# Meta-Análise de Artigos Científicos Segundo Critérios Estatísticos: Um estudo de caso no WSCAD 2018 a 2021

**Bruno L. S. Rech**[1], **Fabiano C. Dicheti**[1], **Gustavo R. Malacarne**[1], **Rodrigo S. Nurmberg**[1], **Thiago S. Elias** [1]

[1] Programa de Pós-Graduação em Ciência da Computação (PPGComp)
Universidade Estadual do Oeste do Paraná (UNIOESTE)
R. Universitária, 1619 - Universitário - Cascavel - PR - Brasil - CEP: 85819-110

{bruno.rech, fabiano.dicheti, gustavo.malacarne}@unioeste.br

{rodrigo.nurmberg, thiago.elias}@unioeste.br

***Abstract.*** *This article presents the systematization of the results of the meta-analysis of publications from the last 4 editions (2018, 2019, 2020 and 2021) of WSCAD according to the categories: statistics, metrics and tests. For that, the manual extraction of texts from PDF files was carried out for the mining of statistical terms. From the sample of 70 publications analyzed, the search words were divided into three large groups: statistical terms, metrics and tests. After carrying out an exploratory analysis of the data, followed by statistical tests, it was observed that after the year 2018 there was no change in the absolute values of terms related to statistics.*

***Resumo.*** *Este artigo apresenta a sistematização dos resultados da meta-analise das publicações das 4 últimas edições (2018, 2019, 2020 e 2021) do WSCAD segundo as categorias: estatística, métricas e testes. Para tanto, realizou-se a extração manual de textos de arquivos PDF para a mineração de termos estatísticos. Da amostra de 70 publicações analisadas, as palavras de busca foram divididas em três grandes grupos: termos estatísticos, métricas e testes. Após realizado uma análise exploratória dos dados, seguido de testes estatísticos, observou-se que após o ano de 2018 não houve nenhuma mudança nos valores absolutos de termos relacionados à estatística.*

## 1. Introdução

A Estatística, desde seu início até os dias de hoje, passou de uma forma de sofisticação do processo de pesquisa, para se tornar um requisito básico, garantidor da confiabilidade dos processos e dos resultados de qualquer estudo científico, essencial na produção e difusão do conhecimento. Por meio da Estatística, um conjunto de equações matemáticas pode ser usado para analisar dados e determinar qual é a influência que cada fator apresenta em um resultado.

Na Computação, com o uso da Estatística, é possível mensurar a análise de desempenho, etapa fundamental na concepção de uma técnica ou algoritmo. A avaliação sob diversas métricas permite que pesquisadores utilizem de meios dispostos convenientemente para atestar o desempenho da solução. Embora aspectos como legibilidade, simplicidade e modularidade de uma solução sejam importantes para a sua manutenibilidade, o desempenho e a acurácia de uma solução é muito relevante para a sua adoção.

Sendo assim, a análise de consistência e coerência dos dados é de suma importância para associar confiabilidade aos resultados apresentados e então levar a obtenção da inferência de comportamentos, interpretações sobre os resultados coletados e as conclusões obtidas.

Como considerado neste artigo, o estudo estatístico a ser realizado será aquele em que a execução terá como base a análise da consistência e coerência dos dados de desempenho obtidos. Sob esse viés, os critérios quantitativos adotados para a investigação foram a análise da frequência absoluta, verificação de termos, métricas ou testes relacionados à estatística, análise da variância e a correlação entre termos de cunho estatístico.

Neste trabalho é elaborada uma meta-análise qualitativa, por meio de processo de mineração de dados, dos artigos dos últimos 4 anos (2018 a 2021) do Simpósio de Sistemas Computacionais de Alto Desempenho (WSCAD). Como resultado, realiza uma síntese sobre os métodos estatísticos e as métricas utilizadas nestas publicações. Em [Osorio 2018], os autores apresentaram a sistematização dos resultados da meta-análise das publicações das primeiras 18 edições do WSCAD (2000 a 2017). Sendo assim, este artigo amplia a análise publicada por aqueles autores, contribuindo com uma investigação casual dos estudos estatísticos nas publicações do simpósio.

O objetivo deste trabalho é fazer um levantamento de como o estudo estatístico está sendo caracterizado nos artigos do simpósio WSCAD, com vistas a trazer indicativos se houve mudanças significativas no uso das terminologias estatísticas nas submissões produzidas em comparação aos anos anteriores, reiterando a análise de [Osorio 2018].

Este trabalho está dividido em 6 seções. A Seção 2 apresenta os trabalhos relacionados ao estudo apresentado neste artigo. Na Seção 3 são apresentados os critérios de análise quantitativa. A Seção 4 apresenta a discussão sobre a metodologia utilizada. Os resultados da coleta realizada são discutidos na Seção 5 e a Seção 6 conclui o trabalho.

## 2. Trabalhos Relacionados

Em 1995 [Tichy et al. 1995] buscaram artigos de pesquisa em Ciência da Computação, do ano de 1993, para determinar se os pesquisadores apoiavam seus resultados com avaliação experimental. Como conclusão, foi possível constatar que há uma falta considerável de apoio para os resultados, sendo que 40% da amostra não possuía nenhuma avaliação. Posteriormente, em 2005, os autores de [Wainer 2007], replicaram a pesquisa de Tichy com 147 artigos publicados, menos da metade do número da pesquisa original. No entanto, verificaram que 33% dos artigos ainda encontravam-se na mesma situação.

Não obstante, estudos mais recentes também enfatizam a importância da experimentação, e constaram a falta dela, em várias áreas da Ciência da Computação. Por exemplo, para os autores [Andujar et al. 2012], a experimentação deve ser melhor compreendida e apreciada como uma metodologia chave em Informática. As ideias básicas do "método experimental" devem ser incluídas no aprendizado de Informática, como também em outros currículos de ciências.

Após isso, [Tedre e Moisseinen 2014] buscaram responder o que os Cientistas da Computação querem dizer quando falam sobre experimentos em Ciência da Computação, debatendo 5 tópicos cruciais. São eles: experimento de viabilidade; experimento de teste; experimento de campo; experimento de comparação; e experimento controlado. Em 2015, [Adler et al. 2015] investigou 183 artigos nos anais do IPIN (*International Conference on*

*Indoor Positioning and Indoor Navigation*). A partir disso, constataram que 95% de todos os artigos relacionados ao IPIN realizaram algum tipo de avaliação.

Portanto, considerando os resultados dos trabalhos mencionados acima, a proposta deste trabalho é complementar e atualizar a obra de [Osorio 2018], em que as publicações do WSCAD foram avaliadas. Essas avaliações foram baseadas na forma como são descritas as análises estatísticas de desempenho e se há algum tipo de validação estatística dos resultados. Os autores analisaram uma amostra total de 426 artigos, dentre as quais, 398 publicações fizeram referência a pelo menos um dos termos pesquisados, correspondendo a 93% do total.

## 3. Critérios de Análise Quantitativa

Considerando a natureza deste trabalho, os artigos publicados no evento WSCAD até a publicação de [Osorio 2018], estão sendo analisados e comparados com as publicações no mesmo evento nos três anos subsequentes. Pode-se afirmar que a presente pesquisa trata-se de uma investigação causal, que segundo [Lakatos e Marconi 2003] visa a descoberta de possíveis relações de causa e efeito, um padrão de comportamento, buscando explicações para semelhanças ou divergências. Sendo assim, os critérios analíticos considerados foram os seguintes:

- Frequência absoluta do uso de termos relativos à estatística em publicações do WSCAD (delimitadas na seção 4);
- Verificar se, nos artigos pesquisados, foi feito o uso de termos, métricas ou testes relacionados à estatística;
- Analise de variância das categorias de termos pesquisados;
- Correlação entre a ocorrência concomitante de termos de cunho estatístico.

## 4. Metodologia

O desenvolvimento do presente trabalho, seguiu os seguintes passos:

- Foram utilizados os dados minerados por [Osorio 2018], para o perídio de 2000 a 2017 do WSCAD, e que estavam disponíveis em repositório público[1];
- Foram minerados dados do WSCAD dos anos de 2018 a 2021;
- Foi feito o tratamento dos arquivos de 2018 a 2021;
- Foram tratadas e conferidas inconsistências de dados, manualmente;
- Os dados minerados, assim que tratados foram salvos no formato de arquivo com valores separados por vírgulas (CSV);
- Os arquivos CSV foram processados nas linguagens de programação Python e R;
- Foram feitas as quantificações comparativas ao trabalho de [Osorio 2018] ;
- Foi calculada, via software R, a analise de variância para três grupos de palavras em função do ano de observação;
- Adicionalmente foi verificado se há correlação entre os dados observados;
- Todos os arquivos de código fonte e CSV foram salvos em repositório público no github[2].

[1]https://github.com/alessanderosorio/SMPE-UFRGS
[2]https://github.com/brunoluizs/ADIMS

Os dados dos trabalhos publicados nos Anais do WSCAD entre os anos de 2000 a 2017 foram extraídos do artigo publicado por [Osorio 2018], já para os dados dos trabalhos publicados entre os anos de 2018 a 2021, os trabalhos foram salvos localmente, numerados e separados por pasta segundo o ano de publicação. Utilizando os mesmos critérios adotados por [Osorio 2018], trabalhos no formato de resumo ou escritos em língua estrangeira foram eliminados do processamento, dessa forma, do total de 136 artigos, restaram 70 publicações para análise.

A extração automatizada, de textos de arquivos PDF, utilizando pacotes Python, se mostrou mais complicada do que o antecipado. Inicialmente, os testes foram efetuados utilizando-se 3 pacotes diferentes, a saber PyPDF2[3], pdfpumbler[4] e PyMuPDF[5], cada um apresentando uma deficiência diferente na leitura de caracteres acentuados. O PyPDF2 inseria espaços, em número variável, entre caracteres acentuados além de deixar outros acentos "soltos" no restante da linha. O pdfpubler, em sua configuração padrão, falhava ao reconhecer o espaçamento entre as palavras. O PyMuPDF, das soluções testadas, foi a que apresentou o melhor desempenho, apesar de não acentuar corretamente as palavras, não inseria outros artefatos no texto.

Assim, seria possível remover a acentuação também dos termos de busca e comparar com o texto não acentuado extraído do artigo, porém isso introduziria outros problemas, por exemplo, na busca pelo termo média, em sua versão não acentuada, gerou correspondências com palavras como imediatamente e "media". Apesar de não parecer um problema, "media" não se referia apenas ao termo estatístico média mas também à tradução, em língua inglesa, da palavra mídia. Se buscássemos uma combinação exata, para evitar correspondência dos termos com palavra maiores e não relacionadas, ainda incorreríamos em problemas como o da correspondência com a palavra inglesa "media", além disso, não seriam geradas correspondências para os plurais dos termos de busca.

Dessa forma, como o número de artigos selecionados era relativamente pequeno, optou-se por conduzir um processo de busca manual, por meio do qual foram realizadas buscas, nos textos dos artigos em arquivos PDF, pelos termos utilizados em análise estatística bem como métricas e testes utilizados para aferição de resultados, assim categorizados neste artigo. Para isso, foi utilizada a função localizar de um programa capaz de ler arquivos PDF[6]. Os termos de busca são apresentados nas Tabelas 1, 2 e 3.

A Tabela 1, contém as palavras de busca referentes aos "termos" estatísticos, bem como suas respectivas siglas. Correspondem à palavras empregadas na rotina estatística.

Na Tabela 2, encontram-se palavras de uso mais restrito, agrupadas sob a referência de "métricas". Tratam-se de palavras empregadas em medidas estatísticas de desempenho no contexto da computação.

Já na Tabela 3, estão os termos de busca e siglas relativos aos testes estatísticos, relacionados à validação estatística, e agrupados sob o nome "testes", que é a palavra de referência no decorrer do trabalho.

[3]https://pypi.org/project/PyPDF2/
[4]https://pypi.org/project/pdfplumber/0.1.2/
[5]https://pymupdf.readthedocs.io/en/latest/
[6]Google Chrome versão 103

**Tabela 1. Termos estatísticos selecionados para coleta de dados**

| Descrição | Chave de Pesquisa |
|---|---|
| Amostra (AM) | amostra |
| Desvio Padrão (DP) | desvio padrão |
| Distribuição Normal (DN) | distribuição normal |
| Frequência (FR) | frequência OR frequência |
| Gaussiana (GA) | gaussiana |
| Intervalo de Confiança (IC) | intervalo de confiança |
| Média (ME) | média |
| Num. Execuções (NE) | número de execuções |
| Num. Iterações (NI) | numero de iterações |
| Teste/Experimento (TE) | teste OR experimento OR simulação |
| Variância (VR) | variância |

Fonte: [Osorio 2018]

**Tabela 2. Métricas selecionadas para coleta de dados**

| Discrição | Chave de Pesquisa |
|---|---|
| Bandwidth (BW) | “bandwidth” OR “largura de banda” |
| BPS (BP) | “bits por segundo” OR “bps” |
| Capacidade Nominal (CN) | “nominal capacity” OR “capacidade nominal” |
| Capacidade Utilizavel (CU) | “usable capacity” OR “capacidade utilizável” |
| Confiabilidade (CO) | Reliability OR Confiabilidade |
| Cost/Performance Ratio (CP) | “cost ratio” OR “performance ratio” |
| Disponibilidade (DI) | availability OR disponibilidade |
| Downtime/Uptime (DU) | downtime OR uptime |
| Eficiência/Acurácia (EA) | eficiência OR eficácia OR accuracy |
| Fator de Estiramento (FE) | “strech factor” OR “fator de estriamento” |
| Tempo Ocioso (TO) | “Idle time” OR “tempo ocisoso” |
| MFLOPS (MF) | MFLOPS |
| MIPS (MI) | MIPS |
| MTTF (MT) | MTTF |
| PPS (PP) | PPS |
| Speed up (SU) | “speedup” OR “speed-up” OR “speed up” |
| Tempo de Reaçao (TR) | “reaction time” OR “tempo de reação” |
| TPS (TP) | TPS |

Fonte: [Osorio 2018]

**Tabela 3. Testes estatísticos selecionados para coleta de dados**

| Descrição | Chave de Pesquisa |
|---|---|
| P-Valor (PV) | "p-valor" OR "p-value" OR "valor p" |
| Teste ANOVA (AN) | anova |
| Teste Chi-quadrado (CH) | "chi-quadrado" OR "qui-quadrado" |
| Teste de Wilcoxon (TC) | wilcoxon signed-rank |
| Teste Exato de Fisher (FI) | "teste exato de fisher" OR "fisher" |
| Teste Kruskal-Wallis (KR) | kruskal-wallis |
| Teste T (TT) | "teste t" OR "teste-t" OR "teste de student" OR "Student" |
| Teste U (TU) | "teste U" OR "mann-whitney" OR "wilcoxon rank-sum" |

Fonte: [Osorio 2018]

## 5. Discussão

Na presente sessão, apresentam-se as comparações entre os trabalhos publicados entre os anos de 2000 a 2017 com os dos anos de 2018 a 2021 no WSCAD. Para possibilitar a análise entre os anos foi utilizada a proporção de artigos com citações em relação total de artigos no ano e não o número de artigos, uma vez que o número de artigos por ano não é o mesmo.

Após a tabulação, os dados foram sumarizados por ano conforme a categoria do termo. As Tabelas 4, 5, 6, 7, 8 e 9 apresentam os resultados de termos estatísticos, métricas e testes. Somente os termos que obtiveram resultados foram sumarizados. Em relação às tabelas 4 e 5, embora, em virtude da extração manual, fosse possível diferenciar a ocorrência do termo estatístico frequência da grandeza física (oscilações por segundo em Hz), para efeitos de comparação dos resultados, o termo foi excluído da contagem, como feito em [Osorio 2018], onde, em virtude da extração automatizada, não era possível diferenciar à grandeza física do termo estatístico. Já os resultados dos termos distribuição normal e gaussiana foram agregados por se tratarem do mesmo objeto.

Pela Tabela 4, constata-se que o total de ocorrência de termos estatísticos na primeira janela de tempo (2000-2017), apresentados na coluna "n" tem média de 29,33 utilizações por ano, com um desvio padrão de 14,28 entre os anos. Na segunda janela (2018-2021), obtém-se, pela Tabela 5, uma média consideravelmente maior, sendo verificadas 40 aparições de termos estatísticos em média para cada ano, com uma menor oscilação, denotada por um desvio padrão de 11,94 pontos. Cabe ressalvar que o número total de observações para a segunda janela de tempo é consideravelmente menor, devido ao menor período temporal e consequentemente ao menor número de artigos analisados.

Sobre o emprego de palavras que remetam a métricas estatísticas, obtém-se da Tabela 6 uma média 25,56 ocorrências por ano, entre 2000 e 2017, contra 30,25 observações/ano no período de 2018 a 2021, decorrente da Tabela 7. Os desvios-padrão para "métricas" apresentaram valores próximos entre as duas janelas de tempo, sendo observados valores de 8,66 e 6,95 pontos para o primeiro e segundo intervalo de tempos respectivamente.

Os maiores valores observados para métricas se deram em 2003, 2004 e em 2019 ("n" = 38), as ocorrências predominantes foram: "EA", "CO" e "SU".

Tanto para o período de 2000 a 2017, quanto de 2018 a 2021, os termos relativos a

**Tabela 4. Citações de Termos Estatísticos de 2000 a 2017**

| Ano | n | AM | DP | DN | GA | IC | ME | NE | NI | TE | VR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 | 7 | - | - | - | - | 2 | - | 1 | 4 | - | - |
| 2001 | 34 | 4 | 1 | - | - | 10 | 1 | 1 | 16 | 1 | - |
| 2002 | 35 | 1 | 2 | - | - | 9 | 1 | 2 | 20 | - | - |
| 2003 | 40 | 2 | 3 | - | - | 11 | 1 | 2 | 21 | - | - |
| 2004 | 53 | - | 4 | - | - | 19 | 1 | 4 | 24 | 1 | - |
| 2005 | 58 | 3 | 6 | - | 1 | 17 | - | 2 | 28 | 1 | - |
| 2006 | 38 | - | 2 | 1 | - | 7 | 1 | 3 | 24 | - | - |
| 2007 | 28 | - | 1 | - | - | 7 | 1 | 1 | 17 | 1 | - |
| 2008 | 41 | 1 | 3 | 1 | 1 | 12 | - | 1 | 22 | - | - |
| 2009 | 18 | - | - | - | - | 4 | - | - | 14 | - | - |
| 2010 | 19 | 2 | - | 1 | - | 5 | - | - | 11 | - | - |
| 2011 | 8 | - | - | - | - | 3 | - | - | 5 | - | - |
| 2012 | 33 | - | 1 | 1 | 1 | 7 | 1 | - | 21 | 1 | - |
| 2013 | 31 | 2 | 2 | 2 | - | 11 | - | - | 14 | - | - |
| 2014 | 33 | - | - | - | - | - | 3 | 2 | 27 | 1 | - |
| 2015 | 16 | - | 1 | 1 | 1 | 1 | - | 1 | 11 | - | - |
| 2016 | 23 | 1 | - | 1 | 1 | 2 | - | 1 | 16 | 1 | - |
| 2017 | 13 | 2 | 1 | - | - | 2 | - | - | 8 | - | - |
| $\mu$ | 29,33 | 1,00 | 1,50 | 0,45 | 0,28 | 7,17 | 0,56 | 1,17 | 16,83 | 0,39 | - |
| $\sigma$ | 14,28 | 1,24 | 1,65 | 0,62 | 0,46 | 5,44 | 0,78 | 1,15 | 7,16 | 0,50 | - |

Fonte: Adaptado de [Osorio 2018]

**Tabela 5. Citações de Termos Estatísticos de 2018 a 2021**

| Ano | n | AM | DP | DN | IC | ME | NE | NI | TE | VR |
|---|---|---|---|---|---|---|---|---|---|---|
| 2018 | 45 | 3 | 4 | - | 2 | 15 | - | 2 | 19 | - |
| 2019 | 53 | 5 | 5 | 2 | - | 14 | - | 3 | 22 | 2 |
| 2020 | 37 | 2 | 7 | 1 | - | 11 | - | - | 16 | - |
| 2021 | 25 | 3 | 2 | - | - | 7 | - | 1 | 12 | - |
| $\mu$ | 40,00 | 3,25 | 4,50 | 0,75 | 0,50 | 11,75 | - | 1,50 | 17,25 | 0,50 |
| $\sigma$ | 11,94 | 1,26 | 2,08 | 0,96 | 1,00 | 3,59 | - | 1,29 | 4,27 | 1,00 |

Fonte: do autor

**Tabela 6. Citações de Métricas de 2000 a 2017**

| Ano | n | Disp. | BW | BP | CP | DU | EA | ID | MF | MI | PP | CO | SU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 | 17 | 5 | 1 | - | - | - | 6 | 1 | - | 1 | - | 1 | 2 |
| 2001 | 29 | 7 | 3 | - | - | - | 9 | - | - | 1 | - | 2 | 7 |
| 2002 | 31 | 5 | 1 | - | - | - | 12 | - | - | 3 | - | 4 | 6 |
| 2003 | 38 | 11 | 2 | - | - | - | 8 | - | 3 | 2 | - | 5 | 7 |
| 2004 | 38 | 10 | - | - | - | - | 14 | - | - | 1 | - | 6 | 7 |
| 2005 | 36 | 10 | 1 | - | - | - | 9 | 1 | - | 1 | - | 4 | 10 |
| 2006 | 15 | 3 | 2 | - | - | - | 4 | - | - | 1 | - | - | 5 |
| 2007 | 17 | 3 | 1 | - | - | - | 3 | 1 | - | 2 | - | - | 7 |
| 2008 | 33 | 6 | 5 | - | - | - | 7 | - | 1 | 3 | - | 3 | 8 |
| 2009 | 23 | 6 | 1 | 1 | - | - | 4 | 1 | 1 | 1 | - | 1 | 7 |
| 2010 | 21 | 5 | - | - | - | 1 | 2 | 1 | - | 5 | - | 2 | 5 |
| 2011 | 8 | 3 | - | - | - | - | 1 | - | - | 1 | - | - | 3 |
| 2012 | 24 | 9 | 2 | 1 | - | - | 4 | - | - | 4 | - | - | 4 |
| 2013 | 31 | 5 | - | 1 | - | 3 | 5 | 1 | - | 3 | - | 2 | 11 |
| 2014 | 32 | 6 | - | - | - | - | 5 | 1 | - | 3 | - | 1 | 16 |
| 2015 | 17 | 1 | - | - | - | - | 3 | 1 | 1 | 2 | 1 | - | 8 |
| 2016 | 28 | 6 | 3 | - | 1 | - | 1 | 4 | 1 | 1 | - | 1 | 10 |
| 2017 | 22 | 5 | 1 | - | - | 1 | - | 2 | - | 1 | - | 1 | 11 |
| $\mu$ | 25,56 | 5,89 | 1,28 | 0,17 | 0,06 | 0,28 | 5,39 | 0,78 | 0,39 | 2,00 | 0,06 | 1,83 | 7,44 |
| $\sigma$ | 8,66 | 2,70 | 1,36 | 0,38 | 0,24 | 0,75 | 3,84 | 1,00 | 0,78 | 1,24 | 0,24 | 1,86 | 3,33 |

Fonte: Adaptado de [Osorio 2018]

**Tabela 7. Citações de Métricas de 2018 a 2021**

| Ano | n | Disp. | BW | BP | CP | DU | EA | ID | MF | MI | PP | CO | SU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2018 | 34 | 5 | 3 | - | 1 | - | 12 | - | - | 2 | - | 3 | 8 |
| 2019 | 38 | 4 | 2 | 1 | - | - | 14 | - | - | - | - | 3 | 14 |
| 2020 | 26 | 3 | 2 | - | - | - | 8 | - | 1 | 1 | - | 4 | 7 |
| 2021 | 23 | 2 | 4 | - | - | - | 10 | - | - | - | - | 2 | 5 |
| $\mu$ | 30,25 | 3,50 | 2,75 | 0,25 | 0,25 | - | 11 | - | 0,25 | 0,75 | - | 3,00 | 8,50 |
| $\sigma$ | 6,95 | 1,29 | 0,96 | 0,50 | 0,50 | - | 2,58 | - | 0,50 | 0,96 | - | 0,82 | 3,87 |

Fonte: do autor

testes estatísticos são os menos frequentes, com média de ocorrência de apenas 0,73 vezes ao ano para os anos anteriores a 2018 expressos na Tabela 8, e média de 2 observações anuais na Tabela 9, que representa o intervalo de 2018 a 2021, com desvios-padrão de 1,10 e 1,63 respectivamente, para a coluna "n" de ambas as tabelas (8 e 9).

**Tabela 8. Citações de Testes de 2000 a 2017**

| Ano | n | PV | TC | KR | TT |
|---|---|---|---|---|---|
| 2001 | 1 | 1 | - | - | - |
| 2002 | - | - | - | - | - |
| 2003 | - | - | - | - | - |
| 2004 | - | - | - | - | - |
| 2005 | - | - | - | - | - |
| 2006 | - | - | - | - | - |
| 2007 | 4 | 1 | - | - | 3 |
| 2008 | 1 | - | - | - | 1 |
| 2009 | 1 | - | - | - | 1 |
| 2010 | - | - | - | - | - |
| 2011 | - | - | - | - | - |
| 2012 | 2 | - | - | - | 2 |
| 2013 | - | - | - | - | - |
| 2014 | - | - | - | - | - |
| 2015 | 1 | - | - | - | 1 |
| 2016 | 1 | - | - | - | 1 |
| 2017 | - | - | - | - | - |
| $\mu$ | 0,73 | 0,13 | - | - | 0,60 |
| $\sigma$ | 1,10 | 0,35 | - | - | 0,91 |

Fonte:Adaptado de [Osorio 2018]

**Tabela 9. Citações de Testes de 2018 a 2021**

| Ano | n | PV | TC | KR | TT |
|---|---|---|---|---|---|
| 2018 | 2 | - | - | 1 | 1 |
| 2019 | 4 | 1 | 1 | - | 2 |
| 2020 | 2 | - | - | - | 2 |
| 2021 | - | - | - | - | - |
| $\mu$ | 2,00 | 0,25 | 0,25 | 0,25 | 1,25 |
| $\sigma$ | 1,63 | 0,50 | 0,50 | 0,50 | 0,96 |

Fonte: do autor

A Tabela 10 apresenta a sumarização dos resultados das Tabelas 4, 6 e 8, enquanto a Tabela 11 apresenta a sumarização dos resultados das Tabelas 5, 7 e 9.

Nas tabelas 10 e 11 é possível observar o número total de artigos no ano (coluna "n"), o número de artigos que contém pelo menos uma ocorrência do termo (colunas "Art."), bem como o número de citações aos termos (colunas "Cit."). Os dados estão agrupados por ano e pelas categorias dos termos. Por exemplo, o ano 2000, foram analisados 9 trabalhos, dos quais: 5 apresentaram termos do grupo "estatística" - perfazendo 7

citações; 7 artigos citaram, ao todo, 17 palavras do grupo “métricas”; não foram citados termos relacionados a “testes”.

**Tabela 10. Distribuição de citações de 2010 a 2017 por tipo do termo e ano**

| Ano | n | Estatística | | Métricas | | Teste | |
|---|---|---|---|---|---|---|---|
| | | Art. | Cit. | Art. | Cit. | Art. | Cit. |
| 2000 | 9 | 5 | 7 | 7 | 17 | - | - |
| 2001 | 23 | 20 | 34 | 16 | 29 | 1 | 1 |
| 2002 | 27 | 22 | 35 | 19 | 31 | - | - |
| 2003 | 32 | 28 | 40 | 23 | 38 | - | - |
| 2004 | 33 | 30 | 53 | 21 | 38 | - | - |
| 2005 | 34 | 31 | 58 | 26 | 36 | - | - |
| 2006 | 28 | 24 | 38 | 12 | 15 | - | - |
| 2007 | 21 | 17 | 28 | 12 | 17 | 2 | 4 |
| 2008 | 28 | 24 | 41 | 17 | 33 | 1 | 1 |
| 2009 | 23 | 14 | 18 | 16 | 23 | 1 | 1 |
| 2010 | 20 | 12 | 19 | 16 | 21 | - | - |
| 2011 | 6 | 5 | 8 | 5 | 8 | - | - |
| 2012 | 28 | 22 | 33 | 17 | 24 | 2 | 2 |
| 2013 | 20 | 17 | 31 | 16 | 31 | - | - |
| 2014 | 39 | 29 | 33 | 23 | 32 | - | - |
| 2015 | 15 | 12 | 16 | 11 | 17 | 1 | 1 |
| 2016 | 21 | 16 | 23 | 17 | 28 | 1 | 1 |
| 2017 | 19 | 9 | 13 | 15 | 22 | - | - |
| $\mu$ | 23,67 | 18,72 | 29,33 | 16,06 | 25,56 | 0,50 | 0,61 |
| $\sigma$ | 8,51 | 8,21 | 14,28 | 5,40 | 8,66 | 0,71 | 1,04 |

Fonte: Adaptado de [Osorio 2018]

**Tabela 11. Distribuição de citações de 2018 a 2021 por tipo do termo e ano**

| Ano | n | Estatística | | Métricas | | Teste | |
|---|---|---|---|---|---|---|---|
| | | Art. | Cit. | Art. | Cit. | Art. | Cit. |
| 2018 | 19 | 19 | 45 | 18 | 34 | 2 | 2 |
| 2019 | 23 | 22 | 53 | 22 | 38 | 3 | 4 |
| 2020 | 16 | 16 | 37 | 13 | 26 | 2 | 2 |
| 2021 | 12 | 12 | 25 | 11 | 23 | - | - |
| $\mu$ | 17,50 | 17,25 | 40,00 | 16,00 | 30,25 | 1,75 | 2,00 |
| $\sigma$ | 4,65 | 4,27 | 11,94 | 4,97 | 6,95 | 1,26 | 1,63 |

Fonte: do autor

Para fins de comparação, das Tabelas 4 a 9, foi elaborada a Tabela 12, onde foi realizada a soma das tabelas elaboradas por [Osorio 2018], de modo que são apresentadas as médias ($\mu$) para cada termo em relação ao período analisado. Na primeira linha, as médias de 2000 a 2018, na segunda linha as médias de 2019 a 2021 e na terceira linha a variação percentual ($\Delta\%$ ) entre as duas médias ($\mu$) em que se observa, para o “n” Total um acréscimo de 21,70% na média de uso de termos dos três grupos de palavras após o ano de 2018.

**Tabela 12. Variação nas médias no "n" de termos citados entre 2000 e 2018 e de 2019 a 2021**

| Período | Estatística | Métricas | Testes | "n" Total |
|---|---|---|---|---|
| 2000 a 2018($\mu$) | 30,16 | 26,00 | 0,81 | 56,97 |
| 2019 a 2021($\mu$) | 38,33 | 29,00 | 2,00 | 69,33 |
| $\Delta$ % | 27,09 | 11,53 | 146,91 | 21,70 |

Fonte: do autor

A Figura 1, apresenta o percentual de artigos com citações para cada um dos três grupos de palavras-chave, em relação ao número de artigos analisados por ano. Pode-se observar que não há qualquer padrão de linearidade ou mudança a partir do ano de 2018. A volatilidade visível na figura é expressa por desvio padrão de 6,33% para "termos", 6,65% para "métricas"e uma volatilidade de 2,51 pontos percentuais para "testes". Observando a Figura 1, em que a linha referente aos "testes" aparece muito inferior aos demais itens, é valido supor que dentre os autores das publicações analisadas, há muito pouco conhecimento avançado em termos de estatística.

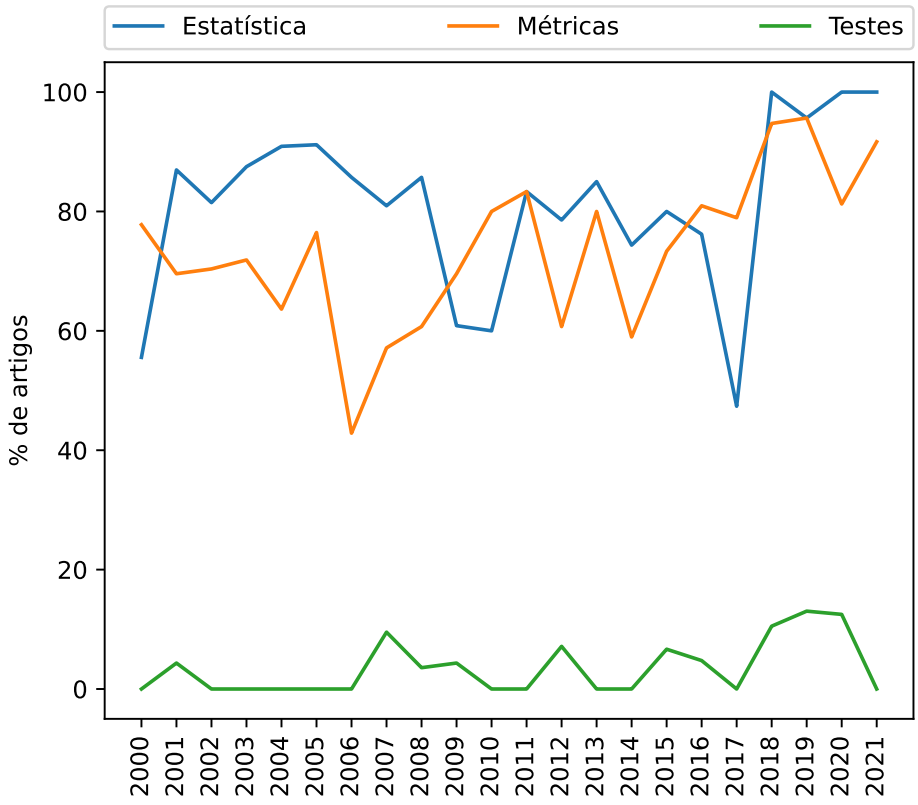


**Figura 1. Proporção de artigos com citações por ano**

Fonte: do autor

As mesmas considerações feitas para os valores percentuais, são válidas quando observamos a Figura 2, que contém a razão entre o número de ocorrência de cada um dos grupos de palavras e o volume de artigos para os respectivos anos, com a ressalva de que, o fato de a linha relativa à "testes" estar mais expressiva, não demonstra o uso dos

mesmos, mas sim, que há poucos trabalhos contendo este tipo específico de refinamento Estatístico.

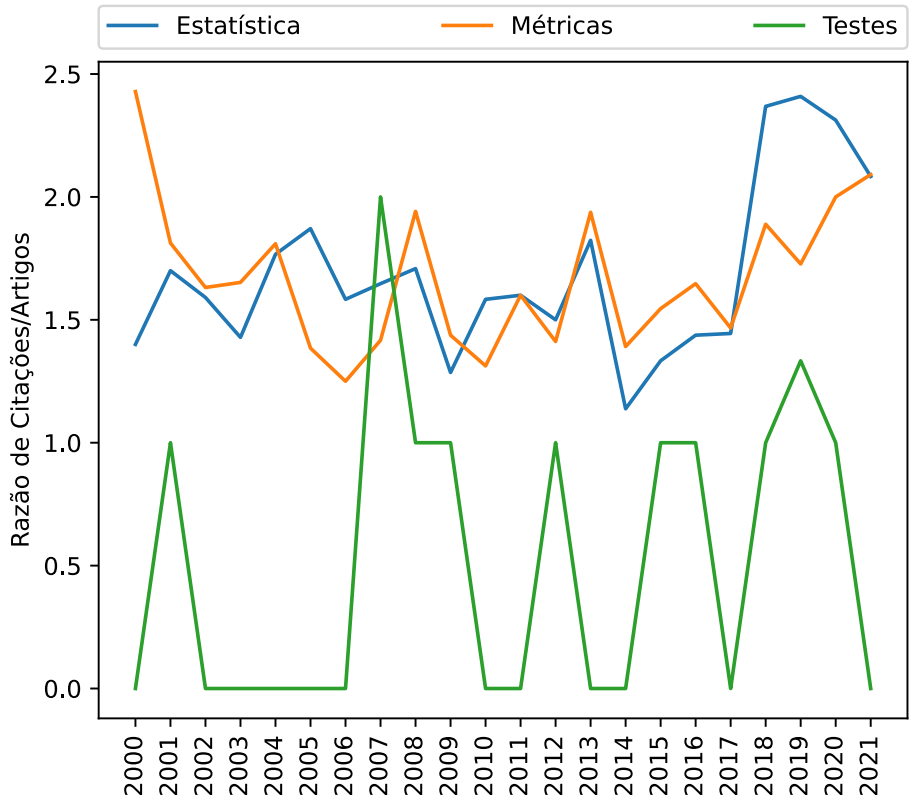


**Figura 2. Razão de citações por artigo ano**
Fonte: do autor

Para que se pudesse analisar visualmente o total de termos utilizados no decorrer dos anos, com o intuito de verificar se houve mudanças ou tendências após a publicação de [Osorio 2018], foi elaborada a Figura 3. São apresentados os respectivos valores totais para cada ano, os quais possibilitam a observação de uma distribuição volátil (desvio padrão de 21,85 pontos) sem haver qualquer tipo de tendência aparente.

Embora haja no gráfico um aclive após o ano de 2018, o mesmo nível de valor pode ser observado entre os anos de 2004 e 2005. Observa-se que no ano de 2021 o valor de ocorrências atingiu nível significativamente baixo, ficando abaixo da média geral do gráfico de 58,54 citações por ano.

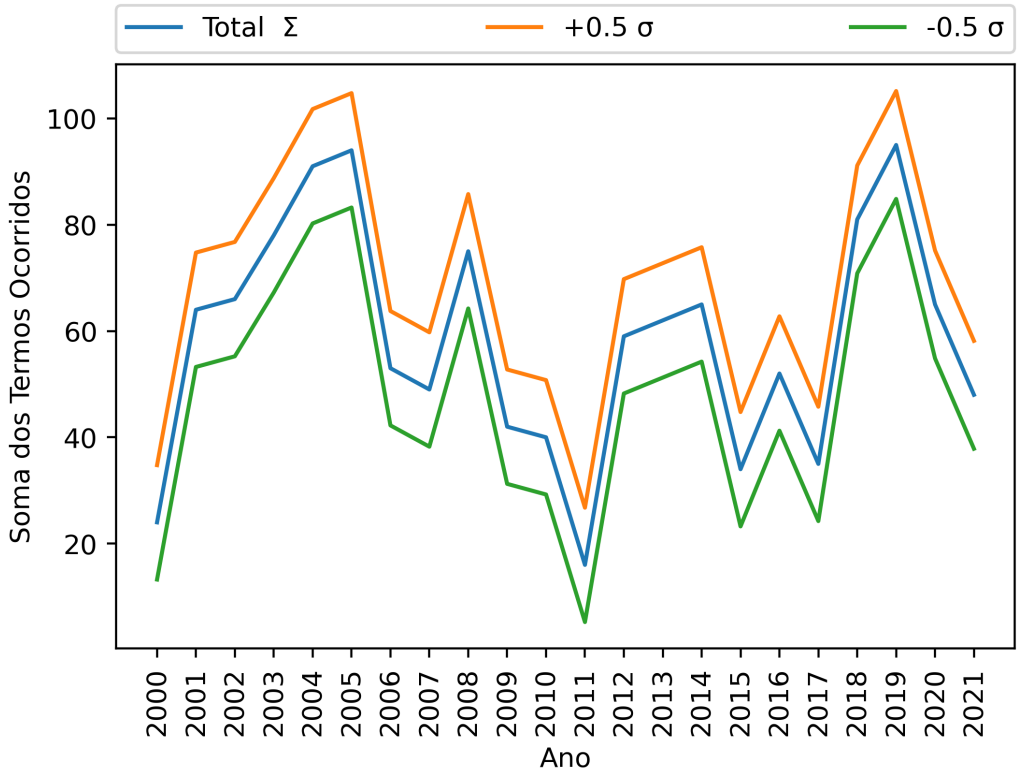


**Figura 3. Soma do "n" de termos por ano, de 2000 a 2021**
Fonte: do autor

### 5.1. Análise de Variância (ANOVA) dos Termos e Teste U

Na Tabela 12 é possível visualizar um aumento de no uso de Termos Estatísticos depois da publicação do trabalho de [Osorio 2018], porém, esta informação é divergente do observado na Figura 3, onde não há qualquer indício de mudança no padrão gráfico de uso dos termos posterior à referida publicação.

Dada a divergência apresentada acima, foram elaboradas as duas hipóteses apresentadas a seguir, para as quais, com base nas Tabelas 10 e 11, que sumarizam os valores totais relativos às ocorrências de termos estatísticos, realizou-se a análise de variâncias (ANOVA), para cada um dos três grupos de termos de busca, utilizou-se um nível de significância $\alpha$ de 0,05.

$H_0$ : *após a publicação de [Osorio 2018] não houve mudança nas médias de utilização de termos estatísticos em publicações do WSCAD*

$H_1$ : *após a publicação de [Osorio 2018] houve mudança nas médias de utilização de termos estatísticos em publicações do WSCAD*

É válido ressaltar que, apesar do ano de divisão entre as tabelas apresentadas ser 2018, como [Osorio 2018] analisou trabalhos até o ano de 2017, para fins de análise de variância, o ano de 2018 foi considerado como "anterior" pois caso a hipótese ($H_1$) de

que o trabalho publicado em 2018 introduzisse mudanças mas médias de utilização de termos estatísticos, esta mudança só seria observada a partir do ano de 2019.

Sendo assim, os dados foram rotulados em duas classes, da seguinte maneira:

- situação = "anterior", para os trabalhos de 2000 até 2018
- situação = "posterior", para os trabalhos de 2019 até 2021

Na Listagem 1, pode-se visualizar os valores obtidos da análise da variâncias, entre as situações anterior e posterior, para cada grupo de palavras, na linguagem de programação "R":

```
alfa <- 5/100  # nivel de significancia: 5% = 0.05

k <- 2 # 2 tratamentos: anterior/posterior
n <- length(df_dados$situacao) # 22 repeticoes: 2000 a 2021

aov_metricas <- aov(metricas ~ situacao, df_dados)
aov_testes <- aov(testes ~ situacao, df_dados)
aov_termos <- aov(termos ~ situacao, df_dados)

summary(aov_metricas)
##             Df Sum Sq Mean Sq F value Pr(>F)
## situacao     1    1.8   1.762   0.062  0.806
## Residuals   20  567.2  28.360

summary(aov_testes)
##             Df Sum Sq Mean Sq F value Pr(>F)
## situacao     1  3.065  3.0654   4.008  0.059
## Residuals   20 15.298  0.7649

summary(aov_termos)
##             Df Sum Sq Mean Sq F value Pr(>F)
## situacao     1   11.1   11.10   0.186  0.671
## Residuals   20 1196.4   59.82
```

**Listagem 1. ANOVA em R.**

Observa-se na Listagem 1, que os p-valores (colunas "Pr(>F)") para as três categorias de termos é maior que o valor de $\alpha$ adotado (0,05) e portanto, aceita-se a hipótese nula ($H_0$), desta forma não se verificou diferença entre as situações anterior e posterior à publicação de [Osorio 2018], para nenhum dos três grupos de palavras analisados.

O grupo "métricas", resultou em uma estatística de 0,806 maior que o valor de 0,05, vide última coluna da linha 12.

Para o grupo "testes", na linha 15, observou-se uma estatística muito próxima de 0,05, mas o valor disposto na linha 17, de 0,059 aponta para a validade da hipótese nula ($H_0$).

Do mesmo modo, para o terceiro grupo "termos" localizado na linha 20 da Listagem 1, observou-se uma estatística de 0.671 na linha 22, também suportando a aceitação

de $H_0$.

Para aplicação da ANOVA, as seguintes pressuposições devem ser satisfeitas:

1. As observações devem ser independentes;
2. As variâncias populacionais devem ser iguais (homoscedasticidade);
3. Normalidade dos resíduos;
4. A distribuição das observações em cada grupo deve ser normal.

Foi utilizado o teste de Levene[7] para homoscedasticidade, todos os p-valores retornados ficaram acima do valor de significância adotado, apontando para igualdade das variâncias, como pode-se observar na Listagem 2.

```
for tipo in ['EST', 'MET', 'TST']:
  antes  = df_anova[df_anova['tratamento'] == 'antes'  ][
     tipo]
  depois = df_anova[df_anova['tratamento'] == 'depois' ][
     tipo]

  _,levene_pvalue = stats.levene(antes, depois, center='
     median')
  print(f"{tipo}: p-value ={levene_pvalue:6.3f}")

# EST: p-value = 0.253
# MET: p-value = 0.976
# TST: p-value = 0.404
```

**Listagem 2. Teste de Levene para homoscedasticidade em Python.**

Aplicou-se o teste de Shapiro-Wilk[8] para verificar a normalidade dos resíduos, cujo código encontra-se na Listagem 3.

```
import statsmodels.formula.api as smf

for tipo in ['EST', 'MET', 'TST']:
    # Prever o tratamento (var categorica) com base no
       n_art de um dado tipo
    formula = f'{tipo} ~ C(tratamento)'
    reg = smf.ols(formula, data=df_anova).fit()

    # Verificar a normalidade dos residuos
    shapiro_test = stats.shapiro(
      reg.resid
    )
    print(f"{tipo}: p-value ={shapiro_test.pvalue:6.3f}")

# EST: p-value = 0.681
```

[7] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.levene.html
[8] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html

```
# MET: p-value = 0.741
# TST: p-value = 0.002
```

**Listagem 3. Teste de Shapiro-Wilk para normalidade dos resíduos em Python.**

Como se pode verificar pela Listagem 3, a normalidade dos resíduos não foi completamente satisfeita, com um p-valor para "testes" = 0,002, possivelmente em decorrência de que em diversos anos não houveram citações de "testes", implicando, assim, em uma série de valores iguais (zero).

Além disso, dada a grande diferença entre o número de amostras para as situações "anterior" e "posterior", bem como, o pequeno número de amostras da situação "posterior", não havia segurança para atestar a normalidade das observações (pressuposição 4)[9].

Por esse motivo, realizou-se também o teste U de Mann-Whitney[10]. Por ser um teste não paramétrico, baseado nas medianas, não assume qualquer tipo de distribuição em relação às amostras. Utilizou-se a função mannwhitneyu[11] do pacote Python ScyPy, e cujo código pode ser visto na Listagem 4:

```
import scipy.stats as stats

for tipo in ["EST", "MET", "TST"]:
  antes  = df_anova[df_anova["situacao"] == "antes" ][tipo]
  depois = df_anova[df_anova["situacao"] == "depois"][tipo]

  teste_u = stats.mannwhitneyu(
    antes,
    depois,
    alternative = "two-sided",
    method="exact"
  )
  print(f"{tipo}: p-value ={teste_u.pvalue:6.3f}")

# EST: p-value = 0.651
# MET: p-value = 0.718
# TST: p-value = 0.226
```

**Listagem 4. Teste U em Python.**

Os p-valores encontrados, linhas 15 a 17 da Listagem 4, por serem maiores que o nível de significância adotado ($\alpha$=0,05), indicam pela aceitação da hipótese nula, não havendo diferença entre as medianas das situações "anterior" e "posterior", ou seja, após a publicação de [Osorio 2018], não houve diferença significativa, estatisticamente, na utilização de termos para qualquer uma das três categorias pesquisadas.

[9]Mesmo motivo pelo qual não se utilizou o teste de Alexander Govern.

[10]O teste H de Kruskal-Wallis não foi utilizado pois pressupõe grupos com pelo menos 5 amostras.

[11]https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html

## 5.2. Análise de Correlação Entre os Termos

Dadas as considerações feitas na seção Metodologia, onde se observou a necessidade do agrupamento dos termos ”distribuição normal” e ”Gaussiana”, bem como, o fato de algumas palavras aparecerem raramente, como uma possível melhoria metodológica para trabalhos futuros, foi calculada a correlação de $\phi k$, desenvolvido por [Baak et al. 2019].

Sendo assim, a observação de haver uma correlação forte entre dois termos de busca, é indicativo de que ambos podem ser computados no mesmo índice, facilitando o processo de contagem, tabulação e análise dos dados, uma vez que, com menos índices nas tabelas a visualização e a interpretação das mesmas se torna mais objetiva.

A escolha do índice de correlações de $PhiK(\phi k)$, deu-se em detrimento da correlação de Pearson expressa na Equação 1, que é a mais popular e praticamente a técnica padrão de estabelecimento de correlações, porém apresenta algumas restrições, sendo aplicável apenas às variáveis contínuas, considerando uma relação linear entre as variáveis, de modo que, é sensível à *outliers*.

$$\rho = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} \tag{1}$$

Em que:

$\sum(x - \bar{x})(y - \bar{y})$, é a covariância entre X e Y;

$\sum(x - \bar{x})^2$, é a variância de X;

$\sum(y - \bar{y})^2$, é a variância de Y.

A correlação de $\phi k$ se diferencia, não só da correlação de Pearson, como de qualquer outra técnica de estabelecimento de causa e efeito, principalmente por admitir variáveis categóricas, ordinais e intervalares, demonstrando dependências não lineares entre as variáveis e no caso de variáveis bivariadas, os resultados computados são iguais aos da técnica de Pearson. É valido ressaltar que o cálculo do coeficiente de $\phi k$ apresenta um custo computacional alto, devido aos cálculos de integral em seu processo.

Interpreta-se o valor de $\chi^2$ encontrado nos dados, como proveniente de uma distribuição normal bivariada com uma quantidade fixa de ruído estatístico e com o parâmetro de correlação $\phi k$ . As relações não lineares são capturados por $\phi k$ através do teste do chi-quadrado para independência de variáveis e no caso de se tratar de uma distribuição normal bivariada, a correlação $\phi k$ aplica (revertendo para) o coeficiente de correlação de Pearson, com variáveis de intervalo categorizadas. O procedimento pode ser estendido para mais variáveis usando uma gaussiana multivariada em vez de uma distribuição bivariada [Baak et al. 2019].

De forma simplificada, pode-se definir o Algoritmo 1, para obtenção do coeficiente de correlação $\phi k$ da seguinte maneira[12]:

[12]Etapa de Pré-processamento está discriminada no Apêndice B.

**Algoritmo 1** - Pseudo-Algoritmo para obtenção do coeficiente de correlação $\phi k$

**Entrada:** Pré-Processamento [Apêndice B]
  **se** as variáveis forem intervalares não categorizadas **então**
    aplicar uma categorização a cada uma
  **finaliza se**
**Entrada:** alocar e preencher uma tabela de contingência para cada par de variáveis, contendo "N" registros de "r" linhas e "k" colunas
**Saída:** Cálculo do $\chi^2$(chi-quadrado) de Pearson
  **para todo** $\chi^2$, Interpretar como proveniente de uma distribuição normal bivariada, sem estatística, de acordo com a Equação 3 **faça**
    **se** $\chi^2 < \chi^2_{ped}$ **então**
      defina $\rho$ como zero
    **senão**
      com $N, r, k$ corrigidos, inverter a função $\chi^2_{b.n.}$ usando o método de Brent [Apêndice A] e resolver numericamente para $\rho$ entre [0,1]
    **finaliza se**
  **finaliza para**
  **retorna** A solução para $\rho$, define o coeficiente de correlação $\phi k$

Se as variáveis forem intervalares não categorizadas, então se aplica uma categorização a cada uma. Um dimensionamento razoável é geralmente específico do caso de uso, entretanto, por padrão, tomam-se 10 espaços de memória[13] por variável;

Deve-se alocar e preencher uma tabela de contingência para cada par de variáveis, contendo "N" registros de "r" linhas e "k" colunas;

Obtém-se o $\chi^2$ (chi-quadrado) de Pearson[14], para a estatística de teste das variáveis qualitativas e nominais. Conforme a Equação 2, se calculam as estimativas de frequência estatisticamente dependentes;

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2}$$

Em que:

$O$ , é o número de observações; $E$ , é a frequência esperada (teórica);

$\chi^2$ , é a estatística de teste cumulativa de Pearson, que assintoticamente se aproxima da distribuição chi-quadrado;

Interpreta-se o $\chi^2$ como proveniente de uma distribuição normal bivariada, sem estatística, de acordo com a Equação 3[15]

[13] Para fins de calculo da integral ou da área de distribuição, considera-se este dimensionamento como as "bins", distância entre as variáveis no eixo X(abscissas).

[14] O teste qui-quadrado de Pearson (ou teste chi-quadrado de Pearson) é um teste estatístico aplicado a dados categóricos para avaliar quão provável é que qualquer diferença observada aconteça ao acaso. É adequado para amostras não pareadas/emparelhadas.

[15] $b.n.$ : leia-se distribuição bivariada normal.

$$X^2_{b.n.}(\rho, N, r, k) = \chi^2_{ped} + \left\{ \frac{\chi^2_{max}(N, r, k) - \chi^2_{ped}}{\chi^2_{b.n.}(1, N, r, K)} \right\} \chi^2_{b.n.}(\rho, N, r, K) \tag{3}$$

Caso $\chi^2 < \chi^2_{ped}$, então se define $\rho$ como zero, caso contrário, com $N, r, k$ corrigidos, inverte-se a função $\chi^2_{b.n.}$ usando o método de Brent[Apêndice A] e se resolve numericamente para $\rho$ entre [0,1];

A solução para $\rho$ define o coeficiente de correlação $\phi k$

Para este trabalho, os dados foram carregados em ambiente Python via CSV (arquivo de valores separados por vírgulas), conforme linha 5 da Listagem 5, e convertidos em dataframe, com o qual se aplicou o calculo da matriz de correlação de $\phi k$ disposto na linha 9.

```
import pandas as pd
import seaborn as sns
import phik

carrega_arquivo = pd.read_csv("WSCAD.csv")

arquivo_tratado = carrega_arquivo.drop(["nome"], axis=1)

matriz_de_correlacao = arquivo_tratado.phik_matrix()

sns.heatmap(matriz_de_correlacao, cmap="crest")

matriz_de_correlacao.to_csv("matriz_phik.csv")
```

**Listagem 5. $\phi k$ em Python.**

**Tabela 13. Correlações significativas 2018 a 2021**

| Termo | CO | DN | IC | ME | MI | TE | TT |
|---|---|---|---|---|---|---|---|
| Confiabilidade | 1,00 | 0,000 | 0,000 | 0,775 | 0,488 | 0,360 | 0,000 |
| Distribuição Normal | 0,000 | 1,00 | 0,000 | 0,143 | 0,000 | 0,283 | 0,750 |
| Intervalo de confiança | 0,000 | 0,000 | 1,00 | 0,000 | 0,807 | 0,000 | 0,000 |
| Média | 0,775 | 0,143 | 0,000 | 1,00 | 0,000 | 0,000 | 0,384 |
| MIPS | 0,488 | 0,000 | 0,807 | 0,000 | 1,00 | 0,826 | 0,000 |
| Teste ou Experimento | 0,360 | 0,283 | 0,000 | 0,000 | 0,826 | 1,00 | 0,000 |
| Teste T | 0,000 | 0,750 | 0,000 | 0,384 | 0,000 | 0,000 | 1,00 |

Fonte: do autor

Na Linha 11 da Listagem 5, está expressa a criação da Figura 4, a qual apresenta um mapa térmico das correlações obtidas pelo índice de $\phi k$. Em primeira análise, pode-se enumerar termos que possivelmente possuam correlação positiva com os demais, dadas as tonalidades mais escuras da figura.

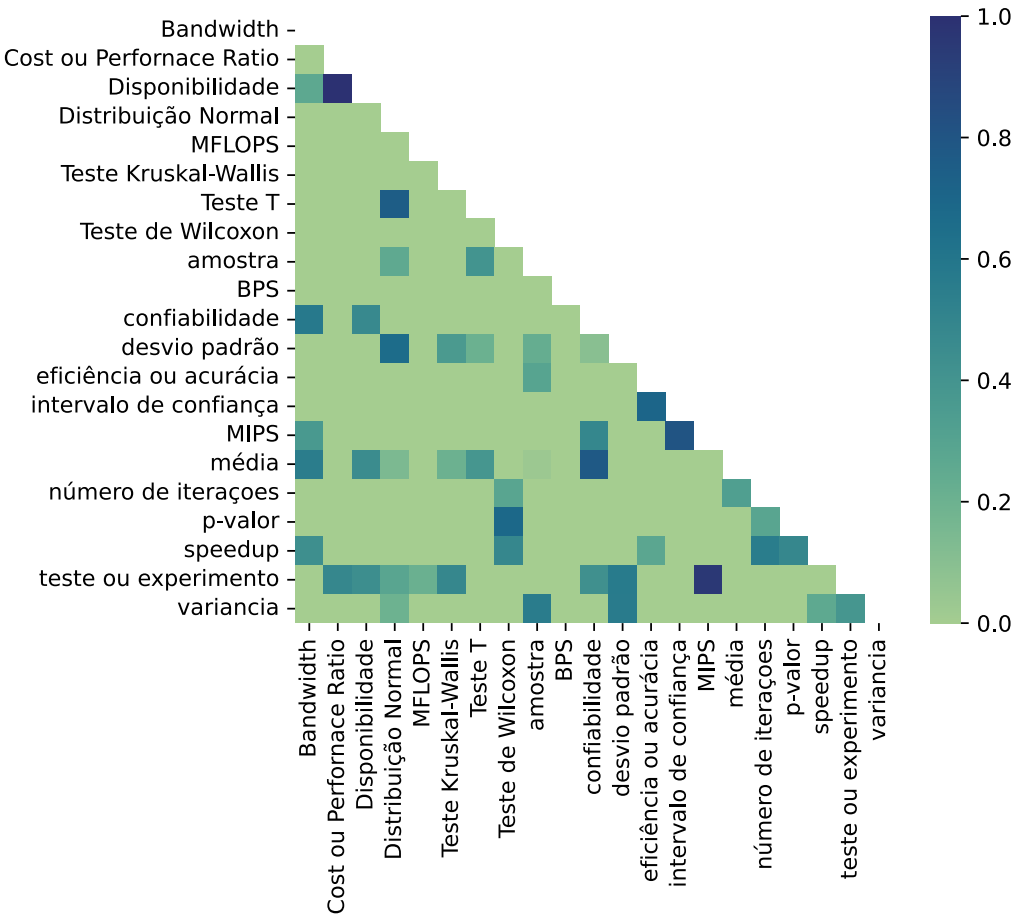


**Figura 4. Mapa térmico de correlações entre as palavras**
Fonte: do autor

Feita a primeira análise com base na Figura 4, pode-se sumarizar os índices realmente significativos, isto é, de termos que apresentem entre si correlações superiores a 0,70. Com base no arquivo produzido na Linha 13, a Tabela 13 exemplifica sete termos que apresentaram as seguintes correlações:

- Confiabilidade x Média, $\phi k = 0{,}775$;
- Distribuição Normal x Teste T, $\phi k = 0{,}750$;
- MIPS x Intervalo de Confiança, $\phi k = 0{,}807$;
- MIPS x Teste ou Experimento, $\phi k = 0{,}826$;
- Intervalo de Confiança x Teste ou Experimento, $\phi k = 0{,}000$.

Observa-se na lista apresentada acima, quatro itens com forte correlação e um último sem nenhuma, logo, possivelmente o termo "Confiabilidade" poderia ter sido computado (agrupado) no mesmo índice do termo "Média", assim como o termo "Distribuição Normal" possivelmente devesse ser sumarizado juntamente ao termo "Teste T".

Um ponto interessante da análise, são os três últimos itens da lista, pois o termo "MIPS" apresenta forte correlação com dois termos distintos, os quais, "Intervalo de Confiança" e "Teste ou Experimento", não possuem nenhuma correlação entre si, ou seja

um índice $\phi k = 0{,}00$, neste caso específico, não seria possível o agrupamento do termo MIPS com nenhum dos dois, exceto o caso de se prever este tipo de situação e se definir previamente um critério objetivo na metodologia da pesquisa.

Salienta-se que o simples cálculo de uma correlação, não é por si só um delimitador absoluto de que por exemplo, palavras sejam agrupadas, mas sim uma ferramenta de suporte metodológico. É necessário que, uma vez estabelecidos os possíveis agrupamentos, o pesquisador faça todas as considerações de contexto antes da sumarização dos termos.

### 5.3. Extração de texto de documentos PDF

A fim de identificar a origem das variações presentes na Tabela 12, partiu-se da hipótese de que a correta extração dos termos de busca depende sobremaneira do programa utilizado para busca dos termos.

Em uma verificação rápida, utilizando 3 leitores de PDF distintos [16], em uma pesquisa manual, foram obtidos números divergentes de ocorrências para os termos de busca. Assim, a diferença de comportamento entre as aplicações utilizadas, em [Osorio 2018] e neste estudo, poderia explicar, pelo menos em parte, a variação encontrada.

Como apontado por [Fenniak 2022] e [Shinyama et al. 2022], a extração de textos de documentos PDF é complicada, por uma série de motivos, como formatação, hifenização, codificação dos caracteres, e também pela forma como o PDF armazena as informações. Diferentemente dos outros formatos de documento, em um arquivo PDF, não existe o conceito de sequência de caracteres. Os textos, apesar de parecerem bem estruturados, na verdade são armazenados como um conjunto de informações de posicionamento de caracteres individuais, dificultando a identificação de palavras e frases, ou de linhas e colunas em objetos como tabelas.

Durante as tentativas, frustradas, de extração automatizada de texto de documentos PDF utilizando pacotes Python, apresentadas na seção 4, observou-se que o maior impeditivo para utilização de tais soluções residia na codificação de caracteres. No padrão Unicode, diversos caracteres podem ser expressos de mais de uma maneira. Por exemplo, o caractere Ç, pode ser expresso como U+00C7 (*LATIN CAPITAL LETTER C WITH CEDILLA*) ou também pela a sequência U+0043 (*LATIN CAPITAL LETTER C*) U+0327 (*COMBINING CEDILLA*).

A possibilidade de representar um mesmo caractere de formas distintas, impacta as buscas que envolvam tais caracteres e foram responsáveis pela maioria das diferenças entre os três programas utilizados. Isso pode ser verificado através de buscas pelos termos, utilizando as codificações alternativas dos caracteres, que assim o permitiam. Ao somar os resultados obtidos para as múltiplas codificações, os resultados entre os programas praticamente se igualaram.

Segundo o padrão Unicode, cada caractere tem duas formas normais, na Forma Normal D (NFD), os caracteres apresentam-se decompostos, na forma de sequências de código, e na Forma Normal C (NFC), primeiro aplica-se a decomposição canônica NFD, e então, combinam-se os códigos para gerar o caractere pré-combinado, com um código único. Adicionalmente, essas formas normais podem ser combinadas com uma tabela de

[16] Google Chrome 104, Foxit Reader para Linux 2.4.4 e Xreader 2.8.3

equivalências K, gerando as formas normais NFKD e NFKC, nas quais além dos respectivos processos de normalização, são realizadas substituições de caracteres equivalentes, mapeados na tabela K, como por exemplo o numeral romano I (U+2160: *ROMAN NUMERAL ONE*) e a letra i maiúscula I (U+0049: *LATIN CAPITAL LETTER I*). Em Python, essas normalizações podem ser feitas utilizando-se a função normalize[17], do módulo unicodedata, passando os argumentos forma (NFD, NFC, NFKD ou NFKC) e o texto Unicode que se deseja normalizar.

Ao normalizar o texto extraído com pacote tika-python[18], que utiliza a ferramenta Java Apache Tika[19] para extração de texto de documentos, foram obtidos excelentes resultados, superando inclusive os três leitores de PDF testados e se aproximando de uma leitura do texto em busca dos termos de pesquisa, estando o desempenho da solução limitado à qualidade das expressões de busca elaboradas. O bom desempenho dessa ferramenta pode ser atestado[20], através da análise manual, pelos autores deste trabalho, de 40 artigos selecionados por amostragem aleatória simples, entre os 397 artigos completos, publicados em português no WSCAD nos anos 2001 e de 2003 a 2021[21].

## 6. Conclusão

Uma base adequada de afirmações qualificadas é fundamental para o desenvolvimento científico e neste contexto, o uso da Estatística deve ser específico e eficaz, fornecendo subsídios para a parametrização e reprodutibilidade de todos os tipos de experimentos. Limitações, ou falhas no rigor estatístico podem comprometer o alcance e a validade dos resultados de uma pesquisa.

Neste trabalho foi apresentado o resultado das pesquisa sobre os "termos estatísticos", "métricas" e "testes estatísticos". Em complemento ao trabalho de [Osorio 2018], foram analisadas 70 publicações adicionais, completas e em português, submetidas nas edições de 2018 a 2021 do WSCAD.

Nas Tabelas 10 e 11, observa-se que os termos mais utilizados são os empregados predominantemente para analise exploratória: "Termos" e "Métricas". Foram poucas as observações de palavras relativas à "Testes Estatísticos", apontando para uma deficiência na aplicação de técnicas de inferência Estatística nas Ciências da Computação, sobretudo nas publicações do evento analisado.

Entre os anos de 2000 a 2017 foram feitas em média 55,5 menções a termos relativos à estatística, com um desvio padrão de 21,52 (Tabelas 4, 6 e 8). Nos anos seguintes, de 2018 a 2021, houve um incremento de 30,18%, com uma media de 72,25 menções por ano e com desvio padrão de 20,29 pontos, porém, esta diferença cai para 21,70% quando se computa o ano de 2018 às tabelas de [Osorio 2018], conforme Tabela 12.

Os anos onde as publicações atingiram a maior contagem de aplicação de termos estatísticos foram 2004, 2005, 2018 e 2019 com 91, 94, 81 e 95 menções respectivamente,

[17] https://docs.python.org/3/library/unicodedata.html#unicodedata.normalize

[18] https://pypi.org/project/tika/

[19] https://tika.apache.org/

[20] Como os resultados dessa análise não puderam ser compilados em tempo hábil, não foram incluídos neste artigo. Porém o código desenvolvido e os dados coletados encontram-se disponíveis no repositório deste trabalho para verificação pelos leitores.

[21] 2000 e 2002 não encontravam-se disponíveis em https://sol.sbc.org.br/index.php/wscad/issue/archive

bem como, os piores anos de acordo com o mesmo critério foram 2000, 2011 e 2017 com 24, 16 e 35 palavras, nessa ordem.

Quanto a haver uma possível relação de causa-efeito entre a publicação de [Osorio 2018] e o aumento ou mudança no uso de Termos da Estatística, observa-se que antes de 2018, para o grupo "termos", as palavras chave: "Num. de Iterações" e "Intervalo de Confiança" eram as mais utilizadas, a partir deste ano, os termos mais utilizados são Média e Teste/Experimento. Para o grupo "métricas", observa-se que, independente da janela de tempo, as palavras mais recorrentes são "Eficiência/Acurácia", "Confiabilidade" e "Speed-Up". Para o grupo "Testes" devido ao baixo índice de uso, não são possíveis considerações de usualidade.

Comparando-se as tabelas 4 e 5, pode-se supor não haver mudança de comportamento atrelada à publicação de [Osorio 2018], visto que, no ano de 2005 ocorreu o maior "n" de 58 ocorrências, seguido pelos anos de 2004 e 2019 ambos com 53 termos. A partir do ano de 2018, verifica-se uma troca do emprego maioritário de "IC" e "NI" por "ME" e "TE", mesmo que o valor de "n" não tenha sofrido alterações, conforme Figura 2 onde as linhas referentes a "Estatística" e "Métricas" se alternam. A suposição de não haver mudança é reforçada pelo padrão gráfico sem tendências apresentado na Figura 3 e negado pelo exposto na Tabela 12 onde ha um aumento de 21,70% apos a publicação do trabalho de [Osorio 2018].

Apesar das divergências encontradas, entre os métodos de extração de dados, elas não invalidam as conclusões desse trabalho ou do trabalho de [Osorio 2018], uma vez que a proporção de citações de termos referentes a testes estatísticos, e consequentemente sua utilização, frente as citações de métricas e outros termos estatísticos permaneceu baixa, indicando que os resultados apresentados não são validados por meio de testes estatísticos.

As suposições elaboradas acimas e descritas neste trabalho como hipóteses ($H_0$ e $H_1$) foram testadas através da ANOVA que aponta para uma não existência de causa e efeito entre a publicação de [Osorio 2018] e um possível aumento no emprego de Termos Estatísticos nas publicações do WSCAD.

Como melhoria metodológica para trabalhos futuros, sugere-se que, primeiro seja elaborada uma lista inicial de palavras e que à estas palavras seja aplicada a técnica de correlação de $\phi k$, para que palavras com correlação forte sejam agrupadas e tabuladas no mesmo índice. Quando correlacionadas as palavras de busca do presente trabalho, pode-se afirmar que há uma intersecção de ocorrência entre as palavras "Confiabilidade" e "Média", bem como, entre "Distribuição Normal" e "Teste T", que em novos trabalhos, ou para fins de refinamento em continuação a este, podem ser agrupadas.

Por fim, conclui-se que não houve qualquer relação de causa-efeito entre o comportamento das publicações do WSCAD e a publicação de [Osorio 2018], visto os scores da ANOVA dos três grupos de palavras. Tampouco se pode estabelecer um padrão previsível de comportamento com o decorrer do tempo, conforme exposto pela Figura 3 onde se visualiza que após o ano de 2018 não houve nenhuma mudança nos valores absolutos de termos relacionados à estatística, os quais podem ser verificados se comparando o valor de "n"= 19 do ano de 2017 (antes da publicação) e "n" = 12 em 2021 (depois da publicação), presentes nas Tabelas 10 e 11 respectivamente.

Possíveis trabalhos futuros incluem o refinamento das expressões de busca, para

que gerem resultados de melhor qualidade, e também a exploração da ferramenta desenvolvida em 5.3.

## Referências


Adler, S., Schmitt, S., Wolter, K., e Kyas, M. (2015). A survey of experimental evaluation in indoor localization research.

Andujar, C., Schiaffonati, V., Schreiber, F. A., Tanca, L., Tedre, M., van Hee, K., e van Leeuwen, J. (2012). The role and relevance of experimentation in informatics.

Baak, M., Koopman, R., Snoek, H., e Klous, S. (2019). A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics. *https://doi.org/10.48550/arXiv.1811.11440*.

Brent, R. P. (1973). *Algorithms for Minimization without Derivatives*. Prentice-Hall by Dover Publications, isbn-10: 0-486-41998-3 edition.

Fenniak, M. (2022). Extract Text from a PDF: Why Text Extraction is hard - PyPDF2 documentation. `https://pypdf2.readthedocs.io/en/latest/user/extract-text.html#why-text-extraction-is-hard`.

Lakatos, E. M. e Marconi, M. d. A. (2003). *Fundamentos de metodologia científica*. Atlas.

Osorio, A. (2018). Meta-análise de artigos científicos segundo critérios estatísticos: Um estudo de caso no wscad. *WSCAD 2018 - XIX Simpósio de Sistemas Computacionais de Alto Desempenho*, 19.

Shinyama, Y., Guglielmetti, P., e Marsman, P. (2022). Converting a PDF file to text - pdfminer.six documentation. `https://pdfminersix.readthedocs.io/en/latest/topic/converting_pdf_to_text.html`.

Tedre, M. e Moisseinen, N. (2014). Experiments in computing: A survey. *TheScientificWorldJournal*, 2014:549398.

The Python Software Foundation, N. (2022). unicodedata — Unicode Database: module documentation. `https://docs.python.org/3/library/unicodedata.html#unicodedata.normalize`.

The Unicode Consortium, C. (2021a). The unicode standard. `https://www.unicode.org/versions/latest/`.

The Unicode Consortium, C. (2021b). Unicode Standard Annex 15: "Unicode Normalization Forms". `https://www.unicode.org/reports/tr15/`.

Tichy, W., Lukowicz, P., Prechelt, L., e Heinz, E. (1995). Experimental evaluation in computer science. *Journal of Systems and Software - JSS*.

Wainer, J. (2007). Métodos de pesquisa quantitativa e qualitativa para a ciência da computação. *Instituto de Computação – UNICAMP*.

## Apêndice A - Método de Brent

O método de analise numérica desenvolvido por [Brent 1973] que combina o método da bissecção, o método da secante de interpolação linear e a interpolação quadrática inversa, pode ser descrito segundo o algoritmo abaixo[22]:

Inicializar: $\delta > 0$ # tolerância para parada

Inicializar: $a, b$ tal que $f(a)f(b) < 0$

se necessário $a$ e $b$ podem ser trocados,$|f(b)| \leq |f(a)|$,

logo $b$ é considerado a melhor solução aproximada.

Inicializar: $c = a$

Para cada iteração: manter $a, b, c$ tal que $b \neq c$ e:

$f(b)f(c) < 0$

$|f(b)| \leq |f(c)|$

$a$ deve ser diferente de $b$ e $c$,

se $a = c$, $a$ é o valor anterior de $b$

1) Se $|b - c| \leq \delta$:

retorna: $b$ como solução aproximada

2) Se não: determinar um ponto de teste $\hat{b}$ para:

se $a = c$ , então: $\hat{b}$ é determinado por: $\hat{b} = \frac{af(b)-bf(a)}{f(b)-f(a)}$ , interpolação linear[23].

se não: $a, b, c$ são diferentes e $\hat{b}$ é definido usando interpolação quadrática inversa:

determinando $\alpha, \beta, \gamma$ de modo que $p(y) = \alpha y^2 + \beta y + \gamma$ satisfaçam

$p(f(a)) = a, p(f(b)) = b$ e $p(f(c)) = c$

então: $\hat{b} = \gamma$

3) Se necessário, $\hat{b}$ pode ser ajustado ou substituído pelo ponto de bissecção.

4) Uma vez finalizado o $\hat{b}$, são usados $a, b, c, \hat{b}$ para determinar novos valores para $a, b, c$ .

[22] Adaptado de [Brent 1973], capítulo 5.

[23] Método de interpolação da Secante.

## Apêndice B - Pré-Processamento para Algoritmo de Cálculo do Índice de Correlações $\phi k$

Primeiro se define a distribuição normal bivariada dada pela Equação 4, com parâmetro de correlação $\rho$ e unidades de distancia, centralizado em torno do ponto de partida, com a variação de [-5,5] para ambas as variáveis.

Usando alocação uniforme para os dois intervalos de variáveis, com $r$ linhas e $k$ colunas.

$$f_{b.n.}(x, y|\bar{x}, \bar{y}, \sigma_x, \sigma_y, \rho) = \left( -\frac{1}{2(1-\rho^2)} \left[ \frac{(x-\bar{x}^2)}{\sigma_x^2} + \frac{(y-\bar{y}^2)}{\sigma_y^2} - \frac{2\rho(x-\bar{x})((y-\bar{y})}{\sigma_x \sigma_y} \right] \right) \quad (4)$$

As frequências observadas $O_{ij}$, são definidas como a probabilidade por intervalo multiplicado pelo número de registros $N$.

As frequências esperadas $E_{ij}$, são definidas como predições da distribuição normal bivariada, com $\rho = 0$ e com $N$ registros de mesmo intervalo.

Então se avalia o $\chi^2$ conforme Equação 5[24]

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (5)$$

Definição da função:

Primeiro se calcula a integral da distribuição normal bivariada sobre a área do intervalo $ij$ dada pela Equação 6, que leva à soma dos intervalos expressa na Equação 7.

$$F_{ij}(\rho) = \int_{area_{ij}} f_{b.n.}(x, y|\rho) dx dy \quad (6)$$

$$\chi^2_{b.n.}(\rho, N, r, k) = N \sum_{i,j}^{k,r} \frac{(F_{ij}(\rho = \rho) - F_{ij}(\rho = 0))^2}{F_{ij}(\rho = 0)} \quad (7)$$

O valor do $\chi^2$ da Equação 7, ignora as flutuações estatísticas nas frequências observadas, e está em função do número de linhas e colunas, $N$, e do valor de $\rho$.

Para levar em consideração o ruido estatístico, introduz-se um suporte relacionado a uma estimativa simples do número de graus de liberdade da amostra bivariada, $n_{sdof}$[25]

$$n_{sdof} = (r-1)(k-1) - n_{vazio}(esperado) \quad (8)$$

Com o numero de linhas $r$ e colunas $k$, considerando que $n_{vazio}(esperado)$ é o número de intervalos vazios da frequência dependente estimada para a amostra, o suporte é definido na Equação 9.

[24] Equação 5 = Equação 2, reescrita para facilitar a leitura.

[25] sdof, leia-se: graus de liberdade da amostra, do inglês: sample degrees of freedom.

$$\chi^2_{sup} = n_{sdof} + C.\sqrt{2n_{sdof}} \tag{9}$$

O ruido do suporte é passível de ajuste através do parâmetro $c$. Por padrão, assume-se $c = 0$. O maior valor possível para $\chi^2$, no teste de contingência é dado pela Equação 10 e depende apenas do numero de registros $N$, de linhas $r$ e colunas $k$. O $\chi^2_{max}$ é atingido quando ha uma dependência de 1 para 1 entre duas variáveis, logo, é dependente do formato da distribuição $p(x, y)$.

$$\chi^2_{max}(N, r, k) = N\ min(r - 1, k - 1) \tag{10}$$

É possível escalar a Equação 7 na Equação 11, para garantir que seja igual ao $\chi^2_{sup}$ para $\rho = 0$ e $\chi^2_{max}$ para $\rho = 1$.

$$\chi^2_{b.n.}(\rho, N, r, k) = \chi^2_{sup} + \left\{ \frac{\chi^2_{max}(N, r, k) - \chi^2_{sup}}{\chi^2_{b.n.}(1, N, r, k} \right\}.\chi^2_{b.n.}(\rho, N, r, k). \tag{11}$$

Esta função é simétrica em $\rho$, e aumenta monotonicamente de $\chi^2_{sup}$ para $\chi^2_{max}$ à medida que $\rho$ vai de zero a um.