da presente dissertação está dividida em dez capítulos. Os Capítulos 2, 3, 4 e 5 estão relacionados ao referencial teórico. O Capítulo 2 abrange a definição e uso de técnicas estatísticas na Engenharia de Software. Já no Capítulo 3 é fundamentada a pesquisa quantitativa e conceitos comumente utilizados em estatística. O Capítulo 4 apresenta os principais testes estatísticos e suas aplicações em diferentes tipos de análises quantitativas. Já no Capítulo 5 são apresentados trabalhos relacionados, que ajudarão a fundamentar a relevância do presente estudo. No Capítulo 6 é apresentada a proposta de pesquisa, com os objetivos do estudo e métodos empregados para atingimento dos resultados propostos pelo estudo. No Capítulo 7 é apresentada a Revisão Sistemática de Literatura, visando entender os tipos de testes estatísticos mais utilizados na Engenharia de Software. No Capítulo 8 é apresentada a construção e avaliação dos fluxogramas com os pesquisadores da área. No Capítulo 9 são apresentados exemplos de casos de uso dos fluxogramas propostos, tendo como base a própria literatura utilizada nessa pesquisa. Para finalizar, no Capítulo 10, são apresentadas as considerações finais do trabalho, incluindo os principais resultados encontrados, limitações do estudo e sugestões de trabalhos futuros.

2. DEFINIÇÃO E USO DE TÉCNICAS ESTATÍSTICAS NA ENGENHARIA DE SOFTWARE

De acordo com Pressman [111], os processos de software são abordagens adaptáveis que possibilitam que a equipe possa selecionar e decidir sobre o conjunto de ações e tarefas que pretende realizar para o desenvolvimento de um software. O objetivo desses processos é a entrega de um software de qualidade, que satisfaça a necessidade da equipe que desenvolveu e do usuário que irá utilizá-lo.

Os requisitos de um sistema mudam ao longo do processo e o impacto de cada mudança também varia no decorrer do tempo. Quando as mudanças ocorrem em processos iniciais, os impactos em custo são pequenos. Porém, conforme o tempo passa os custos podem aumentar, uma vez que recursos foram comprometidos e que podem ter ocorrido mudanças fundamentais e em etapas iniciais do projeto [111].

Os processos de gerenciamento, desenvolvimento, implementação, manutenção e avaliação de sistemas computacionais da Engenharia de Software, através de coleta de dados como os custos de desenvolvimento de software, defeitos de software, resistência na adoção de um novo sistema, qualidade de diferentes procedimentos ou outros tópicos, pode beneficiar o entendimento acerca de um problema de pesquisa e facilitar a criação de leis gerais ou teorias em diferentes áreas da computação [140].

Na Engenharia de Software, os métodos estatísticos podem ser ferramentas amplamente utilizadas em diferentes processos e atividades da área, auxiliando no planejamento, avaliação de melhorias futuras e no desenvolvimento de novos produtos e sistemas computacionais [59].

As pesquisas realizadas em Engenharia de Software não podem se preocupar somente com a coleta de dados, mas também com as análises, resultados e explicações sobre as informações coletadas. Por exemplo, para comprovar a eficácia de um sistema computacional, não é suficiente aplicar o sistema em dois grupos, como por exemplo, em um grupo de teste (grupo A) e o outro de controle (grupo B). Concluir baseado apenas em dados descritivos que a média do grupo A é superior ao grupo B pode ser equivocado. Esse resultado poderia ter ocorrido devido à casualidade, gerando conclusões errôneas que prejudicam a tomada de decisão [144].

Uma das definições muito utilizadas em análises quantitativas é a de experimento. O conceito de experimento abrange a identificação de potenciais variáveis que são manipuladas, mensuradas e com seus efeitos suficientemente controlados para observação de seus resultados em um estudo [84].

Os experimentos exigem etapas de planejamento, realização, análise e avaliação dos resultados obtidos. Essas etapas podem ser descritas através de uma metodologia

conhecida inicialmente como Método de Melhorias, desenvolvida pelo estatístico Walter A. Shewart na década de 30, que abrange conceitos de controle estatístico do processo.

Os conceitos de Shewart foram amplamente utilizados e popularizados na década de 50 por Edwards Deming, abrangendo etapas de planejamento, construção e avaliação de ciclos de desenvolvimento de projetos, e conhecidos como ciclo *PDCA* (do inglês: PLAN - DO - CHECK - ACT ou Adjust).

O ciclo *PDCA* é um método iterativo de controle e melhoria contínua de processos e produtos, cujas etapas podem ser vistas na Figura 2.1.

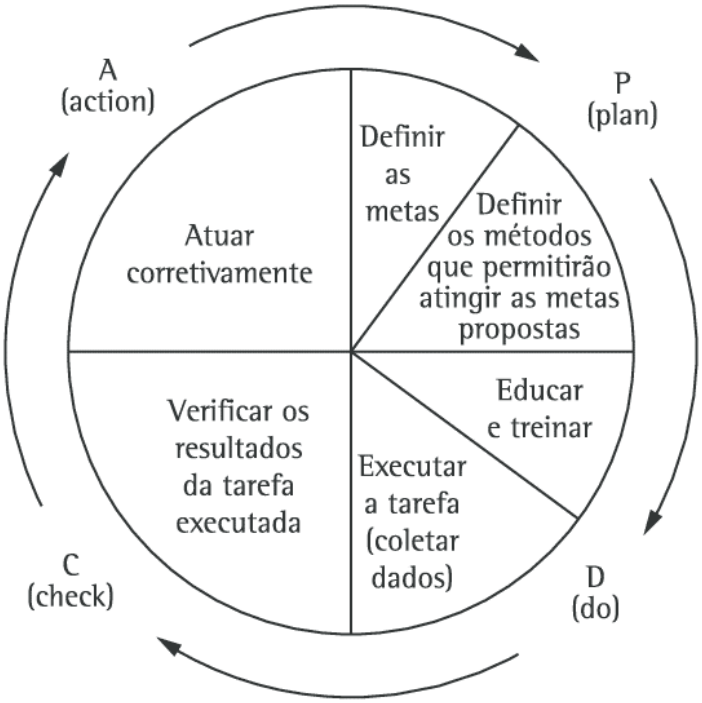

Figura 2.1 – Ciclo PDCA [106]

A primeira fase do processo (*PLAN* = Planejar) consiste em planejar o que será realizado, envolvendo a definição de objetivos, estratégias e ações. Além disso, essa fase consiste em definir os métodos que serão utilizados para atingir os objetivos propostos.

A segunda fase (*Do* = Executar) caracteriza-se pela implementação do processo planejado.

Na terceira fase (*CHECK* = Verificar) ocorre a avaliação do processo. Diferenças entre os resultados planejados e o resultado real alcançado constituem problemas a serem resolvidos. Nesta etapa ocorre a coleta de dados do processo e a comparação dos resultados obtidos com os resultados esperados. A análise dos dados desse processo fornece subsídios relevantes para a continuidade e implementação da próxima etapa.

A quarta fase (*Action* = Agir) consiste em fazer as correções necessárias do processo, com o objetivo de evitar repetições de problemas apresentados em outras fases.

Essa fase envolve a busca por melhoria contínua, propiciando a criação de novos conhecimentos e as atualizações necessárias.

Para o uso de testes estatísticos em Engenharia de Software, de forma semelhante ao ciclo *PDCA*, também são realizadas diversas etapas, que abrangem a delimitação do problema a ser estudado, a coleta de dados a ser realizada, a análise dos dados obtidos e a interpretação dos resultados e conclusões.

Na Figura 2.2, é apresentado um desenho esquemático sobre os processos que são utilizados em casos de estudos quantitativos da Engenharia de Software.

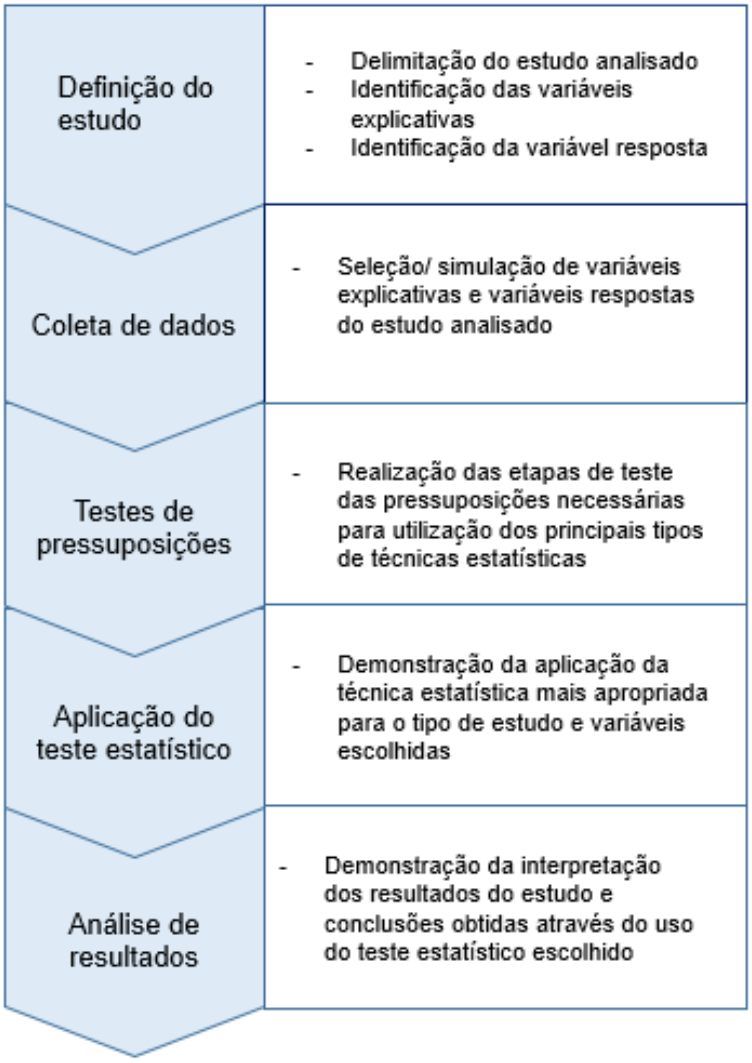

Figura 2.2 – Etapas da exemplificação do uso de testes estatísticos em Engenharia de Software

Dessa forma, nos Capítulos 3 e 4, são apresentados conceitos importantes para utilização de técnicas estatísticas bem como as definições de alguns dos testes estatísticos mais usuais a serem aplicados em Engenharia de Software.

3. PESQUISA QUANTITATIVA E CONCEITOS COMUMENTE UTILIZADOS EM ESTATÍSTICA

Os métodos de pesquisa quantitativos caracterizam-se pelo uso da quantificação das informações para coleta e análise dos dados. Para isso, primeiramente são formuladas diferentes hipóteses sobre os fenômenos estudados. A partir disso, ocorre a coleta dos dados (enfatizando dados numéricos e mensuráveis) que busca informações relevantes para aceitação ou não das hipóteses formuladas. Após a coleta de dados, realiza-se então a análise dos dados, onde as informações são analisadas com apoio de técnicas estatísticas ou matemáticas [31].

As técnicas estatísticas de coleta e análise de dados podem ser utilizadas com o objetivo de descrever e compreender diferentes características e relações entre fenômenos da natureza. Para isso, são utilizados diferentes conceitos para identificar, mensurar e analisar informações relevantes no estudo desses fenômenos [31].

As informações coletadas para estudo em estatística são denominadas como *variáveis*. Existem diferentes tipos de variáveis, qualitativas ou quantitativas. O tipo de análise empregada depende da natureza dessas informações.

Variável resposta e explicativa

Uma das principais definições de variáveis é a diferenciação entre *variável resposta* e *variável explicativa*. Variável resposta é nome dado ao evento que se pretende estudar, tais como desempenho de um software, número de downloads de um aplicativo e outros. Os valores observados para a variável resposta são dependentes e/ou explicados por uma ou mais variáveis explicativas, como por exemplo as diferentes técnicas utilizadas para solução de um problema, os níveis de conhecimento dos desenvolvedores e outros [74].

Amostra e população

Um segundo conceito importante utilizado na Estatística são as definições de *amostra* e *população*. População é definida como a totalidade de objetos, pessoas ou sujeitos sob estudo. Já a amostra caracteriza-se como uma subparcela selecionada da população. Em geral, os experimentos realizados utilizam os conceitos de amostra, uma vez que se torna complexo e por vezes excessivamente caro coletar todos os dados de uma população de interesse. Dessa forma, uma das características fundamentais na análise dos dados de testes estatísticos é a garantia da representatividade da amostra com as características da população estudada [74].

Algumas características individuais coletadas de uma população de interesse tais como peso de um equipamento, tempo de processamento de um software, taxas de per-

formance de uma equipe de desenvolvimento ou outros tipos de informações, podem variar entre indivíduos de um mesmo grupo em um dado instante de tempo e, em um mesmo indivíduo, em instantes diferentes do tempo. Com isso, ordem e regularidade nas medidas coletadas só podem ser estabelecidas quando coletados dados de uma quantidade grande de indivíduos e quando analisadas as informações consolidadas desses indivíduos [95].

Medidas descritivas de dados quantitativos

Após a coleta dos dados, é necessário descrever e resumir as informações afim de apresentar os resultados de forma mais facilitada. Para isso, existem algumas medidas descritivas bastante utilizadas para descrever e resumir os dados de um estudo. Algumas das medidas mais utilizadas são a *média*, a *mediana*, a *moda*, a *variância* e o *desvio-padrão*.

Média

A média é a medida estatística mais conhecida e utilizada para descrever um conjunto de dados. Existem vários tipos de médias (aritmética, harmônica, ponderada e outras), sendo a mais utilizada a média aritmética simples. Essa medida é obtida dividindo-se a soma dos valores das observações pelo número total de observações. Essa medida é muito influenciada por valores extremos (conhecidos como *outliers*) [97].

Mediana

A mediana é o valor central do conjunto de dados. Antes da mediana encontram-se 50% dos dados e após a mediana os outros 50%. A característica dessa medida é que a mesma representa muito mais uma medida de posição do que de grandeza [97].

Moda

A moda é simplesmente o valor mais frequente em um conjunto de dados. Isto é, se um conjunto de dados possuir valores repetidos, aquele que aparece o maior número de vezes representa a moda [97].

Variância e Desvio-padrão

A variância se trata de uma medida de dispersão. Variação ou dispersão é o grau de variação com que os dados numéricos tendem a se distanciar de um valor médio. Neste sentido, as medidas de dispersão indicam o grau de variabilidade demonstrada pelos dados coletados em torno de medidas resumo como a média aritmética. O cálculo da variância é realizado utilizando a soma dos quadrados da diferença entre cada valor e a média aritmética [97].

Apenas calcular a variância como medida de dispersão pode não ser suficiente, pois trata-se de uma medida muito influenciada por valores distantes da média. Além disso, a variância é calculada com os quadrados dos valores, dificultando a interpretação. Com isso, uma alternativa definida foi o cálculo do desvio padrão, que é simplesmente o resultado positivo da raiz quadrada da variância [97].

Eventos e variáveis aleatórias

Quando ocorre a impossibilidade de predizer antecipadamente o resultado de um evento (ocorrido através de sucessivas repetições e sob as mesmas condições), caracteriza-se esse evento como aleatório. A variabilidade presente, nestas condições, é chamada variabilidade aleatória, casual, randômica ou estocástica [95].

Para entender e interpretar os eventos aleatórios, é necessário utilizar conceitos da teoria da probabilidade. Para todo evento aleatório é possível associar uma ou mais variáveis, denominadas como variáveis aleatórias, e para cada variável aleatória (ou conjunto de variáveis aleatórias) é possível encontrar uma função que descreva a sua referida distribuição de probabilidades [95].

Uma variável aleatória pode assumir dois tipos de características: *contínua* ou *discreta*. Uma variável aleatória discreta possui um número finito de opções. Já uma variável aleatória contínua se distribui em um campo infinito de opções. Cada tipo de variável aleatória possui distribuição de probabilidade específica. No caso das variáveis discretas, uma das distribuições de probabilidade mais comuns é a *distribuição Binomial*. Já no caso de variáveis aleatórias contínuas, a mais comum e usual é a *distribuição Normal* [17].

Distribuição Normal

A distribuição Normal, também conhecida como distribuição de Gauss, é a mais conhecida de todas as distribuições de probabilidade contínuas. Trata-se de uma distribuição de probabilidade simétrica, em formato de sino, cujos valores de média, moda e mediana são iguais e encontram-se no centro da curva. Essa distribuição é frequentemente utilizada e usada como base nos testes de hipóteses paramétricos existentes [17, 97].

Teste de Hipóteses

O teste de hipóteses é um dos principais problemas a serem resolvidos através da Estatística. Testar uma hipótese significa que, após realizada uma afirmação sobre uma população estudada, os dados coletados de uma amostra proveniente daquela população são analisados para verificar se eles contrariam ou não a afirmação realizada. Em grande parte das pesquisas, os estudos anteriores ou outro tipo de conhecimento prévio sobre a população estudada é que definem as hipóteses construídas. Dessa forma, o objetivo do teste de hipótese estatístico é, de forma metodológica e através de análise quantitativa de

dados, verificar se os dados da amostra coletada trazem evidências que apoiem ou refutem uma hipótese estatística formulada [97].

Com o teste de hipóteses formulado, se estabelece um nível de significância (erro) para as afirmações e conclusões do teste. Esse nível de significância estabelecido geralmente encontra-se em três níveis: 10%, 5% ou 1%.

Na construção de um teste de hipótese estatístico são construídas duas hipóteses, denominadas como *hipótese nula* (H0) e *hipótese alternativa* (H1). De forma geral, a hipótese nula é construída com base na afirmação que deseja-se refutar e a hipótese alternativa apresenta a possível afirmação a ser confirmada no estudo. Com isso, quando é realizado um teste de diferença entre dois ou mais grupos, geralmente a hipótese nula abrange a afirmação de igualdade entre os grupos, enquanto que a hipótese alternativa abrange a diferença entre dois ou mais grupos analisados.

Com o teste de hipóteses formulado, então o teste estatístico é aplicado. O valor numérico resultante desse teste de hipóteses chama-se *estatística de teste*. O valor da estatística de teste é então comparado com o valor teórico obtido através da distribuição de probabilidade do teste, com um valor pré-fixado do nível de significância.

P-valor

O resultado estatístico do teste de hipóteses é interpretado através de uma medida denominada como *p-valor*. O *p-valor* resulta na probabilidade de que a estatística de teste possua valor extremo em relação ao valor teórico dos dados, sob a condição da hipótese nula ser verdadeira. Escores do *p-valor* tradicionalmente utilizados para rejeitar a hipótese nula são obtidos utilizando como referência um erro máximo tolerável de 0,05 (5%). Nesse caso, em um teste de comparação de grupos, quando a hipótese nula é verdadeira, valores extremos para a estatística de teste são esperados em menos de 5% das vezes. Dessa forma, verifica-se que o teste possui um nível de confiança de 95%. Níveis de erro tradicionalmente utilizados como 10%, 5% e 1%, correspondem respectivamente aos níveis de confiança de 90%, 95% e 99%.

Exemplo de estudo quantitativo utilizando os conceitos estatísticos

Um exemplo de estudo quantitativo é a avaliação do tempo em que desenvolvedores concluem uma tarefa específica utilizando as técnicas A e B. Nesse caso, a *população* de pesquisa são todos os desenvolvedores, sendo a *amostra* uma parcela desses desenvolvedores (aqueles cujos dados compõem o estudo). Esse estudo possui uma *variável explicativa*, que é tipo de técnica utilizada pelos desenvolvedores (A ou B), e a *variável resposta* é o tempo transcorrido até a conclusão da tarefa. Inicialmente é realizada uma análise descritiva dos dados, onde a média de tempo para conclusão da tarefa utilizando a técnica A é de 24 minutos e utilizando a técnica B é de 32 minutos. Para verificar se a diferença das médias de tempo (24 e 32 minutos) é estatisticamente significativa (isto é,

se essa diferença não ocorreu por casualidade observada nessa amostra) é realizado um teste de hipóteses. Nesse caso essas são as seguintes hipóteses:

- **H0**: não há diferenças no tempo de conclusão da tarefa utilizando as técnicas A ou B;
- **H1**: há diferenças no tempo de conclusão da tarefa utilizando as técnicas A ou B.

Através dos resultados obtidos com o uso de uma técnica estatística de comparação de médias, o *p-valor* do estudo foi de 0,012. Dessa forma, como o *p-valor* foi inferior a 0,05, **rejeita-se a H0**. Logo é possível concluir (com um nível de confiança de 95%) que utilizando a técnica A os desenvolvedores concluem a tarefa de forma mais rápida do que se utilizassem a técnica B.

4. PRINCIPAIS TESTES ESTATÍSTICOS E SUAS APLICAÇÕES EM DIFERENTES TIPOS DE ANÁLISES QUANTITATIVAS

Para análise de dados em estudos quantitativos, podem ser utilizadas diferentes técnicas estatísticas. A escolha da técnica mais adequada depende de uma série de fatores tais como: tipo de variável resposta (quantitativa ou categórica), número de variáveis explicativas e variáveis respostas, distribuição de probabilidade dos dados analisados, entre outras.

Os principais testes estatísticos estão subdivididos em dois grupos: *paramétricos* e *não paramétricos*. Os testes paramétricos exigem que sejam aplicados em variáveis respostas contínuas e que sejam cumpridos os pressupostos de normalidade e homogeneidade de variâncias. Já os testes não paramétricos não necessitam de conhecimento a priori sobre a distribuição de probabilidade dos dados e são adaptáveis aos estudos que envolvem variáveis com níveis categóricos ou ordinais [90].

Antes da utilização de qualquer teste estatístico, o primeiro passo a ser realizado são as análises das pressuposições necessárias para utilização das técnicas estatísticas mais usuais. Os testes de verificação da concordância dos dados com as pressuposições fundamentam as bases teóricas de uso de dados testes estatísticos e auxiliam na correta análise dos dados e interpretação dos resultados obtidos [62].

De forma simplificada, as principais suposições a serem testadas para utilização de grande parte dos testes estatísticos são: *normalidade*, *homocedasticidade* e *linearidade* [28].

Normalidade

A *normalidade* é a mais comum das pressuposições da análise estatística. A suposição abrange a informação de que os dados de uma variável contínua, extraídos de determinada amostra estudada, devem possuir distribuição correspondente à distribuição normal. Há diferentes formas de testar essa suposição, que varia desde metodologias descritivas, que analisam visualmente o resultado de alguns gráficos, bem como alguns testes de aderência, tais como o *teste de Kolmogorov-Smirnov (KS)* e o *teste de Shapiro-Wilk (SW)* [28] [82].

Por exemplo, a variável que será testada a suposição de normalidade é a idade dos alunos do curso de Ciência da Computação de uma universidade privada. Dessa forma, utilizando o teste KS para testar a normalidade da variável, a *hipótese nula (H0)* é de que os dados de idade dos alunos são provenientes de uma população com distribuição normal e a *hipótese alternativa (H1)* é de que os dados não são provenientes de uma população com distribuição normal. Utilizando um nível de significância de 5%, após a realização do teste o *p-valor* obtido foi de 0,62. Como esse valor é superior a 0,05, logo não rejeita-se a hipótese

nula. Isto é, não há evidências estatísticas de que os dados não sejam provenientes de uma população com distribuição normal.

Homocedasticidade

A segunda suposição utilizada em técnicas estatísticas é a homocedasticidade, que significa que as variáveis explicativas devem exibir níveis iguais de variâncias ao longo do domínio da variável resposta do estudo. Quando o grau de severidade da heterogeneidade de variância é elevado, os resultados e a significância estatística da análise podem ser comprometidos. Alguns dos testes estatísticos utilizados para testar essa suposição estatística são o *teste de Bartlett* e o *teste de Levene* [28] [117].

Por exemplo, a variável que será testada a suposição de homogeneidade de variâncias é o tipo de ferramenta (1 e 2) utilizada para desenvolvimento de um aplicativo *Web*. Dessa forma, utilizando o teste de Levene para testar a homogeneidade de variâncias, a *hipótese nula (H0)* é de que as ferramentas possuem variâncias semelhantes e a *hipótese alternativa (H1)* é de que as ferramentas não possuem variâncias semelhantes. Utilizando um nível de significância de 5%, após a realização do teste o *p-valor* obtido foi de 0,035. Como esse valor é inferior a 0,05, logo rejeita-se a hipótese nula. Isto é, há evidências estatísticas de que as ferramentas não possuam homogeneidade de variâncias (variâncias semelhantes).

Linearidade

Já o pressuposto de *linearidade* verifica se há uma relação linear entre a variável resposta e as variáveis explicativas. Esse pressuposto pode ser testado a partir de observações gráficas, através de um gráfico denominado como diagrama de dispersão (conforme pode ser visto na Figura 4.1), ou através do resultado do *teste de regressão linear (RL)*, o qual será apresentado na próxima seção.

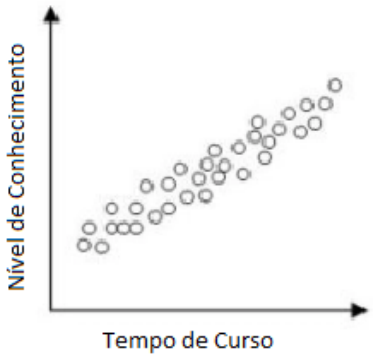

Figura 4.1 – Exemplo de diagrama de dispersão

Através do diagrama de dispersão apresentado na Figura 4.1, é possível verificar o pressuposto de linearidade. Nesse caso, foi testada a linearidade da *variável explicativa*

Tempo de Curso com a *variável resposta* **Nível de Conhecimento**. Como pode ser observado, a relação entre as duas variáveis é semelhante ao formato de uma reta. Nesse caso, o pressuposto de linearidade é atendido.

Nas próximas seções serão apresentados exemplos de testes estatísticos que podem ser utilizados conforme os tipos de variáveis (quantitativas contínuas ou categóricas) em estudos quantitativos. Inicialmente, serão apresentados os testes utilizados quando a variável resposta e a variável explicativa são quantitativas contínuas. Na segunda seção, serão apresentados os testes que podem ser realizados quando a variável resposta é contínua e a variável resposta é categórica. Na terceira seção, serão apresentados os testes estatísticos que podem ser utilizados quando ambas as variáveis (resposta e explicativa) são categóricas. Na última seção, são apresentados outros tipos de testes, como análise de agrupamento, análise de concordância e teste de tamanho do efeito (TDE). Com exceção da última seção, todos os outros testes apresentados são testes que testam hipóteses quando o objetivo é avaliar o efeito de uma variável explicativa no resultado de uma variável resposta.

4.1 Variável resposta e variáveis explicativa quantitativa contínua

Nos casos onde a variável resposta e a variável explicativa de um estudo são quantitativas contínuas, em geral, são utilizadas as técnicas estatísticas como a *análise de regressão*, *correlação* e outras.

Análise de Regressão

O principal objetivo da análise de regressão é estimar os valores de uma *variável resposta* de interesse com base nos valores conhecidos e/ou fixados de *uma ou mais variáveis explicativas*. Dessa forma, a análise de regressão estabelece uma relação funcional de causa e efeito, onde há duas ou mais variáveis envolvidas para descrição de um fenômeno estudado. Para utilizar a técnica de análise de regressão, as variáveis explicativas necessitam ser *independentes*. Dessa forma, os valores obtidos em uma das variáveis explicativas não podem ser dependentes ou influenciados pelos resultados das demais variáveis explicativas. O tipo mais usual de análise de regressão é a *regressão linear*, que ocorre quando há uma relação de dependência linear da variável resposta com as variáveis explicativas. A análise de regressão apresenta algumas pressuposições para sua utilização que são [28]:

- normalidade dos resíduos;
- homocedasticidade;
- ausência de correlação entre os resíduos;

- relação linear entre a variável resposta e as variáveis explicativas;
- ausência de multicolinearidade (alta correlação entre as variáveis explicativas).

Quando as suposições da análise de regressão paramétrica não são atendidos, então é necessário realizar as análises dos dados utilizando uma alternativa não paramétrica. A característica desse tipo de técnica é a ausência (completa ou quase completa) de conhecimento a priori a respeito da distribuição de probabilidade da função que será estimada através da análise dos dados. Dessa forma, como a função é estimada a partir do ajuste de parâmetros livres, a classe de funções que o estimador pode prever torna-se bastante ampla [150].

Correlação

Quando há dependência entre as variáveis explicativas, a análise de regressão não pode ser realizada. Dessa forma, uma das possibilidades a ser realizada é o teste de correlação. A correlação entre duas variáveis apresenta como objetivo mensurar um possível inter-relacionamento entre essas variáveis. O resultado desse teste é uma medida que varia dentro do intervalo fechado de -1 a 1, onde -1 indica uma perfeita correlação negativa ou inversa e 1 indica perfeita correlação positiva ou direta. Valores negativos da correlação indicam que o crescimento de uma das variáveis implica, em geral, no decrescimento da outra variável. Já valores positivos indicam, em geral, o crescimento ou decrescimento concomitante das duas variáveis estudadas. Uma das técnicas de correlação mais conhecidas é a correlação obtida através do *coeficiente de Pearson*, que mensura a correlação linear entre duas variáveis. [109].

Quando a relação entre as variáveis não é linear (como nos casos onde a relação é quadrática, cúbica, exponencial, etc.), ela não será medida adequadamente pelo coeficiente de correlação mais usual, que é o de Pearson. Nesses casos, é possível utilizar outro coeficiente de correlação, como o de coeficiente de Spearman. Esse coeficiente, por utilizar as informações de postos dos dados[1], não pressupõe uma distribuição de probabilidade específica para os dados, podendo ser utilizado nas situações em que a relação entre os pares de dados não é linear [109].

[1] Postos são ordenações (de forma crescente) dos dados. Dessa forma, os testes que usam postos não utilizam os valores brutos dos dados, mas sim a ordenação deles na amostra [126].

4.2 Variável resposta quantitativa contínua e variável explicativa categórica

Nessa seção são apresentados os testes estatísticos mais conhecidos na avaliação de variáveis respostas contínuas, quando a variável explicativa é categórica.

Comparação de dois grupos independentes

Os testes de hipóteses que testam a igualdade entre duas médias são alguns dos testes estatísticos mais comuns, sendo utilizados em diversas situações. Existem diversas alternativas para comparação de médias, sendo o teste mais conhecido e utilizado o teste paramétrico baseado na distribuição *t de Student*. Esse teste baseia-se em algumas pressuposições para o seu uso, tais como a suposição de independência e normalidade dos dados, além da suposição de homogeneidade das variâncias das populações sob estudo. Além disso, o *teste T* não é aplicável quando os dados são qualitativos ordinais. Dessa forma, nos casos onde as suposições de uso do *teste T* não são atendidas, são apresentadas alternativas não paramétricas, como o teste de Mann e Whitney [9].

Comparação de dois ou mais grupos independentes

No caso de comparação de dois ou mais grupos, onde a variável resposta é quantitativa, mas a variável explicativa é categórica, um dos testes mais utilizados é a *Análise de Variância* (ANOVA). Da mesma forma como o *teste T*, esse teste verifica se os diferentes níveis da variável explicativa (também conhecidas como fator na ANOVA) promovem mudanças sistemáticas em uma variável resposta estudada. O caso mais simplificado de utilização da técnica é a ANOVA unifatorial, onde são comparados os diferentes níveis de um fator (variável explicativa) em uma resposta de interesse. A ANOVA também possui pressupostos para sua utilização, tais como normalidade dos erros, homocedasticidade entre os tratamentos e amostras independentes [90, 40].

Caso as suposições da ANOVA não sejam atendidas, é possível realizar um teste não paramétrico correspondente. Uma alternativa nesse caso é a utilização do teste *Kruskal-Wallis*, que trata-se de um *teste T* também utilizado para comparar diferentes níveis de um fator de uma resposta de interesse, porém não exige os pressupostos da ANOVA [90].

Comparação de grupos pareados

Já para dados pareados, onde há dependência entre as variáveis explicativas, o *teste T* a ANOVA convencional não podem ser realizadas. Dados pareados podem ser entendidos como aqueles dados que são mensurados, por exemplo, antes e após um tratamento ser realizados em um mesmo indivíduo (ou unidade amostral), ou ainda, aqueles dados onde dois tratamentos são aplicados em indivíduos (unidades amostrais) muito se-

melhantes entre si. Dessa forma, uma das possibilidades a ser realizada é o *teste T pareado* e a ANOVA para medidas repetidas [9].

A técnica estatística ANOVA com medidas repetidas é utilizada para comparação de três ou mais médias de grupos onde há repetição total dos participantes em cada grupo. Esse tipo de análise geralmente é realizado em duas situações: quando os participantes são submetidos a várias medidas de uma mesma intervenção ao longo do tempo ou quando os participantes são submetidos a mais de uma condição (intervenção) e as respostas mensuradas desejam ser comparadas. Para utilizar a ANOVA de medidas repetidas, algumas suposições devem ser verificadas para que a técnica forneça um resultado válido. As suposições da ANOVA de medidas repetidas são: *ausência de valores extremos* nos grupos analisados, *normalidade da distribuição da variável dependente* nos dois ou mais grupos relacionados e *esfericidade* (as variações das diferenças entre todas as combinações de grupos relacionados devem ser iguais) [58].

No caso onde as suposições desses testes não forem atendidas, da mesma forma como nos testes para amostras independentes, é necessário realizar testes alternativos. Nesse caso, para o *teste T*, a alternativa não paramétrica é o teste de Wilcoxon. Já no caso da ANOVA para medidas repetidas, quando os dados analisados são relacionados e as suposições da ANOVA para medidas repetidas não forem atendidas, uma das opções de teste estatístico a ser realizado é o *teste de Friedman*. Esses testes são opções não paramétricas utilizadas para comparar dados amostrais vinculados, quando por exemplo um mesmo objeto de estudo ou indivíduo é avaliado mais de uma vez. Esse tipo de teste não utiliza diretamente os dados numéricos na análise, mas os postos ocupados por esses dados. Realiza-se a soma dos postos ocupados pelos dos dados em cada um dos grupos e, então, é testada a hipótese de igualdade dos grupos [125, 9].

Teste de comparações múltiplas

Os testes de comparação múltipla entre médias de tratamentos são realizados após a análise de variância (ou o teste não paramétrico equivalente), quando o resultado desse teste é estatisticamente significativo. Quando a análise de variância detecta um efeito significativo (a partir de um determinado nível de significância), de modo que se decide rejeitar a hipótese de nulidade, é necessário verificar onde existe os pares de diferenças entre os grupos (níveis da variável). Os testes de comparação mais comuns são os testes de Tukey, Duncan, Dunnet e o teste LSD, normalmente utilizados para detalhar onde ocorre a diferença entre pares, permitindo mostrar, especificadamente, quais tratamentos diferem, ou não, estatisticamente entre si [39].

4.3 Variável resposta e variável explicativa categórica

Quando a variável resposta e a variável explicativa são categóricas, podem ser utilizados testes estatísticos tais como *teste qui-quadrado* e *teste exato de Fisher*.

Teste Qui-Quadrado de Pearson

O teste qui-quadrado (também conhecido como o teste do qui-quadrado de Pearson) é um dos testes mais conhecidos para testar hipótese de variáveis categóricas (nominais). Esse teste fornece informações não apenas da importância de qualquer diferença observada nas amostras, como fornece informações sobre quais categorias diferem da(s) demais [91].

O teste qui-quadrado é uma estatística não paramétrica, também chamado de teste livre de distribuição. Porém, assim como na maioria dos testes estatísticos, são necessários requisitos para seu uso apropriado, as chamadas “suposições” da estatística. O teste pode ser utilizado nas seguintes condições [91]:

- os dados precisam ser frequências ou contagem de casos;
- os níveis (categorias) das variáveis precisam ser simultaneamente exclusivos;
- cada indivíduo aparece apenas em uma combinação de categorias das variáveis resposta e explicativa;
- os grupos de indivíduos precisam ser independentes;
- O valor de cada combinação de categorias não deve ser inferior a cinco (em pelo menos 80% das combinações) e nenhuma combinação deve ter um tamanho amostral de menos de um caso.

Teste exato de Fisher

O Teste Exato de Fisher é utilizado para avaliar proporções de variáveis categóricas (em tabelas de contingência 2x2 - duas linhas e duas colunas), comparando os resultados de grupos de duas amostras independentes. Esse teste fornece *valor-p* exato e não exige técnica de aproximação. O *valor-p* do teste exato de Fisher é preciso para todos os tamanhos amostrais, enquanto os resultados provenientes do teste qui-quadrado que examina as mesmas hipóteses podem ser imprecisos quando o número de células é pequeno. Dessa forma, em casos de pequenos tamanhos amostrais e comparação de variáveis categóricas com no máximo duas categorias, pode ser usado o teste exato de Fisher [27].

4.4 Alguns outros tipos de análises

Análise de Agrupamento

Existem diferentes modalidades de análise de agrupamento. Todas elas apresentam um objetivo em comum, que é a redução da dimensionalidade dos dados. O objetivo dessa análise é reduzir uma grande quantidade de variáveis a um número menor de fatores/componentes/grupos. Uma das formas de realização de análise de agrupamento é através da análise fatorial. Nessa técnica, os fatores/componentes construídos a partir da análise de agrupamento são combinações lineares das variáveis observadas que explicam/representam a variação das variáveis originais. Dessa forma, através de poucos fatores é possível analisar a variação apresentada por um grande conjunto de variáveis. Para uso da técnica, no geral as variáveis devem ser preferencialmente contínuas ou discretas, mas muitas vezes os pesquisadores usam variáveis ordinais e nominais. Nesse tipo de técnica, por apresentar como objetivo a redução de dados, é necessário que as variáveis incluídas na análise estejam correlacionadas [50].

Um outro tipo de análise de agrupamento, diferente da análise fatorial é a análise de cluster. A análise de cluster/agrupamentos/conglomerados é uma denominação de um conjunto de diferentes técnicas que podem ser utilizadas para agrupar casos em grupos. Dessa forma, diferente da análise fatorial, que agrupa variáveis, a análise de cluster agrupa casos que possuam características similares em um mesmo grupo, e casos com características distintas em grupos diferentes [50].

Análise de Concordância

Um importante critério para avaliação da qualidade de um instrumento/avaliação é a confiabilidade da mensuração. Uma avaliação ou instrumento confiável é aquele que após repetidas mensurações, apresentar uma menor variação. Existem diversas aplicações para utilização desse tipo de informação. Por exemplo, uma das aplicações desse tipo de técnica é verificar o grau de correspondência entre as avaliações independentes de dois ou mais engenheiros de software na avaliação de um mesmo sistema (utilizando os mesmos procedimentos e instrumentos de classificação) [108].

Dessa forma, uma das técnicas utilizadas para esse tipo de análise de informações é o Coeficiente Kappa. Essa medida pode ser definida como uma métrica de associação utilizada para descrever e testar o grau de concordância (confiabilidade e precisão) de diferentes avaliações/classificações. Esse coeficiente, apesar de amplamente utilizado, apresenta limitações pois nem sempre considera aspectos importantes dos dados na estrutura de concordância e discordância das informações. Diversos autores classificaram o coeficiente de diferentes formas para explicar os diferentes níveis de concordância. Para esses

autores, valores maiores que 0,75 representam excelente concordância. Valores abaixo de 0,40 representam baixa concordância e valores situados entre 0,40 e 0,75 representam concordância mediana [108].

Tamanho do Efeito - TDE

Os testes que mensuram tamanhos do efeito (TDE) são análises estatísticas adicionais que dão significado aos testes de hipóteses, enfatizando o poder dos testes estatísticos, e reduzindo o risco de que uma mera variação amostral encontrada no estudo possa ser interpretada como relação real. Existem diferentes tipos de testes para avaliação do TDE. Um dos testes mais utilizados é o *teste d de Cohen* (medida comum de TDE após a aplicação de *teste T* para amostras independentes). Esse teste avalia a diferença entre médias (populacionais) de dois grupos, e é usado quando o tamanho amostral é semelhante e os desvios-padrões populacionais também são similares. Esse teste também pode ser utilizado para amostras pareadas [43].

5. TRABALHOS RELACIONADOS

Neste capítulo são apresentados alguns trabalhos que relatam o uso da estatística em estudos da Engenharia de Software. Dessa forma, algumas das dificuldades apresentadas pelos artigos foram desafios trabalhados durante o estudo.

Statistical power and its subcomponents — missing and misunderstood concepts in empirical software engineering research

Miller et al. [94] apresentaram uma questão fundamental a ser desenvolvida através do presente estudo, que é a dificuldade das pesquisas em Engenharia de Software utilizarem análises e o poder de resultado dos testes estatísticos em seus projetos. Nesse estudo, os autores citam desconhecer muitas pesquisas realizadas em Engenharia de Software que utilizam o poder estatístico como parâmetro fundamental em seus projetos.

Com isso, no texto são apresentadas dificuldades de aplicação dos testes estatísticos em projetos de experimentos em Engenharia de Software e discutem propostas para aplicação da estatística nos processos de design experimental.

No estudo, os autores citam que talvez alguns dos pontos mais críticos na utilização da Estatística em experimentos da Engenharia de Software esteja relacionado com as dificuldades na formulação e avaliação das hipóteses de teste.

Um dos pontos positivos do artigo foi a discussão sobre a importância da Estatística para a Engenharia de Software empírica. O artigo apresentou conceitos importantes na utilização de testes estatísticos como conceitos de amostragem, teste de hipóteses e significância estatística. Porém, um dos pontos negativos do estudo é que foi apresentado apenas um exemplo de experimento na área de Engenharia de Software solucionado através de uma técnica estatística. Dessa forma, os diversos tipos de testes estatísticos que podem ser utilizados em diferentes contextos não foram apresentados e exemplificados.

An Adaptation of Experimental Design to the Empirical Validation of Software Engineering Theories

O trabalho de Juristo e Moreno [75] apresentou dois objetivos importantes relacionados com o uso de técnicas estatísticas em Engenharia de Software. Primeiramente, foi discutida a necessidade de incentivo de projetos de experimentos em Engenharia de Software uma vez que, segundo os autores, a falta de preocupação com o uso de experimentação nos projetos pode atrasar a adoção de novas tecnologias por organizações. Além disso, a introdução de novas tecnologias de software pode ser considerada um risco, pois muitas vezes os processos não são avaliados e validados corretamente, logo suas aplicações podem causar sérias consequências para os usuários e organizações.

No artigo foi apresentado um estudo aprofundado da aplicação de desenho experimental para avaliação empírica em Engenharia de Software. Para isso, foi apresentado o valor de alguns conceitos experimentais para a Engenharia de Software (com um resumo das técnicas de *design* experimental que podem ser aplicadas), bem como adaptação de terminologia de *design* de experimentos para a área.

Um dos pontos positivos do estudo é a apresentação de uma tabela contendo possíveis variáveis explicativas a serem utilizadas em estudos na Engenharia de Software, bem como potenciais variáveis respostas. Porém, uma das limitações do artigo é a apresentação de casos apenas utilizando ANOVA como técnica estatística de análise de dados limitando o número de técnicas possíveis de serem realizadas bem como as limitações e pressuposições para utilização destas técnicas.

Survey research in software engineering: problems and strategies

O estudo de Ghazi et al. [57] revisou a literatura e entrevistou nove pesquisadores da área de Engenharia de Software, com o objetivo de identificar quais são as principais dificuldades e estratégias dos pesquisadores na realização de pesquisas em Engenharia de Software.

Através dos resultados identificados durante as entrevistas com os nove especialistas, foram identificados 24 problemas e 65 estratégias na realização de pesquisas em Engenharia de Software. Os problemas e estratégias são agrupados com base nas fases do processo de pesquisa.

Na definição do público-alvo e do processo de amostragem, o maior problema identificado foi o tamanho de amostra insuficiente. Diante disso, esse foi o problema com maior número de estratégias relacionadas a ele (26 estratégias). Algumas das estratégias apresentadas foram a brevidade da pesquisa (limitando o comprimento da pesquisa), e o uso de redes sociais para recrutar respostas. Além disso, diferentes estratégias de amostragem foram discutidas (como amostragem aleatória e de conveniência).

Na execução de instrumentos de pesquisa, o principal problema observado foi na formulação de questões. Com isso, as estratégias apresentaram recomendações sobre quais tipos de perguntas evitar, bem como apresentaram a necessidade de realização de pré-testes.

No processo de análise de dados, os principais problemas identificados foram com a imprecisão das informações coletadas e dificuldades com a análise de dados. Nesse processo foi identificada a importância de envolver vários pesquisadores na análise de dados.

Dessa forma, o estudo apresenta como aspecto positivo a identificação de diversas dificuldades na realização de pesquisas em Engenharia de Software. Porém, um aspecto negativo foi a falta de exemplos de casos em que ocorreram essas dificuldades. Uma sugestão apresentada pelo estudo para trabalhos futuros foi a definição de conceitos, atividades

e estratégias importantes para pesquisa em Engenharia de Software junto a comunidade da área.

Problems with Statistical Practice in Human-Centric Software Engineering Experiments

O estudo de Kitchenham, Madeyski e Brereton [76] revisou artigos relacionados a 45 experimentos, que envolviam uma total de 1.303 participantes. O objetivo desse estudo era verificar quais questões que contribuam para o mau desempenho da prática estatística em experimentos com humanos na Engenharia de Software.

De acordo com os resultados apresentados pelo estudo, os problemas observados no uso de procedimentos e técnicas estatísticas na Engenharia de Software estavam relacionados com o uso incorreto de terminologia e incompreensão dos princípios estatísticos. Especificamente, as questões consideradas no estudo foram relacionadas ao uso indevido da terminologia estatística, análise incorreta de experimentos de medidas repetidas, uso de testes de comparações múltiplas (*Post-Hoc*), uso de pré-teste para suposições de normalidade e variância e teste de múltiplas hipóteses.

Segundo os autores, o uso incorreto de terminologia estatística pode ser indicativo de uma incompreensão dos métodos estatísticos. Nesse caso, eles apontam uma limitação importante do uso desse tipo de análise na área que é que se os pesquisadores não entendem a terminologia estatística, consequentemente eles terão problemas para entender os conceitos estatísticos apresentados nos livros e nos resultados das análises.

Os autores acreditam que um dos problemas subjacentes aos apresentados é a adoção de desenhos estatísticos complicados que, como consequência, implicam em métodos de análise complicados também. Quanto maior o desenho experimental, maior o tamanho da amostra necessária. No caso de diversas famílias de experimentos, a intenção do estudo é usar um número de repetições adequado para combater os pequenos tamanhos amostrais. Porém, experimentos maiores e mais complexos, mesmo com amostras totais grandes, podem apresentar tamanhos amostrais pequenos em cada subgrupo avaliado, podendo gerar resultados não confiáveis.

Evolution of statistical analysis in empirical software engineering research: Current state and steps forward

O trabalho de Oliveira Neto et al. [38] analisou, através de uma revisão sistemática de literatura, quais são os principais métodos estatísticos utilizados em 15 anos de pesquisa empírica na Engenharia de Software. A análise do trabalho incluiu dados de cinco periódicos conhecidos da Engenharia de Software (TSE, TOSEM, EMSE, JSS e IST) publicados entre 2001 e 2015.

As perguntas de pesquisa do trabalho incluíram a busca de quais são os principais métodos estatísticos usados em pesquisa na Engenharia de Software, até que ponto é possível extrair automaticamente o uso de métodos estatísticos da literatura, se há alguma tendência no uso de técnicas estatísticas na Engenharia de Software e com que frequência os pesquisadores usam resultados estatísticos (como os resultados de significância estatística) para analisar significados práticos.

Como resultado, o estudo revelou que no geral, há uma tendência de aumento do uso da análise estatística de dados quantitativos nos últimos anos. Eles evidenciaram a informação de que, especificamente nos últimos cinco anos, houve aumento do uso de técnicas estatísticas não paramétricas nas publicações, bem como do uso de medidas de tamanho do efeito. Embora tenha sido visto um aumento do uso de testes estatísticos na área, o estudo identificou alguns problemas relacionados ao uso de testes estatísticos como a quantidade de testes paramétricos realizados sem a identificação de realização de testes de suposição de uso (como teste de normalidade, por exemplo).

No trabalho, também foi construído um modelo conceitual de fluxo de trabalho de análise estatística para auxiliar pesquisadores da Engenharia de Software na escolha e uso da análise estatística. Este modelo incluiu um conjunto de diretrizes para aumentar a conscientização sobre as principais armadilhas dos estágios iniciais de uma análise estatística e como apoio para os pesquisadores na interpretação do significado prático de seus resultados. Eles também forneceram sugestões de diferentes técnicas que podem complementar o atual conjunto de ferramentas estatísticas dos pesquisadores da Engenharia de Software, como por exemplo, como usar a análise bayesiana e como utilizar técnicas de imputação para dados ausentes.

6. PROPOSTA DE PESQUISA

Um dos achados mais importantes do estudo de Ghazi et al. [57] (apresentado no Capítulo 5) foi a dificuldade dos pesquisadores com os processos de amostragem de dados. Os pesquisadores relataram dificuldades em obter amostras suficientemente grandes em seus projetos de pesquisa. Esse é um aspecto importante a ser considerado, tendo em vista que um dos principais limitadores em análises estatísticas convencionais é o tamanho amostral (que dificulta a aceitação de pressuposições básicas de diversos tipos de testes estatísticos). Com isso, nesse estudo são abordadas análises alternativas de dados (com estatística não paramétrica), que facilitam análises estatísticas (cujos resultados independem do tamanho amostral coletado).

Nos estudos de Miller et al. [94] e Juristo e Moreno [75], também apresentados no Capítulo 5, os autores buscaram demonstrar que o uso de testes estatísticos são componentes essenciais de qualquer desenho experimental. É necessário conhecer os conceitos estatísticos e avaliar o correto uso dos testes estatísticos, pois mesmo quando os resultados dos testes de hipóteses são significativos, pode haver algum erro no processo de coleta e análise dos dados.

O trabalho de Oliveira Neto et al. [38], que analisou os principais métodos estatísticos utilizados em publicações entre 2001 e 2015, evidenciou o uso frequente de técnicas estatísticas paramétricas como *teste T* e ANOVA. Da mesma forma, o estudo evidenciou o aumento do uso de técnicas não paramétricas nas publicações da área nos últimos anos. Além disso, eles forneceram uma série de diretrizes de perguntas para os pesquisadores se questionarem no processo de escolha do teste estatístico tais como a escolha do tipo de técnica (paramétrica ou não paramétrica), se somente resultados significativos são suficientes para os estudos, entre outras questões.

Com isso, esses estudos contribuíram para o problema e pergunta de pesquisa do presente estudo, pois evidenciam alguns dos desafios enfrentados pelos pesquisadores da área de Engenharia de Software na utilização de testes estatísticos em seus experimentos, bem como um direcionamento de como as técnicas estatísticas estão sendo utilizadas por pesquisadores da área nos últimos anos.

6.1 Problema de Pesquisa

Através dos estudos relacionados, foram identificadas dificuldades na utilização de testes estatísticos em projetos da Engenharia de Software. Existem diversos desafios relacionados com a utilização de análises estatísticas de dados, entre eles dificuldades na coleta de informações (muitas vezes devido aos tamanhos amostrais pequenos), difi-

culdades na identificação e utilização de testes estatísticos, bem como a necessidade de incentivo e conscientização da importância do uso da estatística em diversos projetos da área.

Além disso, há a necessidade da construção de um mapeamento de testes estatísticos para aplicação em Engenharia de Software, tais como a construção de um fluxo simples e replicável.

Dessa forma, alguns dos problemas de pesquisa relacionados ao uso de testes estatísticos na Engenharia de Software abrangem a identificação das técnicas de análise de dados quantitativos comumente realizados na área, bem como o mapeamento e a construção de fluxograma de aplicação de testes estatísticos para análise dados dos estudos da Engenharia de Software.

6.2 Questões de Pesquisa

Visando solucionar os problemas de pesquisa identificados, foram definidas as seguintes questões de pesquisa:

(Q1) *Como são realizadas as análises de dados em estudos realizados na Engenharia de Software?*

(Q2) *Como são utilizados testes estatísticos para análise de dados em Engenharia de Software?*

(Q3) *Quais são as principais dificuldades e limitações no uso de técnicas estatísticas para análise de dados quantitativos na Engenharia de Software?*

(Q4) *Quais são os principais tipos de estudos quantitativos realizados na área de Engenharia de Software e como auxiliar a análise dessas informações utilizando como apoio ferramentas estatísticas?*

6.3 Objetivos

O objetivo geral da pesquisa é *propor um fluxograma de aplicação de testes estatísticos para análise de dados em Engenharia de Software*.

Visando alcançar o objetivo geral proposto, foram definidos os seguintes objetivos específicos:

- Identificar os principais tipos de estudos quantitativos realizados na Engenharia de Software;

- Aplicar o fluxograma para fins de avaliação;
- Exemplificar o uso o fluxograma em estudos da Engenharia de Software.

6.4 Método

Para solucionar as questões de pesquisa e atingir os objetivos propostos, esta pesquisa foi realizada em três etapas:

Etapa 1 - *Identificação dos principais tipos de análise de dados quantitativos utilizados na Engenharia de Software*.

Para essa etapa, através de uma Revisão Sistemática da Literatura (RSL), foram avaliados estudos que utilizaram análise quantitativa de dados, sendo possível identificar características dos problemas de pesquisa da Engenharia de Software. Com isso, foi possível identificar as técnicas estatísticas mais utilizadas pela área, bem como verificar quais são as características mais comuns desses estudos tais como tipos de variáveis explicativas e variáveis respostas, tipo de método utilizado (paramétrico e não paramétrico) e outros.

Etapa 2 - *Construção e avaliação de um fluxograma de aplicação de testes estatísticos em estudos quantitativos realizados na Engenharia de Software*.

Através da identificação dos principais tipos de testes estatísticos e das variáveis explicativas e variáveis respostas comumente utilizadas em problemas da Engenharia de Software, esta etapa visou a construção e o delineamento de um fluxograma de aplicação de testes estatísticos na área de Engenharia de Software.

O propósito dessa etapa do estudo foi refinar e validar o fluxograma com pesquisadores da área de Engenharia de Software, verificando se o fluxograma construído atendia as necessidades de avaliação e análise de dados de problemas comuns na área.

Etapa 3 - *Exemplificação de análises estatísticas realizadas no fluxograma construído e validado.*

Após a construção e avaliação do fluxograma de aplicação de testes estatísticos em Engenharia de Software, a Etapa 3 consiste em selecionar algumas publicações da Engenharia de Software que utilizaram análise de dados quantitativos e apresentar todos os passos de utilização do fluxograma, que variam desde a aplicação de testes de pressupostos para utilização das principais técnicas estatísticas, até a interpretação dos resultados obtidos através da análise selecionada.

Dessa forma, nessa etapa serão apresentadas análises de casos de estudos quantitativos realizados na Engenharia de Software, utilizando o fluxograma de testes estatísticos proposto. Para a análise dos estudos, uma das premissas é de que as etapas de delineamento do estudo, identificação das variáveis (explicativas e resposta) e coleta de dados já tenham sido previamente realizadas. Com isso, o fluxograma de testes estatísticos deve ser utilizado nas etapas de *Teste de pressuposições* e *Aplicação do teste estatístico*, conforme destacado na Figura 6.1.

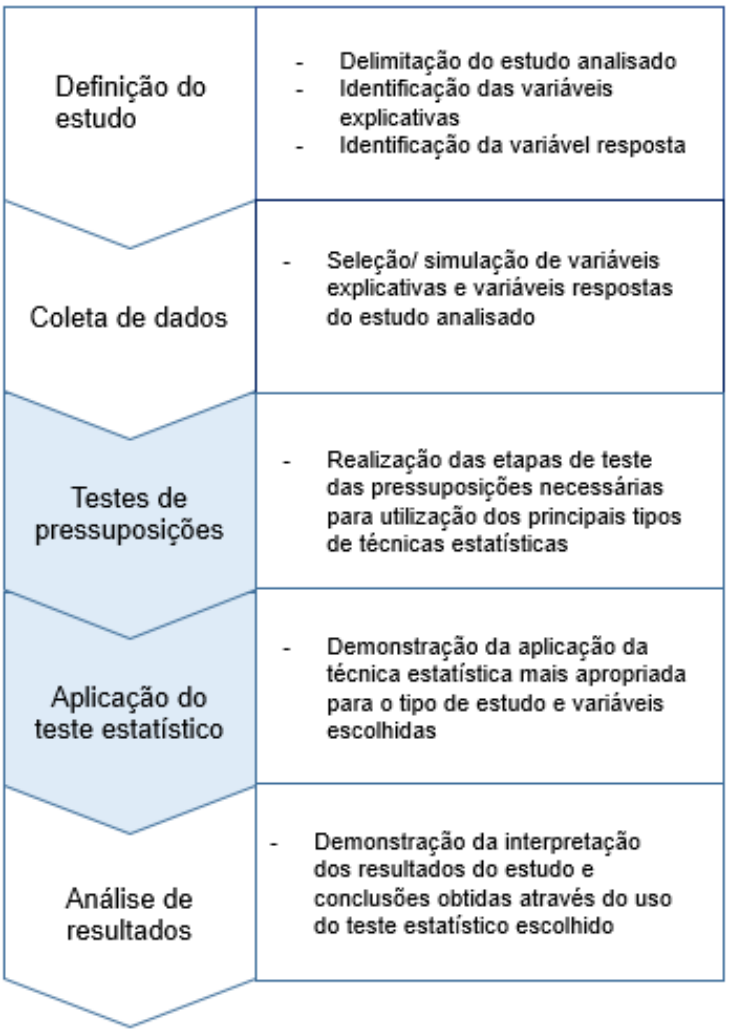

Figura 6.1 – Etapas de aplicação do fluxograma de testes estatísticos em Engenharia de Software

7. REVISÃO SISTEMÁTICA DE LITERATURA

Alguns dos desafios relacionados com a utilização de análises estatísticas de dados na Engenharia de Software (apresentados no Capítulo 5), são as dificuldades na coleta de informações (muitas vezes devido aos tamanhos amostrais pequenos), dificuldades na identificação e utilização de testes estatísticos, bem como a necessidade de incentivo e conscientização da importância do uso da estatística em diversos projetos da área. A justificativa para a escolha da Revisão Sistemática de Literatura (RSL) foi conhecer em profundidade os tipos de estudos quantitativos realizados na área e como são realizadas as análises dos dados desses trabalhos.

Etapa 1 - *Questões de Pesquisa da RSL*.

Visando solucionar os problemas de pesquisa identificados, foram definidas as seguintes questões de pesquisa para serem respondidas pela RSL:

(Q2) *Quais são os principais tipos de análise de dados utilizadas em experimentos da Engenharia de Software?*

(Q3) *Quais os principais tipos de variáveis utilizadas nos experimentos da Engenharia de Software?*

(Q1) *Quais são as principais técnicas estatísticas utilizadas em experimentos da Engenharia de Software?*

Para responder as questões de pesquisa, a RSL foi realizada através das seguintes etapas:

Etapa 2 - *Seleção das bibliotecas digitais*.

Nessa etapa, foram definidas as bibliotecas e fontes de origem da informação, sendo realizada a busca dos artigos científicos publicados no *International Symposium on Empirical Software Engineering and Measurement* (ESEM) no período de cinco anos (2014-2019), disponíveis na página online da IEEE. O ESEM é uma conferência no qual pesquisadores, profissionais e educadores podem relatar e discutir os resultados, inovações, tendências, experiências e preocupações mais recentes em pesquisas empíricas na Engenharia de Software. A conferência se concentra nos processos, design e estrutura de estudos empíricos e nos resultados de estudos específicos. Os estudos recebidos e publicados pela conferência podem variar de experimentos controlados a estudos de campo e de estudos quantitativos até qualitativos. Tendo em vista que o simpósio representa a principal conferência na apresentação de resultados de pesquisa relacionados à engenharia de software

empírica e diante das características dos estudos publicados no ESEM (com bastante ênfase em análises quantitativas), optou-se pela análise das publicações realizadas nessa conferência.

Etapa 3 - *Termos de busca*.

Não foram definidos termos de busca específicos para retorno e avaliação dos artigos, além da sigla da conferência analisada (ESEM). A totalidade dos artigos publicados na conferência e disponíveis no site da IEEE durante o período especificado (600 artigos) foram avaliados. Não foram especificadas palavras-chave para busca dos artigos. Essa decisão ocorreu porque grande parte dos artigos publicados nessa conferência não apresentam descrição de palavras-chave no resumo (principalmente as publicações do ano de 2015). Dessa forma, caso fossem utilizadas palavras-chave, vários estudos importantes para análise poderiam ser excluídos.

Etapa 4 - *avaliação do protocolo de busca de artigos*.

A avaliação do protocolo de busca dos artigos foi realizada a partir da leitura dos artigos selecionados da busca na base de dados e verificando se eles tinham sido publicados na conferência especificada e no período de tempo selecionado (a partir de 2014).

Etapa 5 - *Critérios de Inclusão e Exclusão.*

Com base no problema de pesquisa, foram definidos os seguintes critérios para inclusão e exclusão dos estudos:

Critérios de Inclusão

- artigos publicados no ESEM;
- estudos que utilizem análise quantitativa de dados;
- estudos que descrevam o processo de coleta de dados;
- estudos que descrevam as variáveis analisadas.

Critérios de Exclusão

- artigos não disponíveis online na página da IEEE;
- duplicidade de estudos;
- artigos publicados antes de 2014;

- estudos qualitativos.

Foram selecionados para o estudo os artigos que, além de realizar análise quantitativa dos dados, também apresentassem o processo de coleta de dados e apresentação dos resultados. Foram excluídas da análise as publicações anteriores ao ano de 2014. Essa escolha foi realizada visando abranger problemas de pesquisa mais recentes na área, bem como verificar como estão sendo realizadas as análises estatísticas nos últimos cinco anos na Engenharia de Software e publicados no ESEM. Tendo em vista o objetivo de analisar o processo de coleta e análise de dados quantitativos, foram excluídos da análise os estudos qualitativos.

Etapa 6 - *Seleção dos artigos*

Após a verificação dos artigos disponíveis para consulta na biblioteca digital IEEE, utilizando somente a *string* de busca e sem aplicação de nenhum filtro de dados, foram selecionadas 600 publicações.

Posteriormente, aplicando o filtro de data (identificado no critério de exclusão), foram selecionadas 181 publicações. Essa etapa excluiu do processo as publicações mais antigas, com data anterior ao ano de 2014.

Na etapa seguinte do processo de seleção de artigos, o resumo de cada artigo foi avaliado e, atendendo aos critérios de inclusão e exclusão, esse estudo foi selecionado para análise em profundidade e extração das informações necessárias. Caso o resumo não apresentasse informações suficientes para seleção ou exclusão, foi realizada a leitura do artigo na íntegra, visando não excluir do processo algum artigo relevante de forma errônea. Nessa etapa foram selecionados 110 artigos.

Na última etapa de seleção dos estudos, todos os 110 artigos foram avaliados e foi construída uma tabela de coleta de informações para captar as informações necessárias para responder as questões de pesquisa propostas pela RSL, conforme Figura 7.1.

Etapa 7 - *Análise dos resultados*

Através dos dados obtidos com a revisão de literatura, foi possível identificar como são realizadas as análises estatísticas de dados para profissionais e pesquisadores da área de Engenharia de Software.

Após o processo de seleção dos artigos, as informações necessárias para responder as questões de pesquisa foram consolidadas em uma tabela de dados. Na Figura 7.2, são apresentadas as informações coletadas dos 110 estudos selecionados na RSL. No retorno da busca de dados da biblioteca digital IEEE, retornaram publicações do ESEM

Figura 7.1 – Etapas de seleção dos artigos da RSL

somente dos anos de 2015, 2017 e 2019. A maioria dos estudos selecionados foi do ano de 2017 (47%), seguido do ano de 2019 (29%) e de 2015 (24%).

Dados coletados dos artigos	
Keyword	Palavras - chave do artigo
País	País da maioria dos autores
Origem	Brasil/Exterior
Título	Título do artigo
Ano	Ano de publicação
Autores	Autores da publicação
Amostra	Tamanho da amostra
tipo_amostra	Documentos/objetos ou indivíduos
Tipo Análise	Descritiva/Inferencial
Tipo_Teste	Paramétrico/Não paramétrico
Técnica	Técnica estatística utilizada
Tipo_var_resposta	Tipo de variável resposta (contínua ou categórica)
Tipo_var_explicativa	Tipo de variável explicativa (contínua ou categórica)

Figura 7.2 – Informações coletadas das publicações selecionadas

Dos 110 artigos selecionados, todas as informações apresentadas na Figura 7.2 foram coletadas. Os dados em detalhes dos artigos da RSL encontram-se no Apêndice 1.

Na avaliação das palavras-chave, verificou-se que 29 artigos (26%) não apresentavam palavras-chave no resumo da publicação. Nos outros 81 artigos, não foram identificadas muitas similaridades nos termos de busca das publicações. As palavras-chave mais citadas foram *software engineering* (5 artigos) e *empirical study* (5 artigos).

De acordo com os dados analisados, 77 artigos (70%) coletaram dados de documentos/objetos, enquanto que 33 artigos (30%) coletaram dados de indivíduos. Verificou-se que o problema de amostras pequenas ocorre nos estudos que coletam dados de indivíduos, onde em 14 estudos foram realizadas análises de dados com amostras inferiores a 30 indivíduos. A escolha desse ponto de corte foi definida pois de acordo com o Teorema Central do Limite, a partir de 30 observações, as médias amostrais apresentam uma distribuição de probabilidade que se aproxima da distribuição normal, independentemente da forma da distribuição de probabilidade da amostra em si. Dessa forma, a partir desse tamanho amostral, uma das suposições dos testes paramétricos é atendida [73].

Para classificar a origem do artigo, foi classificado o país com maior número de autores na publicação. Caso houvesse o mesmo número de autores de diferentes nacionalidades, foi considerada a origem do primeiro autor. Dessa forma, de acordo com a Figura 7.3, verifica-se que a maioria dos estudos analisados são de origem estrangeira (78%). Embora o percentual de representatividade seja semelhante, de acordo com os resultados observados na Figura 7.4, verifica-se que a maioria dos estudos de autores brasileiros (55%) realiza análise descritiva dos dados, ou seja, não utiliza teste de hipótese. Já nas publicações de autores estrangeiros, a maioria utilizou estatística inferencial (52%), isto é, realizaram teste de hipóteses para generalização dos resultados da amostra para a população.

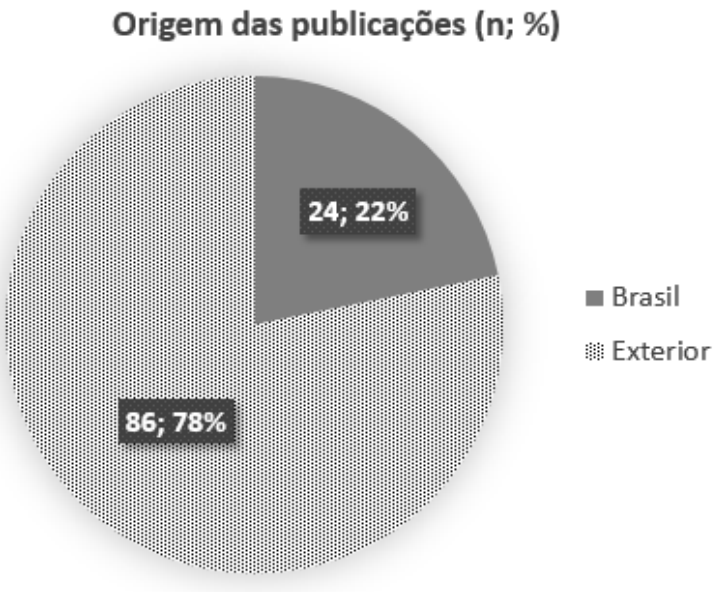

Figura 7.3 – Percentual (%) de origem dos estudos selecionados

Analisando somente estudos que utilizaram estatística inferencial (57 publicações), verifica-se o uso predominante de análise estatística não paramétrica (72%). Segmentando essa informação na origem do estudo (Brasil ou Exterior), de acordo com os dados observados na Figura 7.5, verifica-se que os estudos brasileiros costumam utilizar mais estatística não paramétrica (82%), quando comparado com estudos estrangeiros (68%).

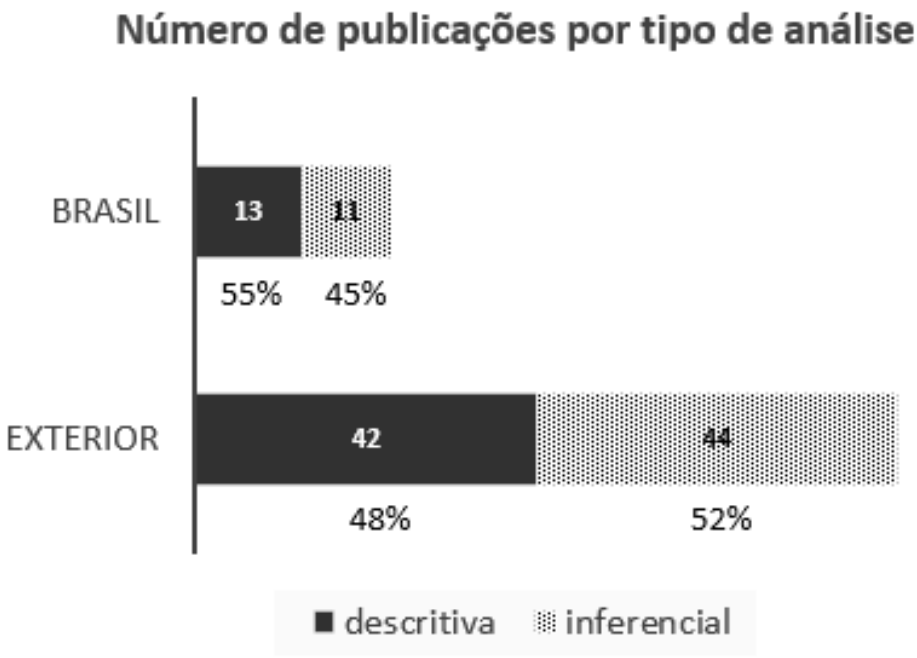

Figura 7.4 – Tipo de análise dos estudos selecionados de acordo com a origem

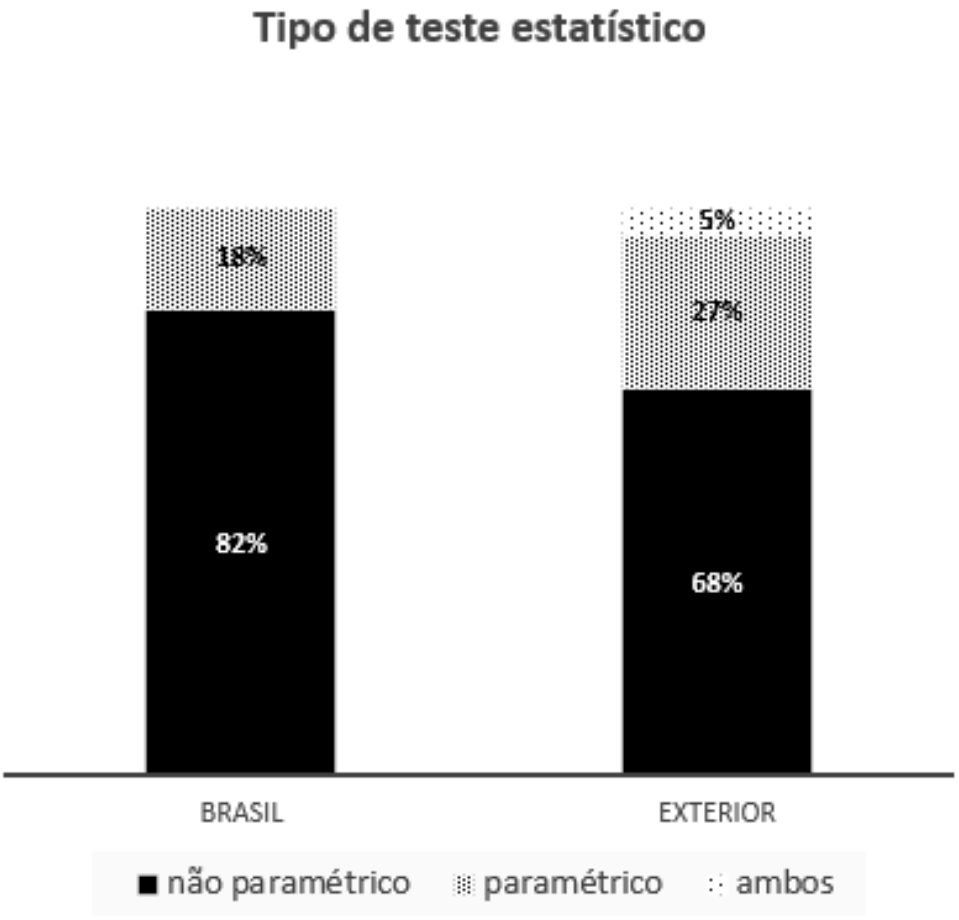

Figura 7.5 – Tipo de metodologia estatística aplicada

Dentre os tipos de variáveis mais frequentes nos estudos, selecionando apenas aqueles que utilizaram estatística inferencial, foi observado que na grande maioria dos artigos (87,72%), a variável resposta era uma informação contínua (Figura 7.6). As variáveis respostas mais frequentes são informações como número de defeitos de um software, indicadores de performance e/ou qualidade de software e outros. Além disso, também verifica-se na Figura 7.6 que o tipo de combinação mais frequente de teste realizado é a combinação de variável resposta contínua e variável(is) explicativa(s) categórica(s) (61,40% dos casos).

Tipo de variável	explicativa			
resposta	categórica	contínua	contínua/categórica	Total Geral
categórica	3 (5,5%)	0 (0%)	0 (0%)	3 (5,5%)
contínua	39 (71%)	3 (5,5%)	8 (14%)	50 (90,5%)
contínua/categórica	1 (2%)	0 (0%)	1 (2%)	2 (4%)
Total Geral	43 (78,5%)	3 (5,5%)	9 (16%)	55 (100%)

Figura 7.6 – Tipo de variáveis (explicativa e resposta

Na análise do tipo de teste estatístico realizado (Figura 7.7), verificou-se que a maioria dos artigos que utilizou estatística inferencial utilizou teste de Wilcoxon.

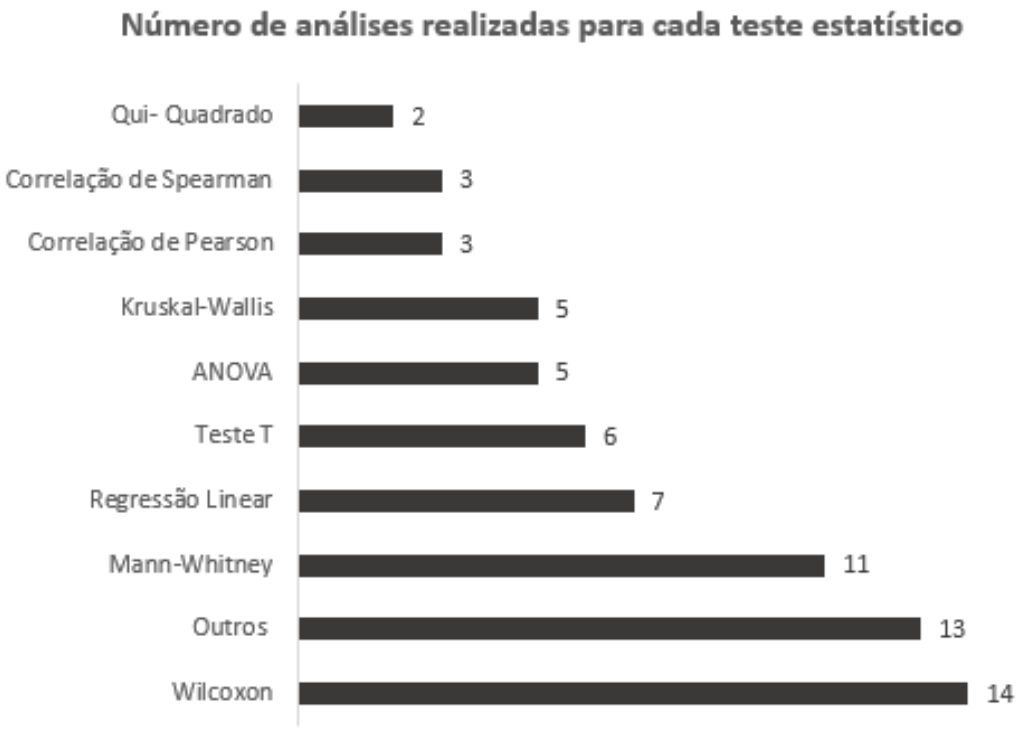

Figura 7.7 – Tipo de teste estatístico realizado

Dessa forma, analisando os resultados obtidos na RSL, verifica-se que o uso de estatística inferencial foi o tipo de abordagem de análise de dados quantitativos utilizada por cerca de metade dos estudos publicados no ESEM nos anos de 2015, 2017 e 2019. Além

disso, verificou-se que nesses casos onde foi utilizada estatística inferencial, os métodos não paramétricos foram muito frequentes, incluindo o uso de testes como Wilcoxon Mann-Whitney e outros.

8. CONSTRUÇÃO E AVALIAÇÃO DO FLUXOGRAMA

8.1 Construção do Fluxograma

De acordo com o estudo e análise dos dados dos artigos elencados na Revisão Sistemática de Literatura (RSL), conforme apresentado no Capítulo 7, foram construídos dois fluxogramas de utilização de testes estatísticos na Engenharia de Software. Esses fluxogramas abrangem os principais testes estatísticos utilizados na análise de dados nas publicações. A fim de facilitar o uso de testes estatísticos na área de Engenharia de Software, foram propostos dois fluxogramas para a utilização de testes estatísticos: (1) quando a variável resposta é contínua e a explicativa é categórica (conforme Figura 8.1); e (2) quando as variáveis explicativa e resposta são contínuas (conforme Figura 8.2).[1]

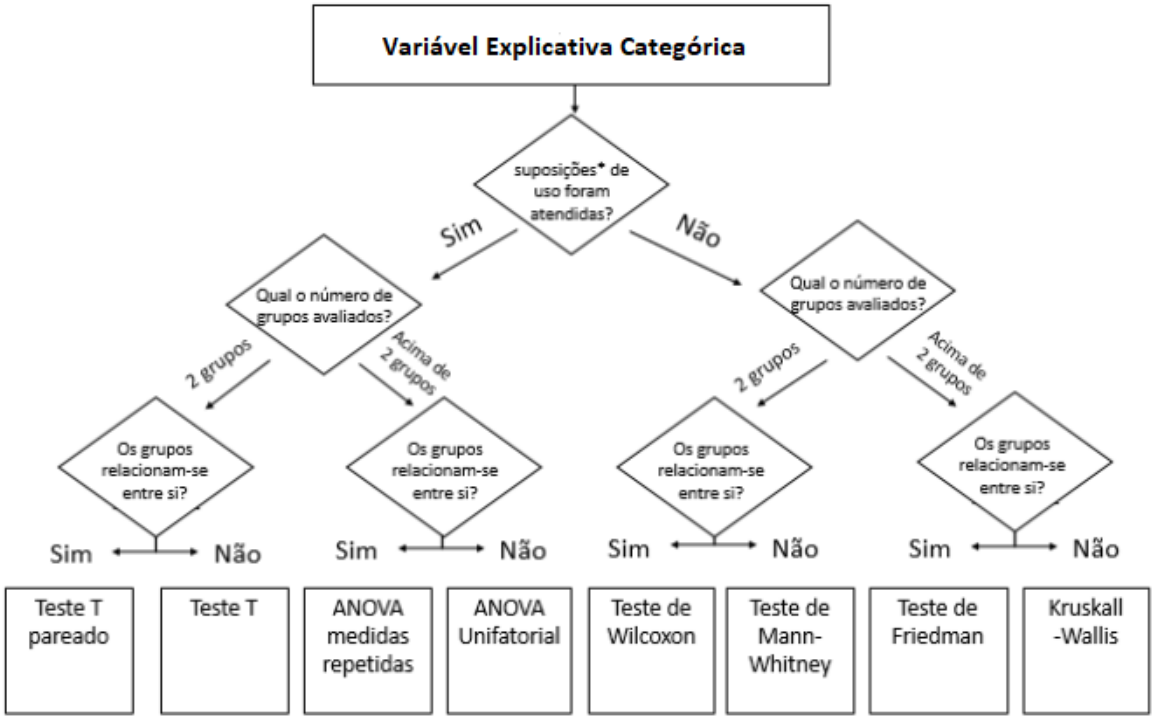

Figura 8.1 – Fluxograma para variável explicativa categórica e variável resposta contínua

Esses primeiros fluxogramas foram construídos somente para variáveis resposta contínuas, que na RSL abrangeram 90,5% dos estudos que utilizaram análise inferencial de dados. Dessa forma, esses foram os fluxogramas validados e apresentados aos pesquisadores da Engenharia de Software, para avaliação e sugestões adicionais, conforme apresentado na seção a seguir.

[1] Note que onde tem-se “suposições* de uso foram atendidas?” nos fluxogramas, isso refere-se aos pressupostos (suposições) dos dados para uso de testes estatísticos, conforme apresentado no Capítulo 4.

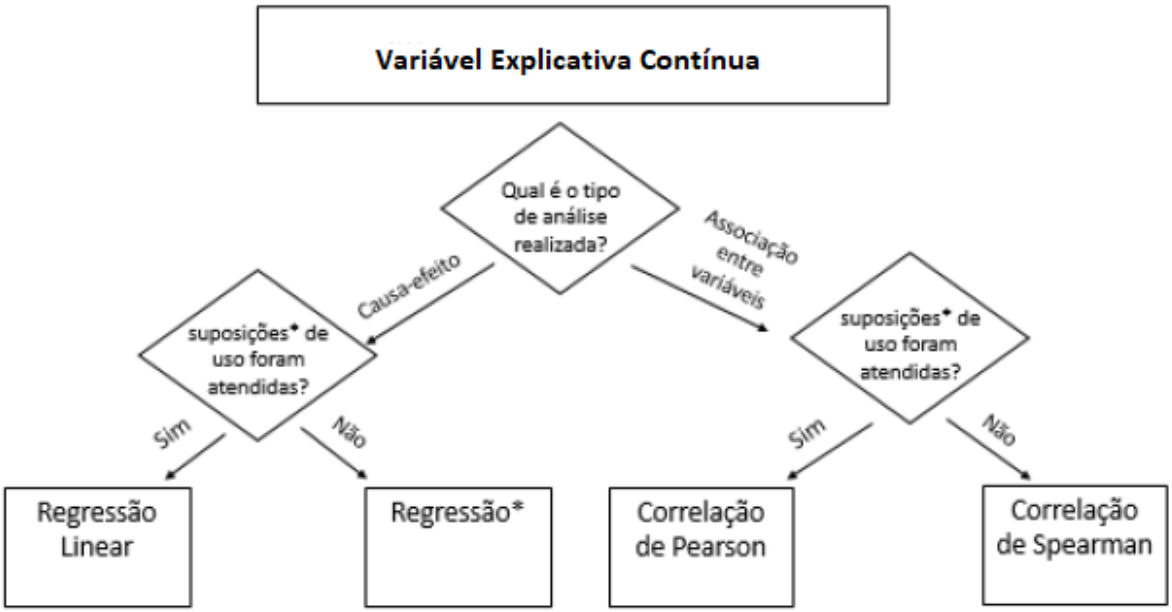

Figura 8.2 – Fluxograma para variável explicativa e resposta contínua

8.2 avaliação dos Fluxogramas

Para avaliação e avaliação do fluxograma, foi executado um painel de especialistas. Esse modelo de avaliação foi proposto por Slocum et al. [127]. Esse painel, através de estudo exploratório, analisa métodos, processos ou modelos buscando pontos fortes, pontos fracos e pontos de melhorias. Para isso, devem ser selecionados especialistas com conhecimento técnico sobre o processo avaliado (principalmente em contextos complexos ou técnicos que exigem um conhecimento muito específico sobre o assunto).

A construção do painel seguiu as recomendações propostas por Slocum et al. [127], e teve como objetivo as seguintes informações:

- Coletar a opinião de pesquisadores da Engenharia de Software sobre o uso de testes estatísticos na área;
- Conhecer os meios de consultas de informações sobre testes estatísticos na Engenharia de Software;
- Reunir sugestões e pontos de melhorias para os fluxogramas;
- Evoluir os fluxogramas, conforme recomendações e sugestões dos especialistas.

Dessa forma, foram selecionados cinco especialistas para avaliação dos fluxogramas propostos. Como o objetivo dessa avaliação não era realizar inferências estatísticas

sobre o uso do fluxograma, mas coletar sugestões e pontos de melhoria para evolução dos fluxogramas, não há problema em trabalhar com amostras pequenas.

Entrevista com os pesquisadores

Os fluxogramas foram apresentados e validados por pesquisadores da área de Engenharia de Software. Esses pesquisadores foram convidados por e-mail pelos autores do estudo e aceitaram contribuir com avaliações e sugestões para o aperfeiçoamento do trabalho. As entrevistas foram realizadas de forma remota, através de um software que permite comunicação pela Internet através de conexões de voz e vídeo.

As entrevistas com os pesquisadores abrangeram uma série de perguntas (Figura 8.3), incluindo informações sobre o uso ou não de testes estatísticos em seus trabalhos, dificuldades no uso da estatística e avaliação dos fluxogramas.

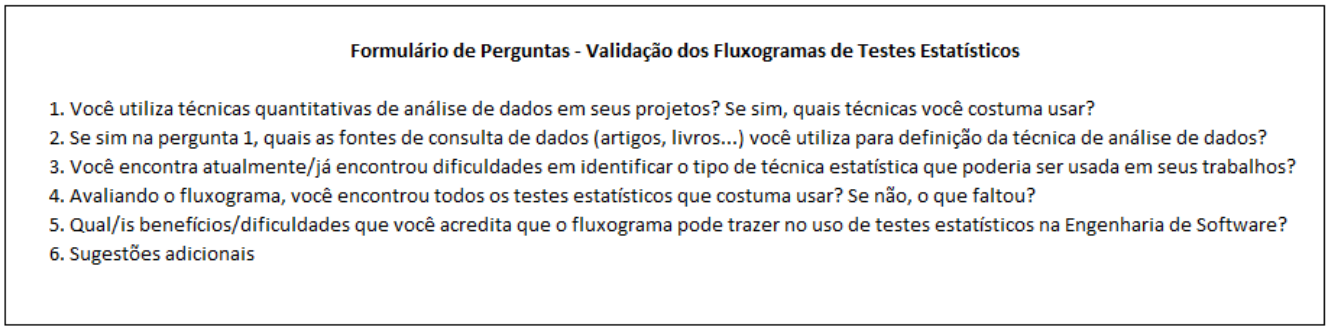

Figura 8.3 – Formulário de questões realizadas aos pesquisadores para avaliação dos fluxogramas de testes estatísticos

Além disso, foram coletadas as seguintes informações sobre os pesquisadores:

- Local de atuação;
- Local de formação;
- Nível de formação acadêmica;
- Tempo transcorrido desde a graduação;
- Tempo transcorrido desde a última formação.

Essas informações foram coletadas visando caracterizar o perfil dos especialistas consultados, abrangendo pesquisadores com diferentes experiências e realidades, o que enriquece o olhar e avaliação das necessidades no uso de testes estatísticos.

As informações demográficas coletadas foram categorizadas da seguinte forma:

- Local de atuação (LA): 1-Brasil, 2-Exterior, 3-Brasil e Exterior;

- Local de formação (LF): 1-Brasil, 2-Exterior, 3-Brasil e Exterior;
- Nível de formação acadêmica (NFA): 1-Graduação, 2-Especialização, 3-Mestrado, 4-Doutorado, 5-Pós-Doutorado;
- Tempo transcorrido desde a graduação (TTG) em anos;
- Tempo transcorrido desde a última formação (TTUF) em anos.

Dessa forma, na Figura 8.4, são apresentados os resultados demográficos dos pesquisadores obtidos nas entrevistas.

Pesquisador	Características				
	LA	LF	NFA	TTG	TTUF
1	1	3	5	12	3
2	1	1	4	17	3
3	3	3	3	13	0
4	1	3	5	14	3
5	3	3	5	20	5

Figura 8.4 – Perfil de formação e atuação dos pesquisadores entrevistados

Através dos resultados apresentados na Figura 8.4, verifica-se que todos os pesquisadores estão atuando no Brasil, sendo que dois deles atuam de forma conjunta no Brasil e Exterior. Além disso, quatro pesquisadores possuem formação mista, com experiência no Brasil e em outros países. Avaliando o nível de formação acadêmica, três pesquisadores possuem pós-doutorado, um deles possui doutorado e o outro possui mestrado (com doutorado em andamento). Todos os pesquisadores possuem mais de dez anos de formação (média de 15,2 anos) e a média de tempo transcorrido da última formação acadêmica é de 2,8 anos.

Quanto ao resultado da primeira pergunta do formulário (Figura 8.3), que se refere ao uso de estatística na Engenharia de Software, os cinco pesquisadores entrevistados afirmaram utilizar testes estatísticos. Alguns dos testes citados pelos pesquisadores abrangem técnicas paramétricas (*teste T*, *Coeficiente de Correlação de Pearson*) e não paramétricas (*Mann-Whitney, Wilcoxon, Kappa, Análise de Cluster*).

Na segunda pergunta do formulário, que verifica as fontes de consulta de informações para uso de testes estatísticos, os pesquisadores relataram que buscam informações em artigos semelhantes ao seu tópico de pesquisa, em artigos publicados na mesma conferência, artigos publicados por autores conhecidos ou através da indicação de algum

coautor. Além disso, foi mencionado o uso de livros de estatística, tutoriais da própria ferramenta de análise (*Minitab*) e o apoio em consultorias estatísticas externas (disponíveis nas universidades em que atuam).

Ao serem questionados sobre as dificuldades encontradas no uso da estatística e na identificação da técnica estatística que poderia ser utilizada nos seus trabalhos (questão 3 da Figura 8.3), todos os pesquisadores afirmaram possuir dificuldades no processo de escolha e utilização de testes estatísticos. O pesquisador 2 (Figura 8.4), afirmou que no início dos seus estudos sobre uso de técnicas estatísticas, não existiam guias de apoio e que o conhecimento estava espalhado. Posteriormente, ao longo do tempo, foram publicados alguns *guidelines* de apoio, como o trabalho da Kitchenham [77]. O mesmo pesquisador também cita dificuldade com alguns conceitos de estatística, como definições de tamanho amostral. Já o pesquisador 3 relatou algumas dificuldades no uso da estatística, como problemas na avaliação dos resultados das análises. Além disso, referenciou dificuldades no uso de questionários com escala Likert, relatando a frase *"não faço idéia de que teste fazer e como fazer"*. Relata ter lido artigos que utilizam respostas da escala Likert como variáveis contínuas e se questionou se essa análise está correta do ponto de vista estatístico. Esse mesmo pesquisador afirma que sempre foi auxiliado por profissionais e consultores da universidade no uso de testes estatísticos, mas que ainda sente falta de orientações práticas (afirma ter tido disciplinas pouco aplicadas de estatística para a Engenharia de Software). Já para o pesquisador 4, os conceitos de população e amostra, bem como planejamento de *survey* são uma grande "dor de cabeça" aos pesquisadores. Esse pesquisador relata que na Engenharia de Software não se tem base confiável de população: *"Não sei quem são os engenheiros de software no mundo"*. Com isso, esse pesquisador relata problemas em identificar perfis de profissionais da área para estratificar as amostras por conveniência, por exemplo.

Resultados do Painel de Especialistas

Seguindo o protocolo de entrevista apresentado na Figura 8.3, foram evidenciados aspectos positivos e negativos relacionados à estrutura proposta, bem como oportunidades de melhorias e sugestões adicionais. Dessa forma, nos próximos tópicos serão descritos os resultados mais relevantes e importantes obtidos no estudo.

- Aspectos Positivos

Um dos aspectos positivos relatado pelos pesquisadores foi a facilidade de compreender o processo de uso do fluxograma através das perguntas norteadoras (necessárias para definição do teste estatístico que pode ser realizado). Outro aspecto positivo foi a simplicidade do fluxograma. Os pesquisadores 3 e 4 afirmaram que o fluxograma pode ser um recurso de consulta simplificado para tomada de decisão sobre que teste aplicar.

Além disso, para todos os pesquisadores, a maioria dos testes que eles afirmam utilizar estavam contemplados no fluxograma.

Para o pesquisador 5, *"o modelo de árvore (utilizado para construção do fluxograma) parece interessante"*. Além disso, todos os pesquisadores afirmaram encontrar dentro do fluxograma de testes estatísticos, algumas das técnicas que eles utilizam ou já utilizaram em suas análises.

Quando foi perguntado sobre os benefícios/dificuldades encontradas com o uso de fluxograma, o pesquisador 1 relatou a seguinte frase: *"Acho que isso (fluxograma) é muito importante para basicamente qualquer pesquisador em Engenharia de Software. Acho que muita gente só roda os testes sem interpretação. Eu mesmo, por exemplo, demorei para saber que tinha teste para saber se a distribuição é normal e que preciso rodar um teste específico depois"*.

- Aspectos Negativos

Para o pesquisador 4, alguns termos utilizados no fluxograma eram desconhecidos para ele: *"Eu verifiquei que variável explicativa é o mesmo que variável independente (fator). Desconhecia esse termo..."*.

O pesquisador 2 afirmou que acredita que somente um guia do uso de testes estatísticos não é suficiente, que o conhecimento já se solidificou nos testes estatísticos mais convencionais. Ele acredita que outros tópicos como séries temporais e análise de sobrevivência poderiam ser abordados. Citou que alguns conceitos são pouco explorados na área como: *"O que é uma variável resposta na engenharia de software?"*. Esse mesmo pesquisador diz que o ideal seria fugir da idéia do *framework* para escolha do teste estatístico e diz que focar na análise dos dados poderia contribuir mais.

Já para o pesquisador 3, são necessárias informações mais detalhadas para definição do teste estatístico, conforme relato do pesquisador: *"Eu acho que o fluxograma em si é bastante claro, mas eu não sei o quanto vai ajudar uma pessoa pouco experiente, porque responder às questões que levam às folhas não é trivial, por exemplo quando o fluxo pergunta se 'suposições de uso foram atendidas', por exemplo... Eu acho que é um começo... mas se eu, por exemplo estivesse com esse fluxograma na mão e isso fosse tudo que eu tinha, não sei se conseguiria confiar em uma decisão. É por isso que eu acho que o fluxograma por si só é difícil de usar... Mas é straightforward... eu conseguiria fazer um bom chute nas respostas para cada casinha do workflow e selecionar um método que talvez fosse o correto. O problema que eu vejo é esse de chutar as respostas, mas isso não é um problema da estatística, isso é um problema da nossa falta de experiência ou da maneira como aprendemos (ou não aprendemos) estatística"*.

Através do relato dos pesquisadores 2 e 3, observamos um distanciamento grande nas necessidades desses pesquisadores. Mesmo com níveis de formação e experiências semelhantes, as necessidades e expectativas com relação ao fluxograma foram diferentes.

Enquanto que para o pesquisador 2 o nível de conhecimento em testes convencionais já é bastante sólido (e então ele sugeriu tópicos mais complexos e profundos como séries temporais e análise de sobrevivência), para o pesquisador 3 alguns conceitos chave para entendimento e bom uso do fluxograma, mesmo no caso de testes convencionais (como apresentados nos fluxogramas), ainda é necessário.

- Oportunidades de Melhorias

Para o pesquisador 1, o fluxograma atende de maneira geral as suas necessidades, mas para ele uma oportunidade de melhoria é a inclusão de testes de tamanho do efeito e testes de concordância, como o teste Kappa.

Para o pesquisador 3, seria ótimo um detalhamento maior do uso de testes estatísticos para a Engenharia de Software. Esse pesquisador afirma que incluir exemplos do uso de testes estatísticos pode contribuir bastante: *"Quando eu aprendi estatística, por exemplo, os exemplos eram todos baseados em ciências naturais... O que não é muito diretamente aplicável (ou mesmo comparável) com a Engenharia de Software"*.

O pesquisador 4 diz que não encontrou todos os tipos de teste estatístico que costuma usar, pois faltaram os testes qui-quadrado e teste de normalidade. Dessa forma, a inclusão desses testes no fluxograma é uma oportunidade de melhoria.

Para o pesquisador 5, poderiam incluir alguns testes adicionais como análise fatorial e processos de reamostragem como *Bootstrap*.

- Sugestões Adicionais

Para o pesquisador 2, uma sugestão é avaliar os problemas do uso da estatística através da análise de artigos. O pesquisador faz o seguinte relato: *"Talvez tenha que olhar mais profundo os problemas específicos (uso de regressão, interpretação dos resultados...) como analisar e não como utilizar"*.

O pesquisador 4 trouxe como sugestão a construção de um framework para auxiliar os pesquisadores com os conceitos de amostra e planejamento de survey. Além disso, esse mesmo pesquisador fez o seguinte relato: *"Boa parte comunidade de pesquisa em Engenharia de Software ainda é relativamente inexperiente com testes estatísticos. Um possível diferencial destas árvores seria orientar sobre as condições (suposições*) necessárias para aplicar cada teste. Por exemplo, tamanhos mínimos de amostra, normalidade. Também sugiro incluir recomendações sobre o processo de amostragem, referente à representatividades de amostras em surveys da Engenharia de Software"*.

Adaptações realizadas no fluxograma

Após a avaliação realizada pelos pesquisadores e com os resultados obtidos na RSL (Capítulo 7), foram realizadas algumas adaptações nos fluxogramas sugeridos. Sendo

assim, visando atender algumas das sugestões realizadas, foram realizadas algumas alterações como a segmentação dos fluxogramas (de acordo com o tipo de variável explicativa e resposta). Dessa forma, foi construído um fluxograma para a combinação variável explicativa categórica com variável resposta contínua (Figura 8.5), um para variável explicativa e variável resposta contínua (Figura 8.6) e um para variável explicativa e variável resposta categórica (Figura 8.7). Nesses fluxogramas, foi realizada a inclusão de alguns testes estatísticos adicionais (qui-quadrado, análise de cluster e análise fatorial), bem como foi incluída uma etapa adicional de testes *Post-Hoc*. O número de artigos que realizou análise com as técnicas apresentadas no fluxograma estão apresentados no próprio fluxograma logo abaixo do nome da técnica[2].

Além disso, foi construído um fluxograma separado para avaliação do tamanho do efeito (*Effect Size*) e para descrever e testar o grau de concordância (confiabilidade e precisão) na classificação (Coeficiente de Kappa), apresentado na Figura Figura 8.8.

[2]Note que onde tem-se “suposições* de uso foram atendidas?” nos fluxogramas, isso refere-se aos pressupostos (suposições) dos dados para uso de testes estatísticos, conforme apresentado no Capítulo 4.

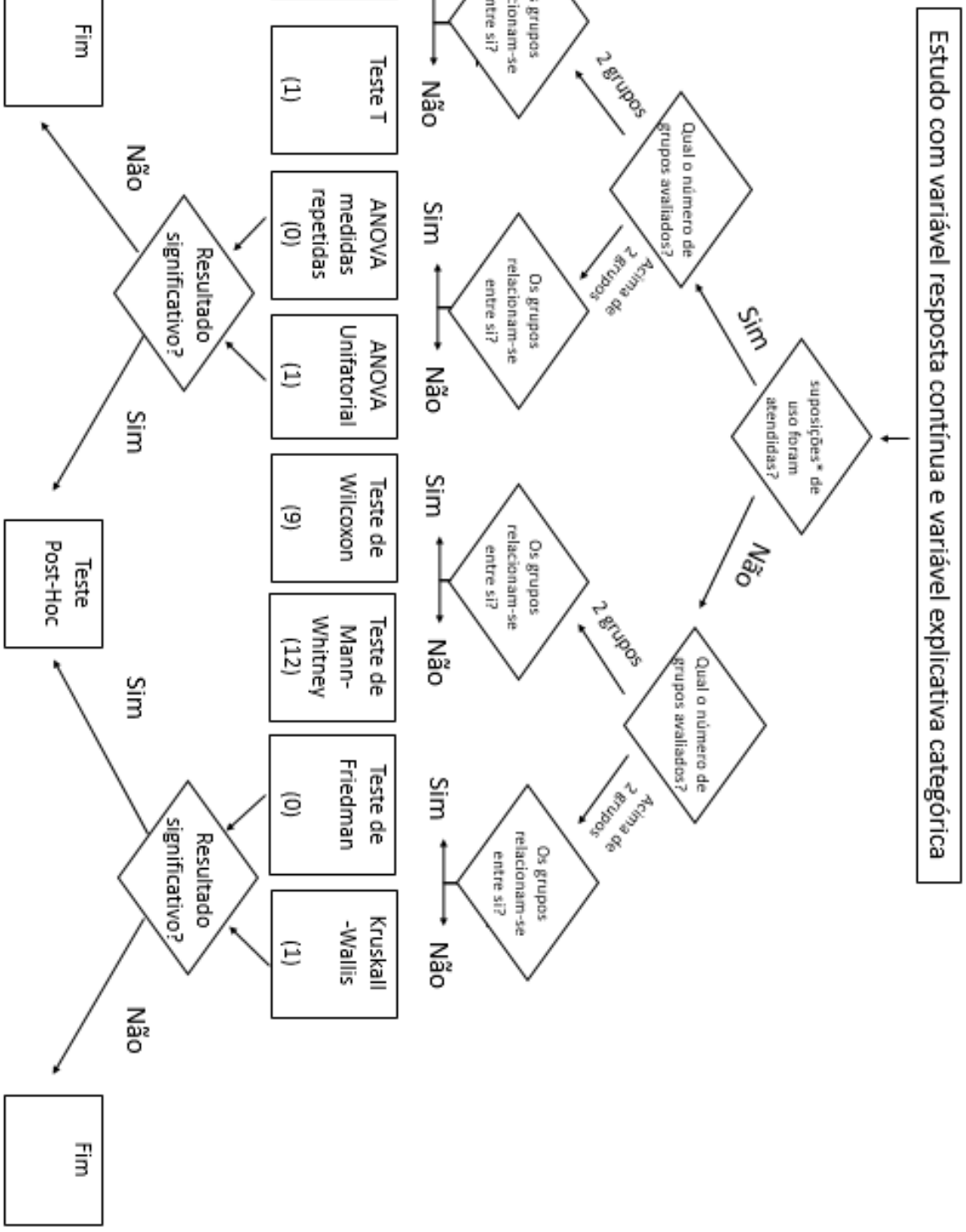

Figura 8.5 – Modelo de fluxograma para ser utilizado quando a variável explicativa é categórica e a variável resposta é contínua

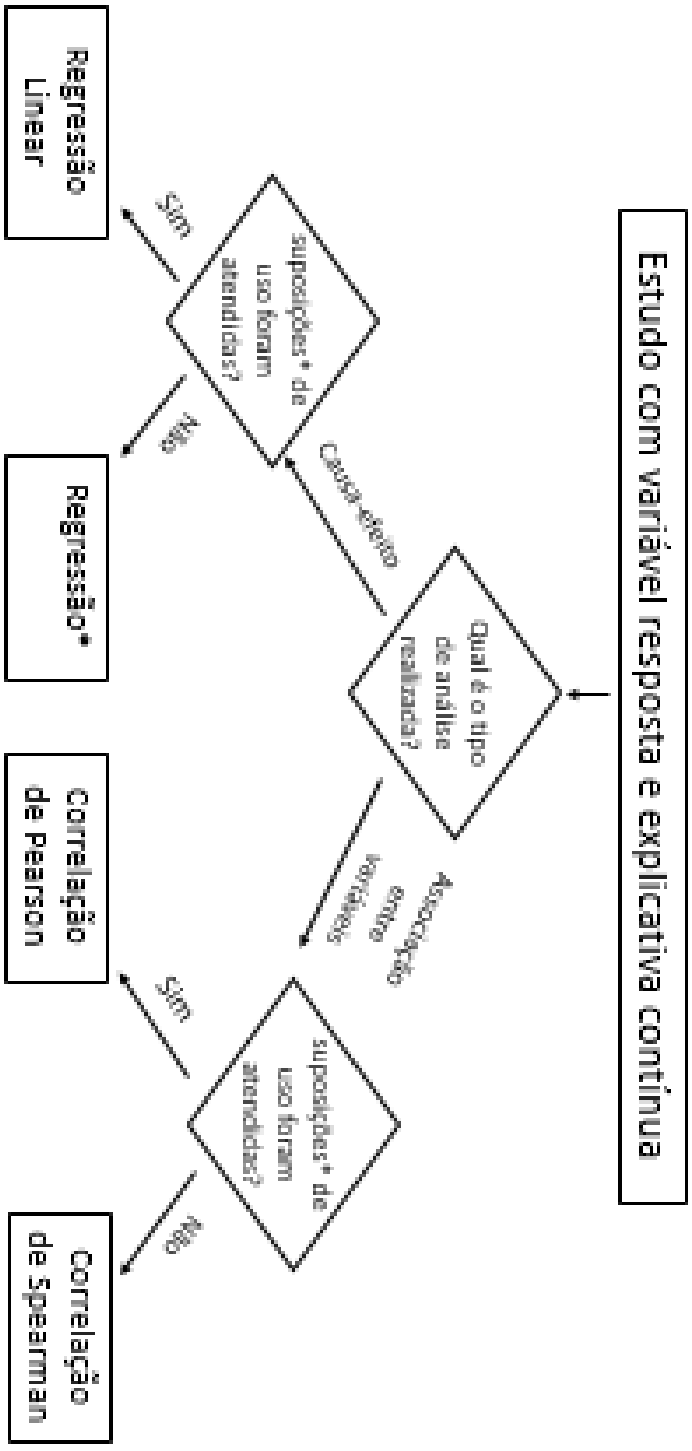

Figura 8.6 – Modelo de fluxograma para ser utilizado quando a variável explicativa e a variável resposta são contínuas

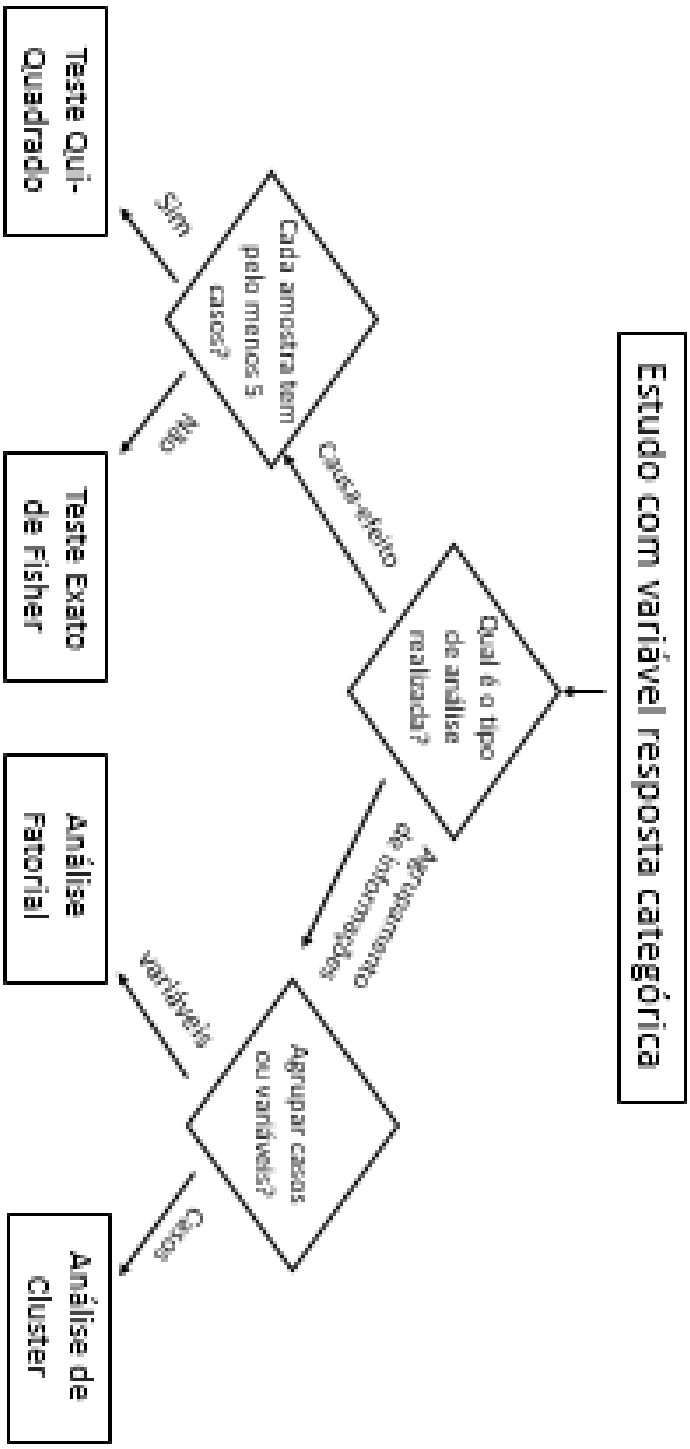

Figura 8.7 – Modelo de fluxograma para ser utilizado quando a variável explicativa e a variável resposta são categóricas

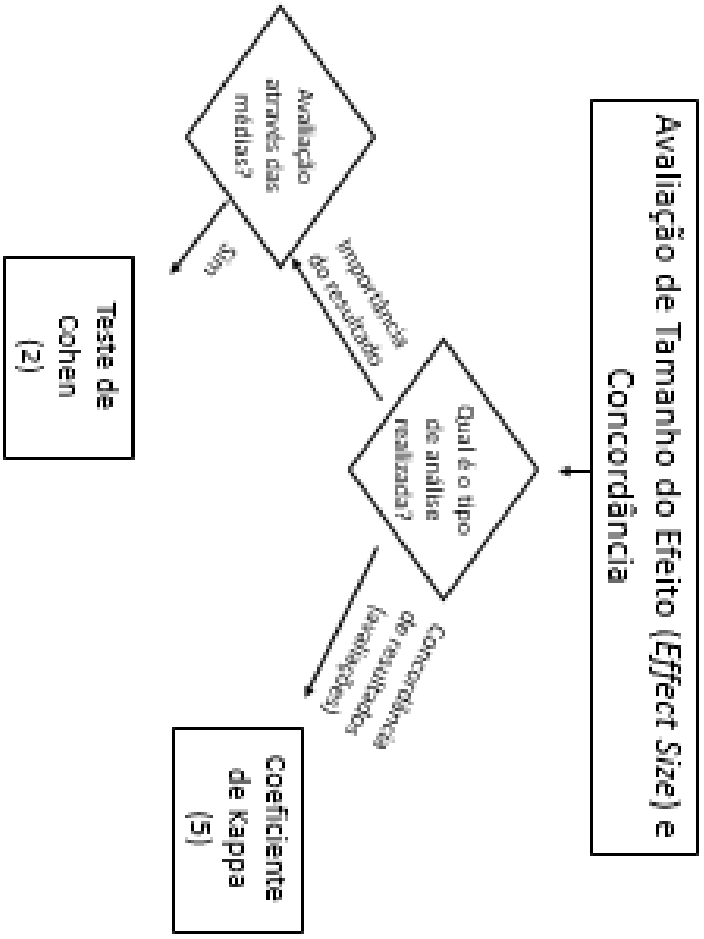

Figura 8.8 – Modelo de fluxograma para utilizado na análise de concordância e tamanho do efeito

9. EXEMPLIFICAÇÃO DE CASOS DE USO DOS FLUXOGRAMAS

Após a construção das versões finais dos fluxogramas de utilização de testes estatísticos na Engenharia de Software (apresentados no Capítulo 8), nesse capítulo serão apresentados alguns exemplos de uso dos fluxogramas para definição do teste estatístico na Engenharia de Software. Os exemplos utilizados para apresentação do uso dos fluxogramas foram construídos a partir de alguns dos artigos analisados na RSL (apresentados no Capítulo 7).

9.1 Exemplo de uso do fluxograma para variável explicativa categórica e variável resposta contínua

Teste T

O estudo de Qin et al. [112] comparou métricas de qualidade de conclusão de tarefas e de satisfação de usuários em versões de aplicativos traduzidas manualmente e através de tradutores automáticos. Para isso, foram recrutados 24 participantes no estudo. No estudo, um participante não recebe versões diferentes de um mesmo aplicativo. Portanto, cada usuário trabalhou em quatro versões diferentes de quatro diferentes aplicativos, com 6 diferentes participantes em cada um dos grupos. Tendo em vista que as variáveis resposta (qualidade de conclusão de tarefas e satisfação de usuários) são variáveis contínuas e que a variável resposta é categórica (tipos de tradução), o fluxograma indicado é o apresentado na Figura 8.5.

A primeira questão abordada nesse fluxograma (Figura 9.1) verifica se as suposições do teste foram atendidas (apresentadas no Capítulo 4).

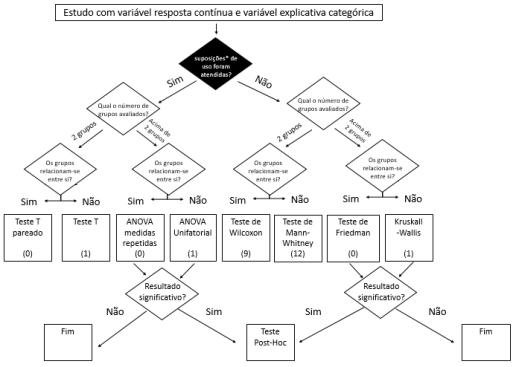

Figura 9.1 – Primeira etapa de escolha do fluxograma utilizado quando a variável explicativa é categórica e a variável explicativa é contínua.

No texto do artigo de Qin et al. [112], não são descritos os resultados dos testes de suposições. Porém, como foi escolhido um teste paramétrico, assume-se que as suposições

foram atendidas. Dessa forma, na Figura 9.2 é apresentado o próximo passo a ser definido na escolha do teste estatístico.

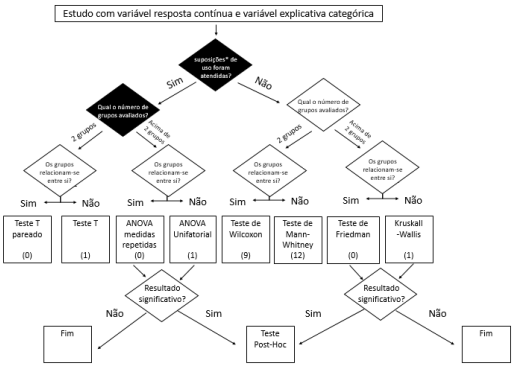

Figura 9.2 – Segunda etapa de escolha do fluxograma utilizado quando a variável explicativa é categórica e a variável explicativa é contínua.

No estudo de Qin et al. [112], os grupos de tradução automática e manual foram comparados em pares. Dessa forma, mesmo sendo quatro grupos de tradução no total, como a avaliação é realizada em pares, são dois grupos avaliados. Com isso, segue na Figura 9.3 a terceira etapa de escolha do teste estatístico.

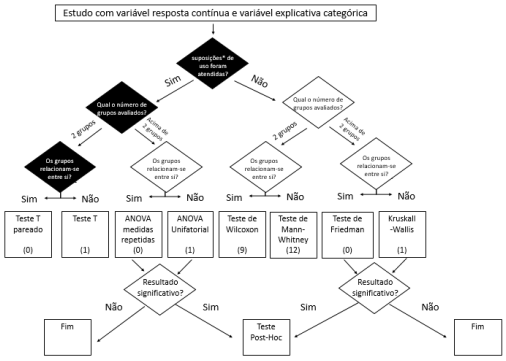

Figura 9.3 – Terceira etapa de escolha do fluxograma utilizado quando a variável explicativa é categórica e a variável explicativa é contínua.

Como são dois grupos analisados, a última pergunta do fluxograma verifica se as amostras são pareadas ou não. Isto é, se um mesmo indivíduo avalia o mesmo aplicativo em diferentes versões ou se esse indivíduo avalia uma mesma versão de mais de um aplicativo. Como cada usuário trabalhou em quatro versões diferentes de quatro diferentes aplicativos, as amostras são independentes. Sendo assim, a escolha final do teste estatístico é o *teste T*, conforme apresentado na Figura 9.4.

Caso os grupos avaliados não fossem independentes (por exemplo, se um mesmo participante tivesse avaliado o mesmo aplicativo em diferentes versões ou se cada participante tivesse avaliado o mesmo tipo de versão de mais de um aplicativo), o resultado final de indicação de teste seria modificado. Nesse caso, o teste final escolhido seria o *teste T* pareado, conforme Figura 9.5.

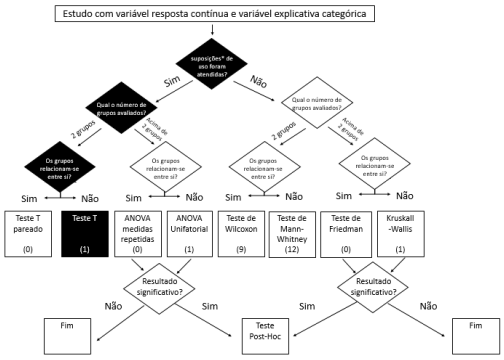

Figura 9.4 – Última etapa de escolha do fluxograma utilizado quando as amostras são independentes (variável explicativa categórica e variável explicativa é contínua).

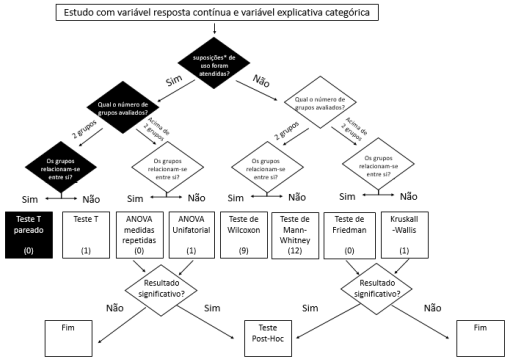

Figura 9.5 – Última etapa de escolha do fluxograma utilizado quando as amostras são pareadas (variável explicativa categórica e variável explicativa é contínua).

ANOVA

No estudo de Rodeghero e McMillan [119] foi realizada uma análise dos padrões do movimento ocular de programadores. Para isso, como variável resposta da análise, foi construída uma métrica de quantificação de similaridade de leitura dos programadores. Como variável explicativa, para entender os tipos de padrões de leitura, foram construídas categorias que agrupavam padrões oculares como “seção 1 à seção 3” e “seção 5 à seção 2” com padrões tais como “da esquerda para direita” e “de cima para baixo”. Dessa forma, como é um estudo com variável resposta contínua e variável explicativa categórica, deve ser utilizado para análise o fluxograma da Figura 8.5.

A primeira questão (Figura 9.6) abrange as suposições de uso de testes paramétricos (apresentadas no Capítulo 4).

Não foram descritos os resultados dos testes no artigo de Rodeghero e McMillan [119]. Porém, como foram realizadas análises com testes paramétricos, as suposições dos modelos devem ter sido atendidas. Sendo assim, a segunda etapa de definição do teste estatístico é apresentada na Figura 9.7.

Como os tipos de padrões oculares mapeados do estudo de Rodeghero e McMillan [119] abrangeram mais de duas categorias, na questão sobre o número de grupos avaliados

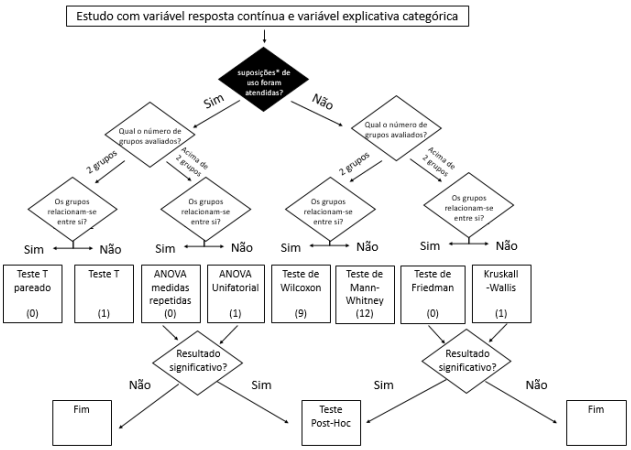

Figura 9.6 – Primeira etapa de escolha do fluxograma utilizado para variável explicativa categórica e variável explicativa contínua.

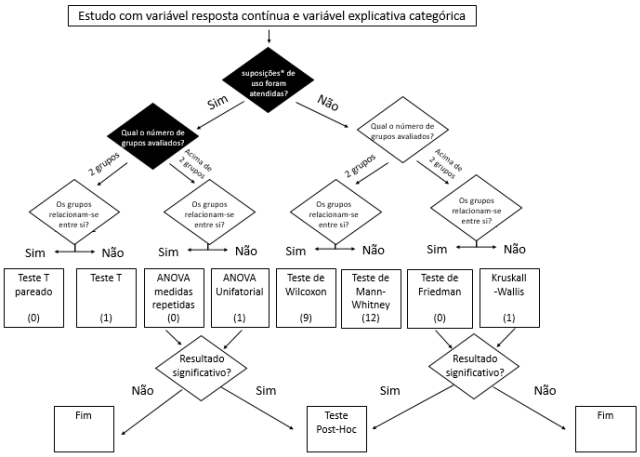

Figura 9.7 – Segunda etapa de escolha do fluxograma utilizado para variável explicativa categórica e variável explicativa contínua.

a reposta é “acima de dois grupos”. O resultado dessa etapa no fluxograma é apresentado na Figura 9.8.

A questão de definição do teste estatístico no fluxograma avalia se os grupos (amostras) são independentes. Como os programadores foram classificados com somente um padrão ocular (não sendo possível classificar um programador em mais de um dos grupos), as amostras avaliadas são independentes. Dessa forma, o teste escolhido foi a ANOVA unifatorial, conforme Figura 9.9.

Caso o padrão ocular dos desenvolvedores tivesse sido avaliado em dois momentos distintos, o resultado no fluxograma da Figura 9.9 teria sido alterado. Nesse caso, o resultado de teste seria a ANOVA para medidas repetidas, conforme Figura 9.10.

Quando são comparados mais de dois grupos, é necessário realizar uma etapa adicional do fluxograma, conforme apresentado na Figura 9.11, onde verifica-se se o resultado do teste é estatisticamente significativo.

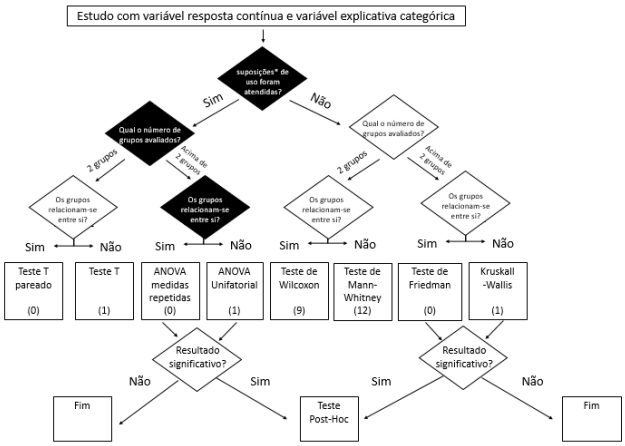

Figura 9.8 – Terceira etapa de escolha do fluxograma utilizado para variável explicativa categórica e variável explicativa contínua.

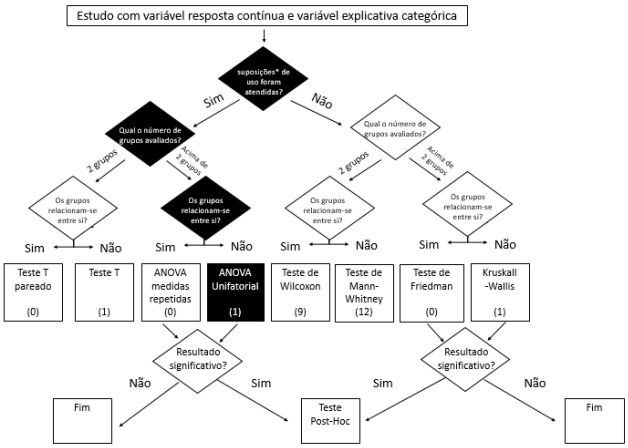

Figura 9.9 – Etapa de escolha do teste estatístico no fluxograma utilizado para variável explicativa categórica e variável explicativa contínua.

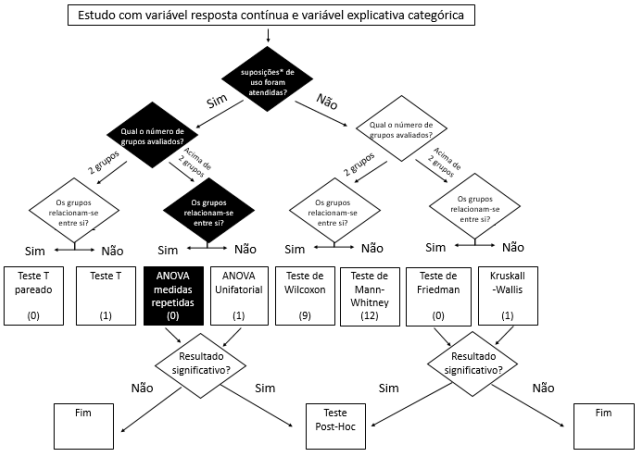

Figura 9.10 – Etapa de escolha do teste estatístico no fluxograma utilizado para variável explicativa categórica e variável explicativa contínua com medidas repetidas.

Caso o resultado seja afirmativo, então é necessário realizar um teste de comparações múltiplas, para avaliar os grupos em pares (conforme apresentado no Capítulo 4).

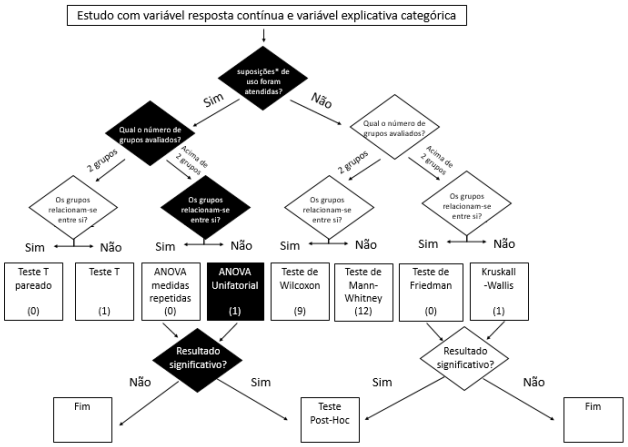

Figura 9.11 – Etapa adicional de escolha do teste estatístico no fluxograma quando são avaliados mais de dois grupos (para variável explicativa categórica e variável explicativa contínua)

Dessa forma, como no estudo de Rodeghero e McMillan [119] o resultado foi significativo, o resultado do fluxograma seria o apresentado na Figura 9.12.

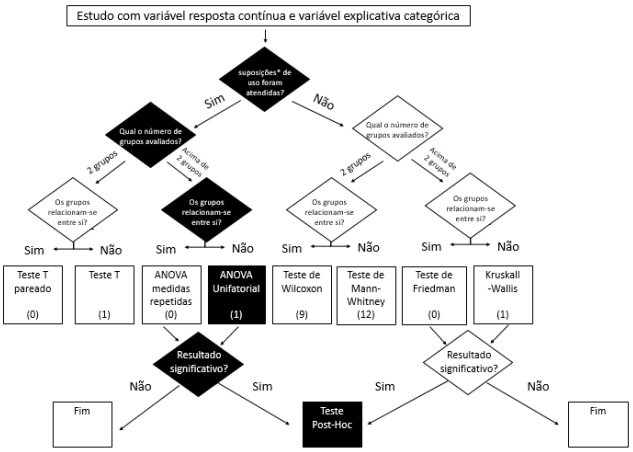

Figura 9.12 – Etapa de definição do teste de comparações múltiplas no fluxograma utilizado para variável explicativa categórica e variável explicativa contínua.

No caso de avaliação de medidas repetidas, conforme apresentado na Figura 9.10, caso os resultados fossem significativos, também seria necessário utilizar testes de comparações múltiplas, conforme o fluxograma da Figura 9.13.

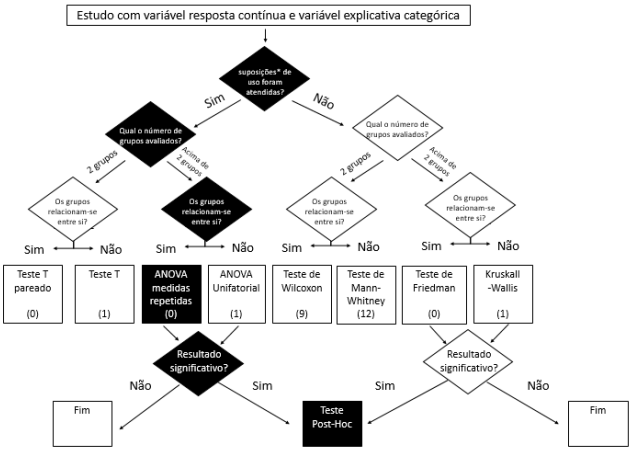

Figura 9.13 – Etapa de definição do teste de comparações múltiplas no fluxograma utilizado para variável explicativa categórica e variável explicativa contínua

Teste de Wilcoxon

O objetivo do estudo de Labunets, Massaci e Tedeschi [78], através de entrevista com 573 desenvolvedores de software, foi investigar o efeito da complexidade das perguntas e da notação de linguagem na compreensão sobre riscos de segurança de projetos de software. Diante disso, uma hipótese testada foi verificar se havia diferença entre questões simples e complexas no nível de compreensão (medida pela precisão e recuperação de respostas) sobre as informações de diferentes modelos de risco. Os mesmos modelos de risco foram comparados através de perguntas simples e perguntas complexas. Como trata-se de um estudo cuja variável resposta é uma avaliação contínua (nível de compreensão) e a variável resposta é categórica (complexidade da pergunta), deve ser utilizado o fluxograma apresentado na Figura 8.5.

A primeira questão do fluxograma (Figura 9.14) é sobre as suposições de uso de testes paramétricos (apresentadas no Capítulo 4).

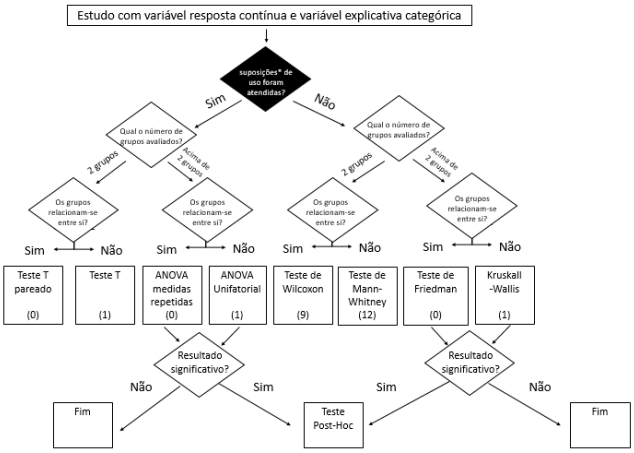

Figura 9.14 – Primeira etapa de escolha do fluxograma utilizado quando a variável explicativa é categórica e a variável explicativa é contínua.

No estudo de Labunets, Massaci e Tedeschi [78] foi utilizado para análise de dados o teste de Wilcoxon, que é um teste não paramétrico. Logo, nesse estudo não podemos assumir que as suposições do modelo tenham sido atendidas, gerando a resposta do fluxograma apresentado na Figura 9.15.

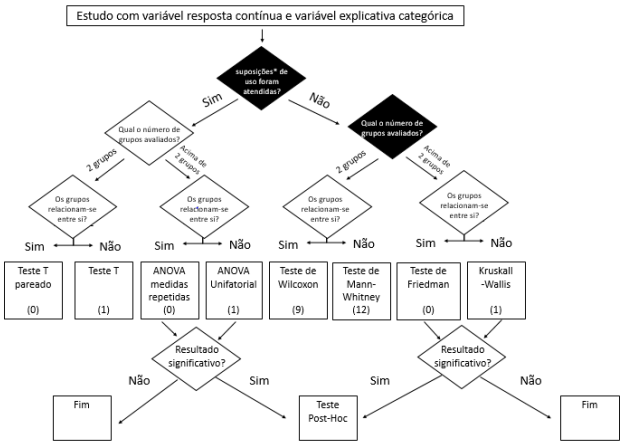

Figura 9.15 – Etapa de escolha do fluxograma quando as suposições de testes paramétricos não são atendidas (variável explicativa é categórica e a variável explicativa é contínua).

A segunda questão analisada no fluxograma é o número de grupos avaliados. Como estudo de Labunets, Massaci e Tedeschi [78] avaliou se havia diferença entre questões simples e complexas no nível de compreensão, o número de grupos avaliados era dois. O resultado dessa etapa do fluxograma pode ser visualizado na Figura 9.16.

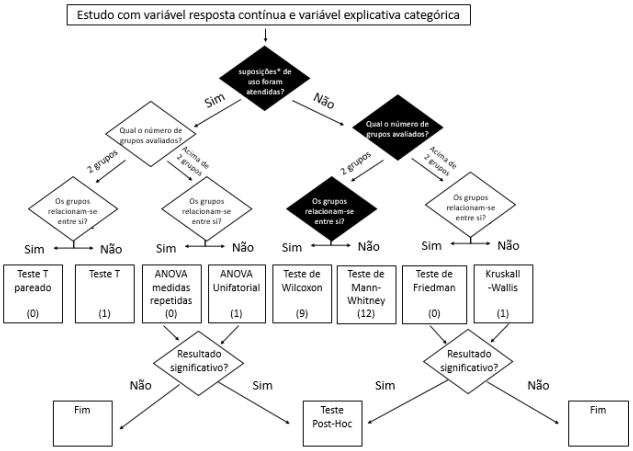

Figura 9.16 – Etapa de definição do número de grupos avaliados no fluxograma de testes não paramétricos (variável explicativa categórica e variável explicativa contínua).

Como são somente dois grupos comparados, a última questão a ser avaliada é se os grupos são pareados ou não. Isto é, se era o mesmo grupo de indivíduos/modelos avaliados em momentos distintos (antes e depois de algum procedimento específico) ou com características muito semelhantes. Como era uma comparação do nível de compreensão dos modelos de risco através de questões simples e de questões complexas, então eram amostras pareadas, gerando como indicação final o teste de Wilcoxon (Figura 9.17).

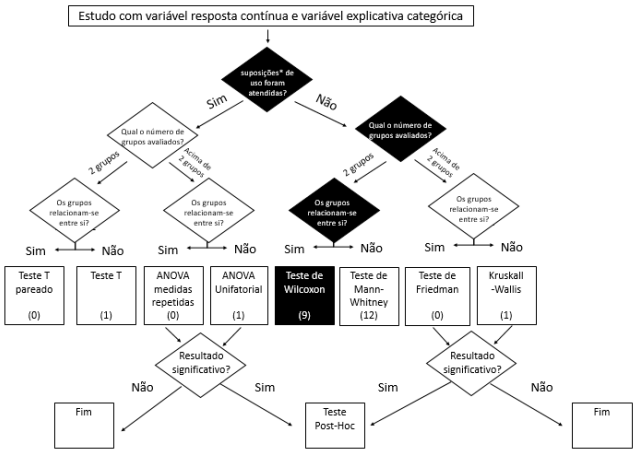

Figura 9.17 – Etapa de definição final do teste estatístico (variável explicativa categórica e variável explicativa contínua)

Teste de Mann-Whitney

No estudo de Islam, Zibran e Nagpal [71], foram estudadas as vulnerabilidades de segurança em códigos clonados e não clonados em 34 sistemas de software de código aberto. Um dos objetivos desse trabalho foi comparar o número de vulnerabilidades de segurança de códigos clonados quando comparados aos códigos não clonados. Sendo assim, esse estudo possui variável resposta quantitativa contínua (número de vulnerabilidades) e variável explicativa categórica (código clonado ou não clonado), sendo então recomendado utilizar o fluxograma apresentado na Figura 8.5.

A primeira questão desse fluxograma (Figura 9.18) verifica se as suposições dos modelos paramétricos foram atendidas. Dessa forma, os primeiros testes realizados devem abranger as suposições de teste apresentadas no Capítulo 4.

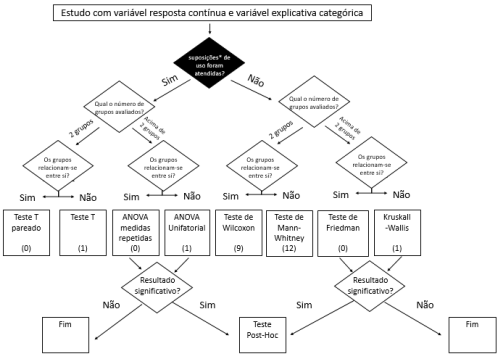

Figura 9.18 – Primeira etapa de escolha do fluxograma utilizado quando a variável explicativa é categórica e a variável explicativa é contínua.

No estudo de Islam, Zibran e Nagpal [71], não foi evidenciado o uso de teste de suposições de dados. Dessa forma, assume-se então que as suposições dos testes paramétricos não foram atendidas. Dessa forma, a partir da resposta de primeira questão, a segunda pergunta abrange o número de grupos avaliados, conforme fluxograma da Figura 9.19.

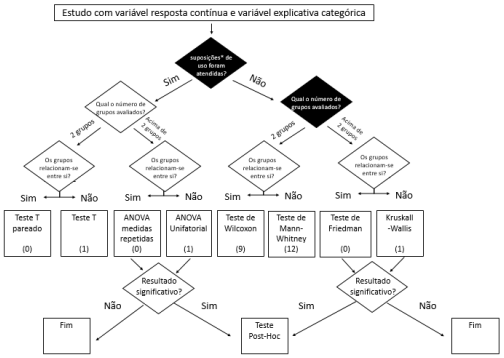

Figura 9.19 – Segunda etapa de escolha do fluxograma utilizado quando a variável explicativa é categórica e a variável explicativa é contínua e as suposições de testes paramétricos não foram atendidas.

A segunda questão do fluxograma aborda o número de grupos avaliados. Como o estudo comparou o número de vulnerabilidades de segurança de códigos clonados e não clonados, o número de grupos avaliados é dois. O resultado dessa etapa de decisão é apresentado na Figura 9.20.

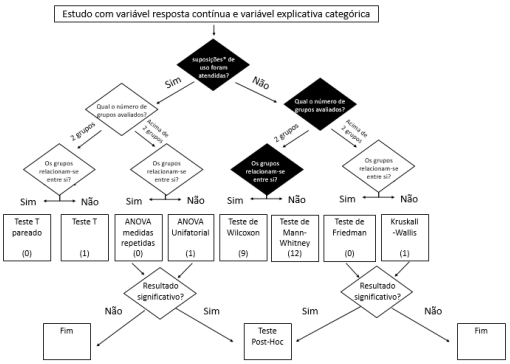

Figura 9.20 – Terceira etapa de escolha do fluxograma utilizado quando são dois grupos avaliados em um teste não paramétrico

Nesse estudo [71], os grupos de códigos clonados e não clonados eram independentes. Isto é, não era o mesmo grupo de código avaliado em momentos distintos (antes e depois de algum procedimento específico) ou com características muito semelhantes. Dessa forma, os testes não devem ser pareados, gerando como definição final o teste de Mann-Whitney, conforme apresentado na Figura 9.21 e apresentado pelo artigo avaliado.

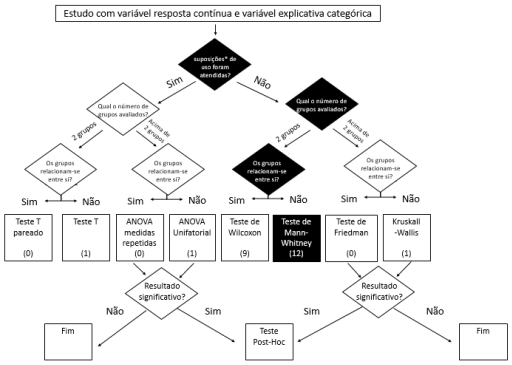

Figura 9.21 – Última etapa de escolha do fluxograma utilizado quando são dois grupos independentes avaliados em um teste não paramétrico

Teste de Kruskal-Wallis e Teste de Friedman

O estudo de Hazhirpasand et al. [66] avaliou a relação entre o desempenho dos desenvolvedores com o número de confirmações que eles realizaram em projetos de software. Para isso, o número de confirmações foi classificado em três categorias distintas: de 2 a 4 confirmações, de 5 a 8 confirmações ou a partir de 9 confirmações. Já o desempenho dos desenvolvedores foi avaliado a partir do número de bugs e confirmações seguras dos desenvolvedores. Dessa forma, como a variável explicativa do estudo é uma informação categórica (número de confirmações em classes), e a variável resposta é uma informação contínua, é necessário utilizar o fluxograma apresentado na Figura 8.5.

A primeira questão a ser respondida nesse fluxograma (Figura 9.22) é se as suposições do uso de testes estatísticos paramétricos foram atendidas (conforme apresentado no Capítulo 4).

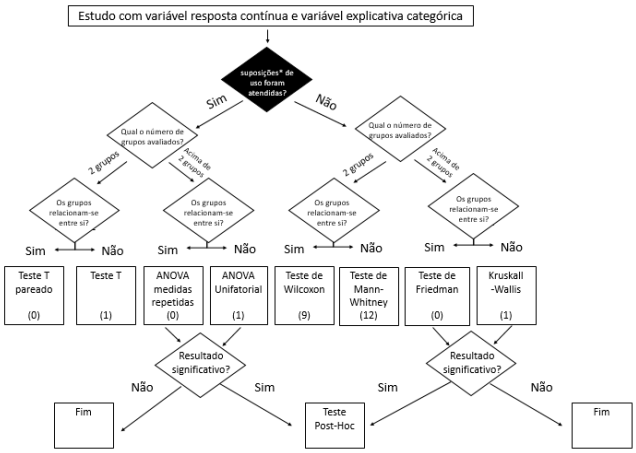

Figura 9.22 – Início do Fluxograma para ser utilizado quando a variável explicativa é categórica e a variável explicativa é contínua.

Conforme apresentado no estudo de Hazhirpasand et al. [66], a suposição de normalidade dos dados do estudo não foi atendida, direcionando a análise para técnicas não paramétricas. A primeira etapa de decisão do fluxograma é apresentada na Figura 9.23.

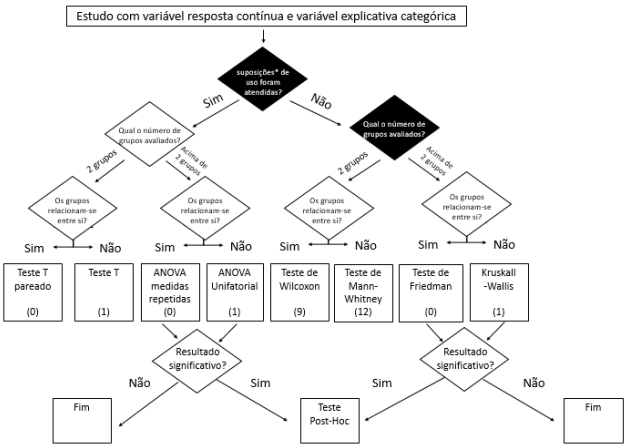

Figura 9.23 – Primeira escolha do fluxograma para ser utilizado quando a variável explicativa é categórica e a variável explicativa é contínua.

A segunda questão a ser respondida no fluxograma é o número de grupos avaliados. Como a variável explicativa (número de confirmações) é classificada em três categorias, a próxima etapa do fluxograma pode ver visualizada conforme resultado apresentado na Figura 9.24.

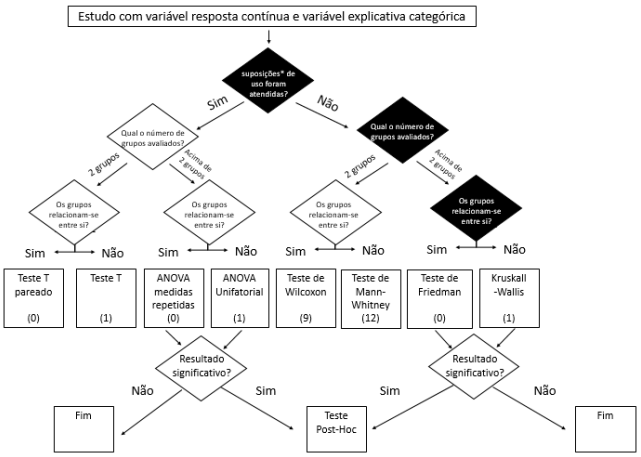

Figura 9.24 – Definição do número de grupos avaliados no fluxograma para ser utilizado quando a variável explicativa é categórica e a variável explicativa é contínua.

A terceira questão a ser respondida é se os grupos avaliados relacionam-se entre si. Essa questão é importante para avaliar se deve ser realizado um teste pareado ou não. Como as categorias da variável explicativa são independentes (não foram os mesmos desenvolvedores avaliados em momentos distintos, por exemplo) a resposta da questão é “Não”, direcionando o fluxograma para o teste a ser realizado, que é o teste de Kruskal-Wallis, conforme Figura 9.25.

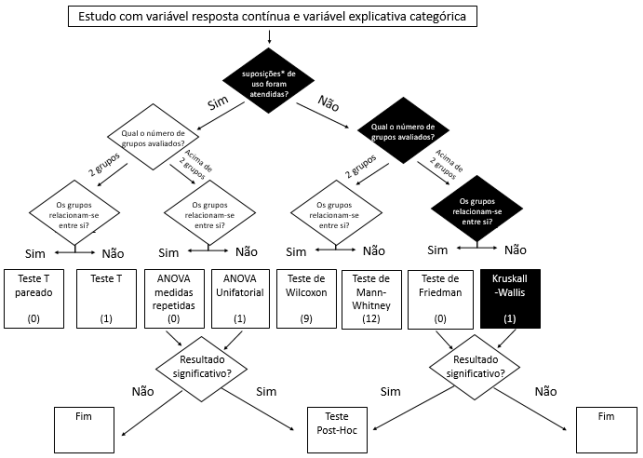

Figura 9.25 – Definição do teste final não paramétrico (com variável explicativa categórica e variável resposta contínua para comparação de mais de dois grupos independentes.

Após a aplicação de um teste estatístico (com variável explicativa categórica e variável resposta contínua) na comparação de mais de dois grupos, é necessário verificar se o resultado do teste é significativo, conforme Figura 9.26. Em caso afirmativo, é necessário realizar uma etapa adicional de testes (testes de comparações múltiplas) para testar em qual(is) par(es) a diferença é estatisticamente significativa, conforme relatado no Capítulo 4.

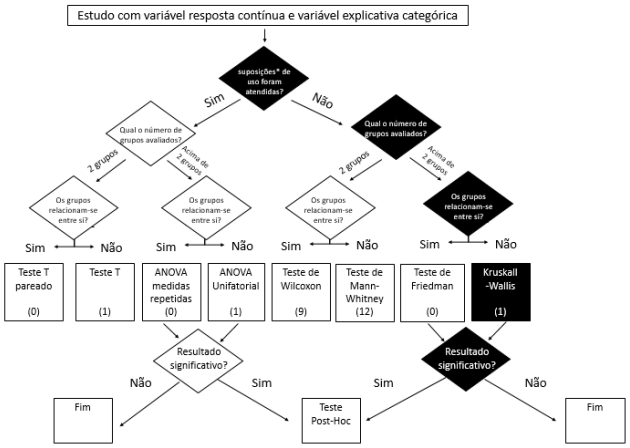

Figura 9.26 – Verificação adicional em testes que avaliam mais de dois grupos (para variável explicativa categórica e variável resposta contínua).

Após a aplicação do teste de Kruskal-Wallis no estudo de Hazhirpasand et al. [66], os resultados foram significativos (*p-valor*<0,001). Dessa forma, como são mais de dois grupos avaliados, é necessário realizar um teste de comparações múltiplas, conforme apresentado na Figura 9.27.

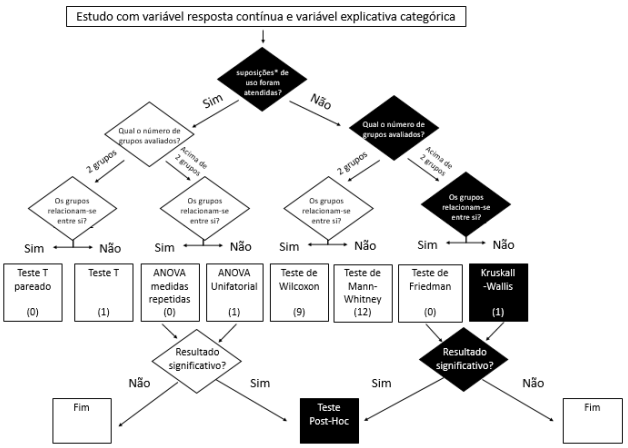

Figura 9.27 – Definição final a ser realizada quando resultado do teste é significativo (em variável explicativa categórica e a variável explicativa é contínua).

Caso os grupos de desenvolvedores do estudo de Hazhirpasand et al. [66] fossem muito semelhantes entre si ou tivessem sido avaliados em tempos distintos (como antes e depois de uma técnica ou procedimento), o fluxograma da Figura 9.25 deveria ser reavaliado. Nesse caso, a melhor escolha de teste não seria o Kruskal-Wallis e sim o teste de Friedman, conforme Figura 9.28.

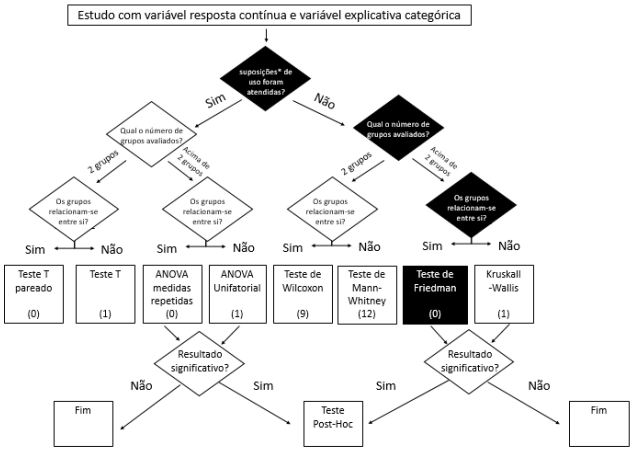

Figura 9.28 – Definição do teste final não paramétrico (com variável explicativa categórica e variável resposta contínua) para comparação de mais de dois grupos pareados.

Assim como no teste de Kruskal-Wallis, após a realização do teste de Friedman é necessário verificar se os resultados são significativos, conforme Figura 9.29.

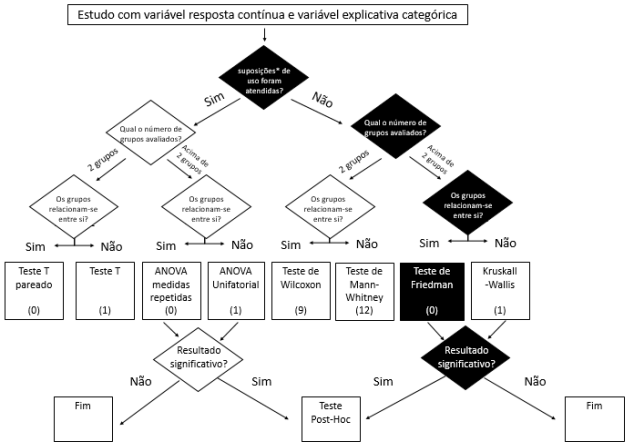

Figura 9.29 – Verificação adicional em testes que avaliam mais de dois grupos (para variável explicativa categórica e variável resposta contínua).

Caso o resultado seja afirmativo (isto é, existe diferença significativa entre os grupos avaliados), é necessário realizar um teste de comparações múltiplas para verificar em qual(is) par(es) ocorre a diferença entre os grupos. Sendo assim, o resultado final da definição dos testes estatísticos está apresentado na Figura 9.30.

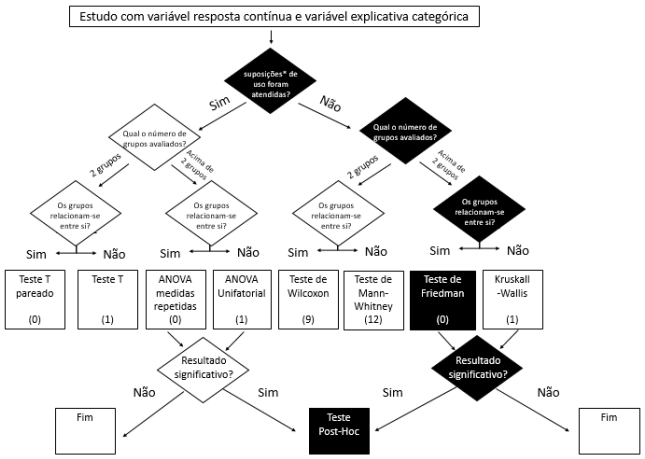

Figura 9.30 – Definição adicional a ser realizada quando resultado do teste pareado é significativo (em variável explicativa categórica e a variável explicativa é contínua).

9.2 Exemplo de uso do fluxograma para variável explicativa e resposta contínua

Teste de Correlação de Pearson e Spearman

O estudo de Rastogi et al. [114] avaliou os fatores que influenciam na jornada de aceleração de profissionais iniciantes para especialistas dentro das empresas. Com isso, para entender os motivos relevantes, foi realizado um estudo analisando os dados extraídos dos sistemas de controle de versão de oito produtos grandes e populares na Microsoft. Foi estudado o tempo necessário para o desenvolvedor realizar o seu primeiro *check-in* na versão de controle do sistema (indicando a primeira contribuição). Além disso, foi avaliado o tempo necessário para esse novo colaborador atingir o mesmo nível de produtividade que os funcionários existentes (em quantidade de *check-ins* realizados).

No estudo, foram entrevistados 411 desenvolvedores, incluindo questões sobre a presença de um mentor, conhecimento prévio das habilidades necessárias e outras. Além disso, foram coletadas algumas outras métricas para avaliar o desempenho de um desenvolvedor, tais como contagem de *check-ins*, quantidade de linhas de código alteradas e número de arquivos alterados. Na análise, foi calculada a relação entre as contagens de check-in em dois momentos (do tempo do primeiro *check-in* com o tempo de aceleração). A mesma análise foi realizada com a quantidade de linhas alteradas e número de arquivos alterados. Para mostrar como seria esse processo de decisão no fluxograma, como em todos

os casos tratam-se de duas variáveis contínuas (variável resposta e variável explicativa), é necessário utilizar o fluxograma apresentado na Figura 8.6.

No fluxograma, a primeira pergunta é se o tipo de análise é de causa-efeito ou associação de variáveis (Figura 9.31).

Figura 9.31 – Início do fluxograma para ser utilizado quando o objetivo é verificar a relação entre duas variáveis contínuas.

Como um dos objetivos do trabalho de Rastogi et al. [114] era verificar se as contagens das variáveis tinham relação nos dois momentos, não tinha uma avaliação de causa-efeito e sim de associação entre as informações. Dessa forma, a Figura 9.32 representa o resultado da primeira escolha a ser realizada no fluxograma para esse estudo.

Figura 9.32 – Resultado da primeira de escolha do fluxograma para ser utilizado quando o objetivo é verificar a relação entre duas variáveis contínuas.

A segunda pergunta para definição do teste estatístico no fluxograma é se as suposições dos testes de associação de variáveis foram atendidas. Isto é, para verificar o relacionamento de duas variáveis contínuas, é necessário verificar se o relacionamento das duas informações é linear ou não. No trabalho de Rastogi et al. [114], não foi especificado a realização de teste de suposição de linearidade das variáveis, sendo escolhido como teste final o teste de Correlação de Spearman, conforme Figura 9.33.

Caso o relacionamento das variáveis fosse linear, a resposta da questão apresentada pelo fluxograma da Figura 9.32 seria afirmativa, sendo então possível utilizar o teste de Correlação de Pearson (conforme Figura 9.34).

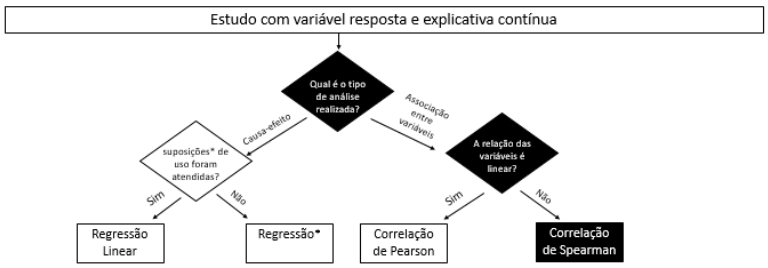

Figura 9.33 – Resultado da segunda escolha do fluxograma para ser utilizado quando o objetivo é verificar a relação entre duas variáveis contínuas (que não apresentam comportamento linear).

Figura 9.34 – Resultado da segunda escolha do fluxograma para ser utilizado quando o objetivo é verificar a relação entre duas variáveis contínuas (que apresentam comportamento linear).

Análise de Regressão

Uma publicação que apresentou variável explicativa e resposta contínua foi o estudo de Spinellis, Louridas e Kechagia [130]. Esse estudo apresentou como objetivo a análise da evolução de Programação C no sistema operacional Unix. Para isso, foram coletados 66 *snapshots* obtidos de um repositório de gerenciamento de software artificial, sendo coletadas diversas métricas quantitativas diferentes de práticas de programação ao longo do tempo. Para análise, as diversas métricas foram analisadas visando identificar tendências longitudinais. Dessa forma, as métricas foram ordenadas por data, sendo realizada uma análise de regressão linear de cada uma dessas informações com os dias corridos desde a primeira *release*. Como tratam-se de duas variáveis contínuas, é necessário utilizar no processo de decisão do teste estatístico o fluxograma apresentado na Figura 8.6.

No fluxograma da Figura 9.35, a primeira pergunta questiona se o tipo de análise é de causa-efeito ou associação de variáveis. Como o objetivo do estudo de Spinellis, Louridas e Kechagia [130] era realizar a análise da evolução de métricas de Programação C no sistema operacional Unix, temos uma relação de causa-efeito, onde quer se analisar o efeito das métricas de programação ao longo do tempo. Sendo assim, a Figura 9.36 representa o resultado da primeira escolha a ser realizada no fluxograma para esse estudo.

Após a primeira definição do fluxograma, a segunda questão abrange as pressuposições do uso de testes estatísticos. Como as variáveis (explicativa e resposta) são

Figura 9.35 – Início do fluxograma para ser utilizado quando o objetivo é verificar a relação entre duas variáveis contínuas.

Figura 9.36 – Resultado da primeira de escolha do fluxograma para ser utilizado quando o objetivo é verificar a relação de causa-efeito entre duas variáveis contínuas.

informações contínuas, as suposições de teste abrangem as definições apresentadas no Capítulo 4 (Seção 4.1). As pressuposições de teste apresentadas nessa seção abrangem os testes de normalidade dos resíduos, homocedasticidade, ausência de correlação entre os resíduos, relação linear entre a variável resposta e as variáveis explicativas e a ausência de multicolinearidade (alta correlação entre as variáveis explicativas). No estudo de Spinellis, Louridas e Kechagia [130] não foram apresentados os resultados dos testes suposições ao longo do texto. Porém, como a análise realizada foi através da técnica de Regressão Linear (conforme Figura 9.37), os resultados dos testes não devem ter inviabilizado o uso da técnica linear.

Figura 9.37 – Resultado da segunda etapa de escolha do fluxograma quando as suposições de teste de modelos com variável resposta e explicativa contínuas são atendidas.

Embora o estudo de Spinellis, Louridas e Kechagia [130] tenha analisado os dados através de regressão linear, não foi realizada uma análise inferencial das informações (nesse caso pode ser dispensável o uso dos testes de suposições dos dados). Porém, essa

prática não é aconselhável (porque mesmo de forma descritiva, é possível interpretar métricas da regressão linear de forma equivocada quando as suposições do modelo linear não são atendidas) [49].

Caso as suposições do modelo linear não tivessem sido atendidas (etapa do fluxograma apresentadas na Figura 9.36) outro tipo de modelo de regressão deveria ser realizado, com técnicas não lineares para análise de variáveis contínuas (explicativa e resposta). Sendo assim, a resposta do fluxograma final mudaria, conforme apresentado na Figura 9.38.

Figura 9.38 – Resultado da segunda etapa de escolha do fluxograma quando as suposições de teste de modelos com variável resposta e explicativa contínuas não são atendidas.

9.3 Exemplo de uso do fluxograma para variável explicativa e resposta categórica

Teste Qui-Quadrado e Teste Exato de Fisher

Além da análise de correlação (apresentada na Seção 9.1), o estudo de Rastogi et al. [114] também avaliou quais fatores os entrevistados avaliavam serem responsáveis por acelerar o desenvolvimento dos novos colaboradores. Para isso, os desenvolvedores de software responderam diversas perguntas em escala Likert (com categorias como *aumento forte*, *aumento moderado*, *diminuição forte*, *diminuição moderada*, *não sei* e *sem efeito*). Para calcular a significância estatística, foram agrupadas categorias originais das perguntas. Foram agrupadas as categorias *aumento forte* e *aumento moderado* sendo construída a categoria *aumento*. Da mesma forma, foram mescladas as categorias *diminuição forte* com *diminuição moderada*, sendo definida a categoria *diminuição*. As outras duas categorias (*não sei* e *sem efeito*) foram consideradas como uma única categoria e instituladas como *sem efeito/não sei*. Alguns dos fatores (variáveis explicativas) avaliados incluiram o conhecimento prévio sobre linguagens de programação, o nível de perguntas proativas realizadas ao gerente, mentor ou equipe, a familiaridade prévia com processos, a manutenção de documentação e listas de tarefas, entre outros. Dessa forma, como ambas as

variáveis do estudo (variável explicativa e resposta) são categóricas, é necessário utilizar o fluxograma da Figura 8.7.

A primeira questão a ser respondida do fluxograma é o tipo de análise a ser realizada (Figura 9.39).

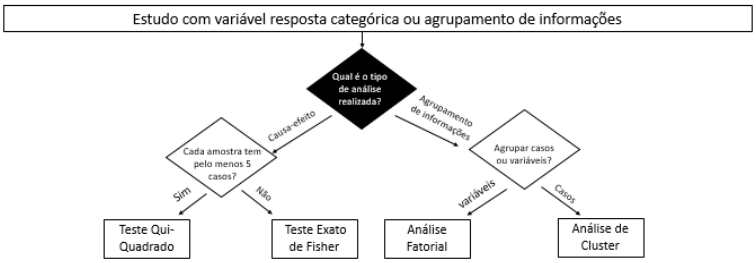

Figura 9.39 – Início do fluxograma para ser utilizado quando a variável resposta é categórica.

Como o objetivo da análise era verificar quais fatores (variáveis explicativas) eram responsáveis por causar o aumento do aceleramento dos novos profissionais, a análise a ser realizada é uma avaliação de causa-efeito, gerando a primeira resposta do fluxograma conforme Figura 9.40

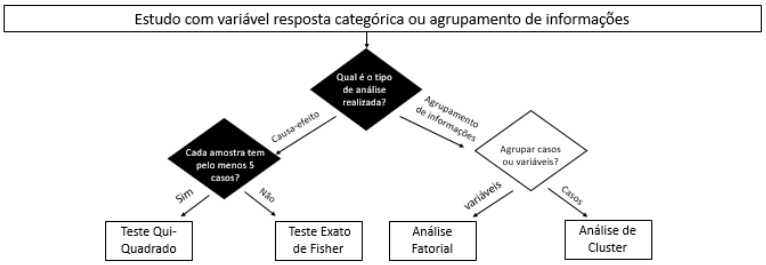

Figura 9.40 – Primeira escolha do fluxograma para ser utilizado quando o objetivo é verificar a relação de causa-efeito entre duas variáveis categóricas.

Posteriormente, para definição do teste a ser realizado, é necessário identificar se tem pelo menos cinco casos (amostras) para cada uma das combinações de categorias de cada fator analisado. No texto do artigo de Rastogi et al. [114] não está descrito o número de respondentes de cada categoria (*aumento*, *diminuição* e *sem efeito/não sei*) em cada fator (variável explicativa) analisado. Porém, como a amostra total foi composta por 411 respondentes, é provável existir mais de cinco casos para cada uma das combinações possíveis. Com isso, a indicação de teste estatístico final é o teste qui-quadrado, conforme Figura 9.41.

Caso o estudo avaliado não apresente pelo menos cinco casos de cada categoria para cada fator (em pelo menos em 80% das possíveis combinações), a resposta da pergunta apresentada no fluxograma da Figura 9.40 é negativa. Nesses casos, não deve ser realizado o teste qui-quadrado, sendo indicado o uso do teste exato de Fisher. Um exemplo do final do fluxograma nesses casos é o resultado apresentado na Figura 9.42.

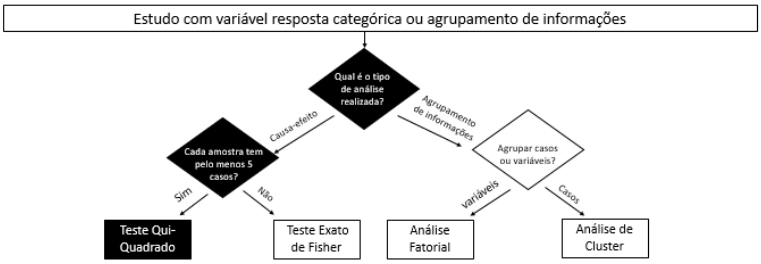

Figura 9.41 – Escolha do fluxograma de variável resposta categórica, quando há pelo menos cinco casos de cada categoria para cada fator avaliado.

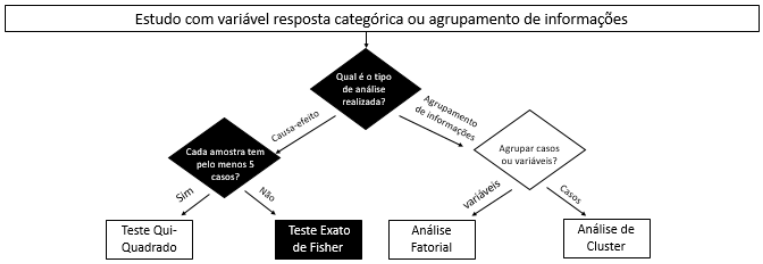

Figura 9.42 – Escolha do fluxograma de variável resposta categórica, quando não há pelo menos cinco casos de cada categoria para cada fator avaliado.

Análise de Cluster e Análise Fatorial

O estudo de Ford et al. [93] teve como objetivo criar perfis de profissionais da Engenharia de Software. Para isso, foram coletadas diversas informações dos perfis de 868 engenheiros de software de uma grande empresa. As informações coletadas desses profissionais envolviam variáveis como tempo na empresa, experiência profissional, nível de autonomia na empresa e outros. Sendo assim, o fluxograma a ser utilizado é o de variável resposta categórica (Figura 8.7), uma vez que o retorno esperado são as categorias (*personas*) de profissionais.

A primeira questão a ser respondida nesse fluxograma é o tipo de análise realizada (Figura 9.43).

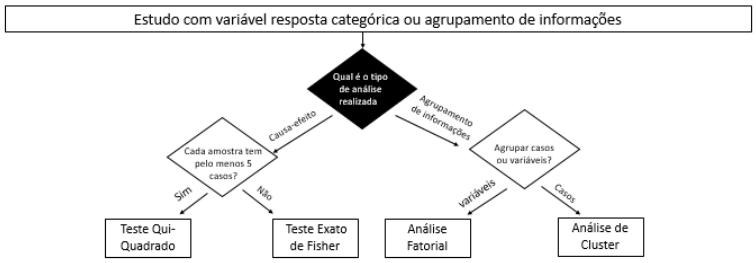

Figura 9.43 – Início do fluxograma para ser utilizado quando a variável resposta é categórica.

Dessa forma, no estudo de Ford et al. [93], o objetivo da análise é o agrupamento de informações, conforme apresentado na Figura 9.44.

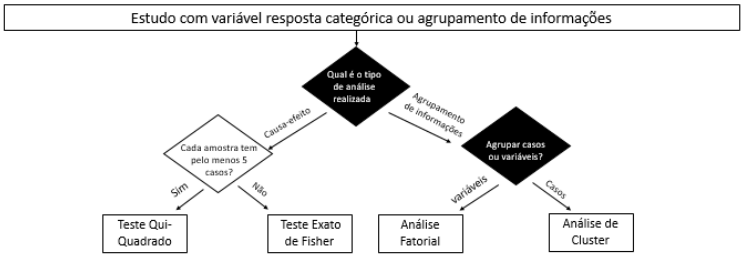

Figura 9.44 – Primeira escolha do fluxograma para ser utilizado quando o objetivo é o agrupamento de informações.

A segunda questão a ser respondida no fluxograma é o tipo de agrupamento que será realizado. No caso do estudo de Ford et al. [93], como o objetivo era construção de *personas*, eram os profissionais que deveriam ser agrupados, resultando no teste de Análise de Cluster (fluxograma final apresentado na Figura 9.45).

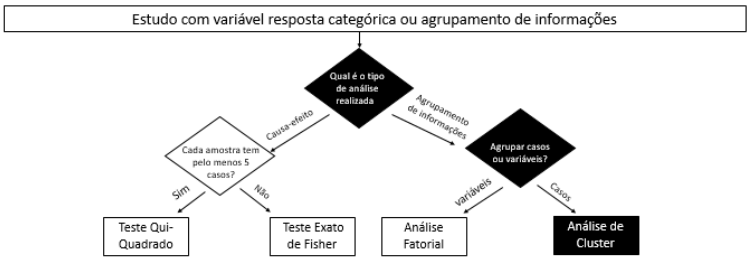

Figura 9.45 – Escolha final do fluxograma de agrupamento de casos.

Caso o objetivo do estudo fosse reduzir o número de características analisadas dos profissionais da Engenharia de Software (visando não precisar armazenar tantas informações, por exemplo), então seria realizado um agrupamento de variáveis e então o teste estatístico realizado seria Análise Fatorial (conforme Figura 9.46).

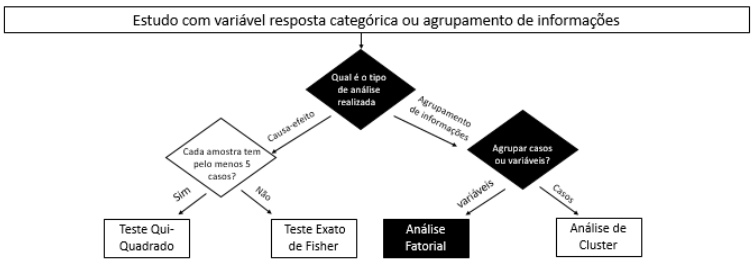

Figura 9.46 – Escolha final do fluxograma de agrupamento de variáveis.

9.4 Exemplo de uso do fluxograma para avaliação de concordância e tamanho do efeito *Effect Size*

Coeficiente de Kappa

O estudo de Poulding et al. [110] teve como objetivo avaliar o impacto das citações de artigos acadêmico. Para isso, foi realizada uma avaliação inicial dos artigos utilizando a taxonomia do comportamento de citação proposta por Bornmann e Daniel. Além disso, foi realizada uma classificação de forma independente as citações em dez trabalhos publicados no ESEM. Para análise, o grau de concordância das classificações realizadas pelos pesquisadores foi avaliado, através do uso do coeficiente de Kappa. Diante disso, a etapa de decisão do fluxograma a ser utilizado para essa análise é apresentada na Figura 9.47 (o mesmo proposto na Figura 8.8).

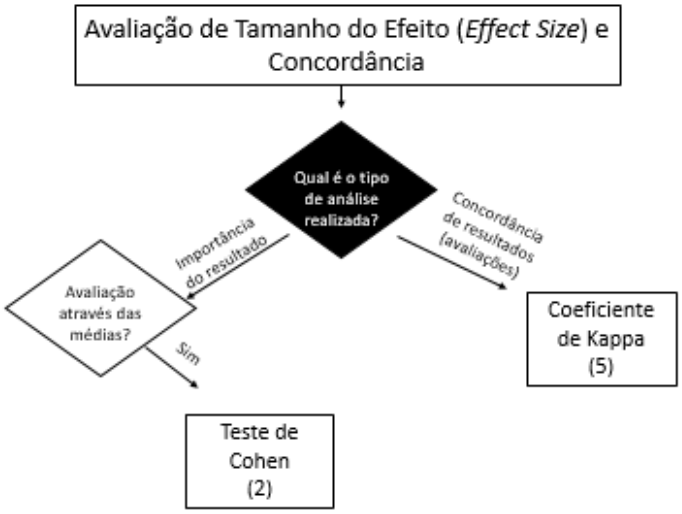

Figura 9.47 – Definição de fluxograma para ser realizado com o objetivo de avaliar consistência de classificações.

A única definição a ser realizada nesse fluxograma é verificar se o tipo de análise é concordância de resultados (avaliações) ou importância de resultados. Como o objetivo era verificar a concordância das informações, o teste estatístico definido é o Coeficiente de Kappa (conforme Figura 9.48).

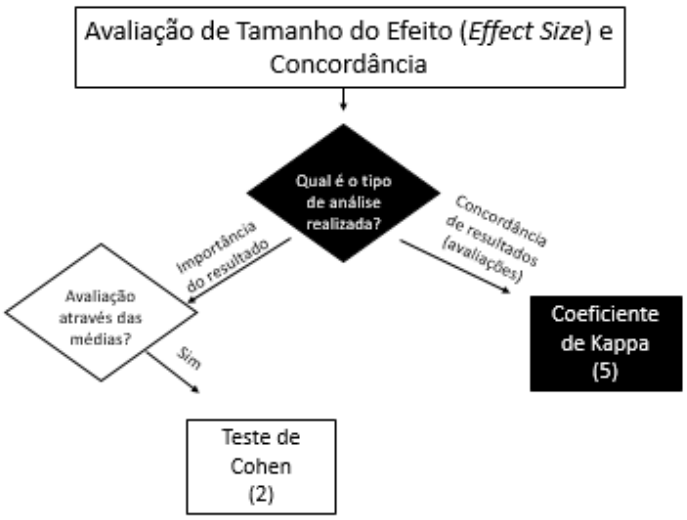

Figura 9.48 – Definição do teste estatístico para ser realizado com o objetivo de avaliar consistência de classificações

Teste de Cohen

O estudo de Griffith e Huvaere [72] teve como objetivo identificar inconsistências na interpretação da qualidade do software através de operacionalização de um padrão em um modelo de qualidade. Para isso, as características de qualidade de dois modelos de qualidade foram avaliados. Para análise de dados, o estudo utilizou a avaliação de tamanho do efeito. Com isso, a primeira etapa a ser respondida no fluxograma da Figura 9.49 é o tipo de análise a ser realizada.

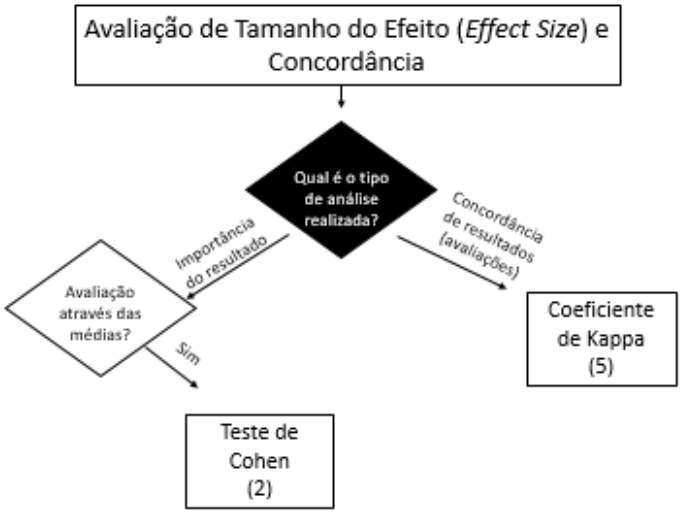

Figura 9.49 – Definição de fluxograma para ser realizado com o objetivo de avaliar importância do resultado.

A segunda questão a ser respondida, é se o tipo de avaliação da importância do resultado é calculada a partir da avaliação das médias (Figura 9.50).

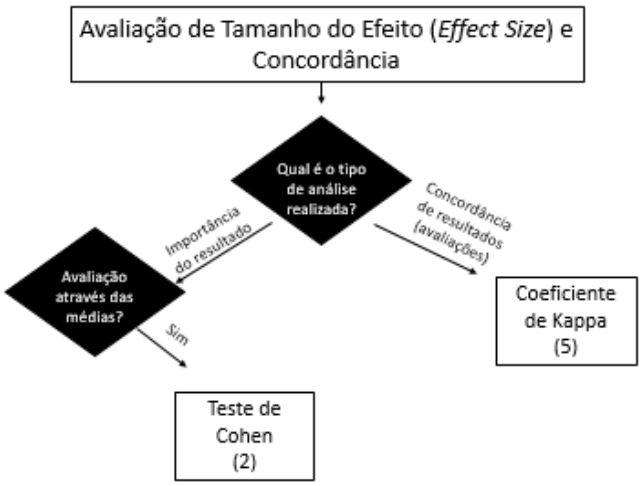

Figura 9.50 – Definição de fluxograma para ser realizado com o objetivo de avaliar importância do resultado.

Como o objetivo do estudo de Griffith e Huvaere [72] era avaliar a importância do resultado (através das características da avaliação de qualidade) através das médias, o tipo de análise indicada é a realização do teste de Cohen, conforme Figura 9.51.

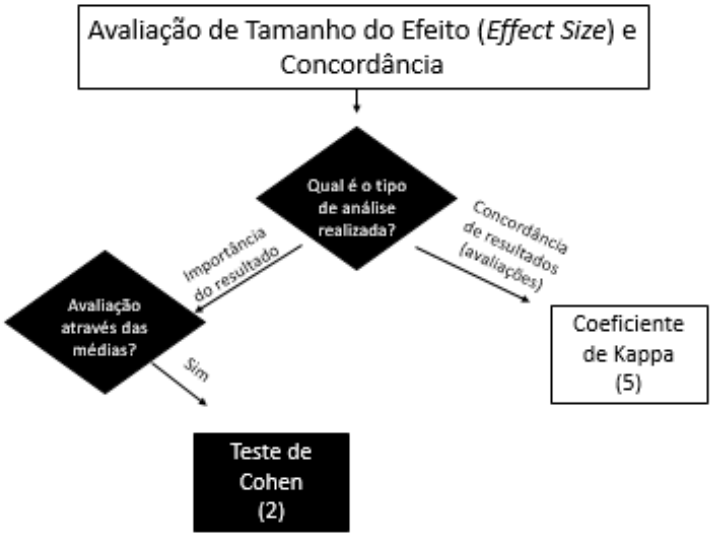

Figura 9.51 – Definição do teste estatístico para ser realizado com o objetivo de avaliar importância do resultado

10. CONSIDERAÇÕES FINAIS

10.1 Revisão dos objetivos de pesquisa

Através dos resultados obtidos com a implementação e exemplificação de fluxogramas de testes estatísticos para a Engenharia de Software, o problema de pesquisa que o presente estudo visou auxiliar foi facilitar o uso de análise estatística de dados para profissionais e pesquisadores da área de Engenharia de Software.

Com isso, os objetivos do estudo incluiam a identificação dos principais estudos quantitativos realizados na Engenharia de Software, construção e avaliação de fluxogramas de testes estatísticos e exemplificação do uso dos fluxogramas em estudos da área.

Além disso, o presente estudo teve a intenção de auxiliar os pesquisadores e profissionais da área de Engenharia de Software a conhecer mais os conceitos de análise de dados estatísticos e conseguir identificar possíveis tipos de técnicas estatísticas para análise dos dados dos seus experimentos.

10.2 Principais resultados encontrados

Os estudos discutidos no Capítulo 5 relataram pouca utilização de técnicas de análise estatística na área de Engenharia de Software. Alguns dos desafios apresentados pelos autores desses estudos foram o pouco conhecimento sobre o tema, além de dificuldades na coleta e análise dos resultados obtidos. Com isso, através de uma revisão sistemática da literatura, um dos resultados obtidos nesta pesquisa foi o maior conhecimento sobre os tipos de análises estatísticas realizadas por pesquisadores na Engenharia de Software.

Nos resultados da Revisão Sistemática de Literatura (RSL), apresentados no Capítulo 7, observou-se que o uso de estatística descritiva ou inferencial foi bem semelhante nos artigos publicados na Engenharia de Software (publicados no ESEM nos anos de 2015, 2017 e 2019). Além disso, nos casos onde foi utilizada estatística inferencial, os métodos não paramétricos foram bastante frequentes, incluindo o uso de testes como Mann-Whitney, Wilcoxon e outros.

Nas publicações do Brasil, a proporção de uso de testes não paramétricos foi maior quando comparados com os artigos publicados no exterior. É interessante de confrontar esses resultados com as entrevistas realizadas com os pesquisadores através do painel com especialistas. Um dos pesquisadores afirmou utilizar muito mais alternativas não paramétricas por receio de cometer algum erro ou interpretação equivocada dos resultados se utilizasse um teste paramétrico. Isso ocorre uma vez que os testes não paramétricos são

mais simples de usar e não exigem uma série de pressuposições que os testes paramétricos exigem. Porém, ao utilizar um teste não paramétrico, como são perdidas informações em decorrência da forma de calcular os resultados, o teste não apresenta tanto poder e precisão de interpretação dos seus resultados.

Diante disso, esse trabalho buscou trazer alguns dos conceitos mais comuns da estatística para serem aplicados na Engenharia de Software, tentando facilitar o entendimento e compreensão do tema para os pesquisadores, visando informar melhor sobre o uso da estatística em estudos quantitativos da área. Além disso, através das sugestões realizadas pelos pesquisadores, os fluxogramas foram alterados, visando atender melhor as necessidades dos pesquisadores. Além disso, através da exemplificação de alguns casos de uso dos fluxogramas, o trabalho visa facilitar a replicação do uso de técnicas estatísticas por profissionais e pesquisadores da área.

10.3 Limitações

Analisando as informações dos pesquisadores entrevistados, verificamos um distanciamento de opiniões sobre o uso dos fluxogramas. Enquanto que para um dos pesquisadores o conhecimento para identificação e uso dos testes estatísticos convencionais já se solidificou, para outro pesquisador entrevistado são necessárias informações mais detalhadas (uma vez que responder às questões que levam as folhas do fluxogramas não é trivial para esse pesquisador). Como esses pesquisadores possuem níveis de formação e experiência semelhantes, não é simples identificar qual é o principal público de pesquisadores na Engenharia de Software que pode se beneficiar do uso dos fluxogramas (se pesquisadores mais iniciantes ou os mais experientes).

Nesse estudo não foram apresentadas e exemplificadas todas as possíveis técnicas estatísticas de análise de dados. Porém, como o objetivo dos fluxogramas é servir como uma ferramenta de apoio para a utilização dos principais testes estatísticos para análise de dados, essa limitação abrange casos mais específicos de análise. Sendo assim, através da RSL, o estudo identificou alguns dos principais tipos de análise de dados quantitativos realizados em Engenharia de Software, incluindo esses resultados nos fluxogramas.

10.4 Benefícios Esperados

Um benefício realizado pelo presente estudo foi o mapeamento de testes estatísticos utilizados por pesquisadores da Engenharia de Software, assim como uma maior identificação das necessidades dos pesquisadores da área no conhecimento de Estatística.

Com a construção dos fluxogramas espera-se auxiliar e orientar os pesquisadores da Engenharia de Software na escolha do teste estatístico. Como o fluxograma foi customizado (visando atender aos resultados encontrados na Revisão Sistemática da Literatura e na avaliação com os pesquisadores), foi possível atender grande parte das necessidades desses pesquisadores, auxiliando na identificação adequada dos testes estatísticos pelos profissionais da área. Entende-se essa customização como uma importante constribuição do estudo para Engenharia de Software, tanto na construção dos fluxogramas como nos exemplos de uso (que abrangem estudos e publicações da área).

10.5 Trabalhos Futuros

Foram encontrados diversos achados no presente estudo, incluindo alguns dos principais tipos de testes estatísticos utilizados na Engenharia de Software, bem como dificuldades encontradas pelos pesquisadores no uso de estatística na área.

Uma das sugestões de trabalhos futuros é a construção de um sistema de recomendação online do uso de testes estatísticos para Engenharia de Software utilizando como escolha e recomendação de uso os fluxogramas de testes estatísticos desse estudo.

Durante as entrevistas com os pesquisadores, surgiram sugestões de melhorias para o presente estudo, assim como sugestões de novos estudos. Um dos pesquisadores sugeriu como trabalho futuro a realização de uma *survey* com estudantes, apresentando um estudo de caso (que poderia ser realizado utilizando análise estatística de dados) e perguntando que tipo de teste eles fariam. Ao mesmo tempo, o pesquisador sugeriu fazer uma *survey* com especialistas na área e confrontar os resultados. O objetivo desse trabalho seria identificar o nível de conhecimento dos profissionais da Engenharia de Software sobre Estatística.

Uma outra sugestão dos pesquisadores entrevistados é a realização de uma análise aprofundada sobre o uso da estatística nos artigos publicados na área, verificando se a análise dos dados está correta e se todas as etapas de coleta e análise dos dados foram realizadas corretamente.

Além disso, uma outra sugestão de trabalho futuro possível seria a análise das grades curriculares da Engenharia de Software, verificando se as dificuldades no uso da estatística estão relacionadas com a quantidade e carga-horária de disciplinas ofertadas nos cursos de graduação e pós-graduação sobre o tema.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Ahmed, I.; Brindescu, C.; Mannan, U. A.; Jensen, C.; Sarma, A. “An empirical examination of the relationship between code smells and merge conflicts”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 58–67.

[2] Ahmed, I.; Mannan, U. A.; Gopinath, R.; Jensen, C. “An empirical study of design degradation: How software projects get worse over time”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–10.

[3] Ajienka, N.; Capiluppi, A.; Counsell, S. “Managing hidden dependencies in oo software: a study based on open source projects”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 141–150.

[4] Alkadhi, R.; Johanssen, J. O.; Guzman, E.; Bruegge, B. “React: an approach for capturing rationale in chat messages”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 175–180.

[5] Alomar, E. A.; Mkaouer, M. W.; Ouni, A.; Kessentini, M. “On the impact of refactoring on the relationship between quality attributes and design metrics”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–11.

[6] Alqahtani, S. S.; Rilling, J. “An ontology-based approach to automate tagging of software artifacts”. In: ACM/IEEE International Symposiumon Empirical Software Engineering and Measurement, 2017, pp. 169–174.

[7] Alshangiti, M.; Sapkota, H.; Murukannaiah, P. K.; Liu, X.; Yu, Q. “Why is developing machine learning applications challenging? a study on stack overflow posts”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–11.

[8] Aman, H.; Amasaki, S.; Sasaki, T.; Kawahara, M. “Empirical analysis of change-proneness in methods having local variables with long names and comments”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–4.

[9] Andrade, P. C. d. R.; Chaves, L. M.; Ferreira, D. F. “Proposta de um teste não-paramétrico de sinal com postos para dados independentes de duas populações”, *Rev. Mat. Estat*, vol. 21–2, Jan-Dez 2003, pp. 07–23.

[10] Anu, V.; Walia, G.; Hu, W.; Carver, J. C.; Bradshaw, G. “Issues and opportunities for human error-based requirements inspections: an exploratory study”. In: ACM/IEEE International Symposiumon Empirical Software Engineering and Measurement, 2017, pp. 460–465.

[11] Avelino, G.; Constantinou, E.; Valente, M. T.; Serebrenik, A. “On the abandonment and survival of open source projects: An empirical investigation”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–12.

[12] Bach, T.; Andrzejak, A.; Pannemans, R.; Lo, D. “The impact of coverage on bug density in a large industrial software project”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 307–313.

[13] Barros, M. d. O.; Gonçalves, V. P. “A function point formulation for the software release planning problem”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–11.

[14] Bennin, K. E.; Keung, J.; Monden, A.; Phannachitta, P.; Mensah, S. “The significant effects of data sampling approaches on software defect prioritization and classification”. In: ACM/IEEE International Symposiumon EmpiricalSoftware Engineering and Measurement, 2017, pp. 364–373.

[15] Bibiano, A. C.; Fernandes, E.; Oliveira, D.; Garcia, A.; Kalinowski, M.; Fonseca, B.; Oliveira, R.; Oliveira, A.; Cedrim, D. “A quantitative study on characteristics and effect of batch refactoring on code smells”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–11.

[16] Bin, Y.; Zhou, K.; Lu, H.; Zhou, Y.; Xu, B. “Training data selection for cross-project defection prediction: which approach is better?” In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 354–363.

[17] Bonafini, F. C. “Estatística”. São Paulo, SP: Pearson, 2012, 176p.

[18] Bosu, A.; Sultana, K. Z. “Diversity and inclusion in open source software (oss) projects: Where do we stand?” In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–11.

[19] Calazans, T. A.; Paldês, R. Á.; Canedo, E. D.; Masson, T. E.; Guimares, F. d. A.; Rezende, M. K.; Braosi, E.; Kosloski, R. A. D. “Quality requirements: Analysis of utilization in the systems of a financial institution”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–6.

[20] Campos, E. C.; Maia, M. A. “Common bug-fix patterns: A large-scale observational study”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement , 2017, 2017, 2017, pp. 404–413.

[21] Carroll, C.; Falessi, D.; Forney, V.; Frances, A.; Izurieta, C.; Seaman, C. “A mapping study of software causal factors for improving maintenance”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–4.

[22] Cartaxo, B.; Borba, P.; Soares, S.; Fugimoto, H. “Improving performance and maintainability of object cloning with lazy clones: An empirical evaluation”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–8.

[23] Cartaxo, B.; Pinto, G.; Fonseca, B.; Ribeiro, M.; Pinheiro, P.; Baldassarre, M. T.; Soares, S. “Software engineering research community viewpoints on rapid reviews”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–12.

[24] Cavalcanti, G.; Accioly, P.; Borba, P. “Assessing semistructured merge in version control systems: A replicated experiment”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–10.

[25] Chagas, A. B.; Melo, F. A.; dos Santos, W. F.; de Oliveira, A. A. N.; Bora, S. M.; da Silva, F. Q. B. “Analysis of the understanding of the concepts of task and skill variety by software engineering professionals”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 217–222.

[26] Chen, T.; Li, Z.; Zhang, Y.; Luo, X.; Wang, T.; Hu, T.; Xiao, X.; Wang, D.; Huang, J.; Zhang, X. “A large-scale empirical study on control flow identification of smart contracts”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–11.

[27] Contador, J. L.; Senne, E. L. F. “Non-parametric tests for small samples of categorized variables: a study”, *Gestão & Produção*, vol. 23–3, Jan-Dez 2016, pp. 588–599.

[28] Corrar, L. J.; Paulo, E.; Dias Filho, J. M. “Análise multivariada: para os cursos de administração, ciências contábeis e economia”. São Paulo, SP: Editora Atlas, 2007, 541p.

[29] Counsell, S.; Arzoky, M.; Destefanis, G.; Taibi, D. “On the relationship between coupling and refactoring: An empirical viewpoint”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–6.

[30] da Silva, L. M.; Tavares, A. T.; Ferreira, V. A.; Costa, A. J.; de Souza, G. I.; Magalhães, C. J.; Da Silva, F. “Autonomy in software engineering: a preliminary study on the influence of education level and professional experience”. In: ACM/IEEE International Symposiumon Empirical Software Engineering and Measurement, 2017, pp. 229–234.

[31] Dalfovo, M. S.; Lana, R. A.; Silveira, A. “Métodos quantitativos e qualitativos: um resgate teórico”, *Revista Interdisciplinar Científica Aplicada*, vol. 2–3, Jan-Dez 2008, pp. 1–13.

[32] Dandolini, J. “Ferramenta de apoio a realização de experimentos em engenharia de software”, Monografia, Universidade Regional de Blumenau, 2006, 91p.

[33] Daneva, M.; Wang, C.; Hoener, P. “What the job market wants from requirements engineers? an empirical analysis of online job ads from the netherlands”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 448–453.

[34] Datta, S.; Bhatt, D.; Jain, M.; Sarkar, P.; Sarkar, S. “The importance of being isolated: An empirical study on chromium reviews”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–4.

[35] de Mello, R. M.; Oliveira, R. F.; Garcia, A. F. “On the influence of human factors for identifying code smells: A multi-trial empirical study”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 68–77.

[36] de Mello, R. M.; Stolee, K. T.; Travassos, G. H. “Investigating samples representativeness for an online experiment in java code search”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–10.

[37] de Mello, R. M.; Uchôa, A.; Oliveira, R.; Oizumi, W.; Souza, J.; Mendes, K.; Oliveira, D.; Fonseca, B.; Garcia, A. “Do research and practice of code smell identification walk together? a social representations analysis”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–6.

[38] de Oliveira Neto, F. G.; Torkar, R.; Feldt, R.; Gren, L.; Furia, C. A.; Huang, Z. “Evolution of statistical analysis in empirical software engineering research: Current state and steps forward”, *Journal of Systems and Software*, vol. 156, Out 2019, pp. 246–267.

[39] de Sousa, C. A.; Junior, M. A. L.; Ferreira, R. L. C. “Avaliação de testes estatísticos de comparações múltiplas de médias”, *Revista Ceres*, vol. 59–3, Mai-Jun 2012, pp. 350–354.

[40] dos Anjos, A. “Planejamento de experimentos i”, Relatório técnico, Departamento de Estatística, UFPR, 2009, 97p.

[41] Echeverría, J.; Pérez, F.; Panach, J. I.; Cetina, C.; Pastor, O. “The influence of requirements in software model development in an industrial environment”. In: ACM/IEEE International Symposium on Empirical Software Engineering andMeasurement , 2017, 2017, pp. 277–286.

[42] Eckhardt, J.; Fernández, D. M.; Vogelsang, A. “How to specify non-functional requirements to support seamless modeling? a study design and preliminary results”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–4.

[43] Espírito Santo, H.; Daniel, F. “Calcular e apresentar tamanhos do efeito em trabalhos científicos (2): Guia para reportar a força das relações [calculating and reporting effect sizes on scientific papers (2): Guide to report the strength of relationships]”, *Portuguese Journal of Behavioral and Social Research*, vol. 3–1, Jul 2017, pp. 53–64.

[44] Fagerholm, F.; Becker, C.; Chatzigeorgiou, A.; Betz, S.; Duboc, L.; Penzenstadler, B.; Mohanani, R.; Venters, C. C. “Temporal discounting in software engineering: A replication study”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–12.

[45] Falessi, D.; Russo, B.; Mullen, K. “What if i had no smells?” In: ACM/IEEE International Symposium on Empirical Software Engineering andMeasurement , 2017, 2017, pp. 78–84.

[46] Fan, Q.; Yu, Y.; Yin, G.; Wang, T.; Wang, H. “Where is the road for issue reports classification based on text mining?” In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 121–130.

[47] Felidré, W.; Furtado, L.; da Costa, D. A.; Cartaxo, B.; Pinto, G. “Continuous integration theater”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–10.

[48] Ferreira, L.; Nogueira, S.; Lima, L.; Fonseca, L.; Ferreira, W. “Initial findings on the evaluation of a model-based testing tool in the test design process”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–6.

[49] Figueiredo Filho, D.; Nunes, F.; da Rocha, E. C.; Santos, M. L.; Batista, M.; Silva Júnior, J. A. “O que fazer e o que não fazer com a regressão: pressupostos e

aplicações do modelo linear de mínimos quadrados ordinários (mqo)", *Revista Política Hoje*, vol. 20–1, Jan-Jul, 2011, pp. 44–99.

[50] Figueiredo Filho, D. B.; da Silva Júnior, J. A.; dos Santos Filho, R. P.; da Rocha, E. C.; da Silva Nascimento, W.; da Silva, M. B.; de Oliveira Silva, L. E. "Happy together: como utilizar análise fatorial e análise de cluster paras mensurar a qualidade das políticas públicas", *Revista Teoria & Sociedade*, vol. 22–2, Jul-Dez 2014, pp. 123–152.

[51] Ford, D.; Zimmermann, T.; Bird, C.; Nagappan, N. "Characterizing software engineering work with personas based on knowledge worker actions". In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 394–403.

[52] Fu, S.; Shen, B. "Code bad smell detection through evolutionary data mining". In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–9.

[53] Gadler, D.; Mairegger, M.; Janes, A.; Russo, B. "Mining logs to model the use of a system". In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement , 2017, 2017, pp. 334–343.

[54] Gallaba, K.; Mesbah, A.; Beschastnikh, I. "Don't call us, we'll call you: Characterizing callbacks in javascript". In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–10.

[55] Ghafari, M.; Eggiman, M.; Nierstrasz, O. "Testability first!" In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–6.

[56] Ghanbari, H.; Besker, T.; Martini, A.; Bosch, J. "Looking for peace of mind?: manage your (technical) debt: an exploratory field study". In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 384–393.

[57] Ghazi, A. N.; Petersen, K.; Reddy, S. S. V. R.; Nekkanti, H. "Survey research in software engineering: problems and strategies", *e-Informatica Software Engineering Journal*, vol. 1–1, Fev, 2017, pp. 1–24.

[58] Girden, E. R. "ANOVA: Repeated measures". Newbury Park, California: Sage Publications, 1991, 88p.

[59] Gomes, M. M. F.; Pessoa, D.; Fernandes, L. A.; SANTOS, J. d. C.; Vasconcelos, A. M. N. "Estatística aplicada à engenharia e áreas afins incentivando meninas do ensino médio nas carreiras de ciências exatas, engenharias e computação". In: Congresso Brasileiro de Educação em Engenharia, 2014, pp. 1–8.

[60] Guzman, E.; Aly, O.; Bruegge, B. “Retrieving diverse opinions from app reviews”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–10.

[61] Guzmán, L.; Vollmer, A. M.; Ciolkowski, M.; Gillmann, M. “Formative evaluation of a tool for managing software quality”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 297–306.

[62] Hair, J. F.; Black, W. C.; Babin, B. J.; Anderson, R. E.; Tatham, R. L. “Análise multivariada de dados”. São Paulo, SP: Bookman Editora, 2009, 688p.

[63] Hassan, F.; Mostafa, S.; Lam, E. S.; Wang, X. “Automatic building of java projects in software repositories: A study on feasibility and challenges”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 38–47.

[64] Hassan, F.; Wang, X. “Change-aware build prediction model for stall avoidance in continuous integration”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 157–162.

[65] Hata, H.; Guo, M.; Babar, M. A. “Understanding the heterogeneity of contributors in bug bounty programs”. In: ACM/IEEE International Symposium on Empirical Software Engineering andMeasurement, 2017, pp. 223–228.

[66] Hazhirpasand, M.; Ghafari, M.; Krüger, S.; Bodden, E.; Nierstrasz, O. “The impact of developer experience in using java cryptography”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–6.

[67] Hebig, R.; Derehag, J.; Chaudron, M. R. “Identifying metrics’ biases when measuring or approximating size in heterogeneous languages”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–4.

[68] Huang, Y.; Zheng, Q.; Chen, X.; Xiong, Y.; Liu, Z.; Luo, X. “Mining version control system for automatically generating commit comment”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 414–423.

[69] Huijgens, H.; Gousios, G.; van Deursen, A. “Pricing via functional size-a case study of a company’s portfolio of 77 outsourced projects”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–10.

[70] Islam, M. R.; Zibran, M. F. “A comparison of dictionary building methods for sentiment analysis in software engineering text”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 478–479.

[71] Islam, M. R.; Zibran, M. F.; Nagpal, A. “Security vulnerabilities in categories of clones and non-cloned code: An empirical study”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 20–29.

[72] Izurieta, C.; Griffith, I.; Huvaere, C. “An industry perspective to comparing the sqale and quamoco software quality models”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 287–296.

[73] Júnior, C. A. M. “Questões em bioestatística: o tamanho da amostra”, *Revista Interdisciplinar de Estudos Experimentais*, vol. 1–1, Jan 2009, pp. 26–28.

[74] Júnior, M. “armadilhas e como evitá-las”, *Boletim do Centro de Biologia da Reprodução. Juiz de Fora*, vol. 26–1/2, Jan-Dez 2007, pp. 105–111.

[75] Juristo, N.; Moreno, A. M. “An adaptation of experimental design to the empirical validation of software engineering theories”. In: Annual NASA Software Engineering Workshop, 1998, pp. 1–10.

[76] Kitchenham, B.; Madeyski, L.; Brereton, P. “Problems with statistical practice in human-centric software engineering experiments”. In: *Evaluation and Assessment on Software Engineering*, 2019, pp. 134–143.

[77] Kitchenham, B. A.; Pfleeger, S. L.; Pickard, L. M.; Jones, P. W.; Hoaglin, D. C.; El Emam, K.; Rosenberg, J. “Preliminary guidelines for empirical research in software engineering”, *IEEE Transactions on software engineering*, vol. 28–8, Ago 2002, pp. 721–734.

[78] Labunets, K.; Massacci, F.; Tedeschi, A. “Graphical vs. tabular notations for risk models: on the role of textual labels and complexity”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 267–276.

[79] Lage, L. C. d. F.; Kalinowski, M.; Trevisan, D.; Spinola, R. “Usability technical debt in software projects: A multi-case study”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–6.

[80] Lami, G.; Biscoglio, I.; Falcini, F. “Investigation on common software process weaknesses in automotive”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–8.

[81] Lee, A.; Carver, J. C. “Are one-time contributors different? a comparison to core and periphery developers in floss repositories”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 1–10.

[82] Leotti, V. B.; Coster, R.; Riboldi, J. “Normalidade de variáveis: métodos de verificação e comparação de alguns testes não-paramétricos por simulação”, *Revista HCPA. Porto Alegre*, vol. 32–2, Abr 2012, pp. 227–234.

[83] Liaskos, S.; Ronse, A.; Zhian, M. “Assessing the intuitiveness of qualitative contribution relationships in goal models: an exploratory experiment”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 466–471.

[84] Lima, K. E. C.; Teixeira, F. M. “A epistemologia e a história do conceito experimento/experimentação e seu uso em artigos científicos sobre ensino das ciências”. In: Encontro Nacional de Pesquisa em Educação em Ciência, 2011, pp. 1–12.

[85] Lima, M.; Ahmed, I.; Conte, T.; Nascimento, E.; Oliveira, E.; Gadelha, B. “Land of lost knowledge: An initial investigation into projects lost knowledge”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–6.

[86] Liu, J.; Zhou, Y.; Yang, Y.; Lu, H.; Xu, B. “Code churn: A neglected metric in effort-aware just-in-time defect prediction”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 11–19.

[87] Malloy, B. A.; Power, J. F. “Quantifying the transition from python 2 to 3: an empirical study of python applications”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 314–323.

[88] Marchezan, L.; Bolfe, G.; Rodrigues, E.; Bernardino, M.; Basso, F. P. “Thoth: A web-based tool to support systematic reviews”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–6.

[89] Marinho, M.; Noll, J.; Richardson, I.; Beecham, S. “Plan-driven approaches are alive and kicking in agile global software development”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–11.

[90] Martins, G. A. “Estatística geral e aplicada”. São Paulo, SP: Editora Atlas, 2005, 421p.

[91] McHugh, M. L. “The chi-square test of independence”, *Biochemia medica*, vol. 23–2, Mai 2013, pp. 143–149.

[92] Menezes, G.; Cafeo, B.; Hora, A. “Framework code samples: How are they maintained and used by developers?” In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–11.

[93] Meyer, A. N.; Zimmermann, T.; Fritz, T. “Characterizing software developers by perceptions of productivity”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 105–110.

[94] Miller, J.; Daly, J.; Wood, M.; Roper, M.; Brooks, A. “Statistical power and its subcomponents—missing and misunderstood concepts in empirical software engineering research”, *Information and Software Technology*, vol. 39–4, Abr 1997, pp. 285–295.

[95] Minayo, M. C. d. S.; Sanches, O. “Quantitativo-qualitativo: oposição ou complementaridade?”, *Cadernos de saúde pública*, vol. 9–3, Jul-Set 1993, pp. 237–248.

[96] Minku, L.; Sarro, F.; Mendes, E.; Ferrucci, F. “How to make best use of cross-company data for web effort estimation?” In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–10.

[97] Morettin, P. A.; Bussab, W. O. “Estatística básica”. São Paulo,SP: Editora Saraiva, 2017, 576p.

[98] Mourão, E.; Kalinowski, M.; Murta, L.; Mendes, E.; Wohlin, C. “Investigating the use of a hybrid search strategy for systematic reviews”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 193–198.

[99] Mubin, A.; Kuai, M. “Identifying software decays: a system usage perspective”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 472–473.

[100] Munaiah, N.; Rahman, A.; Pelletier, J.; Williams, L.; Meneely, A. “Characterizing attacker behavior in a cybersecurity penetration testing competition”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–6.

[101] Mund, J.; Fernandez, D. M.; Femmer, H.; Eckhardt, J. “Does quality of requirements specifications matter? combined results of two empirical studies”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–10.

[102] Munezero, M.; Kojo, T.; Männistö, T. “An exploratory analysis of a hybrid oss company’s forum in search of sales leads”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 442–447.

[103] Nayebi, M.; Farahi, H.; Ruhe, G. “Which version should be released to app store?” In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 324–333.

[104] Neto, E. C.; da Costa, D. A.; Kulesza, U. “Revisiting and improving szz implementations”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–12.

[105] Owhadi-Kareshk, M.; Nadi, S.; Rubin, J. “Predicting merge conflicts in collaborative software development”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–11.

[106] Pacheco, A. P. R.; Salles, B. W.; Garcia, M. A.; Possamai, O. “O ciclo pdca na gestão do conhecimento: uma abordagem sistêmica”, Relatório técnico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, UFSC, 2012, 10p.

[107] Pashchenko, I.; Dashevskyi, S.; Massacci, F. “Delta-bench: differential benchmark for static analysis security testing tools”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 163–168.

[108] Perroca, M. G.; Gaidzinski, R. R. “Avaliando a confiabilidade interavaliadores de um instrumento para classificação de pacientes: coeficiente kappa”, *Revista da Escola de Enfermagem da USP*, vol. 37–1, Mar 2003, pp. 72–80.

[109] Pontes, A. C. F. “Ensino da correlação de postos no ensino médio”. In: Simpósio Nacional De Probabilidade E Estatística, 2010, pp. 26–30.

[110] Poulding, S.; Petersen, K.; Feldt, R.; Garousi, V. “Using citation behavior to rethink academic impact in software engineering”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–4.

[111] Pressman, R.; Maxim, B. “Engenharia de Software - 8ª Edição”. São Paulo,SP: McGraw Hill Brasil, 2016, 933p.

[112] Qin, X.; Holla, S.; Huang, L.; Montijo, L.; Aguirre, D.; Wang, X. “How does machine translated user interface affect user experience? a study on android apps”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 430–435.

[113] Rashid, M.; Ardito, L.; Torchiano, M. “Energy consumption analysis of algorithms implementations”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–4.

[114] Rastogi, A.; Thummalapenta, S.; Zimmermann, T.; Nagappan, N.; Czerwonka, J. “Ramp-up journey of new hires: Tug of war of aids and impediments”. In: ACM/IEEE

International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–10.

[115] Raulamo-Jurvanen, P.; Mantyla, M. V.; Garousi, V. “Citation and topic analysis of the esem papers”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–4.

[116] Ribeiro, H. L.; de Araujo, P. R.; Chaim, M. L.; de Souza, H. A.; Kon, F. “Evaluating data-flow coverage in spectrum-based fault localization”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–11.

[117] Riboldi, J.; Barbian, M. H.; Kolowski, A.; Selau, L. P. R.; Torman, V. “Precisão e poder de testes de homocedasticidade paramétricos e não-paramétricos avaliados por simulação”, *Revista Brasileira de Biomedicina*, vol. 32–3, Jul-Set 2014, pp. 334–344.

[118] Rocha, G.; Castor, F.; Pinto, G. “Comprehending energy behaviors of java i/o apis”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–12.

[119] Rodeghero, P.; McMillan, C. “An empirical study on the patterns of eye movement during summarization tasks”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–10.

[120] Rodriguez, I.; Wang, X. “An empirical study of open source virtual reality software projects”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 474–475.

[121] Rosa, W.; Madachy, R.; Clark, B.; Boehm, B. “Early phase cost models for agile software processes in the us dod”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 30–37.

[122] Santos, R. E.; Magalhães, C. V.; Correia-Neto, J. S.; da Silva, F. Q.; Capretz, L. F.; Souza, R. E. “Would you like to motivate software testers?: ask them how”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 95–104.

[123] Shahin, M.; Babar, M. A.; Zahedi, M.; Zhu, L. “Beyond continuous delivery: an empirical investigation of continuous deployment challenges”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 111–120.

[124] Sharma, T.; Fragkoulis, M.; Spinellis, D. “House of cards: code smells in open-source c# repositories”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 424–429.

[125] Sheldon, M. R.; Fillyaw, M. J.; Thompson, W. D. “The use and interpretation of the friedman test in the analysis of ordinal-scale data in repeated measures designs”, *Physiotherapy Research International*, vol. 1–4, Nov 1996, pp. 221–228.

[126] Siegel, S.; Castellan Jr, N. J. “Estatística não-paramétrica para ciências do comportamento”. Porto Alegre, RS: Artmed Editora, 1975, 448p.

[127] Slocum, N. “Participatory methods toolkit: A practitioner’s manual, joint publication of the king baudouin foundation and the flemish institute for science and technology assessment (viwta) in collaboration with the united nations university–comparative regional integration”, Relatório técnico, Flemish Institute for Science and Technology Assessment, 2005, 2013p.

[128] Soh, Z.; Drioul, T.; Rappe, P.-A.; Khomh, F.; Gueheneuc, Y.-G.; Habra, N. “Noises in interaction traces data and their impact on previous research studies”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–10.

[129] Sommerville, I. “Engenharia de Software - 6ª Edição”. São Paulo,SP: Addison Wesley, 2003, 594p.

[130] Spinellis, D.; Louridas, P.; Kechagia, M. “An exploratory study on the evolution of c programming in the unix operating system”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–4.

[131] Talwadker, R.; Aggarwal, D. “Popcon: Mining popular software configurations from community”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–6.

[132] Tang, X.; Wang, S.; Mao, K. “Will this bug-fixing change break regression testing?” In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–10.

[133] Thiselton, E.; Treude, C. “Enhancing python compiler error messages via stack”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–12.

[134] Travassos, G. H.; Gurov, D.; Amaral, E. “Introdução à engenharia de software experimental”. Rio de Janeiro,RJ: UFRJ, 2002, 52p.

[135] Tsunoda, M.; Amasaki, S. “On software productivity analysis with propensity score matching”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 436–441.

[136] Venson, E.; Alfayez, R.; MF, G. M.; Rejane, F.; Boehm, B. “The impact of software security practices on development effort: An initial survey”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–12.

[137] Verdecchia, R.; Procaccianti, G.; Malavolta, I.; Lago, P.; Koedijk, J. “Estimating energy impact of software releases and deployment strategies: The kpmg case study”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 257–266.

[138] Vetro, A.; Bohm, W.; Torchiano, M. “On the benefits and barriers when adopting software modelling and model driven techniques-an external, differentiated replication”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–4.

[139] Vollmer, A. M.; Martınez-Fernández, S.; Bagnato, A.; Partanen, J.; López, L.; Rodrıguez, P. “Practical experiences and value of applying software analytics to manage quality”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–6.

[140] Wainer, J.; et al.. “Métodos de pesquisa quantitativa e qualitativa para a ciência da computação”, *Atualização em informática*, vol. 1, Jan-Dez 2007, pp. 221–262.

[141] Wang, L.; Yang, Y.; Wang, Y. “Do higher incentives lead to better performance?-an exploratory study on software crowdsourcing”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–11.

[142] Wang, Y. “Language matters”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–10.

[143] Wang, Y. “Characterizing developer behavior in cloud based ides”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 48–57.

[144] Wazlawick, R. “Metodologia de pesquisa para ciência da computação”. Rio de Janeiro, RJ: Elsevier Brasil, 2017, 339p.

[145] Wu, D.; Chen, L.; Zhou, Y.; Xu, B. “An empirical study on c++ concurrency constructs”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–10.

[146] Yaman, S.; Fagerholm, F.; Munezero, M.; Mäenpää, H.; Männistö, T. “Notifying and involving users in experimentation: ethical perceptions of software practitioners”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 199–204.

[147] Yan, M.; Fang, Y.; Lo, D.; Xia, X.; Zhang, X. “File-level defect prediction: Unsupervised vs. supervised models”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2017, pp. 344–353.

[148] Yang, Y.; Saremi, R. “Award vs. worker behaviors in competitive crowdsourcing tasks”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2015, pp. 1–10.

[149] Zafar, S.; Malik, M. Z.; Walia, G. S. “Towards standardizing and improving classification of bug-fix commits”. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–6.

[150] Zuben, F. V. “Regressão paramétrica e não-paramétrica”. Recuperado de http://calhau.dca.fee.unicamp.br, Campinas, SP, 2003, 38p.

APÊNDICE A – ARTIGOS DA RSL

Neste apêndice, apresenta-se os resultados tabulados da Revisão Sistemática da Literatura (RSL) descrita no Capítulo 7, onde tem-se as seguintes colunas com seus respectivos valores:

- **Ref**: indicação da referência bibliográfica estudada;
- **Análise**: I para análise Inferencial e D para Descritiva;
- **Tipo Teste**: NP para Não Paramétrico e P para Paramétrico;
- **Amostra**: AC30 para amostras com mais de 30 elementos e AB30 para amostras que utilizaram valor abaixo de 30 elementos no experimento;
- **Var Expl**: CAT para variável explicativa Categórica e CONT para variável explicativa Contínua;
- **Var Resp**: CAT para variável resposta Categórica e CONT para variável resposta Contínua;
- **Técnica**: KAPPA - Coeficiente de Kappa, CP - Correlação de Pearson, RL - Regressão Linear, KW - Kruskal-Wallis, MW - Mann-Whitney, W - Wilcoxon, CS - Correlação de Spearman, ANOVA - Análise de Variância, AF - Análise Fatorial, AC - Análise de Cluster, COHEN - Teste de Cohen, B - Teste de Brunne, KENDALL - Correlação de Kendall, SK - Scott-Knott, CA - Cochran-Armitage, A - Análise de Covariância (ANCOVA), T - Teste T, QQ - Qui-quadrado, TB - Teste Binomial, EF - Exato de Fisher, MR - Medidas Repetidas, MANOVA - Análise de Variância Multivariada.

Ref	Análise	Tipo Teste	Amostra	Var Expl	Var Resp	Técnica
[115]	D	-	AC30	CAT	CONT	-
[42]	I	NP	AC30	CAT	CONT	KAPPA
[99]	D	-	AC30	CONT	CONT	CP
[130]	D	-	AC30	CONT	CONT	RL
[113]	I	P e NP	AC30	CAT/CONT	CONT	KW/RL
[70]	D	-	AC30	CAT	CONT	-
[120]	D	-	AC30	CAT/CONT	CONT	-
[110]	D	-	AC30	CONT	CONT	KAPPA
[26]	D	-	AC30	CONT	CAT	-
[88]	D	-	AB30	CAT	CONT	-
[118]	I	NP	AC30	CAT	CONT	MW

continua na próxima página

Ref	Análise	Tipo Teste	Amostra	Var Expl	Var Resp	Técnica
[139]	D	-	AB30	CAT	CONT	-
[55]	D	-	AC30	CAT	CONT	-
[89]	D	-	AC30	CAT	CONT	-
[149]	D	-	AC30	CAT	CONT	-
[37]	D	-	AC30	CAT	CAT	-
[131]	D	-	AC30	CAT	CONT	-
[79]	D	-	AC30	CAT	CONT	-
[92]	D	-	AC30	CAT	CONT	-
[104]	D	-	AC30	CAT	CONT	-
[116]	I	NP	AC30	CAT	CONT	W
[66]	I	NP	AC30	CAT	CONT	KW/W/CP
[100]	I	NP	AC30	CAT	CONT	COHEN
[13]	I	NP	AC30	CAT	CONT	MW
[133]	D	-	AB30	CAT	CAT	-
[44]	I	NP	AC30	CAT/CONT	CONT	KW/CP
[85]	D	-	AC30	CAT	CONT	KAPPA
[105]	I	NP	AC30	CONT	CONT	CS
[48]	I	P	AB30	CAT	CONT	ANOVA
[141]	I	P	AC30	CAT/CONT	CONT	ANOVA/RL
[11]	I	NP	AC30	CAT	CONT	MW
[23]	D	-	AC30	CAT	CONT	AF
[18]	I	NP	AC30	CAT	CONT	MW
[15]	I	NP	AC30	CAT	CAT	EF
[47]	D	-	AC30	CAT/CONT	CAT/CONT	-
[103]	D	-	AC30	CAT	CONT	-
[7]	D	-	AC30	CAT/CONT	CONT	KAPPA
[29]	D	-	AC30	CAT	CONT	-
[12]	I	NP	AC30	CAT	CONT	W/MW
[147]	I	NP	AC30	CAT	CONT	W/SK
[1]	I	P	AC30	CAT/CONT	CONT	RL
[101]	D	-	AC30	CONT	CONT	KENDALL
[22]	I	NP	AC30	CAT	CONT	MW
[14]	I	NP	AC30	CAT	CONT	B
[45]	I	P e NP	AC30	CAT	CONT	MW/RL
[123]	D	-	AC30	CAT/CONT	CONT	-
[8]	I	NP	AC30	CAT	CAT	C
[122]	D	-	AC30	CAT	CAT	-

continua na próxima página

Ref	Análise	Tipo Teste	Amostra	Var Expl	Var Resp	Técnica
[**?**]	D	-	AC30	CAT	CONT	-
[96]	I	NP	AC30	CAT	CONT	W
[20]	D	-	AC30	CAT	CONT	-
[135]	I	P e NP	AC30	CAT/CONT	CONT	RL/A
[124]	I	NP	AC30	CONT	CONT	CS
[33]	D	-	AC30	CAT/CONT	CAT/CONT	-
[107]	D	-	AC30	CAT	CONT	-
[71]	I	NP	AC30	CAT	CONT	MW/KW
[80]	D	-	AC30	CAT/CONT	CONT	-
[138]	I	NP	AC30	CONT	CONT	CS
[10]	I	P	AB30	CAT	CONT	T
[4]	D	-	AB30	CAT	CONT	-
[98]	D	-	AB30	CAT	CONT	-
[81]	I	P	AB30	CAT	CONT	ANOVA
[114]	I	NP	AC30	CAT/CONT	CAT/CONT	QQ/CS
[60]	I	NP	AB30	CAT	CONT	W
[25]	I	P	AC30	CONT	CONT	CP
[19]	D	-	AC30	CAT	CONT	-
[5]	I	NP	AC30	CAT	CONT	W
[30]	I	NP	AC30	CAT/CONT	CONT	CS/T
[137]	I	P	AC30	CAT	CONT	ANOVA
[132]	D	-	AC30	CAT	CONT	-
[2]	I	P e NP	AC30	CAT	CONT	T/RL
[136]	D	-	AC30	CAT/CONT	CONT	-
[41]	I	P	AB30	CAT	CONT	MM
[119]	I	P e NP	AB30	CAT	CONT	W/ANOVA
[148]	D	-	AC30	CONT	CONT	CP
[61]	I	NP	AB30	CAT	CONT	W
[53]	D	-	AC30	CAT	CONT	-
[143]	D	-	AC30	CAT/CONT	CAT/CONT	-
[87]	D	-	AC30	CAT/CONT	CAT/CONT	-
[16]	I	NP	AC30	CAT	CONT	SK
[35]	I	P	AB30	CAT	CONT	T
[93]	I	NP	AC30	CAT/CONT	CAT/CONT	AC/EF/MW
[54]	D	-	AC30	CAT/CONT	CONT	-
[46]	I	P	AC30	CONT	CAT/CONT	T/RL
[67]	I	NP	AC30	CAT	CAT	TB

continua na próxima página

Ref	Análise	Tipo Teste	Amostra	Var Expl	Var Resp	Técnica
[51]	D	-	AB30	CAT/CONT	CAT/CONT	AC
[112]	I	P	AB30	CAT	CONT	T
[3]	I	NP	AC30	CAT/CONT	CONT	CS
[68]	D	-	AC30	CAT	CONT	-
[102]	D	-	AC30	CAT	CAT/CONT	-
[64]	D	-	AC30	CAT	CONT	-
[86]	I	NP	AC30	CAT	CONT	W
[36]	I	NP	AB30 e AC30	CAT	CAT/CONT	MW/QQ
[6]	D	-	AC30	CAT	CONT	-
[142]	I	P	AB30	CAT	CONT	ANOVA MR
[69]	D	-	AB30	CAT	CONT	TE
[83]	I	P	AC30	CAT	CONT	MANOVA
[146]	D	-	AC30	CAT	CAT/CONT	CP
[128]	I	NP	AB30	CAT	CONT	W
[34]	I	P	AC30	CAT/CONT	CONT	RL
[21]	D	-	AC30	CAT/CONT	CAT/CONT	-
[65]	D	-	AC30	CAT	CAT	AC
[78]	I	NP	AC30	CAT/CONT	CONT	W/KW
[145]	I	NP	AC30	CAT	CONT	W/MW
[24]	D	-	AC30	CAT	CONT	-
[72]	I	NP	AB30	CAT	CONT	W/C
[63]	D	-	AC30	CAT/CONT	CONT	-
[56]	I	NP	AC30	CAT/CONT	CONT	KENDALL
[121]	I	P	AB30	CAT/CONT	CONT	RL
[52]	D	-	AC30	CAT/CONT	CAT/CONT	-

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br