

Instituto Nacional de Metrologia, Qualidade e Tecnologia – INMETRO

Luiz Henrique da Conceição Leal

**APLICATIVO WEB PARA AVALIAÇÃO DE DESEMPENHO
DE LABORATÓRIOS**

Duque de Caxias – RJ

2022

Luiz Henrique da Conceição Leal

**APLICATIVO WEB PARA AVALIAÇÃO DE DESEMPENHO DE
LABORATÓRIOS**

Tese apresentada ao Programa de Pós-Graduação em Metrologia do Instituto Nacional de Metrologia, Qualidade e Tecnologia como parte dos requisitos para a obtenção do título de Doutor em Metrologia.

Werickson Fortunato de Carvalho Rocha

Orientador

Duque de Caxias – RJ

2022

Luiz Henrique da Conceição Leal

**WEB APPLICATION FOR LABORATORY PERFORMANCE
ASSESSMENT**

Doctoral thesis submitted as partial fulfilment of the requirements for the Degree of Doctor of Science in the Postgraduate Program in Metrology of the National Institute of Metrology, Quality, and Technology.

Werickson Fortunato de Carvalho Rocha

Advisor

Duque de Caxias – RJ

2022

L435a Leal, Luiz Henrique da Conceição.

Aplicativo web para avaliação de desempenho de laboratórios / Luiz Henrique da Conceição Leal. Duque de Caixas, RJ, 2022.
147 f. : il., color.

Tese (Doutorado) – Instituto Nacional de Metrologia, Qualidade e Tecnologia, Programa de Pós-Graduação em Metrologia, 2022.

Orientador: Werickson Fortunato de Carvalho Rocha.

1. Aplicativo web 2. Comparação interlaboratorial 3. Ensaio de proficiência 4. Tipo de resultado 5. Teste de proporção I. Rocha, Werickson Fortunato de Carvalho II. Instituto Nacional de Metrologia, Qualidade e Tecnologia III. Título.

CDD 006.76

Ficha catalográfica elaborada pela Biblioteca do Inmetro

REFERÊNCIA BIBLIOGRÁFICA

LEAL, Luiz Henrique da Conceição. **Aplicativo web para avaliação de desempenho de laboratórios**. 2022. 146f. Tese (Doutorado em Metrologia) – Instituto Nacional de Metrologia, Qualidade e Tecnologia, Duque de Caxias, RJ, 2022.

CESSÃO DE DIREITOS

NOME DO AUTOR: Luiz Henrique da Conceição Leal

TÍTULO DA MONOGRAFIA: Aplicativo web para avaliação de desempenho de laboratórios.


TIPO DE MONOGRAFIA: Tese de Doutorado em Metrologia / 2022.

É concedida ao Instituto Nacional de Metrologia, Qualidade e Tecnologia a permissão para reproduzir e emprestar cópias desta monografia somente para propósitos acadêmicos e científicos. O autor reserva outros direitos de publicação.


Luiz Henrique da Conceição Leal

Aplicativo web para avaliação de desempenho de laboratórios

A presente Tese, apresentada ao Programa de Pós-Graduação em Metrologia do Instituto Nacional de Metrologia, Qualidade e Tecnologia como parte dos requisitos para obtenção do título de Doutor em Metrologia, foi aprovada pela seguinte Banca Examinadora:

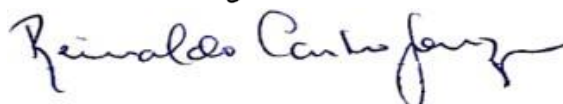
Documento assinado digitalmente
 WERICKSON FORTUNATO DE CARVALHO ROCHA
Data: 08/07/2022 08:28:15-0300
Verifique em <https://verificador.iti.br>

Doutor Werickson Fortunato de Carvalho Rocha– Inmetro
Presidente da Banca Examinadora


Documento assinado digitalmente
 VANDERLEA DE SOUZA
Data: 08/07/2022 11:58:45-0300
Verifique em <https://verificador.iti.br>

Doutora Vanderléa de Souza – Inmetro


Doutor Elcio Cruz de Oliveira – PUC RIO



Doutor Reinaldo Castro Souza – PUC RIO

Documento assinado digitalmente
 CLEBER NOGUEIRA BORGES
Data: 13/07/2022 12:10:58-0300
Verifique em <https://verificador.iti.br>

Doutor Cleber Nogueira Borges – UFLA



Doutor Igor Campos de Almeida Lima – UERJ

Duque de Caxias, 06 de julho de 2022

AGRADECIMENTOS

Primeiramente agradeço a Deus. O meu aprendizado dentro e fora do ambiente acadêmico foi pautado pela sabedoria de Deus que colocou os obstáculos necessários para o meu crescimento pessoal, profissional e espiritual.

À minha esposa Patricia e ao meu filho André pelo apoio e compreensão nos momentos difíceis sempre com muito amor e carinho.

Ao meu orientador Werickson pela confiança, pelo conhecimento compartilhado, pelo apoio durante todas as etapas deste projeto e por tudo que aprendi. Seu suporte foi fundamental para a execução e finalização dessa tese. Serei eternamente grato.

À professora Vanderlea e ao professor Jailton (PPGM/Inmetro) que forneceram dicas valiosas a cerca da produção textual desta tese nos seminários de acompanhamento de projeto.

Aos professores Elcio Cruz de Oliveira (PUC-Rio) e Vinicius Layter Xavier (UERJ) pelas contribuições dadas no exame de qualificação as quais permitiram enriquecer a pesquisa realizada nesta tese.

A todos os excelentes profissionais do Inmetro que, direta e indiretamente, contribuíram para a conclusão do trabalho. Em especial, a: Fernando Gustavo Marques Violante, Lucas Junqueira de Carvalho, Bruno Carius Garrido, Janaína Marques Rodrigues, Eliane Cristina Pires do Rego pelos dados referentes ao ensaio de proficiência para verificar a competência dos participantes no ensaio de quantificação do ácido benzóico em suco de laranja 2ª rodada; a Marcelo Lima Alves pelos dados referentes ao ensaio de proficiência de emissões de automóveis diesel 10ª rodada. Os dados obtidos foram fundamentais para o desenvolvimento desta tese.

À Mancin Marzia (Senior Statistician presso Istituto Zooprofilattico Sperimentale delle Venezie) pelos dados referentes ao artigo “Proposed statistical analysis to evaluate qualitative proficiency testing of Salmonella serotyping”. Sou grato pela gentileza e prestatividade no envio dos dados os quais contribuíram para a análise categórica presente nesta tese.

Ao David Sheen (Physicist, NIST) pelos dados multivariados que foram de grande valia para o desenvolvimento deste trabalho.

Ao meu colega estatístico Gabriel Fonseca Sarmanho (Inmetro) cujas dicas de programação em R foram essenciais para o transcurso desta tese.

Por fim, agradeço a todos que, de algum modo, contribuíram para que este trabalho fosse concluído.

RESUMO

Ensaio de proficiência e comparações interlaboratoriais são ferramentas metrológicas para identificação de diferenças interlaboratoriais que visam avaliar os resultados de diferentes laboratórios, realizados em condições semelhantes, e obter uma avaliação da competência técnica para demonstrar a confiabilidade de seus processos de medição. Os laboratórios participantes, por sua vez, têm a oportunidade de rever seus procedimentos de análise e implementar melhorias em seus processos, se necessário. De modo geral, os ensaios de proficiência, assim como as comparações interlaboratoriais, são mecanismos de busca por medições discrepantes entre os resultados reportados pelos participantes. Os ensaios de proficiência e as comparações interlaboratoriais podem ser segmentados segundo o tipo de resultado reportado: univariado, categórico e multivariado. No caso univariado, a norma ISO 13528:2015 fornece a descrição de métodos estatísticos que propiciam a identificação de diferenças interlaboratoriais. Estes métodos não consideram as diferenças entre as variabilidades dos participantes e a “não homogeneidade” destas variabilidades pode levar a conclusões equivocadas sobre o desempenho dos laboratórios. Neste contexto, sugeriu-se uma nova abordagem de tratamento de dados baseada em análise de variância que permitesse considerar em sua metodologia a “não homogeneidade” da variabilidade entre os participantes. No caso categórico (qualitativo), o resultado reportado pelo laboratório não é uma medição em si, mas um resultado qualitativo. A norma ISO 13528:2015 sugere métricas de avaliação de desempenho baseadas na quantidade de resultados corretos reportados pelos laboratórios. Na presente tese, foram sugeridos testes de proporções como ferramentas adicionais de análise. No contexto multivariado, o participante reporta resultados de múltiplas variáveis. A norma não menciona sobre métodos estatísticos para análise de múltiplas variáveis. Foi sugerido o emprego da técnica de escalonamento multidimensional, para agrupar os participantes, seguida da elipse/elipsóide de confiança robusta para definir quais participantes apresentavam resultados atípicos em relação aos demais. Por fim, desenvolveu-se uma aplicação web em Shiny/R contendo métodos estatísticos que permitem ao provedor do ensaio de proficiência avaliar os resultados reportados pelos laboratórios participantes.

Palavras-chave: aplicativo web; categórico; comparação interlaboratorial; ensaio de proficiência; multivariado; univariado.

ABSTRACT

Proficiency testing and interlaboratory comparisons are metrological tools for identifying interlaboratory differences that aim to evaluate the results of different laboratories, carried out under similar conditions, and to obtain an assessment of technical competence to demonstrate the reliability of their measurement processes. Participating laboratories, in turn, have opportunity to review their analysis procedures and implement improvements in their processes, if necessary. In general, proficiency tests as well as interlaboratory comparisons are outlier detection tools in participants' results. Proficiency tests and interlaboratory comparisons can be splitted according to type of reported result: univariate, categorical and multivariate. In the univariate case, ISO 13528:2015 provides descriptions of statistical methods that allow identifying interlaboratory differences. These methods do not consider the differences between the participants' variability and inhomogeneity of these variabilities may lead to mistaken conclusions about the performance of the laboratories. In this case, a new approach based on analysis of variance was suggested that would allow considering in its methodology the inhomogeneity of the variability among the participants. In the categorical case (qualitative), the result reported by the laboratory is not a measurement, but a qualitative result. The ISO 13528:2015 suggests performance evaluation metrics based on the number of correct results reported by the laboratories. In the present thesis, tests of proportions were suggested as additional analysis tools. In the multivariate context, the participant reports results from multiple variables. The standard does not mention about statistical methods for analyzing multiple variables. The use of the multidimensional scaling technique was suggested to group the participants, followed by the robust confidence ellipse/ellipsoid to define which participants had atypical results in relation to the others. Finally, a web application was developed in Shiny/R containing statistical methods that allow the proficiency testing provider to evaluate the results reported by the participating laboratories.

Keywords: categorical; interlaboratory comparison; multivariate; proficiency testing; univariate; web application.

LISTA DE ILUSTRAÇÕES

Figura 1 – Métodos estatísticos para avaliação de desempenho contidos no aplicativo web. .	21
Figura 2 – Fluxograma para análise de resíduos de um EP/CI univariado.	39
Figura 3 – Fluxograma para análise de dados categóricos.	66
Figura 4 – Estrutura de um mapa auto-organizável.	74
Figura 5 – Topologias.	75
Figura 6 – Aplicativo: <i>Home</i>	83
Figura 7 – Aplicativo: <i>Data sets</i>	85
Figura 8 – Aplicativo: <i>Univariate</i>	86
Figura 9 – Aplicativo: <i>Categorical</i>	87
Figura 10 – Aplicativo: <i>Multivariate</i>	87
Figura 11 – Aplicativo: <i>About</i>	88
Figura 12 – Gráfico dos resíduos do modelo linear normal.	94
Figura 13 – Gráfico dos resíduos do modelo FGLS.	94
Figura 14 – Gráfico dos resíduos do modelo linear normal.	100
Figura 15 – Gráfico dos resíduos do modelo FGLS.	100
Figura 16 – Parte dos resultados reportados por cada participante: EP de sorotipos de salmonela (dados reais).	105
Figura 17 – Intervalo de confiança para o percentual de acertos: EP de sorotipos de salmonela (dados reais).	107
Figura 18 – Quantidade de itens de ensaio enviados aos participantes de um ensaio de proficiência categórico.	112
Figura 19 – Espectro RMN ^1H de cada laboratório participante.	115
Figura 20 – Análise de componentes principais.	116
Figura 21 – Mapas auto-organizáveis de Kohonen	116
Figura 22 – Escalonamento multidimensional e elipse de confiança robusta (MDS-RCE 2D).	117

Figura 23 – Escalonamento multidimensional e elipsóide de confiança robusta (MDS-RCE 3D).	119
---	-----

LISTA DE TABELAS

Tabela 1 – Análise de variância com um fator: conjunto de dados.	42
Tabela 2 – Tabela de análise de variância com um fator.	42
Tabela 3 – Teste de Kruskal-Wallis: conjunto de dados.	44
Tabela 4 – Interpretação do coeficiente de Gower.	55
Tabela 5 – Interpretação do coeficiente kappa de Cohen.	56
Tabela 6 – Coeficiente kappa de Cohen: tabela de frequências.	57
Tabela 7 – Coeficiente kappa de Cohen: tabela de proporções.	57
Tabela 8 – Coeficiente alfa de Krippendorff: tabela de concordância (<i>rik</i>).	58
Tabela 9 – Coeficiente de Gwet: tabela de concordância (<i>rik</i>).	60
Tabela 10 – Índice de Leti: resultados reportados.	62
Tabela 11 – Interpretação do índice de Leti.	62
Tabela 12 – Interpretação do coeficiente de alfa de Krippendorff.	64
Tabela 13 – Interpretação do coeficiente de gama de Gwet.	65
Tabela 14 – Teste Qui-quadrado: tabela de contingência (frequências).	67
Tabela 15 – Teste Qui-quadrado: tabela de contingência (proporções).	68
Tabela 16 – Mapas auto-organizáveis de Kohonen: funções de vizinhança.	77
Tabela 17 – Resultados (em mg/L) dos laboratórios participantes do ensaio de proficiência: ácido benzóico em suco de laranja.	93
Tabela 18 – Resultados (em mg/L) do laboratório de referência: EP de ácido benzóico em suco de laranja.	93
Tabela 19 – <i>Scores</i> e p-valores do modelo FGLS: EP de ácido benzóico em suco de laranja.	95
Tabela 20 – Erro tipo II para as estatísticas de desempenho: EP de ácido benzóico em suco de laranja.	96
Tabela 21 – Comparação entre os resultados obtidos pelos <i>scores</i> e pela metodologia proposta: EP de ácido benzóico em suco de laranja.	97

Tabela 22 – Resultados (em g/km) dos laboratórios participantes do ensaio de proficiência: emissões de NOx em automóvel diesel.....	99
Tabela 23 – <i>Scores</i> : EP de emissões de NOx em automóvel diesel.	101
Tabela 24 – Erro tipo II para as estatísticas de desempenho: EP de emissões de NOx em automóvel diesel.	101
Tabela 25 – Comparações múltiplas duas a duas do laboratório 71 com os demais: EP de emissões de NOx em automóvel diesel.....	102
Tabela 26 – Comparações múltiplas duas a duas do laboratório 163 com os demais: EP de emissões de NOx em automóvel diesel.....	102
Tabela 27 – Comparações múltiplas duas a duas do laboratório 86 com os demais: EP de emissões de NOx em automóvel diesel.....	103
Tabela 28 – Comparação entre os resultados obtidos pelos <i>scores</i> e pela metodologia proposta: EP de emissões de NOx em automóvel diesel.....	104
Tabela 29 – Resultados reportados pelos participantes: EP de sorotipos de salmonela (dados reais).....	106
Tabela 30 – Medidas de concordância para cada um dos participantes: EP de sorotipos de salmonela (dados reais).	108
Tabela 31 – Teste de Cohen (comparações múltiplas): EP de sorotipos de salmonela (dados simulados).	109
Tabela 32 – Interpretação dos resultados obtidos pelas medidas de concordância e pelo teste de Cohen: EP de sorotipos de salmonela (dados simulados).	110
Tabela 33 – Classificação dos resultados pelos métodos de escalonamento multidimensional - elipse de confiança robusta (MDS-RCE 2D) e zscore multivariado.	118
Tabela 34 – Classificação dos resultados pelos métodos de escalonamento multidimensional - elisóide de confiança robusta (MDS-RCE 3D) e zscore multivariado.	120
Tabela 35 – Desvio-padrão robusto.	137
Tabela 36 – Gráficos de controle: base de dados.	143
Tabela 37 – Gráficos de controle: Limites de controle.	144

NOMENCLATURA

As definições a seguir foram extraídas das ISO 13528:2015, ISO-IEC 17043:2010, ISO Guide 30:2015 e ISO Guide 35:2017. A definição de homogeneidade estatística foi extraída de Kutner *et al* (KUTNER *et al*, 2005).

Ensaio de proficiência (EP): Avaliação do desempenho do participante frente a critérios pré-estabelecidos por meio de comparações interlaboratoriais.

Comparação interlaboratorial (CI): Organização, desempenho e avaliação de medições ou ensaios nos mesmos ou em itens similares por dois ou mais laboratórios, de acordo com as condições pré-determinadas.

Valor designado: Valor atribuído a uma propriedade específica de um item de ensaio de proficiência.

Valor de consenso: Valor derivado de uma coleção de resultados em uma comparação interlaboratorial.

- O valor de consenso pode ser obtido a partir de “*experts*” (laboratórios de “referência”) que demonstram competência na determinação do(s) mensurando(s) sob análise, usando métodos validados e conhecidos por serem altamente precisos e comparáveis aos métodos de uso geral.
- O valor de consenso pode ser obtido por meio dos participantes utilizando-se métodos estatísticos descritos na norma ISO 13528:2015 e na IUPAC 2006 levando-se em consideração os efeitos dos *outliers*.

Item de ensaio de proficiência: Amostra, produto, artefato, material de referência, equipamento, padrão, conjunto de dados ou outra informação utilizada pelo ensaio de proficiência. O provedor do EP deve garantir que os itens de ensaio sejam suficientemente homogêneos e estáveis.

Homogeneidade: Na metrologia, de acordo com o ISO Guia 30, homogeneidade refere-se à uniformidade de um valor de propriedade especificado por meio de uma porção definida de um material de referência. Na estatística, homogeneidade refere-se ao pressuposto de que o erro aleatório do modelo estatístico possui variância constante σ^2 .

Estabilidade: refere-se a característica de um material de referência, quando armazenado sob condições especificadas, de manter o valor de uma determinada propriedade dentro de limites especificados por um período de tempo especificado.

Caracterização: determinação dos valores de propriedade ou atributos de um material de referência, conforme parte do processo de produção.

Material de referência: material suficientemente homogêneo e estável com respeito a uma ou mais propriedades especificadas, que foi estabelecido como sendo adequado para o seu pretendido em um processo de medição.

Valor de propriedade: Valor correspondente a uma grandeza representando uma propriedade física, química ou biológica de um material de referência.

SUMÁRIO

SUMÁRIO	15
1. INTRODUÇÃO	18
1.1 OBJETIVO.....	22
1.1.1 Objetivos gerais	22
1.1.2 Objetivos específicos.....	22
1.2 JUSTIFICATIVA	23
2. ENSAIO DE PROFICIÊNCIA UNIVARIADO	26
2.1 FUNDAMENTAÇÃO TEÓRICA	26
2.1.1 Notação.....	27
2.1.2 Estatísticas de desempenho.....	28
2.1.3 Estimadores robustos.....	29
2.1.4 Curva de Horwitz	34
2.1.5 Estatísticas de desempenho alternativas	34
2.2 ABORDAGEM PROPOSTA.....	36
2.2.1 Métodos paramétricos	41
2.2.2 Métodos não paramétricos	44
2.2.3 Mínimos quadrados generalizados	46
2.2.4 P-valor ajustado	48
2.2.5 Análise de resíduos	49
2.2.6 Eficiência dos <i>scores</i>	51
3. ENSAIO DE PROFICIÊNCIA COM DADOS CATEGÓRICOS	53
3.1 FUNDAMENTAÇÃO TEÓRICA	53
3.1.1 Resultados reportados em escala nominal	53
3.1.2 Resultados reportados em escala ordinal.....	61

3.2 ABORDAGEM PROPOSTA.....	65
3.2.1 Grande número de amostras	66
3.2.2 Pequeno número de amostras	68
4. ENSAIO DE PROFICIÊNCIA MULTIVARIADO	70
4.1 FUNDAMENTAÇÃO TEÓRICA	70
4.1.1 Análise de componentes principais	70
4.1.2 Mapas auto-organizáveis de Kohonen	73
4.1.3 Zscore multivariado.....	78
4.2 ABORDAGEM PROPOSTA.....	79
4.2.1 Escalonamento multidimensional	80
4.2.2 Elipse de confiança robusta	80
4.2.3 Elipsóide de confiança robusta	81
5. APLICATIVO WEB	82
5.1 INTRODUÇÃO.....	82
5.2 APLICATIVO	83
5.3 MÓDULOS	84
5.4 PACOTES	88
6. RESULTADOS	91
6.1 ENSAIO DE PROFICIÊNCIA UNIVARIADO.....	91
6.1.1 Ensaio de Proficiência em Sucos (2ª Rodada): Ácido Benzoico em Suco de Laranja	91
6.1.2 Ensaio de Proficiência de Emissões de Automóveis (10ª Rodada): Automóvel Diesel	98
6.2 ENSAIO DE PROFICIÊNCIA CATEGÓRICO.....	104
6.2.1 Ensaio de Proficiência de Sorotipos de Salmonela (dados reais)	105
6.2.2 Ensaio de Proficiência de Sorotipos de Salmonela (dados simulados)	108

6.2.3 Tamanho amostral para dados categóricos	110
6.3 ENSAIO DE PROFICIÊNCIA MULTIVARIADO.....	114
6.3.1 Comparação interlaboratorial de RMN em peixe	114
7. CONCLUSÃO	121
REFERÊNCIAS BIBLIOGRÁFICAS.....	124
APÊNDICE	134
A. DESVIO PADRÃO ROBUSTO	134
B. ESTIMATIVA ROBUSTA MULTIVARIADA	138
C. HOMOGENEIDADE/ESTABILIDADE	140
C.1 Homogeneidade.....	140
C.2 Estabilidade	141

1. INTRODUÇÃO

Proficiência é a qualidade do que é proficiente; competência, capacidade, mestria. É o domínio num determinado campo; capacidade, habilitação. É a consecução de bons resultados; aproveitamento, proficuidade.

Ensaio de proficiência (EP) é o uso de comparações interlaboratoriais (CI) com o objetivo de avaliar a habilidade de um laboratório em realizar um determinado ensaio ou medição de modo competente e demonstrar a confiabilidade dos resultados gerados, ou seja, é uma ferramenta para a determinação do desempenho de laboratórios na execução de ensaios ou calibrações por meio de uma comparação interlaboratorial [1].

Comparações interlaboratoriais referem-se à organização, realização e avaliação de medições ou ensaios nos mesmos ou em itens similares por dois ou mais laboratórios, de acordo com as condições predeterminadas [2-5]. Basicamente os ensaios de proficiência diferem das comparações interlaboratoriais no que se refere ao uso dos métodos de medição. Nos ensaios de proficiência, os participantes podem utilizar diferentes métodos de medição [4] ao passo que nas comparações interlaboratoriais todos os participantes devem utilizar os mesmos métodos de medição. Além disso, nas comparações interlaboratoriais avalia-se o método de medição, ou seja, o desempenho do método ao passo que nos ensaios de proficiência avalia-se o laboratório (desempenho do laboratório).

Os ensaios de proficiência (e as comparações interlaboratoriais) possibilitam aos laboratórios participantes: (i) evidenciar a obtenção de resultados confiáveis; (ii) monitorar continuamente seus processos; (iii) identificar erros sistemáticos nos ensaios; (iv) tomar ações corretivas e/ou preventivas quando identificados problemas sistemáticos dos ensaios; (v) reavaliar controles internos; (vi) especificar características de desempenho, qualidade e validação de métodos e tecnologias empregadas pelo participante; (vii) padronizar os procedimentos de ensaio entre os participantes o que permite o reconhecimento dos resultados de medição em nível nacional e internacional. Isso faz com que a participação em um EP seja considerada pré-requisito para a acreditação [6, 7].

De acordo com o site do INMETRO, acreditação é o reconhecimento formal da competência do laboratório em atender requisitos previamente definidos e realizar suas

atividades com confiança. O desempenho em atividades de ensaio de proficiência pode ser levado em consideração no processo de acreditação de laboratórios. Acreditação é importante para gerar confiança nos serviços ofertados pelos laboratórios. É importante destacar que os requisitos para acreditação variam de acordo com o escopo de atuação do laboratório. A Coordenação Geral de Acreditação do INMETRO fornece documentos orientativos (DOQ-CGCRE) que auxiliam os laboratórios na implementação dos requisitos de acreditação [2].

Segundo o vocabulário internacional de metrologia (VIM, 2012) [8], a metrologia é a ciência das medições e suas aplicações. É a ciência que engloba todos os aspectos teóricos e práticos da medição, qualquer que seja a incerteza de medição e o campo de aplicação [8]. Em outras palavras, a metrologia é a ciência das medições por meio da calibração de instrumentos e da realização de ensaios cuja finalidade é prover confiabilidade, credibilidade, universalidade e qualidade aos resultados das medições.

Ensaio de proficiência é uma atividade metrológica que contribui para o aumento da credibilidade dos resultados das medições dos laboratórios o que facilita o comércio internacional e previne barreiras técnicas. Aprimorar os procedimentos de avaliação de desempenho descritos na norma ISO 13528:2015 (a qual encontra-se em revisão) pode propiciar aos laboratórios aperfeiçoar os serviços prestados à indústria. Medições confiáveis agregam valor e melhoram a qualidade dos produtos, o que contribuiu para tornar a indústria brasileira mais competitiva no comércio internacional.

A análise de dados aplicada aos ensaios de proficiência será segmentada, na presente tese, em 3 categorias: (i) análise de dados univariados; (ii) análise de dados categóricos; (iii) análise de dados multivariados. No que se refere aos ensaios de proficiência com dados univariados, a norma ISO 13528:2015 fornece a descrição de métodos estatísticos para avaliar o desempenho dos laboratórios participantes. Estes métodos são denominados de estatística de desempenho ou *scores*. Os procedimentos descritos na ISO 13528:2015 podem ser aplicados para demonstrar que os resultados de medição obtidos pelos laboratórios, órgãos de inspeção e indivíduos atendem aos critérios especificados para um desempenho aceitável.

Devido a possibilidade dos participantes de um EP poderem utilizar diferentes métodos de medição, a dispersão dos resultados reportados por cada laboratório pode

variar consideravelmente (heterocedasticidade) e isso pode conduzir a conclusões equivocadas sobre o desempenho dos participantes mesmo utilizando os métodos estatísticos descritos na norma. É recomendável que uma eventual presença de heterocedasticidade seja levada em consideração na análise dos resultados reportados pelos participantes. A presente tese propõe métodos estatísticos baseados em análise de variância para avaliar o desempenho dos laboratórios participantes. Esta técnica possibilita considerar a heterocedasticidade em sua análise.

Dados categóricos (qualitativos) são aqueles que assumem um número limitado e fixo de valores ou categorias associando a cada unidade observacional uma categoria baseada em alguma característica qualitativa [9-11]. Os dados categóricos são geralmente expressos por meio de frequências, proporções e/ou tabelas de contingência [9, 12].

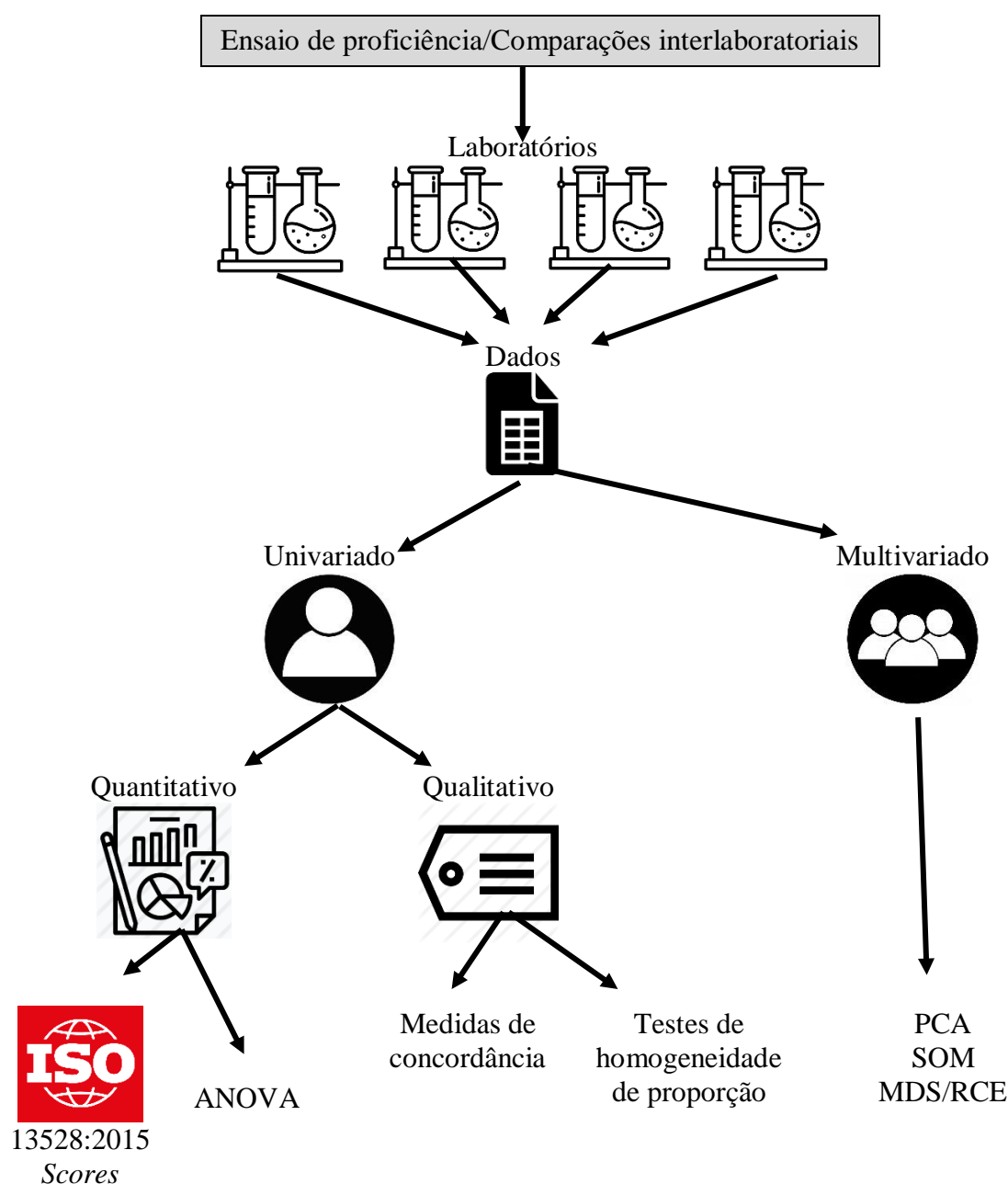
Há ensaios de proficiência em que o resultado reportado pelo participante não é um valor numérico, mas uma classificação do item de teste e, nestes casos, deseja-se avaliar o desempenho dos participantes com base em algum índice de correspondência entre o resultado reportado e o valor designado (classificação esperada). Este contexto refere-se aos ensaios de proficiência com dados categóricos. A norma ISO 13528:2015 menciona, mas não estabelece critérios, sobre métodos estatísticos de avaliação de desempenho para este tipo de EP. Na presente tese são sugeridos procedimentos de análise de dados categóricos baseados em testes de proporção como ferramentas para avaliar o desempenho dos laboratórios participantes.

A análise multivariada, de forma bem geral, refere-se a todos os métodos estatísticos que analisam simultaneamente múltiplas variáveis em cada indivíduo ou objeto sob investigação. Qualquer análise simultânea de duas ou mais variáveis pode ser, de certo modo, considerada como análise multivariada. A norma ISO 13528:2015 descreve procedimentos de avaliação de desempenho quando os laboratórios participantes reportam resultados referentes a uma única variável. Nesta tese é proposto o emprego conjugado de duas técnicas multivariadas como métrica de avaliação de desempenho.

Por fim, a análise estatística de dados pode demandar um tempo considerável de análise do pesquisador/técnico metrologista quando são utilizadas plataformas convencionais de análise. O desenvolvimento de programas amigáveis de análise de dados pode auxiliar o especialista em metrologia no desenvolvimento da sua atividade

tanto em um ensaio de proficiência quanto numa comparação interlaboratorial. Na presente tese é apresentada uma aplicação web desenvolvida para este fim. A figura 1 apresenta um desenho esquemático a cerca dos métodos estatísticos que estão contidos no aplicativo web.

Figura 1 – Métodos estatísticos para avaliação de desempenho contidos no aplicativo web.



Fonte: elaboração própria.

A figura 1 contém os métodos estatísticos propostos na presente tese para avaliar o desempenho dos laboratórios por meio da análise de variância (ANOVA), dos testes de homogeneidade de proporção e do método de escalonamento

multidimensional combinado com elipse/elipsóide de confiança robusta (MDS/RCE). Os *Scores* (Figura 1) referem-se às estatísticas de desempenho. Por fim, PCA refere-se à análise de componentes principais e SOM refere-se à mapas auto-organizáveis de Kohonen (Figura 1).

Cabe mencionar que o desenvolvimento de uma aplicação web está em consonância com o processo de transformação digital do governo federal brasileiro que visa disponibilizar soluções digitais nas atividades finalísticas de suas autarquias (Art 5º da Lei 14.129, de 29 de março de 2021).

1.1 OBJETIVO

1.1.1 Objetivos gerais

De acordo com norma ABNT NBR ISO/IEC 17025:2017 um laboratório deve monitorar a validade de seus resultados, sempre que possível, por meio da participação em ensaios de proficiência e comparações interlaboratoriais. A garantia da validade dos resultados busca assegurar que os sistemas de medição do laboratório atendam à requisitos predefinidos de acordo com a necessidade específica de cada produto e/ou processo [5]. O propósito desta tese é disponibilizar um software e propor métodos estatísticos que possam aprimorar o processo de avaliação de desempenho dos laboratórios que participam de ensaios de proficiência e comparações interlaboratoriais e, por conseguinte, contribuir para a garantia da validade dos resultados obtidos pelos participantes.

O presente trabalho tem os seguintes objetivos: (a) desenvolver e disponibilizar uma aplicação web em Shiny/R que contenha métodos estatísticos de avaliação de desempenho de laboratórios participantes de EP/CI; (b) sugerir métodos estatísticos para avaliar o desempenho de laboratórios participantes de ensaios de proficiência e comparações interlaboratoriais.

1.1.2 Objetivos específicos

No que tange os ensaios de proficiência univariados, tem-se como objetivo propor um novo enfoque na avaliação de desempenho dos participantes baseado na análise de variância paramétrica e não paramétrica considerando, por meio da análise

de resíduos, pressupostos de normalidade e homocedasticidade (variabilidade dos resíduos constante).

No que concerne ao ensaio de proficiência com dados categóricos, o objetivo é analisar o desempenho dos participantes a partir da proporção de resultados reportados que estejam em consonância com o resultado esperado. Neste caso, propõe-se testes globais de igualdade de proporções seguidos, quando necessário, de comparações múltiplas duas a duas para identificar em quais pares de laboratórios há diferença estatisticamente significativa entre as proporções.

Por fim, no contexto multivariado, o objetivo específico consiste em propor o emprego da técnica de escalonamento multidimensional para avaliar a similaridade/dissimilaridade entre as medições realizadas pelos participantes. Em conjunto com este método, sugere-se o emprego da elipse/elipsóide de confiança robusta para identificar quais laboratórios apresentam resultados discrepantes em relação dos demais.

1.2 JUSTIFICATIVA

Ao se trabalhar com ensaios de proficiência que contenham dados univariados, cabe destacar os *scores* (denominados estatísticas de desempenho) descritos na norma ISO 13528:2015. Nestas métricas de avaliação de desempenho, a saber, *zscore*, *z'score*, *zeta score* e E_n score, vale ressaltar que as três primeiras têm como pressuposto a normalidade dos resultados reportados pelos participantes. Cabe ainda observar que o *zscore* é tanto um método de padronização de dados, quando estes provêm de diferentes magnitudes, quanto um método de detecção de *outliers* (nos casos em que os dados provêm de uma distribuição normal) [13]. Adicionalmente, existe a possibilidade de haver diferenças significativas na variabilidade dos dados reportados pelos participantes (heterocedasticidade) devido ao fato dos laboratórios poderem reportar seus resultados a partir de diferentes métodos de medição. Embora a norma ISO 13528:2015 mencione que as estatísticas de desempenho *zscore*, *z'score* e *zeta score* têm pressupostos de normalidade, não há menção sobre o pressuposto de homocedasticidade nem sobre seus eventuais impactos na avaliação de desempenho.

O pressuposto de normalidade dos resultados reportados é, na prática, pouco comum de se observar e a heterocedasticidade pode dificultar a análise dos dados e

eventualmente conduzir o provedor do EP/CI a conclusões equivocadas no que se refere ao desempenho dos laboratórios. É recomendável buscar ferramentas que permitam contornar essas limitações. Análise de resíduos auxilia na definição do modelo mais adequado para avaliar o desempenho dos laboratórios participantes de um ensaio de proficiência. Essa metodologia fornece uma abordagem complementar que possibilita lidar com as limitações supracitadas e oferece uma ferramenta que auxilia na avaliação da competência técnica dos participantes.

No que se refere à análise de dados categóricos, há ensaios de proficiência em que se deseja avaliar a capacidade do participante em classificar corretamente um determinado item de ensaio como, por exemplo, identificar corretamente o sorotipo bacteriano recebido pelo provedor do ensaio. Cabe destacar que sorotipo se refere a um grupo de micro-organismos que pertencem a uma mesma espécie microbiana.

No ensaio de proficiência com dados categóricos, o laboratório reporta as classificações de cada item de ensaio e avalia-se a taxa de acerto de cada participante. No contexto de análise de dados categóricos, a norma ISO 13528:2015 sugere métricas baseadas na quantidade de resultados corretos. Essas medidas podem fornecer um indicador de concordância entre o valor reportado pelo participante e o resultado, entretanto não fornecem elementos para identificar se algum laboratório, eventualmente, apresenta resultados estatisticamente diferente dos demais participantes. Além disso, a norma não estabelece critérios para classificar o desempenho dos participantes quando os resultados reportados estão em escala categórica.

O fato da norma não estabelecer critérios objetivos de avaliação de desempenho, no contexto de análise de dados categóricos, permite sugerir abordagens para esta finalidade. Métodos não paramétricos baseados em testes de proporção podem fornecer informações acerca do desempenho dos participantes. Testes globais de homogeneidade de proporção seguidos, quando o caso, de comparações múltiplas fornecem uma ferramenta para avaliar o desempenho dos participantes.

Por fim, há ensaios de proficiência em que o resultado reportado pelo participante é um espectro de uma amostra fornecida pelo provedor do ensaio. Cabe destacar que espectroscopia é uma técnica bastante utilizada para análises da estrutura química, grupos funcionais, ou composição de uma amostra. No contexto da análise

estatística de dados, o espectro reportado pelo laboratório participante do ensaio de proficiência pode ser considerado como um resultado multivariado.

Métodos estatísticos de análise multivariada podem fornecer ferramentas que auxiliem na interpretação deste tipo de resultado. A norma ISO 13528:2015 não faz menção sobre métodos estatísticos que permitam a avaliação de desempenho quando os resultados reportados pelos participantes são provenientes de múltiplas variáveis como no caso da espectroscopia. Este contexto abre a possibilidade para sugerir técnicas estatísticas multivariadas que possam auxiliar na avaliação de desempenho de resultados provenientes de múltiplas variáveis.

2. ENSAIO DE PROFICIÊNCIA UNIVARIADO

2.1 FUNDAMENTAÇÃO TEÓRICA

ISO (*International Organization for Standardization*) é uma entidade internacional independente e não-governamental responsável pelo desenvolvimento e disseminação de normas internacionais voluntárias de padronização para produtos, processos, procedimentos e serviços. No Brasil, a ISO é representada pela Associação Brasileira de Normas Técnicas (ABNT).

Pode-se dizer que o objetivo da ISO é aprovar, promover e desenvolver normas técnicas, testes e certificações que permitam a comparabilidade dos resultados e que facilitem o comércio internacional de bens e serviços reduzindo barreiras técnicas. Dentre as normas da entidade, cabe destacar as séries: (i) ISO 9000, sistemas de qualidade nas organizações; (ii) ISO 14000, gestão ambiental; (iii) ISO 22000, segurança alimentar.

No âmbito da metrologia científica, vale destacar o ISO Guide 35:2017 o qual fornece orientação sobre questões técnicas, explica os conceitos e apresenta métodos estatísticos para avaliação de homogeneidade, estabilidade e caracterização para a certificação de materiais de referência [14]. Essa norma é um guia aplicável a produção de material de referência. A norma ISO 13528:2015 estabelece que o provedor do ensaio de proficiência deve garantir que os itens de ensaio enviados aos participantes sejam suficientemente homogêneos e estáveis. O ISO Guide 35:2017 fornece elementos para esta finalidade.

Além do guia supracitado, a norma ISO 13528:2015 descreve métodos estatísticos para avaliação de desempenho de laboratórios participantes de ensaios de proficiência. A norma estabelece critérios para que provedores de ensaios de proficiência possam avaliar os resultados reportados pelos laboratórios participantes e fornece recomendações sobre a interpretação destes resultados. Os procedimentos da norma podem ser aplicados para demonstrar que os resultados de medição atendem a critérios específicos de desempenho considerado aceitável. Desempenho é a capacidade de um laboratório produzir resultados fidedignos. A norma elenca quatro estatísticas de desempenho (*performance statistics*) para auxiliar na interpretação e avaliação dos resultados dos laboratórios: zscore, z'score, zeta score e E_n score.

2.1.1 Notação

No que se refere as estatísticas de desempenho, a presente tese utiliza a notação adotada na norma ISO 13528:2015 conforme a seguir:

x_i : resultado do i -ésimo laboratório participante de um ensaio de proficiência (média dos valores reportados)

$u(x_i)$: incerteza-padrão combinada do i -ésimo laboratório participante

$U(x_i)$: incerteza expandida do i -ésimo laboratório participante

x_{pt} : valor designado (valor de consenso)

$u(x_{pt})$: incerteza-padrão combinada do valor designado

$U(x_{pt})$: incerteza expandida do valor designado

σ_{pt} : desvio-padrão do ensaio de proficiência

De acordo com o vocabulário internacional de metrologia (VIM, 2012) [8], tem-se as seguintes definições: (i) incerteza-padrão é a incerteza de medição expressa na forma de um desvio-padrão; (ii) incerteza-padrão combinada é a incerteza-padrão obtida ao se utilizarem incertezas-padrão individuais associadas às grandezas de entrada em um modelo de medição; (iii) incerteza expandida é o produto de uma incerteza-padrão combinada por um fator de abrangência.

Deve-se observar que o fator de abrangência, geralmente simbolizado por k , depende do tipo de distribuição de probabilidade da grandeza de saída e da probabilidade de abrangência escolhida [8].

Segundo a norma ISO 13528:2015, incerteza-padrão combinada do valor designado é obtida por meio da seguinte equação

$$u(x_{pt}) = \sqrt{u_{char}^2 + u_{hom}^2 + u_{trans}^2 + u_{stab}^2} \quad \text{Equação 1}$$

Em que u_{char} denota a incerteza-padrão devido à etapa de caracterização, u_{hom} denota a incerteza-padrão obtida por meio do estudo de homogeneidade, u_{trans} denota a incerteza-padrão devido à instabilidade causada pelo transporte dos itens de teste e u_{stab} denota a incerteza-padrão obtida por meio do estudo de estabilidade [4].

Por fim, cabe destacar que a incerteza-padrão combinada do i -ésimo laboratório participante $u(x_i)$ pode ser obtida por meio de dois tipos de fontes de incerteza: (i) incerteza tipo A cuja componente da incerteza de medição é obtida por meio de uma análise estatística dos valores medidos, obtidos sob condições definidas de medição; (ii) incerteza tipo B cuja componente da incerteza de medição é obtida por meios diferentes daquele adotado na incerteza do tipo A [8].

2.1.2 Estatísticas de desempenho

Os ensaios de proficiência (EP) consistem em uma comparação interlaboratorial para avaliar a capacidade dos laboratórios em realizar um determinado ensaio ou medição. Os provedores do EP distribuem partes de um material homogêneo para cada um dos participantes, que o analisam sob condições específicas pré-determinadas e relatam os resultados para o provedor do ensaio o qual compila os resultados e informa aos participantes as conclusões obtidas, geralmente na forma de uma pontuação relativa à precisão do resultado (estatística de desempenho) [15].

O zscore representa uma medida da distância do resultado apresentado por um específico laboratório em relação ao valor designado do ensaio de proficiência e, portanto, serve para verificar se o resultado da medição de cada participante está em conformidade com o valor designado. O zscore é calculado por

$$z_i = \frac{x_i - x_{pt}}{\sigma_{pt}} \quad \text{Equação 2}$$

De acordo com a norma ISO 13528:2015, o zscore é interpretado do seguinte modo: se $|z_i| \leq 2$ o resultado do laboratório é considerado satisfatório, se $2 < |z_i| < 3$ o resultado do laboratório é considerado questionável e se $|z_i| \geq 3$ o resultado do laboratório é considerado insatisfatório [4].

O z'score é utilizado nos casos em que a incerteza do valor designado é considerada significativa. Também conhecida como zscore modificado esta estatística apresenta os mesmos valores críticos do zscore e pode ser obtida a partir da equação

$$z'_i = \frac{x_i - x_{pt}}{\sqrt{\sigma_{pt}^2 + u^2(x_{pt})}} \quad \text{Equação 3}$$

A interpretação deste parâmetro análoga ao zscore: se $|z'_i| \leq 2$ o resultado do laboratório é considerado satisfatório, se $2 < |z'_i| < 3$ o resultado do laboratório é considerado questionável e se $|z'_i| \geq 3$ o resultado do laboratório é considerado insatisfatório [4].

O zeta score é utilizado quando a incerteza do laboratório e a incerteza do valor designado são fornecidas. Esta estatística apresenta os mesmos valores críticos do zscore é obtida a partir da equação

$$\zeta = \frac{x_i - x_{pt}}{\sqrt{u^2(x_i) + u^2(x_{pt})}} \quad \text{Equação 4}$$

Se $|\zeta| \leq 2$ o resultado do laboratório é considerado satisfatório, se $2 < |\zeta| < 3$ o resultado do laboratório é considerado questionável e se $|\zeta| \geq 3$ o resultado do laboratório é considerado insatisfatório [4].

O E_n score serve para verificar se o resultado da medição de cada participante está em conformidade com o valor designado considerando os resultados das medições e suas respectivas incertezas. Esta estatística é calculada por

$$E_n = \frac{x_i - x_{pt}}{\sqrt{U^2(x_i) + U^2(x_{pt})}} \quad \text{Equação 5}$$

O E_n score é interpretado do seguinte modo: Se $|E_n| < 1$ o resultado do laboratório é considerado satisfatório, caso contrário se $|E_n| \geq 1$ o resultado do laboratório é considerado insatisfatório [4].

2.1.3 Estimadores robustos

Há situações em que valor designado e o desvio-padrão do EP não estão disponíveis e, neste caso, faz-se necessário estimá-los a partir dos resultados reportados pelos participantes. Neste contexto, diz-se que as estimativas foram obtidas a partir de valores de consenso entre os participantes.

Eventualmente, os resultados fornecidos pelos participantes podem conter valores atípicos os quais podem impactar nas estimativas do valor designado (valor de

consenso) e do desvio-padrão do EP. As técnicas de estatística robusta são utilizadas para minimizar a influência de resultados extremos.

2.1.3.1 M-estimador de Huber

O M-estimador de Huber (Algoritmo A, H15 ou Huber proposal 2) fornece uma medida de posição e uma medida de dispersão, ambas robustas, estimadas a partir do seguinte processo iterativo [4]:

Passo 1. Calculam-se as estimativas iniciais

$$\hat{\mu} = \text{mediana}(\mathbf{x}) \quad \text{Equação 6}$$

$$\hat{\sigma} = k \cdot \text{MAD}(\mathbf{x}) \quad \text{Equação 7}$$

Em que o valor de k depende da distribuição do conjunto de dados (apêndice A da presente tese) e

$$\text{MAD}(\mathbf{x}) = \text{mediana}(|x_i - \text{mediana}(\mathbf{x})|) \quad \text{Equação 8}$$

$$\mathbf{x} = (x_1, x_2, \dots, x_n) \quad \text{Equação 9}$$

Passo 2. Estabelecem-se novos valores x_i^* :

$$x_i^* = \begin{cases} \hat{\mu} - c\hat{\sigma} & , \quad \text{se } x_i < \hat{\mu} - c\hat{\sigma} \\ x_i & , \quad \text{se } \hat{\mu} - c\hat{\sigma} \leq x_i \leq \hat{\mu} + c\hat{\sigma} \\ \hat{\mu} + c\hat{\sigma} & , \quad \text{se } x_i > \hat{\mu} + c\hat{\sigma} \end{cases} \quad \text{Equação 10}$$

Em que $1 \leq c \leq 2$ [16, 17]. O valor usual de $c = 1,5$ [4, 17, 18] é adequado para 5 % a 10 % de valores discrepantes no conjunto de dados [17]. Este estimador foi desenvolvido sob o pressuposto de que a proporção de valores atípicos não ultrapasse 10 % das observações [17].

Passo 3. Obtêm-se as novas estimativas de $\hat{\mu}$ e $\hat{\sigma}$ a partir dos novos valores x_i^* :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i^* \quad \text{Equação 11}$$

$$\hat{\sigma} = \sqrt{\frac{1}{\beta} \cdot \frac{\sum_{i=1}^n (x_i^* - \hat{\mu})^2}{n-1}} \quad \text{Equação 12}$$

Em que

$$\beta = \theta + c^2(1 - \theta) - 2c \cdot \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{c^2}{2}\right\} \quad \text{Equação 13}$$

Sendo $\theta = P(|Z| < c)$ e $Z \sim N(0; 1)$. As novas estimativas de $\hat{\mu}$ e $\hat{\sigma}$ são utilizadas para obter novos valores no passo 2. Este procedimento é repetido até que os valores de $\hat{\mu}$ e $\hat{\sigma}$ converjam [4].

2.1.3.2 M-estimador bponderado de Tukey

O M-estimador bponderado de Tukey (*Tukey's biweight* ou *bisquare M-estimator*) fornece uma medida de posição robusta estimada a partir do seguinte processo iterativo:

Passo 1. Calcula-se a estimativa inicial $\hat{\mu}_0 = \text{mediana}(\mathbf{x})$ [19].

Passo 2. Obtêm-se a estimativa ponderada

$$\hat{\mu}_{t+1} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \quad \text{Equação 14}$$

Em que

$$w_i = h\left(\frac{x_i - \hat{\mu}_t}{\hat{\sigma}}\right) \quad \text{Equação 15}$$

$$h(u) = \begin{cases} \left(1 - \frac{u^2}{c^2}\right)^2, & \text{se } |u| \leq c \\ 0, & \text{se } |u| > c \end{cases} \quad \text{Equação 16}$$

Sendo que $c = 4,685$ [20, 21] e $\hat{\sigma} = k \cdot \text{MAD}(\mathbf{x})$ [17, 19] em que o valor de k depende da distribuição do conjunto de dados (apêndice A da presente tese) e

$$\mathbf{x} = (x_1, x_2, \dots, x_n) \quad \text{Equação 17}$$

Passo 3. A nova estimativa $\hat{\mu}_{t+1}$ é utilizada para o cálculo dos pesos w_i no passo 2. Este procedimento é repetido até que o valor de $\hat{\mu}_{t+1}$ convirja ($|\hat{\mu}_{t+1} - \hat{\mu}_t| < \varepsilon$).

2.1.3.3 M-estimador de Hampel

O M-estimador Hampel (*Hampel's three-part M-estimator*) fornece uma medida de posição robusta estimada a partir do seguinte processo iterativo:

Passo 1. Calcula-se a estimativa inicial $\hat{\mu}_0 = \text{mediana}(\mathbf{x})$ [4, 22].

Passo 2. Obtêm-se a estimativa ponderada

$$\hat{\mu}_{t+1} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \quad \text{Equação 18}$$

Em que

$$w_i = h\left(\frac{x_i - \hat{\mu}_t}{\hat{\sigma}}\right) \quad \text{Equação 19}$$

$$h(u) = \begin{cases} 1 & , \quad |u| \leq a \\ a/u & , \quad a < |u| \leq b \\ a(c-u)/[u(c-b)] & , \quad b < |u| \leq c \\ 0 & , \quad |u| > c \end{cases} \quad \text{Equação 20}$$

Sendo que $\hat{\sigma} = k \cdot \text{MAD}(\mathbf{x})$ [4, 22] em que o valor de k depende da distribuição do conjunto de dados (apêndice A da presente tese) e

$$\mathbf{x} = (x_1, x_2, \dots, x_n) \quad \text{Equação 21}$$

Por fim, a , b e c são parâmetros de ajuste (*tuning parameters*) [4]. Segundo Thode (2002) valores adequados para estes parâmetros são $1,5 \leq a \leq 2,5$, $2 \leq b \leq 4$ e $5 \leq c \leq 15$ [17] sendo ainda recomendado que $c - b \geq 2a$ [17, 21].

Passo 3. A nova estimativa $\hat{\mu}_{t+1}$ é utilizada para o cálculo dos pesos w_i no passo 2. Este procedimento é repetido até que o valor de $\hat{\mu}_{t+1}$ convirja ($|\hat{\mu}_{t+1} - \hat{\mu}_t| < \varepsilon$) [4].

2.1.3.4 Algoritmo S

No planejamento de experimentos com um único fator há p grupos (níveis), cada um com n_i observações ($i = 1, \dots, p$) em que y_{ij} representa a j -ésima observação do i -ésimo grupo e w_i é a i -ésima medida de dispersão. O algoritmo S é utilizado para obter uma medida de variabilidade robusta em que a medida de dispersão w_i pode ser o desvio-padrão ou a amplitude dos dados. Após definir a medida de variabilidade a ser empregada a estimativa robusta é obtida a partir do seguinte processo iterativo:

Passo 1. Calcula-se a estimativa inicial

$$w^* = \text{mediana}(w_i). \quad \text{Equação 22}$$

Passo 2. Calcula-se

$$\Psi = \eta \cdot w^* \quad \text{Equação 23}$$

Em que

$$\eta = \sqrt{\frac{q}{v}} \quad \text{Equação 24}$$

$$P\{\chi_v^2 \leq q\} = 1 - \alpha \quad \text{Equação 25}$$

O valor v é o grau de liberdade associado a w_i e obtido do seguinte modo:

$$v = \begin{cases} 1 & , \text{se } w_i \text{ é a amplitude} \\ n - 1 & , \text{se } w_i \text{ é o desvio padrão} \end{cases} \quad \text{Equação 26}$$

Passo 3. Estabelecem-se novos valores

$$w_i^* = \begin{cases} \Psi & , \text{se } w_i > \Psi \\ w_i & , \text{caso contrário} \end{cases} \quad \text{Equação 27}$$

Passo 4. Obtém-se a nova estimativa w^* a partir dos novos valores w_i^* :

$$w^* = \xi \cdot \sqrt{\frac{\sum_{i=1}^p (w_i^*)^2}{p}} \quad \text{Equação 28}$$

Em que

$$\xi = \frac{1}{\sqrt{Z + \alpha \cdot \eta^2}} \quad \text{Equação 29}$$

$$Z = P\{\chi_{v+2}^2 \leq v \cdot \eta^2\} \quad \text{Equação 30}$$

A nova estimativa w^* é utilizada para obter o novo valor no passo 2. Este processo é repetido até que o valor de w^* convirja [4].

2.1.4 Curva de Horwitz

A curva de Horwitz é um modelo geral para estimar o desvio-padrão do ensaio de proficiência por meio da concentração de um determinado analito (substância ou componente químico, em uma amostra, que é alvo de análise). A função proposta por Horwitz [4, 23] foi modificada posteriormente por Thompson [4, 23] que passou a considerar os diferentes níveis de concentração do analito. A curva é obtida conforme equação a seguir:

$$\sigma_{pt} = \begin{cases} 0,22c & , se \ c < 1,2 \cdot 10^{-7} \\ 0,02c^{1-\frac{\log 2}{2}} & , se \ 1,2 \cdot 10^{-7} \leq c \leq 0,138 \\ 0,01\sqrt{c} & , se \ c > 0,138 \end{cases} \quad \text{Equação 31}$$

Em que c é o nível de concentração do analito expresso em fração mássica (mg/kg) [4].

A equação de Horwitz é aplicável somente a métodos analíticos que expressem o mensurando como fração mássica não sendo aplicável a analitos empíricos, analitos indefinidos ou propriedades físicas [23].

2.1.5 Estatísticas de desempenho alternativas

Um método para avaliar desempenho dos participantes de um ensaio de proficiência é o Q-scoring o qual baseia-se no viés relativo (e não no valor padronizado como o z-score). O Q-scoring é calculado por

$$Q = \frac{x_i - x_{pt}}{x_{pt}} \quad \text{Equação 32}$$

Em que x_{pt} é o valor designado. A desvantagem deste método é que a significância dos resultados não é diretamente comparável [24, 25]. Um modo de interpretar o Q-score é definir o percentual aceitável de desvio do valor designado [26].

O erro quadrático médio relativo (QMER, sigla em inglês para *relative quadratic mean error*) é a raiz quadrada da soma do viés quadrático, em relação ao valor designado, com a incerteza-padrão do i -ésimo laboratório participante, sendo esta comparada com a incerteza expandida do valor designado, ou seja,

$$QMER = \frac{\sqrt{(x_i - x_{pt})^2 + n_i \cdot u^2(x_i)}}{U(x_{pt})} \quad \text{Equação 33}$$

Em que n_i é a quantidade de valores reportados pelo i -ésimo laboratório participante de um ensaio de proficiência. Em outras palavras, n_i é o número de replicatas. Esta medida serve para mensurar a competência analítica do laboratório. Se $QMER \leq 5$ o resultado do laboratório é considerado satisfatório caso contrário ($QMER > 5$) o resultado do laboratório é considerado “sem sucesso” (*not successful*) [27-30].

O valor designado pode ser obtido por meio do valor de consenso calculado a partir dos resultados reportados pelos laboratórios participantes do ensaio de proficiência [4]. Isso envolve estimar um parâmetro de locação, geralmente a média robusta obtida a partir do M-estimador de Huber descrito na norma ISO 13528:2015. Uma limitação dessa abordagem é que pode haver uma correlação entre o resultado individual e o valor designado quando há um pequeno número de participantes ($n \leq 15$). Essa correlação pode subestimar o desvio-padrão do ensaio. Neste contexto, é recomendável utilizar o zscore corrigido

$$z_i^c = \frac{x_i - x_{pt}}{\sigma_{pt} \sqrt{1 - 1/n}} \quad \text{Equação 34}$$

Em que n é o número de laboratórios participantes [7, 31-33].

Métodos robustos são empregados para mitigar a influência de resultados atípicos. O M-estimador de Huber, por exemplo, fornece estimativas robustas para o cálculo do z-score. Outra forma de avaliar, quando há resultados discrepantes, o

desempenho dos laboratórios participantes de um ensaio de proficiência consiste no zscore robusto o qual é calculado a partir da equação

$$z_i^R = \frac{x_i - \text{mediana}(\mathbf{x})}{NIQR(\mathbf{x})} \quad \text{Equação 35}$$

Em que $\mathbf{x} = [x_1, x_2, \dots, x_n]$ e n é o número de laboratórios participantes. Nesta métrica, o valor designado é estimado pela mediana dos resultados dos participantes e o desvio-padrão do EP é estimado a partir do intervalo interquartil normalizado (*normalized interquartile range* – NIQR), ou seja,

$$NIQR(\mathbf{x}) = \frac{IQR(\mathbf{x})}{z_{0,75} - z_{0,25}} \quad \text{Equação 36}$$

Em que $z_{0,75}$ e $z_{0,25}$ são, respectivamente, os quantis 75 % e 25 % da distribuição normal padrão [34, 35]. Por fim, $IQR(\mathbf{x}) = Q_3 - Q_1$ é o intervalo interquartil sendo Q_1 e Q_3 o primeiro e o terceiro quartil respectivamente.

2.2 ABORDAGEM PROPOSTA

Os ensaios de proficiência, assim como as comparações interlaboratoriais, são ferramentas metrológicas para identificação de diferenças interlaboratoriais. As diferenças observadas fornecem elementos para avaliação da competência técnica dos participantes e, quando necessário, revisão de seus procedimentos e implantação de melhorias em seus processos.

Propõe-se que as eventuais diferenças interlaboratoriais observadas em um ensaio de proficiência sejam avaliadas por meio de métodos estatísticos baseados em análise de variância e comparações múltiplas. Análise de variância (ANOVA) é uma técnica que possibilita comparar três ou mais grupos de interesse e investigar a existência ou não de diferenças significativas entre os grupos estudados [36]. Testes de comparações múltiplas são empregados quando a ANOVA identifica diferença entre os grupos. Estes testes permitem identificar, dois a dois, quais grupos diferem entre si.

Na análise de variância, a maneira sob a qual as hipóteses podem ser testadas dependem de como as observações foram obtidas. Nos casos em que os grupos são escolhidos especificamente para o experimento testam-se hipóteses concernentes às

médias e as conclusões aplicam-se somente aos níveis considerados. Esta é a análise de variância para modelos de efeitos fixos. Se os grupos representam uma amostra aleatória de uma grande população as hipóteses testadas referem-se à variabilidade. Neste caso, tem-se a análise de variância para modelos com efeitos aleatórios [36]. Os ensaios de proficiência (EP) consistem em um problema de análise de variância para modelos de efeitos fixos, pois os laboratórios participantes escolhidos especificamente para o EP.

Na ANOVA, há pressupostos a cerca do modelo que precisam ser levados em consideração. Nesta modelagem assume-se que os erros seguem uma distribuição normal com média 0 (zero) e variância constante σ^2 . Em outras palavras, pressupõe-se que não há diferença estatisticamente significativa entre as variabilidades dos grupos considerados no estudo [36]. O primeiro caso refere-se ao pressuposto de normalidade e o segundo de homocedasticidade dos resíduos. Estes pressupostos da análise de variância precisam ser verificados por meio da análise de resíduos proposta na figura 2.

A figura 2 apresenta um fluxograma com etapas sugeridas para realizar a análise dos resíduos e assim definir o modelo mais adequado para avaliar o desempenho de laboratórios participantes de um ensaio de proficiência univariado. Na primeira etapa, recomenda-se verificar o pressuposto de normalidade dos resíduos por meio do gráfico dos resíduos padronizados versus os valores estimados pressupondo o modelo linear normal. O gráfico fornece informações visuais sobre o padrão dos resíduos, entretanto recomenda-se utilizar testes formais para checar os pressupostos do modelo. Sugere-se o teste de Shapiro-Wilk para verificar o pressuposto de normalidade dos resíduos (Figura 2a). Optou-se por este teste para avaliar a normalidade, pois apresenta maior poder do teste (probabilidade de rejeitar a hipótese nula quando esta é realmente falsa) para um dado nível significância α quando comparado aos testes de Kolmogorov-Smirnov, Lilliefors, Cramer-von Mises, Anderson-Darling, D'Agostino-Pearson, Jarque-Bera e qui-quadrado [37, 38]. O teste de Shapiro-Wilk original é aplicável para um conjunto entre 3 e 50 observações. A implementação de Royston expande para um conjunto de até 5000 observações [38].

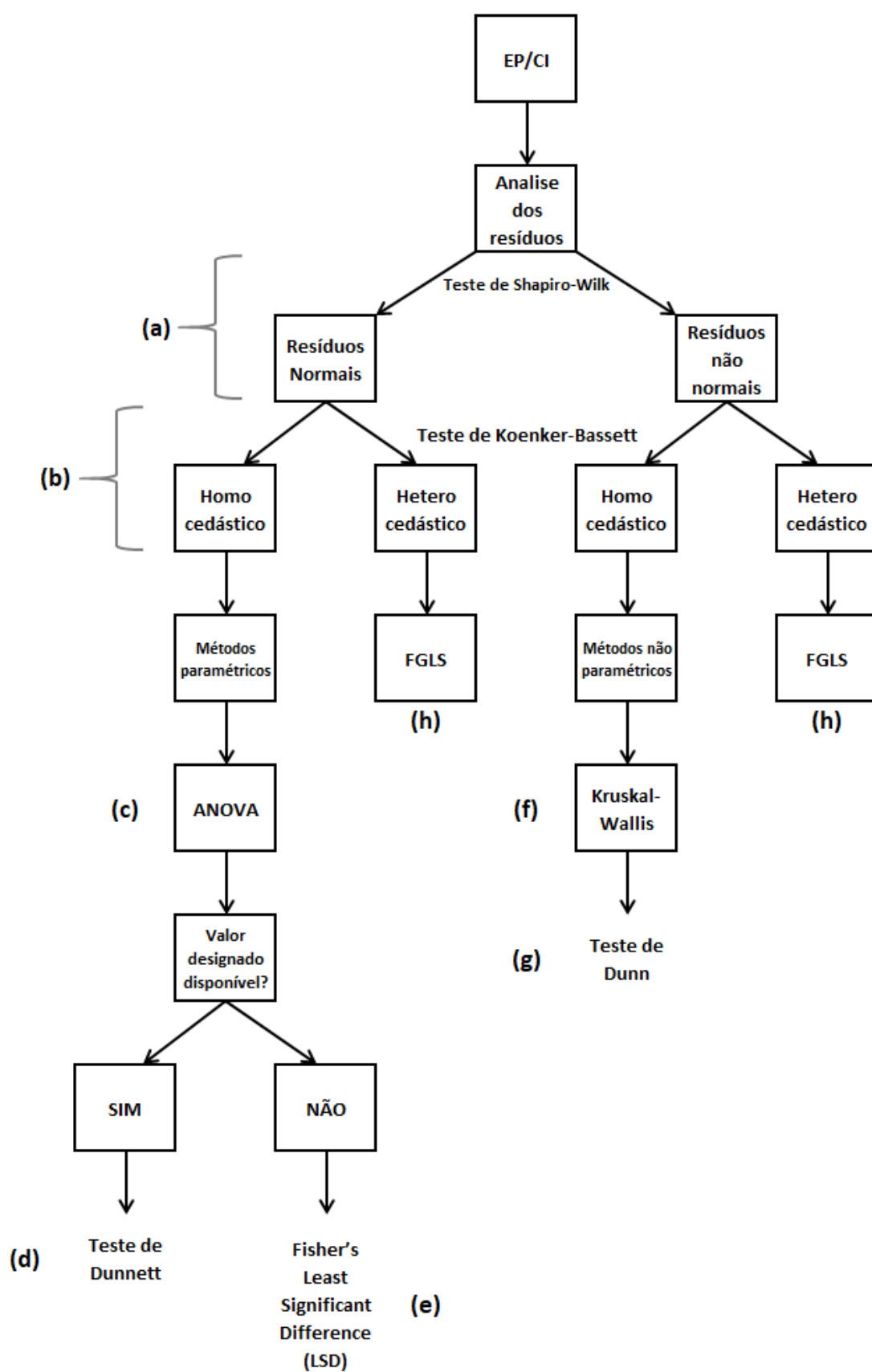
O pressuposto de homocedasticidade é a segunda etapa sugerida na análise de resíduos. Pressupõe-se que os resíduos estejam distribuídos aleatoriamente em torno de 0 (zero). Caso os resíduos apresentem um “padrão assemelhado à um funil” há

indícios da presença de heterocedasticidade. Recomenda-se aplicar o teste de Koenker-Bassett para checar o pressuposto de homocedasticidade dos resíduos (Figura 2b). Optou-se por este teste por ser aplicável mesmo quando o termo do erro do modelo original não for normalmente distribuído [39].

Cabe ainda destacar que a metodologia proposta se baseia na escolha do modelo adequado por meio da análise de resíduos. Os métodos não paramétricos de Levene e Fligner-Killeen são realizados a partir da transformação dos dados originais e testam a homogeneidade da variância entre os grupos ao passo que o teste de Koenker-Bassett é específico para avaliar homocedasticidade dos resíduos de um modelo linear normal. Além disso, no teste de Koenker-Bassett a homocedasticidade é avaliada a partir da significância estatística do coeficiente angular da regressão dos quadrados dos resíduos do modelo original contra os valores estimados do regressando elevados ao quadrado. Essa significância é obtida a partir do teste F [39] o que faz com que o método de Koenker-Bassett apresente maior poder do teste que os métodos de Levene e Fligner-Killeen.

As duas primeiras etapas da análise de resíduos auxiliam a definir o modelo mais adequado para avaliar os resultados reportados pelos participantes. Nos casos em que os resíduos são normais e homocedásticos, recomenda-se adotar a análise de variância paramétrica (Figura 2c). O teste F da ANOVA indica a existência ou não de diferença estatisticamente significativa entre as medições dos laboratórios. Se houver diferença, testes de comparações múltiplas são recomendados para identificar quais medições diferem.

Figura 2 – Fluxograma para análise de resíduos de um EP/CI univariado.



Fonte: elaboração própria.

Quando houver um valor designado disponível, propõe-se utilizar o teste de Dunnett (Figura 2d). Este método verifica se as medições de cada participante diferem do valor designado. Se não houver um valor designado disponível, sugere-se aplicar o teste de diferença mínima significativa (*Fisher's Least Significant Difference – LSD*) (Figura 2e). Nesta abordagem serão realizados testes dois a dois com todas as combinações de participantes no intuito de investigar se algum(ns) participante(s) difere(m) estatisticamente dos demais.

Pode ocorrer que os resultados dos laboratórios não apresentem resíduos normais, mas homocedásticos. Nesta situação é recomendável utilizar métodos não paramétricos. O teste de Kruskal-Wallis (Figura 2f) possibilita checar a hipótese de que as medições reportadas não diferem entre si. Caso difiram, sugere-se o teste de Dunn (Figura 2g) para identificar quais diferenças são significativas. Este teste fornece conclusões para ensaios de proficiência com ou sem o valor designado disponível.

Heterocedasticidade é um problema que pode surgir nos resultados reportados pelos laboratórios. A norma ISO 13528:2015 abre a possibilidade para que cada participante utilize seu próprio método de medições e isso pode gerar uma “não-homogeneidade” na variabilidade dos resultados. Isso pode afetar os testes F da ANOVA e de Kruskal-Wallis e conduzir a conclusões equivocadas a cerca do desempenho dos laboratórios participantes do EP nesta abordagem proposta. Se a análise dos resíduos indicar a presença de heterocedasticidade nos resíduos, o fluxograma proposto na figura 2 indica a necessidade de utilizar modelos estimados por mínimos quadrados generalizados (*Feasible Generalized Least Squares – FGLS*) (Figura 2h).

O teste F do modelo FGLS fornece conclusões a cerca da diferença ou não entre as medições dos participantes. Nos casos em que se verifica divergência entre os resultados reportados o teste t desta modelagem permite identificar quais diferenças são estatisticamente significativas. O teste t do modelo FGLS é aplicável a ensaios de proficiência com ou sem o valor designado disponível.

Nesta abordagem proposta, se o teste global (testes F da ANOVA ou modelo FGLS ou teste de Kruskal-Wallis) indicar que não há diferença estatisticamente significativa entre os grupos todos os laboratórios participantes são classificados como satisfatórios. Caso o teste global indique que há diferença, sugere-se aplicar os testes de comparações múltiplas.

Nos ensaios de proficiência em que houver um valor designado disponível e os testes de comparações múltiplas (testes de Dunnett ou Dunn ou t do modelo FGLS) indicarem que há diferença entre os valores reportados pelo participante e o valor designado este laboratório terá seu resultado classificado com insatisfatório no contexto da abordagem proposta. Se as comparações múltiplas não indicarem a existência de diferença estatisticamente significativa, o laboratório será classificado como satisfatório.

Se não houver um valor designado disponível, os testes de comparações múltiplas (testes de LSD ou Dunn ou t do modelo FGLS) serão empregados para verificar dois a dois todos os pares de laboratórios participantes. Caso os testes indiquem que os resultados de um determinado laboratório diferem de todos os demais, este participante terá seu resultado classificado como insatisfatório pela metodologia proposta. Os laboratórios que não diferirem entre si terão seus resultados classificados como satisfatórios.

P-valor é uma estatística muito utilizada para sintetizar o resultado de um teste de hipóteses (testes globais e de comparações múltiplas). Formalmente, o p-valor é definido como a probabilidade de se obter uma estatística de teste igual ou mais extrema quanto àquela observada em uma amostra, assumindo-se verdadeira a hipótese nula.

Em teste de comparações múltiplas há um potencial acúmulo de erros de decisão, ou seja, à medida que o número de testes aumenta a probabilidade de se cometer pelo menos um erro do tipo I também aumenta [40-46]. O p-valor ajustado é um procedimento recomendado para controlar o aumento desta probabilidade em testes de comparações múltiplas. Nos ensaios de proficiência, o erro do tipo I significa classificar indevidamente os resultados reportados pelo participante como insatisfatório.

2.2.1 Métodos paramétricos

2.2.1.1 Análise de variância

A análise de variância com um fator para modelos com efeitos fixos trata de experimentos em que um único fator, com p tratamentos (no contexto desta tese, laboratórios) ou níveis, é investigado sendo que, em cada um dos tratamentos ou níveis, há n_i observações ($i = 1, 2, \dots, p$) [36].

Tabela 1 – Análise de variância com um fator: conjunto de dados.

Tratamento	Observações				Totais	Médias
1	y_{11}	y_{12}	\dots	y_{1n_1}	$y_{1\bullet}$	$\bar{y}_{1\bullet}$
2	y_{21}	y_{22}	\dots	y_{2n_2}	$y_{2\bullet}$	$\bar{y}_{2\bullet}$
\vdots	\vdots	\vdots		\vdots	\vdots	\vdots
p	y_{p1}	y_{p2}	\dots	y_{pn_p}	$y_{p\bullet}$	$\bar{y}_{p\bullet}$
-	-	-		-	$y_{\bullet\bullet}$	$\bar{y}_{\bullet\bullet}$

Fonte: Montgomery, 2020 [36].

Em que y_{ij} é a ij -ésima observação, $y_{i\bullet}$ é o total do i -ésimo tratamento, $\bar{y}_{i\bullet} = y_{i\bullet}/n_i$ é a média do i -ésimo tratamento, $y_{\bullet\bullet}$ é o total geral de todas as observações, $\bar{y}_{\bullet\bullet} = y_{\bullet\bullet}/N$ é a média geral de todas as observações e $N = \sum_{i=1}^p n_i$. O tratamento estatístico destes dados é realizado por meio da tabela de análise de variância (ANOVA) [36].

Tabela 2 – Tabela de análise de variância com um fator.

Fonte de variação	Graus de liberdade	Soma de quadrados	Média quadrática	F_{obs}
Entre os tratamentos	$p - 1$	$SS_{trat} = \sum_{i=1}^p n_i (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$	$MS_{trat} = \frac{SS_{trat}}{p - 1}$	$\frac{MS_{trat}}{MS_E}$
Erro (dentro dos tratamentos)	$N - p$	$SS_E = \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\bullet})^2$	$MS_E = \frac{SS_E}{N - p}$	
Total	$N - 1$	$SS_T = \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\bullet\bullet})^2$		

Fonte: Montgomery, 2020 [36].

Os p tratamentos são escolhidos especificamente para o experimento e testam-se as hipóteses a respeito das médias desses tratamentos e as conclusões aplicam-se somente aos níveis de fatores considerados [36].

$$\begin{cases} (H_0) \mu_1 = \mu_2 = \dots = \mu_p \\ (H_1) \mu_i \neq \mu_j \text{ para pelo menos um par } (i, j) \end{cases}$$

Se $F_{obs} > F_{p-1; N-p; \alpha}$ rejeita-se a hipótese nula com um nível de confiança de $1 - \alpha$ em que $F_{p-1; N-p; \alpha}$ é o quantil da distribuição Fisher-Snedecor. Nos casos em que a hipótese nula é rejeitada, sugere-se investigar em quais pares (i, j) as médias diferem [36].

2.2.1.2 Teste de Dunnett

Na análise de variância com um fator para modelos com efeitos fixos, quando a hipótese nula, $H_0: \mu_1 = \mu_2 = \dots = \mu_p$, é rejeitada conclui-se que há diferença estatisticamente significativa entre as médias dos tratamentos, entretanto, não é possível especificar quais médias diferem. Comparações adicionais e análises entre grupos de tratamentos podem ser úteis para identificar quais pares de médias diferem, estatisticamente, entre si.

Em muitos experimentos, um dos tratamentos é um controle (nesta tese é denominado de laboratório de referência, ou seja, o laboratório do qual provém o valor designado) e o pesquisador geralmente está interessado em compará-lo com os $p - 1$ tratamentos restantes. Neste caso, utiliza-se o método de Dunnett no qual se testa a hipótese nula $H_0: \mu_i = \mu_p$ contra a hipótese alternativa $H_1: \mu_i \neq \mu_p$ em que $i = 1, \dots, p - 1$ [36]. A hipótese nula é rejeitada se

$$|\bar{y}_{i\bullet} - \bar{y}_{p\bullet}| > d_\alpha(p - 1, f) \sqrt{MS_E \left(\frac{1}{n_i} + \frac{1}{n_p} \right)} \quad \text{Equação 37}$$

Em que $d_\alpha(p - 1, f)$ é um valor tabelado (Montgomery, 2020, apêndice, tabela VI, [36]).

No método de Dunnett é importante observar que quando se comparam tratamentos com um controle, é recomendável utilizar mais observações no controle e

igual número de observações nos demais tratamentos de tal forma que $n_p/n = \sqrt{p}$ [36].

2.2.1.3 Diferença mínima significativa (LSD)

No teste LSD (*Fisher's Least Significant Difference*) investiga-se para quais pares (i, j) a média μ_i difere de μ_j . Em outras palavras, deseja-se testar a hipótese nula $H_0: \mu_i = \mu_j$ contra a hipótese alternativa $H_1: \mu_i \neq \mu_j$ em que $i \neq j$ [36]. A hipótese nula é rejeitada se

$$|\bar{y}_{i\bullet} - \bar{y}_{j\bullet}| > t_{N-p; 1-\alpha/2} \cdot \sqrt{MS_E \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \quad \text{Equação 38}$$

Optou-se por este método por apresentar maior poder do teste que os métodos de Tukey (*Tukey's Honestly Significant Difference – HSD*), Duncan e Newman–Keuls [47-49].

2.2.2 Métodos não paramétricos

2.2.2.1 Kruskal-Wallis

O teste de Kruskal-Wallis é utilizado para testar a hipótese de que k conjuntos de dados independentes foram extraídos de uma mesma população.

Tabela 3 – Teste de Kruskal-Wallis: conjunto de dados.

Tratamento	Posto	Tratamento	Posto	...	Tratamento	Posto
1	(R_{ij})	2	(R_{ij})	...	k	(R_{ij})
y_{11}	1	y_{12}	5	...	y_{1k}	3
\vdots		\vdots			\vdots	
y_{n_11}	N	y_{n_22}	2	...	y_{n_kk}	4
	R_1		R_2			R_k

Fonte: elaboração própria.

Deve-se observar que em cada um dos k tratamentos (laboratórios) há n_i elementos.

As hipóteses a serem testadas são

$$\begin{cases} (H_0) \text{ não existe diferença entre os tratamentos} \\ (H_1) \text{ existe diferença entre os tratamentos} \end{cases}$$

A estatística de teste é definida por:

$$K = \frac{\frac{12}{N(N+1)} \cdot \left(\sum_{i=1}^k \frac{R_i^2}{n_i} \right) - 3 \cdot (N-1)}{1 - \frac{\sum_{i=1}^k (d_i^3 - d_i)}{(N^3 - N)}} \quad \text{Equação 39}$$

Em que $N = n_1 + \dots + n_k$, R_i é a soma dos postos do i -ésimo tratamento, $R_i = \sum_{j=1}^{n_i} R_{ij}$, d_1 é a quantidade de valores y_{ij} empatados de menor valor, d_2 é a quantidade de valores y_{ij} empatados do segundo menor valor e assim sucessivamente. Se $K > \chi_{k-1}^2$ rejeita-se a hipótese nula [50].

Recomenda-se utilizar o teste de Kruskal-Wallis, em detrimento à ANOVA com um fator, quando não é possível supor a normalidade dos resíduos ou nos casos em que há pequenas amostras [50].

2.2.2.2 Teste de Dunn

Nos casos em que a hipótese nula de igualdade entre os k dos tratamentos é rejeitada no teste de Kruskal-Wallis, faz-se necessário investigar para quais pares de tratamentos (i, j) há diferença estatisticamente significativa (comparações múltiplas). Neste contexto, testa-se a hipótese nula H_0 : não existe diferença entre os tratamentos i e j , em que $i \neq j$ contra a hipótese alternativa H_1 : existe diferença entre os tratamentos i e j . O método de Dunn possibilita testar as hipóteses para todos os pares de tratamentos ou comparar um tratamento controle com os $k - 1$ tratamentos restantes. Neste último, a hipótese nula a ser testada é H_0 : não existe diferença entre os tratamentos 0 e j em que 0 representa o tratamento controle [51]. Tem-se a seguir os critérios de rejeição da hipótese nula:

Para EP/CI sem valor designado disponível o critério de rejeição da hipótese nula é definido pela equação 21.

$$|\bar{R}_{i\bullet} - \bar{R}_{j\bullet}| \geq z_{\alpha/k(k-1)} \cdot \sqrt{\left(\frac{N(N+1)}{12} - B\right) \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \quad \text{Equação 40}$$

Para EP/CI com valor designado disponível o critério de rejeição da hipótese nula é definido pela equação 22.

$$|\bar{R}_{0\bullet} - \bar{R}_{j\bullet}| \geq z_{\alpha/k(k-1)} \cdot \sqrt{\left(\frac{N(N+1)}{12} - B\right) \cdot \left(\frac{1}{n_0} + \frac{1}{n_j}\right)} \quad \text{Equação 41}$$

Em que

$$B = \frac{\sum_{i=1}^k (d_i^3 - d_i)}{12 \cdot (N - 1)} \quad \text{Equação 42}$$

2.2.3 Mínimos quadrados generalizados

O método dos mínimos quadrados generalizados “factíveis” (*Feasible Generalized Least Squares* – FGLS) é aplicável quando a variância dos erros não é constante (heterocedasticidade). Nesta abordagem, os parâmetros do modelo de regressão são estimados a partir da equação matricial

$$b = (X^T \hat{\Omega}^{-1} X)^{-1} X^T \hat{\Omega}^{-1} Y \quad \text{Equação 43}$$

Em que $\hat{\Omega}$ é uma matriz diagonal sendo e os resíduos e \hat{Y} os valores estimados.

$$e = Y - \hat{Y} \quad \text{Equação 44}$$

$$\hat{Y} = Xb \quad \text{Equação 45}$$

$$\hat{\Omega} = \begin{bmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{bmatrix} \quad \text{Equação 46}$$

No método FGLS obtém-se a soma de quadrados generalizada

$$SQG = (Y - Xb)^T \hat{\Omega}^{-1} (Y - Xb) \quad \text{Equação 47}$$

A partir de SQG estima-se os resíduos padronizados t_i e a matriz de variância e covariância dos parâmetros $s^2(b)$.

$$t_i = e_i / \sqrt{SQG \cdot w_i / (n - p)} \quad \text{Equação 48}$$

$$s^2(b) = (SQG / (n - p)) \cdot (X^T \hat{\Omega}^{-1} X)^{-1} \quad \text{Equação 49}$$

Em que p é o número de parâmetros [52, 53]. Por fim,

$$\sqrt{SQG / (n - p)} \quad \text{Equação 50}$$

É o erro padrão residual [53, 54].

2.2.3.1 Teste F

O teste F é calculado a partir da solução do sistema de equações

$$chol(s^2(b)) \cdot \mathbf{F} = \mathbf{b} \quad \text{Equação 51}$$

Em que \mathbf{F} é um vetor $p \times 1$, \mathbf{b} é o vetor de parâmetros e $chol(s^2(b))$ é a decomposição de Choleski da matriz de variância-covariância dos parâmetros $s^2(b)$. Na regressão linear simples tem-se que $\mathbf{F}^T = [f_0 \quad f_1]$ [53, 55].

O valor F observado para o intercepto é $F_0 = (f_0)^2$ o qual é comparado com o valor tabelado $F_{1;(n-p);\alpha}$. Rejeita-se a hipótese nula $H_0: \beta_0 = 0$ se $F_0 > F_{1;(n-p);\alpha}$ em que $F_{1;(n-p);\alpha}$ é o quantil da distribuição F de Snedecor com 1 e $(n - p)$ graus de liberdade [53, 55].

O valor F observado para o coeficiente angular é

$$F_1 = \frac{(f_1)^2}{(p - 1)} \quad \text{Equação 52}$$

O qual é comparado com o valor tabelado $F_{(p-1);(n-p);\alpha}$. Rejeita-se a hipótese nula $H_0: \beta_1 = 0$ se $F_1 > F_{(p-1);(n-p);\alpha}$ em que $F_{(p-1);(n-p);\alpha}$ é o quantil da distribuição F de Snedecor com $(p - 1)$ e $(n - p)$ graus de liberdade [53, 55].

Em ambos os testes α é o nível de significância pré-estabelecido.

2.2.3.2 Teste t

O teste t estabelece, separadamente, a significância de cada parâmetro do modelo. Em outras palavras, testa-se a hipótese nula $H_0: \beta_k = 0$ contra a hipótese alternativa $H_1: \beta_k \neq 0$ por meio da estatística de teste

$$t_{obs} = \frac{b_k}{s(b_k)} \quad \text{Equação 53}$$

Rejeita-se a hipótese nula se $|t_{obs}| > t_{n-p}$ [53, 55]. No âmbito da metodologia proposta, se o valor designado está disponível a hipótese nula testada é $H_0: x_i = x_{PT}$. Se o valor designado não estiver disponível a hipótese nula consiste em $H_0: x_i = x_j$ em que $(i; j)$ é o par de laboratórios objeto de análise.

2.2.4 P-valor ajustado

Benjamini e Hochberg (1995) sugerem que a taxa de falsa descoberta (proporção esperada de erros do tipo I entre todas as hipóteses nulas rejeitadas) pode ser apropriada para controlar a taxa de erro tipo I em muitos testes de comparação múltipla. O procedimento de ajuste de p-valor proposto por Benjamini-Yekutieli é válido para estruturas de dependência arbitrárias, em que não há restrição quanto ao seu uso, e é baseado no controle da taxa de falsa descoberta e, neste contexto, será adotado na presente tese como método de ajuste de p-valor em comparações múltiplas. Sejam $p_{(1)} \leq \dots \leq p_{(m)}$ p-valores não ajustados ordenados associados as hipóteses nulas $H_{(1)}, \dots, H_{(m)}$. Os p-valores ajustados pelo procedimento de Benjamini-Yekutieli são

$$q_{(i)} = \min\{1; \min[m \cdot c(m) \cdot p_{(i)} / i]\} \quad \text{Equação 54}$$

Em que

$$c(m) = \sum_{i=1}^m 1/i \quad \text{Equação 55}$$

Send que $i = 1, \dots, m$ [56-59].

2.2.5 Análise de resíduos

Nos experimentos com um fator para modelos com efeitos fixos, pressupõe-se que o erro aleatório é normalmente distribuído e possui variância σ^2 constante para todos os níveis do fator (homogeneidade). Se o pressuposto de homogeneidade é violado, o erro do tipo I (rejeitar a hipótese nula quando esta é verdadeira) pode ser maior que o antecipado. No que concerne à normalidade dos resíduos, um desvio significativo deste pressuposto pode afetar o teste global.

Na abordagem proposta para avaliar o desempenho dos laboratórios participantes de um ensaio de proficiência, recomenda-se verificar os pressupostos da análise de variância por meio do gráfico dos resíduos e de testes formais. Os métodos sugeridos são os testes de Shapiro-Wilk (Figura 2a) e de Koenker-Bassett (Figura 2b) para checar, respectivamente, os pressupostos de normalidade e homocedasticidade dos resíduos.

Conforme já mencionado, o método de Shapiro-Wilk foi adotado por apresentar maior poder do teste quando comparado a outros métodos [37, 38]. Optou-se pelo método de Koenker-Bassett, pois é específico para avaliar homocedasticidade de resíduos mesmo quando estes não apresentam distribuição normal [39]. Além disso, vale ressaltar que os métodos não paramétricos de homogeneidade de variância (Levene e Fligner-Killeen) são realizados a partir da transformação dos dados originais e apresentam menor poder do teste quando comparados ao teste de Koenker-Bassett.

2.2.5.1 Teste de Shapiro-Wilk

Este método detecta desvios da normalidade provenientes da assimetria, curtose ou de ambas, e baseia-se na estatística de teste W [38, 60].

$$W = \frac{(\sum_{i=1}^n a_i x_{(i)})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{Equação 56}$$

Em que

$$a_i = \frac{m_i}{\sqrt{\phi}} \quad \text{para } 2 < i < n - 1 \quad \text{Equação 57}$$

$$a_{n-1} = c_{n-1} + 0,042981y - 0,293762y^2 - 1,752461y^3 + 5,682633y^4 - 3,582663y^5$$

Equação 58

$$a_n = c_n + 0,221157y - 0,147981y^2 - 2,071190y^3 + 4,434685y^4 - 2,706056y^5$$

Equação 59

Sendo que

$$m_i = \Phi\left(\frac{i - \frac{3}{8}}{n + \frac{1}{4}}\right) \quad \text{Equação 60}$$

$$\phi = \begin{cases} \frac{m^T m - 2m_n^2}{1 - 2a_n^2} & \text{se } n \leq 5 \\ \frac{m^T m - 2m_n^2 - 2m_{n-1}^2}{1 - 2a_n^2 - 2a_{n-1}^2} & \text{se } n > 5 \end{cases} \quad \text{Equação 61}$$

$$c_i = (m^T m)^{-1/2} m_i \quad \text{Equação 62}$$

$$y = \frac{1}{\sqrt{n}} \quad \text{Equação 63}$$

Cabe destacar que: (i) $\Phi(\cdot)$ é a função de distribuição acumulada da distribuição normal padrão; (ii) o vetor $a = (a_1, \dots, a_n)$ é anti-simétrico, ou seja, $a_{(n+1-i)} = -a_{(i)} \forall i$ e $a_{(n+1)/2} = 0$ quando n é ímpar; (iii) $x_{(i)}$ é a i -ésima estatística de ordem [38, 60].

Se $W > W_{tab}$ a hipótese de normalidade é rejeitada. Os valores de W_{tab} podem ser obtidos por meio do código descrito no artigo de Royston (ROYSTON, 1995) [60]. O teste de Shapiro-Wilk é aplicável para amostras que contenham até 5000 observações ($n \leq 5000$) [38, 60].

2.2.5.2 Teste de Koenker-Bassett

Neste método estima-se a regressão dos quadrados dos resíduos (e_i^2) do modelo original $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{p-1} x_{i,p-1} + \varepsilon_i$ contra os valores estimados

do regressando elevados ao quadrado (\hat{y}_i^2), ou seja, $e_i^2 = \alpha + \beta \hat{y}_i^2 + v_i$. Se a hipótese nula $H_0: \beta = 0$ não é rejeitada pode-se concluir que não há indícios da presença de heterocedasticidade nos dados. A significância estatística do teste é obtida a partir do teste F [39] o que faz com que o método de Koenker-Bassett apresente maior poder do teste que os métodos de Levene e Fligner-Killeen.

2.2.6 Eficiência dos *scores*

Nesta tese é sugerida uma métrica que possibilite mensurar o grau de eficácia das estatísticas de desempenho (*scores*) em avaliar o desempenho dos laboratórios participantes de um ensaio de proficiência. Essa métrica proposta é o erro do tipo II.

Testes estatísticos de hipóteses paramétricas têm por finalidade refutar ou validar hipóteses científicas postuladas sobre um fenômeno observável. A hipótese que se deseja testar como válida é denominada de hipótese nula (H_0). Esta é comparada com uma hipótese alternativa (H_1).

As estatísticas de desempenho são um tipo de teste de hipóteses em que a hipótese nula consiste em verificar se o valor reportado pelo participante é igual ao valor designado ($H_0: x_i = x_{PT}$). A hipótese alternativa refere-se à divergência destes valores, ou seja, $H_1: x_i \neq x_{PT}$.

Ao se tomar uma decisão a favor ou contra uma hipótese existem dois tipos de erros que se pode cometer. Em testes de hipóteses, o erro do tipo I refere-se à probabilidade de rejeitar a hipótese nula quando esta é verdadeira (α). A probabilidade de aceitar a hipótese nula quando esta é falsa é denominada de erro do tipo II (β). Propõe-se medir a eficácia das estatísticas de desempenho por meio do erro tipo II.

Nos ensaios de proficiência, o erro do tipo II é a probabilidade de considerar os resultados, reportados pelo laboratório participante, como satisfatórios quando é provável que não sejam satisfatórios. Esta probabilidade é calculada por

$$\beta = \Phi(z_{1-\alpha/2} - \delta/\sigma) - \Phi(-z_{1-\alpha/2} - \delta/\sigma) \quad \text{Equação 64}$$

Em que $z_{1-\alpha/2}$ é o quantil da distribuição normal padrão e $\Phi(\cdot)$ é a função de distribuição da normal padrão [61]. Cabe destacar que σ é desvio-padrão considerado

no cálculo do *score* adotado como medida de avaliação de desempenho e δ é a diferença entre o valor reportado pelo laboratório participante e o valor designado.

O erro do tipo II é uma proposta para medir a capacidade que uma estatística de desempenho tem de classificar corretamente os resultados dos laboratórios participantes de um ensaio de proficiência. Valores baixos de β fornecem indícios de que o *score* escolhido como métrica de desempenho apresenta probabilidade significativa de classificar corretamente os resultados reportados.

3. ENSAIO DE PROFICIÊNCIA COM DADOS CATEGÓRICOS

3.1 FUNDAMENTAÇÃO TEÓRICA

Variáveis quantitativas são aquelas que podem ser expressas por um valor numérico. Se a variável assume um número finito ou infinito enumerável de valores, diz-se que a variável quantitativa é discreta. Variáveis contínuas são aquelas que assumem um número infinito e não enumerável de valores. Os *scores* (zscore, z'score, zeta score e E_n score) são as métricas de avaliação de desempenho quando os resultados reportados pelos participantes são variáveis quantitativas [4].

Variáveis categóricas, ou qualitativas, são aquelas em que o resultado não é um valor numérico, mas uma categoria. Há ensaios de proficiência em que o resultado reportado pelo participante consiste em uma variável categórica. Se há uma ordenação entre as categorias da variável categórica está é dita ordinal. Variável categórica nominal é aquela em que não é possível ordenar as categorias.

Medidas de concordância são métricas que permitem aferir o grau de uniformidade nas classificações dadas por dois ou mais avaliadores a cerca de uma variável categórica específica. Nesta seção serão apresentadas medidas de concordância como métricas de avaliação de desempenho dos resultados reportados pelos laboratórios participantes.

Nos ensaios de proficiência com dados categóricos em que os resultados reportados pelos laboratórios estão em escala nominal são discutidos os coeficientes de Gower [4, 62] (mencionado na norma ISO 13528:2015), kappa de Cohen [63-72] (sugerido na literatura), alfa de Krippendorff [73] e gama de Gwet [73-75]. Nos casos em que os resultados são reportados em escalas ordinais são apresentados o Índice de Leti [76] e os coeficientes de kappa de Cohen ponderado [77], alfa de Krippendorff [73] e gama de Gwet [73-75].

3.1.1 Resultados reportados em escala nominal

3.1.1.1 Coeficiente de Gower

A norma ISO 13528:2015 menciona sobre métricas para avaliar o desempenho dos participantes de ensaio de proficiência cujos resultados reportados são variáveis categóricas: (i) proporção de resultados corretos; (ii) total de resultados corretos; (iii) medida de distância baseada na diferença entre o resultado reportado e o resultado esperado. Com relação a esta última o coeficiente de Gower é citado pela norma como medida de distância [4].

O coeficiente de Gower fornece o grau de similaridade entre dois objetos [62]. Nos ensaios de proficiência com dados categóricos, este coeficiente pode quantificar a similaridade entre o resultado reportado pelo laboratório participante e o resultado esperado.

Se dois objetos i e j podem ser comparados na k -ésima categoria, o escore designado s_{ijk} é igual a zero ($s_{ijk} = 0$) quando i e j são considerados diferentes e igual a um ($s_{ijk} = 1$) quando há algum grau de concordância ou similaridade. Há casos em que não é possível comparar os objetos i e j em uma determinada categoria. A quantidade $\delta_{ijk} = 1$ indica quando é possível comparar os objetos i e j na k -ésima categoria e $\delta_{ijk} = 0$ indica quando não é possível. A similaridade S_{ij} entre i e j é definida como o escore médio calculado sobre todas as comparações possíveis [62].

$$S_{ij} = \frac{\sum_{k=1}^v s_{ijk} \cdot \delta_{ijk}}{\sum_{k=1}^v \delta_{ijk}} \quad \text{Equação 65}$$

Em que v é o número de categorias que são objeto de análise. O coeficiente de Gower S_{ij} varia entre 0 e 1: $S_{ij} = 1$ indica que os objetos i e j não diferem em nenhuma das categorias enquanto $S_{ij} = 0$ indica divergência em todas as categorias [62]. O coeficiente de Gower ponderado é definido por:

$$SW_{ij} = \frac{\sum_{k=1}^v s_{ijk} \cdot w_k \cdot \delta_{ijk}}{\sum_{k=1}^v w_k \cdot \delta_{ijk}} \quad \text{Equação 66}$$

Em que w_k são os pesos [62]. Os pesos são definidos de modo discricionário pelo provedor do ensaio. Nesta tese é sugerida uma forma de interpretar os resultados reportados pelos laboratórios participantes de um ensaio de proficiência com dados categóricos em que a métrica de avaliação é o coeficiente de Gower. A classificação

proposta na tabela 4 baseou-se na interpretação das estatísticas de desempenho descritas na norma ISO 13528:2015.

Tabela 4 – Interpretação do coeficiente de Gower.

Coeficiente de Gower	Classificação
$0 \leq S_{ij} < 0,33$	Insatisfatório
$0,33 \leq S_{ij} < 0,67$	Questionável
$0,67 \leq S_{ij} \leq 1$	Satisfatório

Fonte: elaboração própria.

3.1.1.2 Coeficiente kappa de Cohen

Mancin *et al.* (2015) propuseram o coeficiente kappa de Cohen como método para avaliar o desempenho dos participantes dos ensaios de proficiência em que os resultados reportados pelos participantes constituem variáveis categóricas [63]. O coeficiente capta o nível de consonância entre as classificações realizadas por dois avaliadores. Landis e Koch (1977) propuseram uma classificação do coeficiente [64] que pode ser considerada como critério de avaliação do resultado do laboratório.

O coeficiente kappa de Cohen

$$k = \frac{P_{obs} - P_{esp}}{1 - P_{esp}} \quad \text{Equação 67}$$

Mede o grau de concordância entre dois avaliadores, que classificaram n itens em c categorias mutuamente exclusivas [65], em que $P_{obs} = \sum_{x=1}^c p_{xx}$ é a proporção de concordância observada e $P_{esp} = \sum_{x=1}^c p_{.x} \cdot p_{x.}$ é a proporção de concordância esperada. O coeficiente varia entre -1 e 1 em que $k > 0$ indica que há concordância entre os avaliadores, $k < 0$ indica que há discordância, $k = 1$ indica total concordância e $k = 0$ significa que proporção de concordância observada é igual a proporção de concordância esperada (concordância esperada pelo acaso) [65-68].

O coeficiente kappa de Cohen pode ser interpretado do seguinte modo: $k < 0$ indica discordância (*Poor*) entre os avaliadores, $0 \leq k < 0,21$ indica uma concordância fraca (*Slight*), $0,21 \leq k < 0,41$ indica uma concordância regular (*Fair*),

$0,41 \leq k < 0,61$ indica uma concordância moderada (*Moderate*), $0,61 \leq k < 0,81$ indica uma concordância substancial (*Substantial*), $0,81 \leq k < 1$ indica uma concordância “quase perfeita” (*Almost perfect*) entre os avaliadores [64, 69-71]. Uma proposta de interpretação do coeficiente, tomando-se por base a classificação da norma ISO 13528:2015, pode ser vista na tabela 5.

Tabela 5 – Interpretação do coeficiente kappa de Cohen.

Coeficiente kappa	Grau de concordância ^(a)	Classificação ^(b)
$k < 0$	Discordância	Insatisfatório
$0 \leq k < 0,21$	Fraca	
$0,21 \leq k < 0,41$	Regular	Questionável
$0,41 \leq k < 0,61$	Moderada	
$0,61 \leq k < 0,81$	Substancial	Satisfatório
$0,81 \leq k < 1$	Quase perfeita	

^(a) Grau de concordância proposto por Landis e Koch (64).

^(b) Proposta de interpretação baseada na classificação da norma ISO 13528:2015.

Fonte: elaboração própria.

Para avaliar a significância estatística do coeficiente kappa de Cohen, testa-se a hipótese nula $H_0: k = 0$ contra a hipótese alternativa $H_1: k \neq 0$ utilizando-se a estatística de teste

$$z = \frac{k}{\sqrt{\text{Var}(k)}} \quad \text{Equação 68}$$

Se $|z| > z_{1-\alpha/2}$ rejeita-se a hipótese nula em que $z_{1-\alpha/2}$ é o quantil da distribuição normal padrão e α é o nível de significância [63, 68, 72]. A variância da estatística k é dada pela seguinte equação:

$$\begin{aligned} \text{Var}(k) = & \\ = & \frac{1}{n(1 - P_{esp})^2} \left(\sum_{x=1}^c p_{x\cdot} p_{\cdot x} [1 - (p_{x\cdot} + p_{\cdot x})]^2 + \sum_{\substack{x=1 \\ x \neq y}}^c \sum_{y=1}^c p_{x\cdot} p_{y\cdot} (p_{\cdot y} + p_{x\cdot})^2 - P_{esp}^2 \right) \end{aligned} \quad \text{Equação 69}$$

Em que $P_{esp} = \sum_{x=1}^c p_{\cdot x} \cdot p_{x\cdot}$, $p_{\cdot x} = \sum_{i=1}^c p_{ij}$ e $p_{x\cdot} = \sum_{j=1}^c p_{ij}$.

Tabela 6 – Coeficiente kappa de Cohen: tabela de frequências.

Níveis do fator A	Níveis do fator B				Total (Σ)
	1	2	...	c	
1	n_{11}	n_{12}	...	n_{1c}	$n_{1\cdot}$
2	n_{21}	n_{22}	...	n_{2c}	$n_{2\cdot}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
c	n_{c1}	n_{c2}	...	n_{cc}	$n_{c\cdot}$
Total (Σ)	$n_{\cdot 1}$	$n_{\cdot 2}$...	$n_{\cdot c}$	n

Fonte: elaboração própria.

Tabela 7 – Coeficiente kappa de Cohen: tabela de proporções.

Níveis do fator A	Níveis do fator B				Total (Σ)
	1	2	...	c	
1	p_{11}	p_{12}	...	p_{1c}	$p_{1\cdot}$
2	p_{21}	p_{22}	...	p_{2c}	$p_{2\cdot}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
c	p_{c1}	p_{c2}	...	p_{cc}	$p_{c\cdot}$
Total (Σ)	$p_{\cdot 1}$	$p_{\cdot 2}$...	$p_{\cdot c}$	1

Fonte: elaboração própria.

Em que n_{ij} é a frequência observada do i -ésimo nível do fator A e j -ésimo nível do fator B (Tabela 6), $p_{ij} = n_{ij}/n$, $p_{\cdot j} = \sum_{i=1}^c p_{ij}$ e $p_{i\cdot} = \sum_{j=1}^c p_{ij}$ (Tabela 7).

3.1.1.3 Coeficiente alfa de Krippendorff (nominal)

O coeficiente alfa

$$\alpha = \frac{p_a - p_e}{1 - p_e} \quad \text{Equação 70}$$

Mede o grau de concordância entre avaliadores em que p_a é a proporção de concordância observada ponderada e p_e é a proporção de concordância esperada ponderada. A proporção de concordância observada ponderada é obtida por

$$p_a = \left(1 - \frac{1}{n\bar{r}}\right)p'_a + \frac{1}{n\bar{r}} \quad \text{Equação 71}$$

Em que

$$p'_a = \frac{1}{n} \sum_{i=1}^n p_{a|i} \quad \text{Equação 72}$$

$$p_{a|i} = \sum_{k=1}^q \frac{r_{ik}(\bar{r}_{ik+} - 1)}{\bar{r}(r_{i+} - 1)} \quad \text{Equação 73}$$

$$\bar{r}_{ik+} = \sum_{l=1}^q w_{kl} r_{il} \quad \text{Equação 74}$$

Sendo que r_{ik} representa o número de avaliadores que classificaram o elemento (item) i na categoria k .

Tabela 8 – Coeficiente alfa de Krippendorff: tabela de concordância (r_{ik}).

		Categorias				
		1	2	...	q	Total
elementos (itens)	1	r_{11}	r_{12}	...	r_{1q}	r_{1+}
	2	r_{21}	r_{22}	...	r_{2q}	r_{2+}
	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
	n	r_{n1}	r_{n2}	...	r_{nq}	r_{n+}

Fonte: elaboração própria.

A proporção de concordância esperada ponderada é calculada a partir da equação

$$p_e = \sum_{k,l}^n w_{kl} \pi_k \pi_l \quad \text{Equação 75}$$

Em que

$$\pi_k = \frac{1}{n} \sum_{i=1}^n \frac{r_{ik}}{\bar{r}} \quad \text{Equação 76}$$

Por fim,

$$\bar{r} = \frac{1}{n} \sum_{i=1}^n r_{i+} \quad \text{Equação 77}$$

Os pesos para cada par de categorias $(k; l)$ variam entre 0 e 1 e decrescem à medida que a discordância entre duas categorias aumenta. Nos casos em que dois avaliadores classificam um elemento (item) na mesma categoria k , considera-se concordância total e o peso w_{kk} associado ao par $(k; k)$ assume valor 1. Caso os dois avaliadores classifiquem o elemento (item) em categorias distintas ($k \neq l$) tem-se que $w_{kl} < 1$ representa a concordância parcial entre os avaliadores [73].

Para variáveis categóricas nominais os pesos são calculados do seguinte modo:

$$w_{kl} = \begin{cases} 1, & \text{se } k = l \\ 0, & \text{se } k \neq l \end{cases} \quad \text{Equação 78}$$

3.1.1.4 Coeficiente de Gwet (nominal)

Em casos específicos, o coeficiente α (alfa de Krippendorff), assim como o coeficiente k (kappa de Cohen), pode apresentar um valor relativamente baixo mesmo que haja um significativo grau de concordância entre os avaliadores [73]. Devido a esta limitação Gwet (2008) [74] propôs o coeficiente

$$\gamma = \frac{p_a - p_e}{1 - p_e} \quad \text{Equação 79}$$

Em que

$$p_a = \frac{1}{n} \sum_{i=1}^n p_{a|i} \quad \text{Equação 80}$$

$$p_{a|i} = \sum_{k=1}^q \frac{r_{ik}(\bar{r}_{ik+} - 1)}{r_{i+}(r_{i+} - 1)} \quad \text{Equação 81}$$

$$\bar{r}_{ik+} = \sum_{l=1}^q w_{kl} r_{il} \quad \text{Equação 82}$$

Sendo que r_{ik} representa o número de avaliadores que classificaram o elemento (item) i na categoria k .

Tabela 9 – Coeficiente de Gwet: tabela de concordância (r_{ik}).

		Categorias				
		1	2	...	q	Total
elementos (itens)	1	r_{11}	r_{12}	\cdots	r_{1q}	r_{1+}
	2	r_{21}	r_{22}	\cdots	r_{2q}	r_{2+}
	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
	n	r_{n1}	r_{n2}	\cdots	r_{nq}	r_{n+}

Fonte: elaboração própria.

$$p_e = \frac{T_w}{q(q-1)} \sum_{k=1}^q \pi_k (1 - \pi_k) \quad \text{Equação 83}$$

$$T_w = \sum_{k=1}^q \sum_{l=1}^q w_{kl} \quad \text{Equação 84}$$

$$\pi_k = \frac{1}{n} \sum_{i=1}^n \frac{r_{ik}}{r_{i+}} \quad \text{Equação 85}$$

Para variáveis categóricas nominais os pesos são calculados do seguinte modo:

$$w_{kl} = \begin{cases} 1, & \text{se } k = l \\ 0, & \text{se } k \neq l \end{cases} \quad \text{Equação 86}$$

Para os pesos acima mencionados, tem-se que $T_w = q$ em que q representa o número de categorias.

De acordo com Gwet (2008), o coeficiente γ apresenta menor viés na estimação do grau de concordância do que o coeficiente k (kappa de Cohen) o qual tende a subestimar a concordância [74].

Por fim, cabe destacar que o coeficiente Gwet pode ser considerado uma generalização do coeficiente de Brennan-Prediger [75].

3.1.2 Resultados reportados em escala ordinal

3.1.2.1 Índice de Leti

A dispersão de uma variável categórica ordinal pode ser medida por meio do índice de Leti, o qual é definido por

$$D_i = 2 \sum_{k=1}^K F_k^{(i)} (1 - F_k^{(i)}) \quad \text{Equação 87}$$

Em que K é o número de níveis da variável categórica e $F_k^{(i)}$ é calculado por

$$F_k^{(i)} = \frac{1}{n_R} \sum_{j=1}^{n_R} I_{(X_{ij} \leq k)} \quad \text{Equação 88}$$

Sendo n_R o número de avaliadores [76]. O numerador de $F_k^{(i)}$ representa o número de avaliadores que apresentaram escore menor ou igual a k no i -ésimo item avaliado ($i = 1, \dots, n_T$). Cabe destacar que $D_i = 0$ se, e somente se, todos os níveis da variável categórica são iguais (ausência de dispersão). O valor máximo do índice de Leti (D_{max}) é obtido quando todas as observações estão concentradas nos dois níveis extremos da variável categórica (máxima dispersão) [76]. Em outras palavras, tem-se que

$$D_{max} = \begin{cases} \frac{K-1}{2} & \text{se } n \text{ é par} \\ \frac{K-1}{2} \left(1 - \frac{1}{n_R^2}\right) & \text{se } n \text{ é ímpar} \end{cases} \quad \text{Equação 89}$$

É possível definir uma medida de dispersão normalizada no intervalo $[0; 1]$ dada por $d_i = D_i / D_{max}$. Por fim, tem-se que

$$d = \frac{1}{D_{max}} \left(\frac{1}{n_T} \sum_{i=1}^{n_T} D_i \right) \quad \text{Equação 90}$$

Quando todos os níveis da variável categórica são iguais (ausência de dispersão), tem-se que $d = 0$. Por outro lado, quando todas as observações estão concentradas nos dois níveis extremos da variável categórica (máxima dispersão), tem-se que $d = 1$ [76].

Tabela 10 – Índice de Leti: resultados reportados.

	Avaliadores			
	1	2	...	n_R
1	X_{11}	X_{12}	...	X_{1n_R}
2	X_{21}	X_{22}	...	X_{2n_R}
\vdots	\vdots	\vdots	\ddots	\vdots
n_T	X_{n_T1}	X_{n_T2}	...	$X_{n_T;n_R}$

Fonte: elaboração própria.

No contexto de ensaios de proficiência categóricos, os avaliadores referem-se aos laboratórios participantes e na tabela 11 há uma proposta de interpretação do índice de Leti. A interpretação sugerida na tabela 11 baseia-se na classificação dos *scores* definida na norma ISO 13528:2015.

Tabela 11 – Interpretação do índice de Leti.

Índice de Leti (d)	Classificação
$0 \leq d < \frac{1}{2k-2}$	Satisfatório
$\frac{1}{2k-2} \leq d < \frac{1}{k-1}$	Questionável
$\frac{1}{k-1} \leq d \leq 1$	Insatisfatório

Fonte: elaboração própria.

Se a classificação do item do ensaio está circunscrita a 3 níveis (baixo, moderado e severo, por exemplo), tem-se que $k = 3$ e o laboratório é considerado satisfatório se $0 \leq d < 0,25$, questionável se $0,25 \leq d < 0,5$ e insatisfatório se $0,5 \leq d \leq 1$.

3.1.2.2 Coeficiente kappa de Cohen ponderado

O coeficiente kappa de Cohen é recomendado para variáveis categóricas nominais. Para variáveis categóricas ordinais, o grau ou a magnitude da discordância é considerado e, neste caso, recomenda-se utilizar o coeficiente kappa de Cohen ponderado [77]. O coeficiente ponderado é estimado por:

$$k = 1 - \frac{\sum_{i=1}^c \sum_{j=1}^c w_{ij} p_{ij}}{\sum_{i=1}^c \sum_{j=1}^c w_{ij} e_{ij}} \quad \text{Equação 91}$$

Em que p_{ij} é a proporção observada, $e_{ij} = p_{.j} \cdot p_{i.}$ é a proporção esperada e w_{ij} são os pesos [77]. A variância da estatística k é dada pela seguinte equação:

$$Var(k) = \frac{1}{n(1-P_{esp})^2} \left(\left(\sum_{i=1}^c \sum_{j=1}^c e_{ij} \cdot \left(\Omega_{ij} - (\omega_{i.} + \omega_{.j}) \right)^2 \right) - P_{esp}^2 \right) \quad \text{Equação 92}$$

Em que

$$\Omega_{ij} = 1 - \frac{w_{ij} - \min(\mathbf{w})}{\max(\mathbf{w}) - \min(\mathbf{w})} \quad \text{Equação 93}$$

Sendo que \mathbf{w} é a matriz de pesos w_{ij} .

$$P_{esp} = \sum_{i=1}^c \sum_{j=1}^c e_{ij} \cdot \Omega_{ij} \quad \text{Equação 94}$$

$$\omega_{i.} = \sum_{j=1}^c p_{.j} \cdot \Omega_{ij} \quad \text{Equação 95}$$

$$\omega_{.j} = \sum_{i=1}^c p_{i.} \cdot \Omega_{ij} \quad \text{Equação 96}$$

O teste de hipótese $H_0: k = 0$ versus $H_1: k \neq 0$ é realizado de maneira análoga [53]. No que tange a avaliação do desempenho dos participantes sugere-se adotar os mesmos critérios descritos na tabela 5.

3.1.2.3 Coeficiente alfa de Krippendorff (ordinal)

O cálculo do coeficiente para variáveis categóricas ordinais difere apenas no modo como os pesos w_{kl} são calculados.

$$w_{kl} = \begin{cases} 1, & \text{se } k = l \\ 1 - \frac{\# \{(i; j), \min(k; l) \leq i < j \leq \max(k; l)\}}{w_{max}}, & \text{se } k \neq l \end{cases} \quad \text{Equação 97}$$

Vale salientar que $\# \{(i; j), \min(k; l) \leq i < j \leq \max(k; l)\}$ representa o número de pares $(i; j)$ em que $i < j$ [73].

Em ambos os casos (nominal e ordinal), o coeficiente α de Krippendorff varia entre -1 e 1 ($-1 \leq \alpha \leq 1$). Krippendorff (2004) sugere que $\alpha \geq 0,667$ indica uma concordância satisfatória (*acceptable*) [78, 79]. O coeficiente fornece o grau de concordância mesmo quando há valores ausentes e/ou quando há um pequeno número de amostras [79]. Vale destacar que o autor do método não define qual número de amostra é considerado pequeno. No capítulo 6, seção 6.2.3, da presente tese são apresentadas algumas considerações concernentes à tamanhos amostrais no contexto de ensaios de proficiência categóricos. Tem-se a seguir a seguinte sugestão de interpretação dos resultados reportados (tanto em escala nominal quanto ordinal) pelos laboratórios participantes:

Tabela 12 – Interpretação do coeficiente de alfa de Krippendorff.

Coeficiente de Krippendorff	Classificação
$-1 \leq \alpha < 0,33$	Insatisfatório
$0,33 \leq \alpha < 0,67$	Questionável
$0,67 \leq \alpha \leq 1$	Satisfatório

Fonte: elaboração própria.

3.1.2.4 Coeficiente de Gwet (ordinal)

O cálculo do coeficiente para variáveis categóricas ordinais difere apenas no modo como os pesos w_{kl} são calculados.

$$w_{kl} = \begin{cases} 1, & \text{se } k = l \\ 1 - \frac{\# \{(i; j), \min(k; l) \leq i < j \leq \max(k; l)\}}{w_{max}}, & \text{se } k \neq l \end{cases} \quad \text{Equação 98}$$

Vale salientar que $\# \{(i; j), \min(k; l) \leq i < j \leq \max(k; l)\}$ representa o número de pares $(i; j)$ em que $i < j$ [73].

O coeficiente γ de Gwet para variável categórica nominal (*Gwet's AC1*) e ordinal (*Gwet's AC2*) varia entre -1 e 1 em que $\gamma = -1$ indica total discordância entre os avaliadores e $\gamma = 1$ indica total concordância [80]. O coeficiente fornece o grau de concordância mesmo quando há valores ausentes [73]. O coeficiente γ pode ser interpretado do seguinte modo: $\gamma \geq 0,667$ indica uma concordância satisfatória (*acceptable*) [81]. Tem-se a seguir a seguinte sugestão de interpretação dos resultados reportados, tanto em escala nominal quanto ordinal, pelos laboratórios participantes:

Tabela 13 – Interpretação do coeficiente de gama de Gwet.

Coeficiente de Gwet	Classificação
$-1 \leq \gamma < 0,33$	Insatisfatório
$0,33 \leq \gamma < 0,67$	Questionável
$0,67 \leq \gamma \leq 1$	Satisfatório

Fonte: elaboração própria.

3.2 ABORDAGEM PROPOSTA

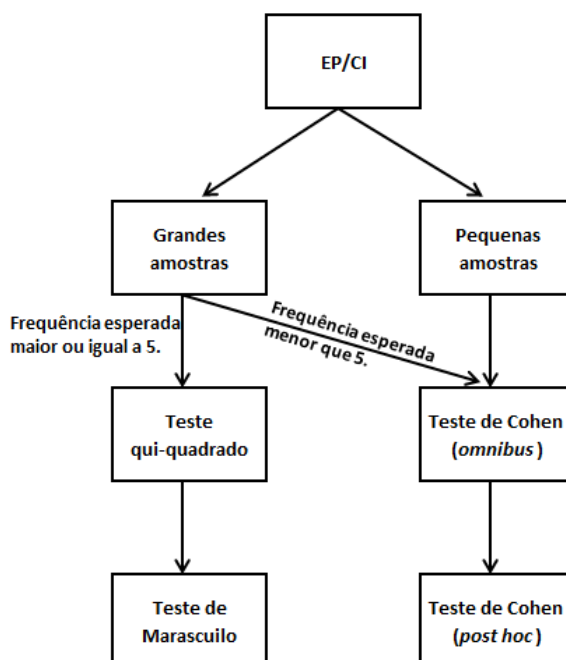
Os ensaios de proficiência podem ser considerados como uma ferramenta para identificar se algum laboratório participante apresenta resultados de medição que diferem, estatisticamente, dos demais participantes ou diferem de um valor designado. No contexto de dados categóricos, esses resultados podem ser expressos e analisados a partir da proporção de classificações em consonância com o resultado esperado. Testes de hipóteses de homogeneidade de proporção podem fornecer elementos que permitam identificar se o participante do ensaio apresenta uma capacidade de medição díspare dos seus pares. Neste caso, os testes de homogeneidade de proporção de qui-quadrado/Cohen (*omnibus* testes) e Marascuilo/Cohen (*post hoc* testes) fornecem subsídios para identificar percentuais atípicos.

A figura 3 apresenta um fluxograma com os métodos estatísticos sugeridos para avaliar o desempenho de laboratórios participantes de um ensaio de proficiência com dados categóricos. Os testes de qui-quadrado e Marascuilo são recomendados para grandes amostras [82-84]. O teste de Cohen (*omnibus* e *post hoc*) é recomendado para

pequenas amostras [83]. Vale destacar que os autores dos métodos não definem qual número de amostra é considerado pequeno. No capítulo 6, seção 6.2.3, da presente tese são apresentadas algumas considerações concernentes à tamanhos amostrais no contexto de ensaios de proficiência categóricos. Adicionalmente, a referida seção também apresenta a justificativa pela opção do teste de qui-quadrado em detrimento ao teste de exato de Fisher para avaliar homogeneidade de proporção (seção 6.2.3).

É importante observar que o tamanho da amostra se refere à quantidade de itens de ensaio enviados aos participantes. Por fim, o teste de qui-quadrado apresenta como restrição adicional que as frequências esperadas devem ser maiores ou iguais a 5 [82].

Figura 3 – Fluxograma para análise de dados categóricos.



Fonte: elaboração própria.

3.2.1 Grande número de amostras

3.2.1.1 Teste Qui-Quadrado

É um teste não paramétrico que pode ser empregado para investigar a aderência dos dados a uma distribuição de probabilidade teórica, verificar a independência entre atributos e/ou variáveis (associação) e, por fim, avaliar a homogeneidade das

proporções em k tratamentos independentes. Neste método busca-se comparar as frequências amostrais observadas (O_{ij}) com as frequências teóricas esperadas (E_{ij}) [82].

No contexto em que se deseja testar a hipótese de homogeneidade de proporções em k tratamentos, a estatística de teste é dada por

$$\chi_{obs}^2 = \sum_{i=1}^k \sum_{j=1}^m \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \text{Equação 99}$$

Em que $E_{ij} = O_{.j} \cdot O_{i.} / O_{..}$, k é o número de linhas e m é o número de colunas. Se $\chi_{obs}^2 > \chi_{(m-1)(k-1);1-\alpha}^2$ (quantil da distribuição qui-quadrado com $(m-1)(k-1)$ graus de liberdade e nível de significância α) rejeita-se a hipótese nula [82]. No contexto de ensaios de proficiência, o tratamento refere-se ao laboratório participante.

Tabela 14 – Teste Qui-quadrado: tabela de contingência (frequências).

	Itens de ensaio				Total
	1	2	...	m	
Laboratório 1	O_{11}	O_{12}	...	O_{1m}	$O_{1.}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
Laboratório k	O_{k1}	O_{k2}	...	O_{km}	$O_{k.}$
Total	$O_{.1}$	$O_{.2}$...	$O_{.m}$	$O_{..}$

Fonte: elaboração própria.

Valores esperados menores que 5 ($E_{ij} < 5$) e amostras pequenas podem afetar a aproximação da distribuição qui-quadrado da estatística χ_{obs}^2 fazendo com que esta não seja suficientemente boa [82].

A hipótese nula de homogeneidade pode ser escrita como H_0 : as proporções em cada uma das m categorias são as mesmas para os k tratamentos. Em outras palavras tem-se que:

$$\begin{cases} (H_0) p_1 = p_2 = \dots = p_k \\ (H_1) p_i \neq p_j \text{ para algum } i \neq j \end{cases}$$

Tabela 15 – Teste Qui-quadrado: tabela de contingência (proporções).

	Itens de ensaio			
	1	2	...	m
Laboratório 1	p_{11}	p_{12}	...	p_{1m}
\vdots	\vdots	\vdots	\ddots	\vdots
Laboratório k	p_{k1}	p_{k2}	...	p_{km}

Fonte: elaboração própria.

3.2.1.2 Teste Marascuilo

Nos casos em que a hipótese nula de igualdade das proporções ($H_0: p_1 = p_2 = \dots = p_k$) é rejeitada, faz-se necessário investigar para quais pares (i, j) a proporção p_i difere da proporção p_j , sendo $i \neq j$. O teste de Marascuilo permite testar simultaneamente as diferenças de todos os pares de proporções e a hipótese nula $H_0: p_i = p_j$ é rejeitada em favor da hipótese alternativa $H_1: p_i \neq p_j$ para o par (i, j) quando:

$$|p_i - p_j| > \sqrt{\chi_{(k-1); 1-\alpha}^2} \cdot \sqrt{\frac{p_i(1-p_i)}{n_i} + \frac{p_j(1-p_j)}{n_j}} \quad \text{Equação 100}$$

Em que $n_i = O_i$ é o total da linha na tabela de contingência e $\chi_{(k-1); 1-\alpha}^2$ é o quantil da distribuição qui-quadrado com $(k-1)$ graus de liberdade e nível de significância α [85-87]. Este método não é recomendado para pequenas amostras [83, 84].

3.2.2 Pequeno número de amostras

3.2.2.1 Teste de Cohen

O teste de Cohen constitui uma alternativa ao teste de qui-quadrado e ao teste de Marascuilo nos casos em que o tamanho amostral é pequeno. No que tange as comparações múltiplas, este método apresenta uma vantagem adicional, pois é aplicável nos casos em que $p_i = 0$ ou $p_i = 1$ ($i = 1, 2, \dots, k$) [83, 84]. A hipótese de

homogeneidade de proporções em k tratamentos é verificada a partir da estatística de teste:

$$\chi^{2*} = \sum_{j=1}^k n_j (\phi_j - \phi_0)^2 \quad \text{Equação 101}$$

Em que

$$\phi_0 = \sum_{j=1}^k \frac{n_j \phi_j}{n} \quad \text{Equação 102}$$

$$\phi_j = \begin{cases} 2 \cdot \arccos\left(\sqrt{1/4n_j}\right) & \text{para } p_j = 0 \\ 2 \cdot \arccos(\sqrt{p_j}) & \text{para } 0 < p_j < 1 \\ \pi - 2 \cdot \arccos\left(\sqrt{1/4n_j}\right) & \text{para } p_j = 1 \end{cases} \quad \text{Equação 103}$$

$$n = \sum_{j=1}^k n_j \quad \text{Equação 104}$$

Se $\chi^{2*} > \chi_{k-1;1-\alpha}^2$, rejeita-se a hipótese nula de homogeneidade de proporções em k tratamentos $H_0: p_1 = p_2 = \dots = p_k$ [83, 84]. Nas comparações múltiplas a hipótese nula $H_0: p_i = p_j$ é rejeitada em favor da hipótese alternativa $H_1: p_i \neq p_j$ se:

$$|\phi_i - \phi_j| > \sqrt{\chi_{(k-1);1-\alpha}^2} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \quad \text{Equação 105}$$

Em que $i \neq j$ e $\chi_{k-1;1-\alpha}^2$ é o quantil da distribuição qui-quadrado com $(k - 1)$ graus de liberdade e nível de significância α [83, 84].

4. ENSAIO DE PROFICIÊNCIA MULTIVARIADO

Análise multivariada, de uma forma bem geral, refere-se a todos os métodos estatísticos que analisam simultaneamente múltiplas variáveis em cada indivíduo ou objeto sob investigação. No contexto de ensaios de proficiência/comparações interlaboratoriais (EP/CI), o participante tem como resultado de medição, por exemplo, um espectro de uma determinada amostra fornecida pelo provedor do ensaio. A seção 6.3.1 da presente tese contém um exemplo deste tipo de dado multivariado (espectro).

No âmbito na análise multivariada de dados, existem técnicas que possibilitam agrupar objetos com base em suas próprias características cuja finalidade é tentar identificar eventuais padrões ou estruturas. Em outras palavras, esses métodos buscam, a partir de alguma medida de similaridade, agrupar esses objetos em aglomerados homogêneos. No contexto de EP/CI, esses objetos referem-se aos laboratórios participantes. Neste capítulo serão discutidas algumas técnicas de agrupamentos de objetos e será proposta uma metodologia de classificação dos resultados reportados pelos participantes.

4.1 FUNDAMENTAÇÃO TEÓRICA

4.1.1 Análise de componentes principais

A análise de componentes principais busca explicar a estrutura de variância e covariância por meio de combinações lineares das variáveis originais. Esta técnica tem como objetivos principais a redução da dimensão dos dados e a interpretação de um determinado fenômeno em estudo. Embora p componentes principais sejam necessárias para reproduzir a variabilidade total do sistema, muito desta variabilidade é devida a um pequeno número $k < p$ de componentes principais. Então as k componentes principais podem substituir as p variáveis iniciais e os dados originais, consistindo em n medidas em p variáveis, são reduzidos a um conjunto de dados consistindo em n medidas em k componentes principais [88].

$$\begin{array}{c}
\text{Dados} \\
\text{originais}
\end{array}
\begin{array}{c}
1 \\
2 \\
\vdots \\
n
\end{array}
\begin{array}{c}
\text{Variáveis} \\
1 \quad 2 \quad \cdots \quad p \\
\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}
\end{array}
\quad \text{Equação 106}$$

A partir da matriz \mathbf{X} obtem-se o vetor de médias $\boldsymbol{\mu}$, a matriz de variância e covariância $\boldsymbol{\Sigma}$ e a matriz de correlação $\boldsymbol{\rho}$.

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \quad \text{Equação 107}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \quad \text{Equação 108}$$

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} \quad \text{Equação 109}$$

Seja \mathbf{X} a matriz de dados originais com vetor de médias $\boldsymbol{\mu}$ e matriz de covariância $\boldsymbol{\Sigma}$ em que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ e $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_p$ são, respectivamente, os autovalores e autovetores associados a $\boldsymbol{\Sigma}$. Define-se a i -ésima componente principal do vetor \mathbf{X}_j por $\mathbf{Y}_i = \mathbf{e}_i^T \mathbf{X}_j$ em que $\mathbf{X}_j^T = [x_{j1} \ x_{j2} \ \cdots \ x_{jp}]$ e $\mathbf{e}_i^T = [e_{i1} \ e_{i2} \ \cdots \ e_{ip}]$ sendo que $i = 1, 2, \cdots, p$ e $j = 1, 2, \cdots, n$ [88].

$$\begin{cases} Y_1 = e_{11}x_{j1} + e_{12}x_{j2} + \cdots + e_{1p}x_{jp} \\ Y_2 = e_{21}x_{j1} + e_{22}x_{j2} + \cdots + e_{2p}x_{jp} \\ \vdots \\ Y_p = e_{p1}x_{j1} + e_{p2}x_{j2} + \cdots + e_{pp}x_{jp} \end{cases} \quad j = 1, 2, \cdots, n \quad \text{Equação 110}$$

A proporção P_i da variabilidade total devida (explicada) pela i -ésima componente principal é

$$P_i = \lambda_i / \sum_{i=1}^p \lambda_i \quad \text{Equação 111}$$

Se $P_1 + \dots + P_k \geq 80\%$, em que $k < p$, sugere-se utilizar as k componentes principais, pois estas explicam um grande percentual da variabilidade do sistema sem muita perda de informação. Por fim, o coeficiente de correlação entre a i -ésima componente principal Y_i e a t -ésima variável X_t é

$$\rho(Y_i; X_t) = e_{it} \cdot \sqrt{\lambda_i} / \sqrt{\sigma_{tt}} \quad \text{Equação 112}$$

Em que $i, t = 1, 2, \dots, p$ sendo $\mathbf{X}_t^T = [x_{1t} \ x_{2t} \ \dots \ x_{nt}]$. Essas correlações frequentemente ajudam a interpretar as componentes principais [88].

Nos casos em que as variáveis originais não podem ser diretamente empregadas (por exemplo, devido a escalas diferentes de mensurações empregadas), pode-se obter as componentes principais a partir dos vetores padronizados

$$\mathbf{Z}_t = \frac{\mathbf{X}_t - \mu_t}{\sigma_{tt}} \quad \text{Equação 113}$$

Em que $t = 1, 2, \dots, p$ sendo $\mathbf{Z}_t^T = [z_{1t} \ z_{2t} \ \dots \ z_{nt}]$. Em notação matricial $\mathbf{Z} = V^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ [88].

$$V^{-1/2} = \begin{bmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sqrt{\sigma_{pp}}} \end{bmatrix} \quad \text{Equação 114}$$

Nesse caso, com variáveis padronizadas, a matriz de variância-covariância e a matriz de correlação tornam-se idênticas. Sendo assim, cabe ressaltar que é possível determinar as componentes principais usando a matriz de covariância ou a matriz de correlação de \mathbf{X} , pois esta última é a matriz de covariância dos dados padronizados. A i -ésima componente principal é definida por $Y_i = \mathbf{e}_i^T \mathbf{Z}_j$ em que $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ e $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p$ são, respectivamente, os autovalores e autovetores associados a \mathbf{p} sendo $\mathbf{Z}_j^T = [z_{j1} \ z_{j2} \ \dots \ z_{jp}]$ [88].

$$\begin{cases} Y_1 = e_{11}z_{j1} + e_{12}z_{j2} + \dots + e_{1p}z_{jp} \\ Y_2 = e_{21}z_{j1} + e_{22}z_{j2} + \dots + e_{2p}z_{jp} \\ \vdots \\ Y_p = e_{p1}z_{j1} + e_{p2}z_{j2} + \dots + e_{pp}z_{jp} \end{cases} \quad j = 1, 2, \dots, n \quad \text{Equação 115}$$

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \dots & z_{1p} \\ z_{21} & z_{22} & \dots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \dots & z_{np} \end{bmatrix} \quad \text{Equação 116}$$

A proporção P_i da variabilidade total devida (explicada) pela i -ésima componente principal é $P_i = \lambda_i/p$ e o coeficiente de correlação entre a i -ésima componente principal Y_i e a t -ésima variável padronizada Z_t é $\rho(Y_i; Z_t) = e_{it} \cdot \sqrt{\lambda_i}$ em que $i, t = 1, 2, \dots, p$ [88].

No contexto de ensaio de proficiência (e comparações interlaboratoriais) multivariado as p variáveis representam os laboratórios participantes.

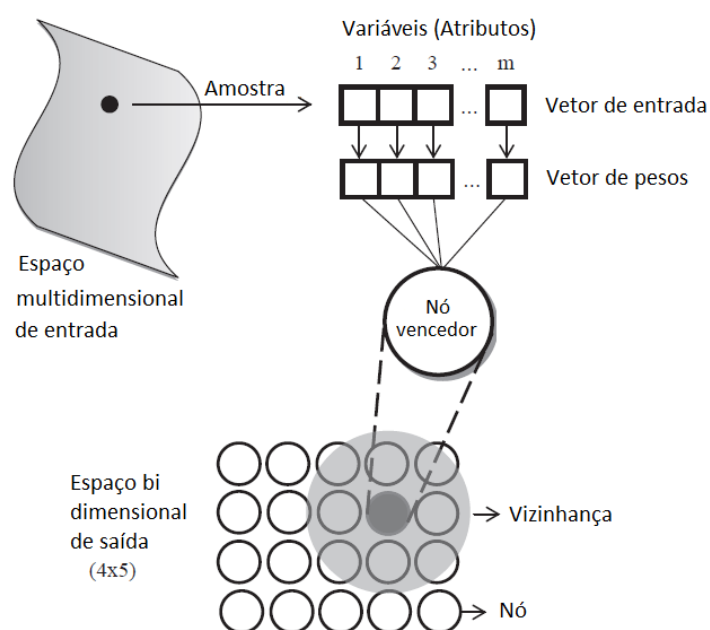
4.1.2 Mapas auto-organizáveis de Kohonen

4.1.2.1 Introdução

Os mapas auto-organizáveis (*self-organizing map* – SOM) são uma técnica de aprendizado não supervisionado cujo principal objetivo é reduzir a dimensão dos dados preservando suas relações [89]. Em outras palavras, é um tipo de rede neural que busca agrupar elementos que apresentam padrão semelhante em relação à determinadas características mensuradas [90]. Um conjunto de dados multidimensional pode ser projetado em um espaço bidimensional cujos elementos podem ser agrupados preservando as informações essenciais [91].

Os mapas consistem em duas camadas de nós: camada de entrada e camada de saída. Diferentemente de outras redes neurais, a camada de entrada conecta-se diretamente com a camada de saída sem camadas intermediárias. Os nós da camada de entrada denotam os atributos (características) ou variáveis contidas nos dados de entrada. Cada parte dos dados de entrada é representada por um vetor de entradas m -dimensional $x = (x_1; x_2; \dots; x_m)'$ cujos elementos indicam os valores do atributo em um conjunto de dados específico [89, 90-92].

Figura 4 – Estrutura de um mapa auto-organizável.



Fonte: extraído de Asan e Ercan, 2012, p. 303 [89].

Se há grandes diferenças entre os valores dos atributos é necessário a normalização dos dados de modo a evitar a influência de um atributo em particular. Um dos métodos mais comuns é a transformação zscore a qual converte o valor de cada atributo em um escore padrão com média 0 e desvio-padrão 1. Em cada atributo subtrai-se a média e divide-se pelo desvio-padrão [89, 90-92].

$$\frac{x_j - \mu_{x_j}}{\sigma_{x_j}}$$

Equação 117

Outros métodos podem ser empregados: (i) dividir o valor de cada atributo pelo máximo de todos os valores; (ii) comprimento unitário em que todos os atributos serão redimensionados para o mesmo comprimento; (iii) transformação min-max em que a escala de valores varia entre 0 e 1; (iv) transformação logarítmica (apropriada para valores exponencialmente distribuídos) [89, 90-92].

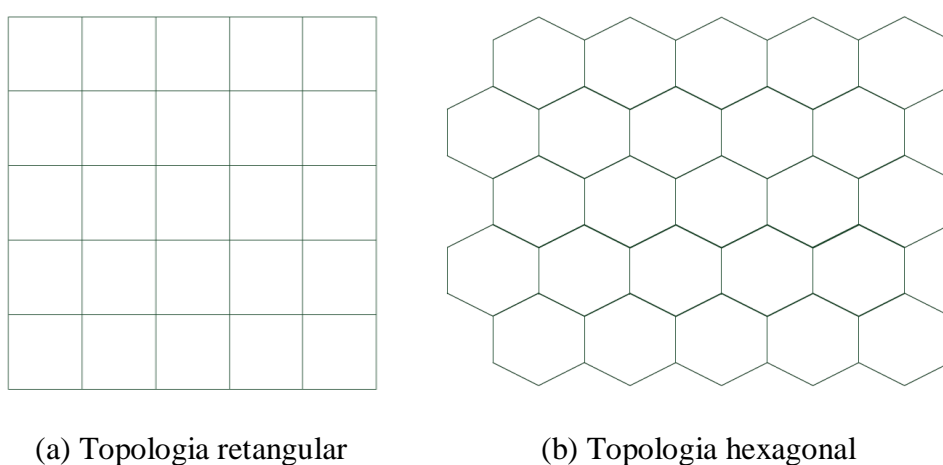
4.1.2.2 Topologia

Nos mapas auto-organizáveis de Kohonen os nós das camadas de saída (camadas de Kohonen) são organizados na forma de uma arquitetura topológica. Essa estrutura predefinida geralmente é organizada como uma grade que consiste em linhas e

colunas. O número de nós na camada de saída denota o número máximo de clusters e influencia na acurácia e capacidade de generalização do mapa. Geralmente um mapa bidimensional fornece a melhor representação dos clusters [89, 90-92].

A topologia é organizada em uma grade retangular ou hexagonal. Na topologia retangular cada nó interno tem 4 vizinhos ao passo que a topologia hexagonal tem 6 (Figura 5). Esta última é mais recomendada porque permite maior número de vizinhos [89, 90-92].

Figura 5 – Topologias.



Fonte: extraído de Asan e Ercan, 2012, p. 304 [89].

4.1.2.3 Pesos

Os pesos $w_i = (w_{i1}; w_{i2}; \dots; w_{im})'$ são links que conectam os nós de entrada aos nós de saída e são atualizados por meio de um processo de aprendizagem em que $i = 1, 2, \dots, n$ sendo n o número de nós de saída. Como os mapas auto-organizáveis são uma técnica de aprendizado não supervisionado os nós de saída competem entre si para tornarem-se ativos. Somente o nó cujo vetor de pesos é mais similar ao vetor de entrada será ativado (*“winner node”*). Para achar este nó, as distâncias entre os dados de entrada (x) e todos os vetores de pesos (w_i) são calculados usando diferentes métodos de medição tais como distância de Manhattan, distância de Chebyshev, distância euclidiana ou distância de Mahalanobis. A distância euclidiana resulta em uma melhor representação visual porque fornece uma exibição mais isotrópica [89, 90-

92]. A distância euclidiana entre a amostra x , escolhida aleatoriamente do conjunto de dados de entrada, e todos os vetores de pesos na interação t é calculada por

$$d_i(t) = \|x(t) - w_i(t)\| = \sqrt{\sum_{j=1}^m (x_{tj} - w_{tji})^2} \quad \text{Equação 118}$$

Em que $\|\cdot\|$ é a norma euclidiana e $i = 1, 2, \dots, n$. O vetor de pesos w_i de cada nó de saída i tem a mesma dimensão do vetor de entrada x [89, 90-92]. O “nó vencedor” (“*winner node*”) c na interação t é determinado usando-se o critério de distância euclidiana mínima:

$$c(t) = \arg \min_i \{\|x(t) - w_i(t)\|\} \quad \text{Equação 119}$$

O vetor de pesos do “vencedor” e suas unidades vizinhas no espaço de saída são ajustados para se tornarem mais representativos das características do espaço de entrada [89, 90-92]. Para um dado vetor de pesos $w_i(t)$ do “neurônio vencedor” i na interação t é atualizado pelo vetor de pesos $w_i(t+1)$ na interação $t+1$ conforme equação abaixo:

$$w_i(t+1) = w_i(t) + \alpha(t)h_{ci}(t)[x(t) - w_i(t)] \quad \text{Equação 120}$$

Em que $\alpha(t)$ é a taxa de aprendizagem que controla a taxa de mudança dos vetores de pesos ($0 \leq \alpha(t) \leq 1$). A taxa de aprendizagem decresce gradualmente com uma função da interação na etapa t . Por fim, $h_{ci}(t)$ é a função de vizinhança em que i representa o índice da unidade vizinha [89, 90-92].

Tabela 16 – Mapas auto-organizáveis de Kohonen: funções de vizinhança.

Função	$h_{ci}(t)$
Gaussiana	$\exp\left(-\frac{d_{ci}^2}{2\sigma^2(t)}\right)$
Bubble	$\mathbf{1}(\sigma(t) - d_{ci})$
Cutgauss	$\exp\left(-\frac{d_{ci}^2}{2\sigma^2(t)}\right) \cdot \mathbf{1}(\sigma(t) - d_{ci})$
Epanechnikov	$\max\{0; 1 - (\sigma(t) - d_{ci})^2\}$

Fonte: Asan e Ercan, 2012 [89].

Em que $\sigma(t)$ representa a largura ou raio efetivo da vizinhança na interação t e d_{ci}^2 representa a distância lateral entre o “neurônio vencedor” c e o “neurônio excitado” i . Por fim, $\mathbf{1}(x)$ é uma função em que $\mathbf{1}(x) = 0$ se $x < 0$ e $\mathbf{1}(x) = 1$ se $x \geq 0$ [89, 90-92].

4.1.2.4 Algoritmo

Os mapas auto-organizáveis de Kohonen são um processo iterativo constituído das seguintes etapas:

Etapa 1. Definir: (i) a dimensão do espaço de saída; (ii) o vetor de pesos iniciais $w_i(0)$; (iii) a taxa de aprendizagem $\alpha(0)$; (iv) o raio da vizinhança $\sigma(0)$; e (v) o número máximo de interações (T).

Etapa 2. Selecionar aleatoriamente um vetor de entrada $x(t)$ dos dados de treinamento.

Etapa 3. Calcular a distância euclidiana entre o vetor de entrada e cada vetor de pesos do nó de saída de modo a encontrar o nó $c(t) = \arg \min_i \{\|x(t) - w_i(t)\|\}$.

Etapa 4. Atualizar os pesos: $w_i(t+1) = w_i(t) + \alpha(t)h_{ci}(t)[x(t) - w_i(t)]$. Para o “nó vencedor” tem-se que $h_{ci}(t) = 1$.

Etapa 5. Ajustar o tamanho da vizinhança e a taxa de aprendizagem.

Etapa 6. Retornar a etapa 2 até que a mudança nos pesos seja menor que um limite pré-definido ou o número máximo de interações (T) seja alcançado [89, 92].

4.1.3 Zscore multivariado

Sheen *et al.* (2017) propuseram um algoritmo para avaliar os resultados dos laboratórios que participaram de uma comparação interlaboratorial que teve por finalidade investigar a eficácia da metabolômica RMN ^1H em gerar conjuntos de dados comparáveis a partir de amostras derivadas do meio ambiente. Os resultados reportados pelos participantes foram espectros de RMN ^1H [93].

No algoritmo proposto por Sheen *et al.* (2017) os espectros devem ser agrupados de modo que o conglomerado S_k consista em múltiplos espectros da k -ésima amostra fornecidos pelos participantes. A partir destes agrupamentos calcula-se a matriz de distância interespectral D_k cujos elementos são as distâncias $D_{ij,k} = d(s_{i,k}, s_{j,k})$ em que $s_{i,k}$ é o espectro do laboratório i e $s_{j,k}$ é o espectro do laboratório j , ambos pertencentes ao conglomerado S_k . Na metodologia proposta, $d(\cdot)$ é uma medida de distância multivariada. Neste artigo os autores sugerem as seguintes medidas de distância: Euclidiana, Mahalanobis, Hellinger, Kullback-Leibler, Jensen-Shannon e Jeffreys [93].

Com base nos valores $D_{ij,k}$ calcula-se a distância média $\hat{D}_{i,k}$.

$$\hat{D}_{i,k} = \frac{1}{n} \sum_j D_{ij,k} \quad \text{Equação 121}$$

Cabe destacar que $\hat{D}_{i,k}$ representa a distância média do espectro $s_{i,k}$ para outro espectro no conglomerado S_k . No método sugerido, os valores $\hat{D}_{i,k}$ precisam ser ajustados a uma determinada distribuição de probabilidade para cada laboratório i . Após esta etapa obtém-se a matriz Z em que Z_i é o vetor Z-score do i -ésimo laboratório. O vetor Z_i é obtido por

$$Z_{i,k} = C^{*-1} \left(C_k(\hat{D}_{i,k}) \right) \quad \text{Equação 122}$$

Em que C_k é a função de distribuição acumulada após ser ajustada ao conglomerado k e C^* é a correspondente função de distribuição padrão [93].

Nesta abordagem, a etapa seguinte consiste em realizar a análise de componentes principais (PCA) na matriz Z . No modelo PCA tem-se que $T = ZP^t$ em que T é a matriz dos escores das componentes principais e P é a matriz das cargas fatoriais. As L componentes principais mais significativas (geralmente $L = 2$) são identificadas obtendo-se $T_L = ZP_L^t$. Para cada participante calcula-se a norma Euclidiana $\|T_{i,L}\|$ e ajusta-se essas distâncias estatísticas a uma nova distribuição de probabilidade com função de distribuição \hat{C} [93].

Esta nova distribuição tem um Z-score associado a ela. Sheen *et al.* (2017) denominaram de escore estatístico projetado (\hat{Z}_i). Esse escore é calculado utilizando-se

$$\hat{Z}_k = \hat{C}^{*-1}(\hat{C}(\|T_{i,L}\|)) \quad \text{Equação 123}$$

Se algum valor \hat{Z}_i ficar fora do intervalo de confiança de 95%, o laboratório i correspondente é considerado um outlier e removido do conjunto de dados. O processo é repetido até que nenhum conjunto de dados fique fora do intervalo de confiança de 95% [93].

4.2 ABORDAGEM PROPOSTA

Nos EP/CI busca-se identificar diferenças interlaboratoriais por meio de métricas definidas pelo provedor do ensaio. Escalonamento multidimensional é uma técnica que permite revelar estruturas “escondidas” em um conjunto de dados multivariado [94]. Em outras palavras, é um método que permite visualizar a similaridade/dissimilaridade entre objetos (laboratórios neste caso) os quais são representados como pontos em um espaço bi (ou tri) dimensional. A proximidade/distância entre os pontos representa a similaridade/dissimilaridade entre objetos. O objetivo do escalonamento multidimensional é que as distâncias correspondam o mais próximo possível as similaridades/dissimilaridades [95].

Elipse/elipsóide de confiança robusta é um método estatístico para detectar informações discrepantes em um conjunto de dados multivariado. A presente tese tem como proposta combinar as técnicas de escalonamento multidimensional e elipse/elipsóide de confiança robusta com o intuito de identificar se algum ou alguns

participantes produzem resultados que destoam dos demais e a partir desta informação classificar o desempenho dos laboratórios participantes do ensaio.

4.2.1 Escalonamento multidimensional

Seja n o número de diferentes objetos (laboratórios no contexto de EP/CI) e δ_{ij} a dissimilaridade entre os objetos i e j . As coordenadas são reunidas na matriz $\mathbf{X}_{n \times p}$ em que p é a dimensionalidade da solução a ser especificada. Sendo assim, a linha i de $\mathbf{X}_{n \times p}$ fornece as coordenadas do objeto i . Seja $d_{ij}(\mathbf{X})$, a distância euclidiana (mais comumente usada [96, 97]) entre as linhas i e j , definida como

$$d_{ij}(\mathbf{X}) = \left(\sum_{s=1}^p (x_{is} - x_{js})^2 \right)^{1/2} \quad \text{Equação 124}$$

que é a menor distância entre os pontos i e j . É importante destacar que a distância euclidiana resulta em uma melhor representação visual porque fornece uma exibição mais isotrópica [89]. O objetivo do escalonamento multidimensional é encontrar uma matriz $\mathbf{X}_{n \times p}$ tal que $d_{ij}(\mathbf{X})$ seja igual a δ_{ij} tanto quanto possível [94, 95]. A matriz $\mathbf{X}_{n \times p}$ é obtida minimizando a equação 71.

$$\sigma^2(\mathbf{X}) = \sum_{i=2}^n \sum_{j=1}^{i-1} w_{ij} \left(\delta_{ij} - d_{ij}(\mathbf{X}) \right)^2 \quad \text{Equação 125}$$

Em que w_{ij} é um peso definido. O problema de minimização de $\sigma^2(\mathbf{X})$ é bastante complexo e faz-se necessário a utilização de algoritmos iterativos para encontrar a matriz $\mathbf{X}_{n \times p}$ que minimiza $\sigma^2(\mathbf{X})$. O procedimento mais utilizado para a solução deste problema de minimização é o algoritmo SMACOF [94, 95].

4.2.2 Elipse de confiança robusta

A elipse de confiança robusta é construída a partir da equação matricial

$$\hat{\mu}_{rob} + \left(\sqrt{2 \cdot F_{2;(n-1);(1-\alpha)}} \right) \cdot U \cdot Q \quad \text{Equação 126}$$

Em que $\hat{\mu}_{rob} = [\bar{x}_{rob} \ \bar{y}_{rob}]$ é o vetor de médias robustas, $F_{2;(n-1);(1-\alpha)}$ é o quantil da distribuição Fisher-Snedecor com nível de confiança de $(1 - \alpha) \%$ e $U = [\cos(\mathbf{a}) \ \sin(\mathbf{a})]$ é o círculo unitário (U é uma matriz $m \times 2$) sendo que $\mathbf{a} = [a_1 \ \dots \ a_m]$ é um vetor de tamanho m ($0 \leq \mathbf{a} \leq 2\pi$). Por fim, tem-se que $Q = chol(S_{rob})$ é a decomposição de Choleski da matriz de variância-covariância robusta S_{rob} [53]. O algoritmo para estimar $\hat{\mu}_{rob}$ e S_{rob} está disponível no apêndice B da presente tese

4.2.3 Elipsóide de confiança robusta

A elipsóide de confiança robusta é construída a partir de equação análoga a elipse de confiança robusta:

$$\hat{\mu}_{rob} + \left(\sqrt{3 \cdot F_{3;(n-1);(1-\alpha)}} \right) \cdot U \cdot Q \quad \text{Equação 127}$$

Em que $\hat{\mu}_{rob} = [\bar{x}_{rob} \ \bar{y}_{rob} \ \bar{z}_{rob}]$ é o vetor de médias robustas, $F_{3;(n-1);(1-\alpha)}$ é o quantil da distribuição Fisher-Snedecor com nível de confiança de $(1 - \alpha) \%$ e $U = [\cos(\boldsymbol{\theta})\sin(\boldsymbol{\varphi}) \ \sin(\boldsymbol{\theta})\sin(\boldsymbol{\varphi}) \ \cos(\boldsymbol{\varphi})]$ é a esfera de raio 1 (U é uma matriz $m \times 3$) sendo que $\boldsymbol{\theta} = [\theta_1 \ \dots \ \theta_m]$ é um vetor de tamanho m ($0 \leq \boldsymbol{\theta} \leq 2\pi$) e $\boldsymbol{\varphi} = [\varphi_1 \ \dots \ \varphi_m]$ é um vetor de tamanho m ($0 \leq \boldsymbol{\varphi} \leq \pi$). A matriz $Q = chol(S_{rob})$ é a decomposição de Choleski da matriz de variância-covariância robusta S_{rob} [53].

5. APLICATIVO WEB

5.1 INTRODUÇÃO

Nesse capítulo é apresentado o software *Web Application for Proficiency Testing Provider* (WAPT), desenvolvido neste trabalho. O software, que consiste em uma aplicação web, foi desenvolvido na plataforma Shiny/R e contém métodos estatísticos de avaliação de desempenho de laboratórios participantes de ensaio de proficiência e comparações interlaboratoriais.

R é um software livre para computação estatística e construção de gráficos. O software é mantido pela *R Foundation for Statistical Computing*. Trata-se de uma linguagem de programação completa, com orientação a objetos e suporte a interação com outras linguagens [53, 98].

O software R é gratuito, possui livre distribuição e código fonte aberto e é mantido por uma equipe internacional de pesquisadores. O código fonte do R está disponível sob a licença GNU GPL [53, 99].

O R é expansível com o uso dos pacotes, que são bibliotecas contendo funções destinadas a alguma finalidade específica. Os pacotes encontram-se disponíveis no site CRAN (*Comprehensive R Archive Network*) e são identificados por uma string alfanumérica [53, 98].

O Shiny é um sistema para desenvolvimento de aplicações web usando o R, um pacote do R e um servidor web. O código de uma aplicação Shiny permite estruturar tanto a interface com o usuário quanto o processamento de dados, geração de visualizações e modelagem, isto é, programa-se tanto o *user side* quanto o *server side* de uma só vez [53, 100].

No *server side*, processam-se requisições e dados do cliente, estruturam-se e enviam-se páginas web, interage-se com banco de dados etc. No *user side*, criam-se interfaces gráficas a partir dos códigos recebidos pelo servidor. É onde enviam-se e recebem-se as informações do *server side* [53, 101].

Neste contexto, ao rodar o código, cria-se um servidor que envia páginas web, recebe informações do usuário e processa os dados, utilizando apenas o R [101]. O

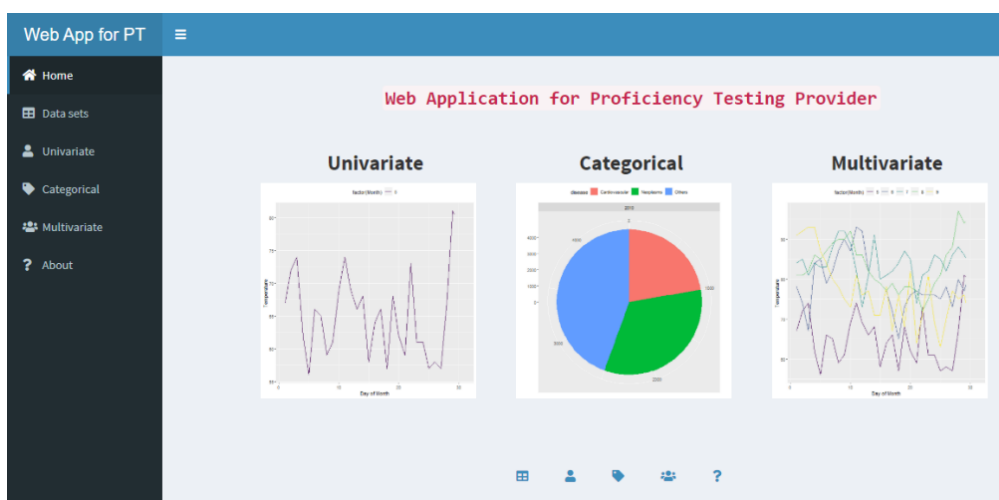
pacote *shiny* do R possui internamente um servidor web básico, geralmente utilizado para aplicações locais, permitindo somente uma aplicação por vez [53].

5.2 APLICATIVO

Web Application for Proficiency Testing Provider (WAPT) é um aplicativo web interativo e amigável para auxiliar provedores a avaliar o desempenho de laboratórios participantes de ensaios de proficiência e comparações interlaboratoriais. Teve-se por objetivo desenvolver uma aplicação web *point-and-click* para facilitar o uso dos métodos estatísticos relacionados na presente tese.

No presente momento, o software encontra-se na fase *release candidate* (versão pronta para ser lançada em que todas as funcionalidades foram especificadas, implementadas e testadas, ou seja, apresenta funções, interface e desempenho sem erros consideráveis). Foram necessários, aproximadamente, 3 anos entre testes de interface, algoritmos computacionais e *layout* de apresentação de resultados para alcançar a fase atual de desenvolvimento do software. Tem-se por objetivo deixá-lo disponível para uso via web no endereço eletrônico do INMETRO (<https://www.gov.br/inmetro/pt-br>).

Figura 6 – Aplicativo: *Home*.



Fonte: elaboração própria.

O aplicativo está estruturado em seis módulos (Figura 6). O módulo *Data sets* fornece arquivos para download com exemplos de *layout* de dados. O módulo *Univariate* possui métodos estatísticos para tratar medidas (réplicas verdadeiras) de

uma única variável. O módulo *Categorical* apresenta técnicas estatísticas quando os resultados reportados são uma classificação sobre o item de teste. O módulo *Multivariate* lida com múltiplas variáveis de cada participante. Por fim, o módulo *About* fornece ajuda para usuários iniciantes.

5.3 MÓDULOS

O modulo *Data sets* contém alguns exemplos de conjuntos de dados que apresentam o formato de entrada dos dados em cada um dos módulos do aplicativo. Os conjuntos de dados *Benzoic acid*, *NOx*, *Calcium*, *Captopril*, e *Metronidazole* foram gentilmente cedidos pelo INMETRO.

Alguns conjuntos de dados foram disponibilizados para permitir o usuário explorar todas as funcionalidades do software: *Small sample*, *Large sample*, *Ordinal*, *Genotoxicity*, e *Balanced data*.

É importante destacar que *Small sample* e *Large sample* refere-se à quantidade de itens de ensaio enviados a cada um dos laboratórios participantes. Considera-se pequena amostra (*Small sample*) quando o número de itens de ensaio enviados a cada laboratório está entre 57, inclusive, e 133, exclusive, unidades ($57 \leq n < 133$). Por outro lado, grande amostra (*Large sample*) consiste em 133 ou mais itens de ensaios enviados aos participantes ($n \geq 133$). A figura 18 do capítulo 6, seção 6.2.3, ilustra os limites retro citados.

Estes conjuntos de dados (*Small sample* e *Large sample*) podem utilizados nos testes de homogeneidade de proporção contidos no painel *Proportion tests* do aplicativo (Figura 9).

O conjunto de dados *Ordinal* se refere a um conjunto de dados categórico (qualitativo) em que os resultados reportados são provenientes de uma variável qualitativa ordinal. Este conjunto de dados foi concebido para permitir ao usuário utilizar as medidas de concordância ordinais disponíveis no painel *Agreement measures* do aplicativo (Figura 9).

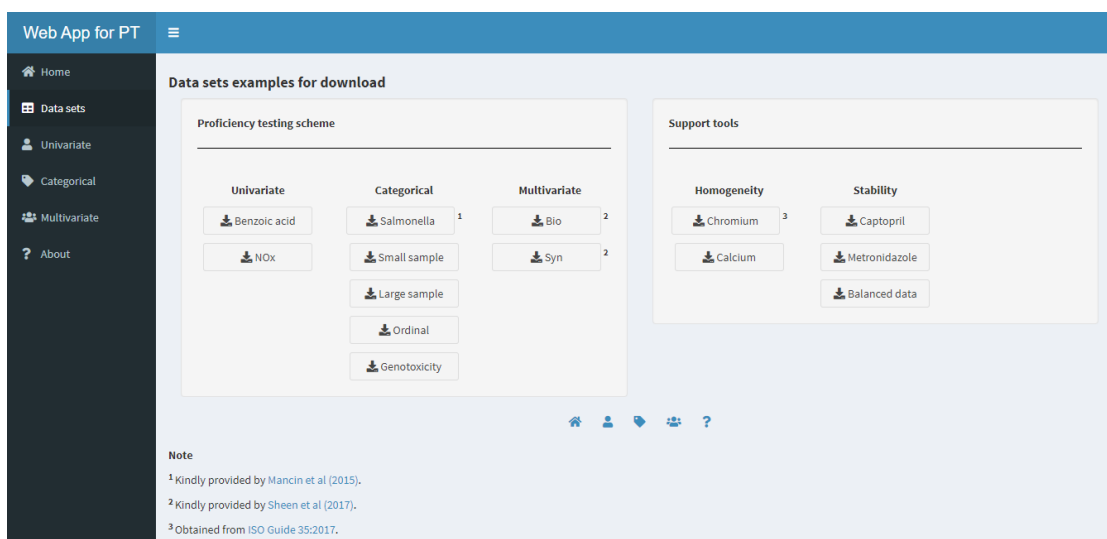
Genotoxicity consite em um conjunto dados o qual cada item de ensaio possui diferente nível de dificuldade para ser analisado e, neste contexto, atribui-se pesos aos itens de ensaio. Este conjunto de dados permite utilizar o coeficiente de Gower em que

se estabelece pesos para cada item de ensaio. Esta funcionalidade está disponível no no painel *Agreement measures* do aplicativo (Figura 9). Por fim, *Balanced data* é um conjunto de dados balanceado (mesmo número de medições em cada linha) para ser utilizado no painel *Support tools*, subpainel *Stability*, do aplicativo (Figura 8).

Marzia Mancin (2015) do Istituto Zooprofilattico Sperimentale delle Venezie gentilmente cedeu o conjunto de dados *Salmonella* [63] e David A. Sheen (2017) do National Institute of Standards and Technology [93] gentilmente cedeu os dados *Bio* (*biological*) e *Syn* (*syntenic*).

Por fim, o conjunto de dados *Chromium* (*in soil*) foi extraído da norma ISO Guide 35:2017 (Figura 7).

Figura 7 – Aplicativo: *Data sets*.



Fonte: elaboração própria.

No módulo *Univariate* há quatro painéis. Para que o software realize os cálculos é necessário carregar previamente os dados no painel *Load data*. Neste painel é automaticamente gerada uma tabela com as estatísticas descritivas de cada participante. Caso haja interesse o usuário pode fazer download desta tabela por meio do botão *download table* (Figura 8).

O segundo painel deste módulo é o *Performance statistics*. Neste painel há dois subpainéis: *Assigned value* e *Scores*. Caso o valor designado e/ou o desvio-padrão do ensaio de proficiência não estejam disponíveis é possível obtê-los no subpainel *Assigned value* o qual contém os métodos de estimação robusta (descritos na seção

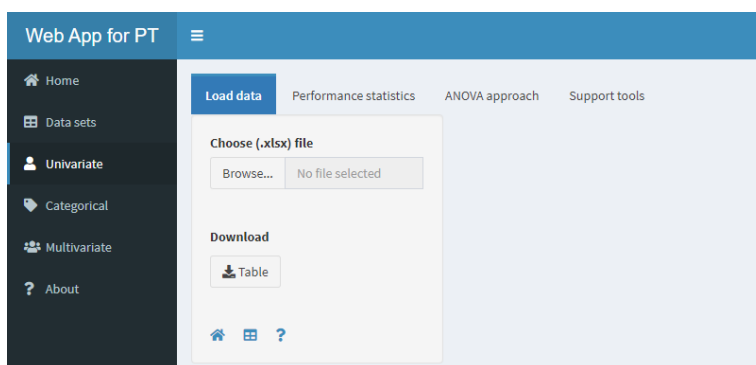
2.1.3 da presente tese) e a curva de Horwitz (seção 2.1.4). O subpainel *Scores* contém as estatísticas de desempenho presentes na norma ISO 13528:2015 e as estatísticas de desempenho alternativas mencionadas na seção 2.1.5.

No terceiro painel, *ANOVA approach*, pode-se utilizar a metodologia proposta nesta tese (Figura 2) para avaliação de desempenho dos laboratórios participantes. O gráfico dos resíduos fornece ao usuário elementos para definir o modelo mais adequado para analisar os resultados reportados.

O último painel, *Support tools*, contém métodos estatísticos para avaliar a homogeneidade e estabilidade do item de teste a ser enviados aos laboratórios participantes do ensaio de proficiência. O apêndice C da presente tese contém uma breve descrição dos métodos disponíveis neste painel.

Todos os painéis possuem botões de *download* em que é possível baixar um relatório, tabela ou gráfico dos resultados.

Figura 8 – Aplicativo: *Univariate*.



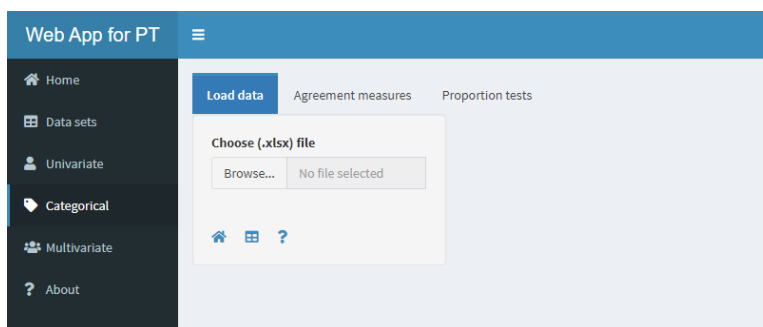
Fonte: elaboração própria.

No módulo *Categorical* (Figura 9) há três painéis: *Load data* (em que os dados são previamente carregados), *Agreement measures* e *Proportion tests*. O painel *Agreement measures* contém as medidas de concordância para resultados reportados em escala nominal e ordinal. Cabe destacar que o aplicativo apresenta as classificações do desempenho dos participantes sugeridas na seção 3.1 da presente tese.

O painel *Proportion tests* disponibiliza os métodos estatísticos concernentes à metodologia proposta na seção 3.2 desta tese para avaliar desempenho de laboratórios participantes de ensaios de proficiência cujos resultados reportados são categóricos. Este painel contém dois subpainéis, *Large sample* e *Small sample*, que se referem à

quantidade de itens de ensaio enviados aos participantes (Figura 3). No capítulo 6, seção 6.2.3, da presente tese são apresentadas algumas considerações sobre tamanhos amostrais considerados “grandes” e “pequenos”.

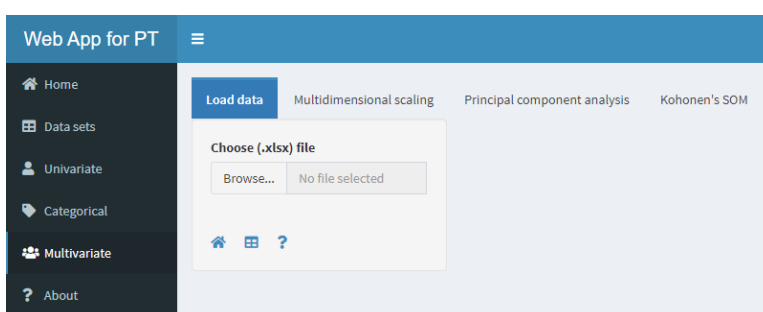
Figura 9 – Aplicativo: *Categorical*.



Fonte: elaboração própria.

O módulo *Multivariate* (Figura 10) contém quatro painéis. Após carregar os dados no painel *Load data*, o usuário pode escolher uma das técnicas multivariadas disponíveis no aplicativo. O painel *Multidimensional scaling* contém a metodologia proposta na seção 4.2 para avaliar o desempenho dos participantes quando os resultados reportados são provenientes de múltiplas variáveis (espectro no contexto da metrologia). Conforme já mencionado, a abordagem sugerida consiste no uso conjugado das técnicas multivariadas de escalonamento multidimensional e elipse de confiança robusta. O usuário pode baixar o relatório, a tabela ou o gráfico dos resultados obtidos. Este módulo também possui as técnicas de análise de componentes principais (*Principal component analysis*) e mapas auto-organizáveis de Kohonen (*Kohonen's SOM*) como ferramentas adicionais de análise (seções 4.1.1 e 4.1.2, respectivamente, da presente tese).

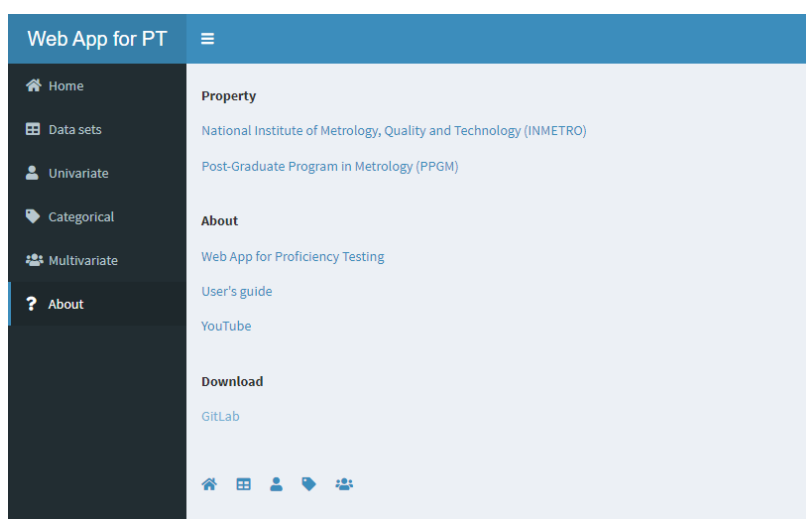
Figura 10 – Aplicativo: *Multivariate*.



Fonte: elaboração própria.

O módulo *About* contém informações sobre o aplicativo além de material para auxiliar usuários iniciantes (Figura 11). O software é de propriedade do Instituto Nacional de Metrologia, Qualidade e Tecnologia (INMETRO) e do Programa de Pós-graduação em Metrologia (PPGM). O link *Web App for Proficiency Testing* contém uma breve descrição do aplicativo. Este módulo disponibiliza um guia do usuário com instruções de uso do programa e um link para um canal do YouTube com vídeos tutoriais. Por fim, este módulo contém um link para o repositório GitLab em que é possível fazer download do aplicativo, para execução local, e acessar o código-fonte.

Figura 11 – Aplicativo: *About*.



Fonte: elaboração própria.

5.4 PACOTES

O software foi construído a partir de funções criadas pelo desenvolvedor do aplicativo e pacotes do R disponíveis no site CRAN. No que tange as bibliotecas do R, o aplicativo contém os seguintes pacotes:

(i) Pacotes para disponibilizar os recursos HTML do aplicativo: *shiny* (R Core Team, 2021), *shinydashboard* (Winston Chang e Barbara Borges Ribeiro, 2021), *Rmarkdown*, *knitr* e *DT* (Yihui Xie, 2022), *htmlwidgets* (Yihui Xie, J. J. Allaire, Joe Cheng, Carson Sievert, Kenton Russell e Ellis Hughes, 2021), *shinycssloaders* (Andras Sali, Luke Hass e Dean Attali, 2020) e *shinyBS* (Eric Bailey, 2022).

(ii) Pacote para upload de conjunto de dados e download de tabelas com resultados: *xlsx* (Adrian Dragulescu e Cole Arendt, 2020)

(iii) Pacotes para os estimadores robustos descritos na seção 2.1.3 da presente tese: O pacote *MASS* (Brian Ripley, Bill Venables, Douglas M. Bates, Kurt Hornik, Albrecht Gebhardt e David Firth, 2022) contém o M-estimador de Huber (algoritmo A) e o pacote *metrology* (Stephen L R Ellison, 2018) contém o Algoritmo S.

(iv) O pacote *stats* (R Core Team, 2021) contém: a análise de variância, o teste de Kruskal-Wallis, o teste de Shapiro-Wilk e o p-valor ajustado descritos na seção 2.2; o teste de qui-quadrado descrito na seção 3.2.1.1 e o escalonamento multidimensional descrito na seção 4.2.1 da presente tese.

(v) Testes de comparações múltiplas descritos na seção 2.2: O pacote *agricolae* (Felipe de Mendiburu, 2021) contém o teste de diferença mínima significativa (LSD); os pacotes *multcomp* (Torsten Hothorn, Frank Bretz, Peter Westfall, Richard M. Heiberger, Andre Schuetzenmeister e Susan Scheibe, 2022) e *nCDunnett* (Siomara Cristina Broch e Daniel Furtado Ferreira, 2015) permitem obter o teste de Dunnett e o pacote *PMCMRplus* (Thorsten Pohlert, 2022) contém o teste de Dunn.

(vi) O pacote *nlme* (José Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, Siem Heisterkamp, Bert Van Willigen, Johannes Ranke e R Core Team, 2022) contém o método de mínimos quadrados generalizados descrito na seção 2.2.3 da presente tese.

(vii) Medidas de concordância descritas na seção 3.1 da presente tese: O pacote *irr* (Matthias Gamer, 2019) contém o coeficiente de kappa, o pacote *FD* (Etienne Laliberté, Pierre Legendre e Bill Shipley, 2022) contém o coeficiente de Gower e o pacote *irrCAC* (Kilem L. Gwet, 2019) contém os coeficientes alfa de Krippendorff e de Gwet. Por fim, o pacote *expss* (Gregory Demin, 2022) auxilia na construção de tabelas de concordância.

(viii) Os pacotes *car* (John Fox, Sanford Weisberg, Brad Price, Daniel Adler, Douglas Bates, Gabriel Baud-Bovy, Ben Bolker, Steve Ellison, David Firth, Michael Friendly, Gregor Gorjanc, Spencer Graves, Richard Heiberger, Pavel Krivitsky, Rafael Laboissiere, Martin Maechler, Georges Monette, Duncan Murdoch, Henric Nilsson, Derek Ogle, Brian Ripley, William Venables, Steve Walker, David Winsemius, Achim Zeileis e R Core Team, 2022) e *pracma* (Hans W. Borchers, 2022) auxiliam,

respectivamente, na construção da elipse (seção 4.2.2) e da elipsoide (seção 4.2.3) de confiança robustas.

(ix) O pacote *scatterplot3d* (Uwe Ligges, 2018) permite a construção de gráficos em três dimensões para o relatório do aplicativo.

(x) Plotly é uma empresa de computação que desenvolve e fornece ferramentas online de análise e visualização de dados. A empresa disponibiliza interfaces para Python, R, MATLAB, Perl, Julia, Arduino e REST. A biblioteca de gráficos do R, *plotly* (Carson Sievert, Chris Parmer, Toby Hocking, Scott Chamberlain, Karthik Ram, Marianne Corvellec, Pedro Despouy e Salim Brüggemann, 2021), cria gráficos interativos com qualidade de publicação.

(xi) O pacote *kohonen* (Ron Wehrens e Johannes Kruisselbrink, 2022) contém os mapas auto-organizáveis de Kohonen descritos na seção 4.1.2 da presente tese.

(xii) O pacote *qcc* (Luca Scrucca, Greg Snow e Peter Bloomfield, 2017) permite construir os gráficos de controle descritos no apêndice B.2.1 da presente tese.

6. RESULTADOS

6.1 ENSAIO DE PROFICIÊNCIA UNIVARIADO

Os métodos de *scores*, descritos na presente tese, e a abordagem proposta baseada em análise de variância foram comparadas utilizando-se dois ensaios de proficiência (EP). O primeiro EP refere-se à quantificação de ácido benzóico em suco de laranja. Neste EP o valor designado foi disponibilizado pelo provedor do ensaio. O segundo EP trata de emissões veiculares em motores diesel. O valor designado foi obtido a partir do valor de consenso entre os participantes. Foram estimados o valor designado e o desvio-padrão robusto utilizando-se o M-estimador de Huber.

6.1.1 Ensaio de Proficiência em Sucos (2ª Rodada): Ácido Benzóico em Suco de Laranja

O ácido benzóico é um dos conservantes mais utilizados e permitidos pela legislação brasileira (e pelo Codex Alimentarius) para bebidas, incluindo os sucos de frutas. A concentração máxima permitida é de 1000 mg.kg⁻¹. O objetivo do ensaio de proficiência foi avaliar o desempenho dos laboratórios de análise de alimentos e bebidas na determinação do ácido benzóico em suco de laranja.

O item de teste consistia em suco de laranja comercial fortificado com uma massa conhecida de ácido benzóico. A concentração de ácido benzóico no item de ensaio encontrava-se na faixa de 100 a 1000 mg/L.

Todas as amostras foram analisadas quanto à técnica validada: análise por injeção em fluxo com espectrometria de massa acoplada (FIA-MS) pela técnica de espectrometria de massa de diluição isotópica. Curva analítica foi utilizada para quantificar as amostras [102].

A homogeneidade das amostras foi avaliada em 14 frascos selecionados aleatoriamente. De cada amostra (frasco) foi retirada uma sub-amostra de aproximadamente 1 mL. As amostras foram consideradas suficientemente homogêneas, pois o desvio-padrão entre as amostras ($s_s = 4,6$ mg/L) foi menor que $0,3\sigma_{pt} = 12,9$ mg/L conforme recomendado pela norma ISO 13528:2015 [4].

O estudo de estabilidade de curta duração foi realizado em 8 semanas nas temperaturas de 4 °C, 20 °C e 50 °C. Os modelos de regressão linear simples demonstraram a estabilidade do ácido benzóico em suco de laranja para as temperaturas de (4 ± 2) °C, (20 ± 3) °C e (50 ± 2) °C por 8, 6 e 5 semanas, respectivamente, ao nível de confiança de 95 %. No estudo de estabilidade de longa duração, realizado na temperatura de (-26 ± 3) °C pelo período de 7 meses, a concentração de ácido benzóico foi considerada estável ao nível de confiança de 95 %.

O valor designado $x_{pt} = 721$ mg/L é a média dos valores obtidos no estudo de homogeneidade (aproximadamente 692 mg.kg⁻¹) conduzido pelo INMETRO (laboratório de referência) e o desvio-padrão para avaliação de proficiência $\sigma_{pt} = 43,1$ mg/L foi obtido pela curva de Horwitz:

$$\sigma_{PT} = ((0,02(k \cdot 10^{-6})^{0.8495})10^6)d \quad \text{Equação 128}$$

Em que k a média dos valores obtidos no estudo de homogeneidade em mg.kg⁻¹ ($k = 692$) e $d = 1,042144$ g/cm³ é a densidade. A incerteza expandida do valor designado é $U(x_{pt}) = 74$ mg/L e o fator de abrangência é 4,3 o que fornece a incerteza-padrão do valor designado:

$$u(x_{pt}) = 74/4,3 \cong 17,2 \text{ mg/L} \quad \text{Equação 129}$$

Os resultados dos laboratórios participantes e do laboratório de referência são apresentados a seguir.

Tabela 17 – Resultados (em mg/L) dos laboratórios participantes do ensaio de proficiência: ácido benzóico em suco de laranja.

Lab	Medição1	Medição2	Medição3
04	125,7	125,7	125,7
27	722,8	721,1	721,4
39	830	806	782
41	529,7	527,8	529,8
44	605	599	602,6
59	593	593,5	592,7
61	802,5	798,9	800,1
63	676,3	675,5	680,2
69	733,2	711,4	711,5
77	632,1	645,7	654,4
83	723,2	722,2	717,3
88	715,9	751,8	761,1
98	711,8	714,6	712,9

Fonte: elaboração própria.

Tabela 18 – Resultados (em mg/L) do laboratório de referência: EP de ácido benzóico em suco de laranja.

Lab	Ampola 3	Ampola 9	Ampola 13	Ampola 21	Ampola 27	Ampola 32	Ampola 39
INMETRO	725,0845	720,3478	724,9162	729,1148	720,1843	727,1244	716,774

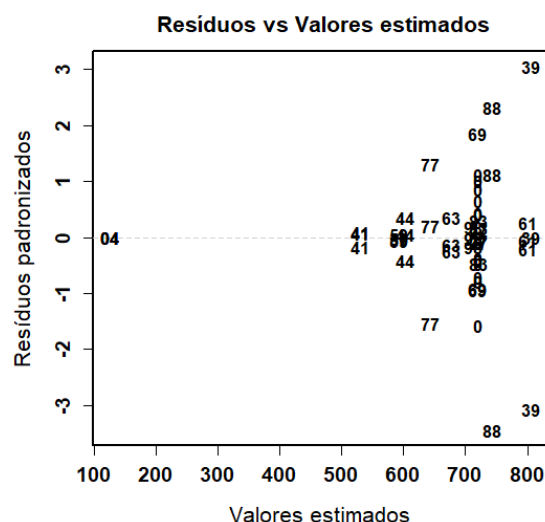
Lab	Ampola 46	Ampola 52	Ampola 58	Ampola 65	Ampola 76	Ampola 83	Ampola 88
INMETRO	731,4292	717,3313	717,5582	713,8603	706,6274	730,503	714,5047

Fonte: elaboração própria.

Na abordagem proposta para avaliar o desempenho dos participantes, sugere-se realizar a análise de resíduos (Figura 2) para identificar qual modelo é mais apropriado para analisar os dados. O gráfico dos resíduos, sob o modelo linear normal, apresenta indícios da presença de heterocedasticidade (Figura 12). O teste de Shapiro–Wilk indica que os resíduos não apresentam distribuição normal (p-valor 5,204E-05) e o

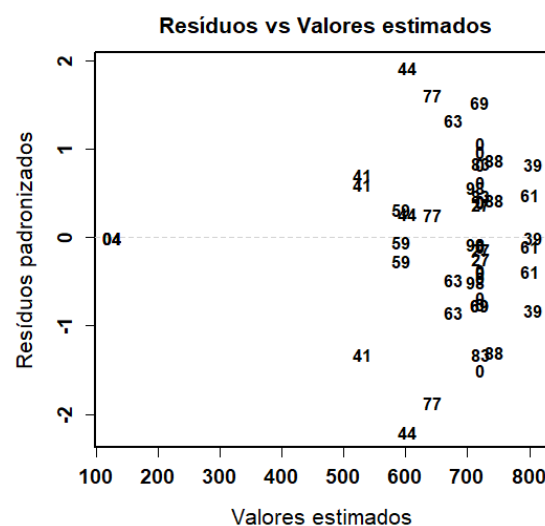
teste de Koenker–Bassett corrobora a análise gráfica dos resíduos indicando a presença de heterocedasticidade (p-valor 0,02982) considerando um nível de significância de 5 %. A adoção do modelo estimado por mínimos quadrados generalizados (FGLS) melhora o padrão dos resíduos (Figura 13). Este modelo atende aos pressupostos de normalidade (p-valor 0,9226) e homocedasticidade (p-valor 0,7125) dos resíduos.

Figura 12 – Gráfico dos resíduos do modelo linear normal.



Fonte: elaboração própria.

Figura 13 – Gráfico dos resíduos do modelo FGLS.



Fonte: elaboração própria.

O valor designado e sua respectiva incerteza, além do desvio-padrão para avaliação de proficiência, possibilitam calcular o zscore, o z'score, o Q-scoring e o zscore corrigido os quais podem ser comparados ao modelo estimado por mínimos

quadrados generalizados (FGLS). No modelo FGLS verificou-se, a partir do teste F , que os resultados reportados pelos participantes diferem entre si (p -valor $< 0,0001$). Utilizou-se o teste t do modelo FGLS para verificar quais os valores reportados pelos participantes não apresentaram diferença significativa, em média, em relação ao valor designado (hipótese nula). A hipótese nula de igualdade da média entre os valores reportados pelos participantes e o valor designado é rejeitada se o p -valor ajustado pelo método de Benjamini-Yekutieli é menor que o nível de significância estabelecido. O nível de significância a ser adotado para avaliar o desempenho dos participantes é uma prerrogativa do provedor do ensaio. Os resultados são apresentados na tabela 19.

Tabela 19 – Scores e p-valores do modelo FGLS: EP de ácido benzóico em suco de laranja.

Lab	Zscore	z'score	Q-scoring ^(a)	zscore corrigido	FGLS p-valor
4	-13,8084	-12,8244	-0,8257	-14,3722	3,41E-61
41	-4,4512	-4,1340	-0,2662	-4,6330	5,55E-42
59	-2,9675	-2,7560	-0,1774	-3,0887	5,32E-35
44	-2,7556	-2,5593	-0,1648	-2,8682	8,06E-34
77	-1,7845	-1,6574	-0,1067	-1,8574	6,65E-19
63	-1,0129	-0,9407	-0,0606	-1,0542	2,66E-17
98	-0,1833	-0,1702	-0,0110	-0,1907	0,05
69	-0,0534	-0,0496	-0,0032	-0,0555	1
83	-0,0023	-0,0022	-0,0001	-0,0024	1
27	0,0178	0,0165	0,0011	0,0185	1
88	0,5088	0,4725	0,0304	0,5295	0,343
61	1,8441	1,7126	0,1103	1,9194	4,71E-23
39	1,9716	1,8311	0,1179	2,0521	6,47E-05

^(a) Se o desvio for acima de 10 % (0,1) o resultado é considerado insatisfatório.

Fonte: elaboração própria.

A probabilidade de considerar os resultados reportados pelos laboratórios participantes como satisfatórios quando é provável que não sejam satisfatórios (erro tipo II) fica em torno de 60 % podendo chegar a 94 % no caso da estatística Q-scoring. O zscore corrigido apresenta o menor erro devido ao fato de ponderar o resultado pelo

número de participantes. O fator de ponderação mostra-se relevante devido ao pequeno o número de participantes (menos que 15) [7].

Tabela 20 – Erro tipo II para as estatísticas de desempenho: EP de ácido benzóico em suco de laranja.

Estatística de desempenho	Erro tipo II
zscore corrigido	0,5664804
Zscore	0,5791999
z'score	0,6026412
Q-scoring	0,9422521

Fonte: elaboração própria.

Na abordagem proposta (modelo FGLS) quando o p-valor ajustado é maior ou igual ao nível de significância α estabelecido não se pode rejeitar a hipótese nula, sendo assim considera-se que o resultado do laboratório é satisfatório, pois não há diferença estatisticamente significativa entre os valores reportados (em média) e o valor designado. Quando ocorre o inverso, ou seja, o p-valor ajustado é menor que α os resultados reportados pelos participantes são considerados insatisfatórios, pois diferem (em média) estatisticamente do valor designado.

Adotando-se um nível de significância de 5 % ($\alpha = 0,05$) pode-se observar na tabela 21 que: (i) apenas 5 laboratórios (98, 69, 83, 27 e 88) têm seus resultados classificados satisfatórios pelo modelo FGLS; (ii) resultados classificados como satisfatórios e questionáveis pelos métodos de *scores* são classificados como insatisfatórios pelo método FGLS. As divergências de interpretação observadas na tabela 21 podem decorrer da significativa probabilidade de ocorrência do erro tipo II nos métodos dos *scores*.

Tabela 21 – Comparação entre os resultados obtidos pelos *scores* e pela metodologia proposta: EP de ácido benzóico em suco de laranja.

Lab	<i>scores</i>				Metodologia proposta
	zscore	z'score	Q-scoring ¹	zscore corrigido	FGLS
4	Insatisfatório	Insatisfatório	Insatisfatório	Insatisfatório	Insatisfatório
41	Insatisfatório	Insatisfatório	Insatisfatório	Insatisfatório	Insatisfatório
59	Questionável	Questionável	Insatisfatório	Insatisfatório	Insatisfatório
44	Questionável	Questionável	Insatisfatório	Questionável	Insatisfatório
77	Satisfatório	Satisfatório	Insatisfatório	Satisfatório	Insatisfatório
63	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Insatisfatório
98	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
69	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
83	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
27	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
88	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
61	Satisfatório	Satisfatório	Insatisfatório	Satisfatório	Insatisfatório
39	Satisfatório	Satisfatório	Insatisfatório	Questionável	Insatisfatório

¹ Se o desvio for acima de 10 % (0,1) o resultado é considerado insatisfatório.

Fonte: elaboração própria.

Na tabela comparativa (tabela 21) observa-se que os métodos dos *scores* não foram adequados para avaliar o desempenho dos laboratórios, pois não apresentam ferramentas para tratar a heterocedasticidade presente nos dados. Pode-se destacar, por exemplo, o laboratório 63 que apresentou resultado satisfatório nos *scores* considerados, mas insatisfatório pela metodologia proposta. Isto decorre da não normalidade e heterocedasticidade dos dados. As estatísticas de desempenho não possuem ferramentas para lidar com estas características dos dados. Nestes casos é recomendável utilizar a metodologia proposta.

6.1.2 Ensaio de Proficiência de Emissões de Automóveis (10ª Rodada): Automóvel Diesel

A poluição do ar constitui uma ameaça à saúde, aumentando a ocorrência de doenças respiratórias e diminuindo a qualidade de vida. Os veículos automotores são potenciais emissores desse tipo de poluição, pois as emissões dos gases de combustão carregam diversas substâncias tóxicas que, em alguns casos, em contato com o sistema respiratório, podem produzir vários efeitos negativos sobre a saúde. Os veículos, motor a diesel, são responsáveis por emissões na atmosfera de, além de outros gases, NOx. O ensaio de proficiência teve a finalidade de avaliar o desempenho dos laboratórios na determinação da quantidade dos compostos presentes nas emissões veiculares.

O item de ensaio foi um veículo modelo Chevrolet Cobalt LTZ, motor 1.3L Diesel, transmissão manual de 5 velocidades e inércia equivalente de 1304 kg. Cada laboratório participante usou o seu próprio combustível (Diesel S-10 B0 padrão conforme Norma ABNT NBR 8689:2012). O NOx foi analisado conforme Norma ABNT NBR 6601:2012.

Os laboratórios reproduziram a curva de desaceleração em dinamômetro informada pelo laboratório de emissão da General Motors do Brasil e drenaram o combustível do tanque, para depois, abastecer com 25 L e realizar todos os ensaios previstos no ensaio de proficiência.

O laboratório da General Motors do Brasil realizou ensaios de estabilidade no início e no fim do ciclo. Pode-se afirmar que, ao nível de confiança de 95 %, não houve diferença estatisticamente significativa entre os resultados obtidos no início e no fim do ciclo. Neste contexto, o veículo se manteve íntegro durante a realização do ensaio de proficiência.

O valor designado foi obtido por meio do valor de consenso entre os resultados reportados pelos participantes. Inicialmente foram estimadas a média ($\hat{\mu}$) e o desvio-padrão robusto ($\hat{\sigma}$) por meio do M-estimador de Huber. As medições médias dos laboratórios que diferiam da média robusta em dois desvios robustos ($x_i < \hat{\mu} - 2\hat{\sigma}$ ou $x_i > \hat{\mu} + 2\hat{\sigma}$) foram consideradas discrepantes e retiradas do conjunto de dados. As estimativas robustas foram recalculadas pelo M-estimador de Huber com os valores remanescentes (média dos laboratórios). Obteve-se como valor de consenso (valor

designado) $x_{pt} = 0,45$ g/km e desvio-padrão para avaliação de proficiência $\sigma_{pt} = 0,021$ g/km. Os resultados reportados pelos participantes estão descritos na tabela 22.

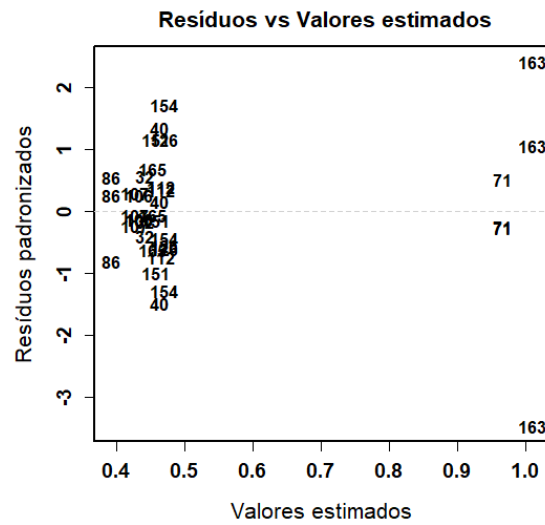
Tabela 22 – Resultados (em g/km) dos laboratórios participantes do ensaio de proficiência: emissões de NOx em automóvel diesel.

Lab	Medição1	Medição2	Medição3
32	0,435	0,452	0,439
40	0,438	0,467	0,488
71	0,9763523	0,9629032	0,96255
86	0,379	0,403	0,398
106	0,432	0,439	0,432
107	0,425	0,428	0,434
112	0,474	0,473	0,454
126	0,491	0,461	0,46
151	0,441	0,479	0,456
154	0,448	0,463	0,501
163	1,03	1,054	0,95
165	0,454	0,444	0,467

Fonte: elaboração própria.

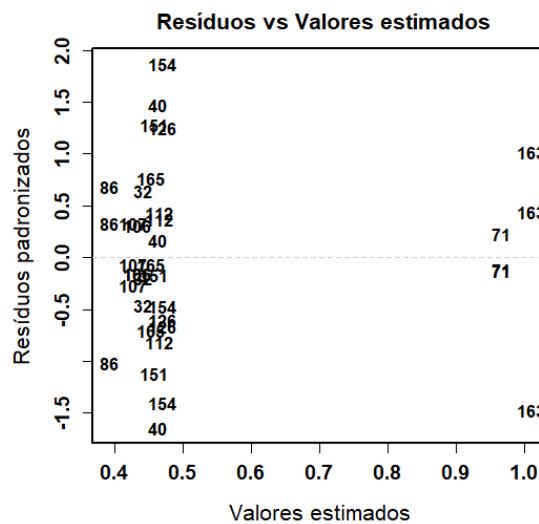
Na abordagem proposta (Figura 2), os resíduos podem ser considerados aproximadamente normais (teste de Shapiro-Wilk, p-valor 0,04897); entretanto, o gráfico dos resíduos do modelo linear normal (Figura 14) apresenta indícios de heterocedasticidade. O teste de Koenker–Bassett (p-valor 0,001874) indica a presença de heterocedasticidade considerando um nível de significância de 5 %. O modelo FGLS apresenta melhor padrão de resíduos (Figura 15) e os testes formais indicam a adequação do modelo aos pressupostos da análise da variância: (i) Shapiro-Wilk, p-valor 0,9325; (ii) Koenker–Bassett, p-valor 0,9619.

Figura 14 – Gráfico dos resíduos do modelo linear normal.



Fonte: elaboração própria.

Figura 15 – Gráfico dos resíduos do modelo FGLS.



Fonte: elaboração própria.

Neste ensaio de proficiência foram calculados o zscore, o Q-scoring, o zscore corrigido e o zscore robusto conforme tabela 23.

Tabela 23 – *Scores*: EP de emissões de NOx em automóvel diesel.

Lab	zscore	Q-scoring ^(a)	zscore corrigido	zscore robusto
86	-2,73829	-0,128082207	-2,8600461	-3,0067173
107	-1,04797	-0,049018475	-1,0945712	-1,4335205
106	-0,79522	-0,03719586	-0,830575	-1,1982761
32	-0,43188	-0,020200853	-0,4510804	-0,8601123
165	0,184219	0,008616769	0,1924105	-0,2867041
151	0,35799	0,016744817	0,3739079	-0,1249736
40	0,626544	0,029306344	0,6544039	0,1249736
112	0,752923	0,035217651	0,786402	0,2425958
126	0,926693	0,043345699	0,9678995	0,4043263
154	0,926693	0,043345699	0,9678995	0,4043263
71	24,46163	1,144182922	25,5493408	22,3085996
163	26,54995	1,241863208	27,7305191	24,2522255

^(a) Se o desvio for acima de 10 % (0,1) o resultado é considerado insatisfatório.

Fonte: elaboração própria.

O erro tipo II fica em torno de 68 % podendo chegar a 92 %. O zscore corrigido apresenta o menor erro devido ao pequeno o número de participantes (menos que 15) o que torna o fator de ponderação não negligenciável [7].

Tabela 24 – Erro tipo II para as estatísticas de desempenho: EP de emissões de NOx em automóvel diesel.

Estatística de desempenho	Erro tipo II
zscore corrigido	0,6762724
zscore robusto	0,6773564
zscore	0,683517
Q-scoring	0,9209691

Fonte: elaboração própria.

No modelo FGLS verificou-se, a partir do teste *F*, que os resultados reportados pelos participantes diferem entre si (p-valor < 0,0001). O teste *t* do modelo FGLS foi empregado em todas as combinações duas a duas de laboratórios e os pares de laboratórios que apresentaram diferença estatisticamente significativa (71, 163 e 86)

estão relacionados nas tabelas 25 a 27. Cabe mencionar que “*Rej. H0*” significa que os resultados reportados pelo laboratório *i* diferem dos resultados reportados pelo laboratório *j*. Todos os p-valores foram ajustados pelo método de Benjamini-Yekutieli.

Tabela 25 – Comparações múltiplas duas a duas do laboratório 71 com os demais: EP de emissões de NOx em automóvel diesel.

Par	t_{obs}	p-valor	Decisão
32-71	21,83	4,03E-16	Rej. H0
40-71	20,79	1,23E-15	Rej. H0
71-86	-24,12	8,18E-17	Rej. H0
71-106	-22,19	1,85E-16	Rej. H0
71-107	-22,44	1,85E-16	Rej. H0
71-112	-20,67	3,42E-16	Rej. H0
71-126	-20,49	3,42E-16	Rej. H0
71-151	-21,05	3,08E-16	Rej. H0
71-154	-20,49	3,42E-16	Rej. H0
71-165	-21,23	3,07E-16	Rej. H0

Fonte: elaboração própria.

Tabela 26 – Comparações múltiplas duas a duas do laboratório 163 com os demais: EP de emissões de NOx em automóvel diesel.

Par	t_{obs}	p-valor	Decisão
32-163	22,12	4,03E-16	Rej. H0
40-163	21,15	1,23E-15	Rej. H0
106-163	22,46	2,77E-16	Rej. H0
107-163	22,69	2,14E-16	Rej. H0
112-163	21,04	1,41E-15	Rej. H0
126-163	20,88	1,71E-15	Rej. H0
151-163	21,40	9,24E-16	Rej. H0
154-163	20,88	1,71E-15	Rej. H0
163-165	-21,56	2,13E-16	Rej. H0
86-163	24,24	4,09E-17	Rej. H0

Fonte: elaboração própria.

Tabela 27 – Comparações múltiplas duas a duas do laboratório 86 com os demais: EP de emissões de NOx em automóvel diesel.

Par	t_{obs}	p-valor	Decisão
32-86	-4,03	5,39E-03	Rej. H0
40-86	-5,76	6,91E-05	Rej. H0
86-106	3,42	7,44E-03	Rej. H0
86-107	2,99	1,92E-02	Rej. H0
86-112	5,96	2,53E-05	Rej. H0
86-126	6,23	1,62E-05	Rej. H0
86-151	5,33	8,68E-05	Rej. H0
86-154	6,23	1,62E-05	Rej. H0
86-165	5,04	1,54E-04	Rej. H0

Fonte: elaboração própria.

A partir do p-valor das comparações múltiplas duas a duas observa-se que: (i) os laboratórios 71 e 163 diferem dos demais, mas não entre si; (ii) o laboratório 86 difere de todos os demais. Pode-se considerar os laboratórios 71, 86 e 163 como insatisfatórios. Insatisfatório neste contexto significa que o resultado do laboratório é discrepante em relação ao demais.

Tabela 28 – Comparação entre os resultados obtidos pelos *scores* e pela metodologia proposta: EP de emissões de NOx em automóvel diesel.

Lab	<i>scores</i>				Metodologia proposta
	zscore	Q-scoring ^(a)	zscore corrigido	zscore robusto	FGLS
86	Questionável	Insatisfatório	Questionável	Insatisfatório	Insatisfatório
107	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
106	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
32	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
165	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
151	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
40	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
112	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
126	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
154	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
71	Insatisfatório	Insatisfatório	Insatisfatório	Insatisfatório	Insatisfatório
163	Insatisfatório	Insatisfatório	Insatisfatório	Insatisfatório	Insatisfatório

^(a) Se o desvio for acima de 10 % (0,1) o resultado é considerado insatisfatório.

Fonte: elaboração própria.

Na tabela 28 cabe destacar que laboratório 86 foi classificado como questionável pelos métodos de zscore e zscore corrigido e insatisfatório pelos métodos de Q-scoring e zscore robusto. Essa divergência de conclusões pode ser decorrente da probabilidade do erro tipo II, nos métodos dos *scores*, estar próxima de 70 %. Na metodologia proposta as comparações múltiplas indicam que a média dos resultados reportados pelo laboratório 86 difere estatisticamente da média dos resultados de todos os demais participantes. Essa diferença indica que este laboratório apresenta resultados atípicos (discrepantes) em relação aos demais, em média, e seu resultado é classificado como insatisfatório.

6.2 ENSAIO DE PROFICIÊNCIA CATEGÓRICO

6.2.1 Ensaio de Proficiência de Sorotipos de Salmonela (dados reais)

Salmonela é uma bactéria que contamina alimentos e é a principal causa de intoxicação alimentar no mundo. A correta identificação do sorotipo é essencial para definir o tratamento mais adequado [63]. O Istituto Zooprofilattico Sperimentale delle Venezie conduziu um ensaio de proficiência em que 19 sorotipos de salmonela foram enviados a 12 participantes. O objetivo do ensaio era verificar a capacidade do laboratório participante de identificar corretamente o sorotipo enviado.

As cepas de salmonella foram inoculadas em um tubo de ágar nutriente de acordo com os procedimentos padrão utilizados pelo provedor do ensaio para obter 15 cópias clonais de cada cepa. Cada cópia, de cada cepa, foi sorotipada novamente pelo provedor do ensaio, imediatamente antes de serem preparadas e enviadas aos laboratórios participantes [63].

Foi solicitado aos laboratórios utilizar o método de sorotipagem realizado rotineiramente. O método de sorotipagem para a realização dos testes de aglutinação é específico para a marca de antissoros utilizada e está especificado nas instruções fornecidas por cada fabricante comercial de soro. As cepas de salmonella foram identificadas de acordo com o esquema Kauffmann-White [63].

Figura 16 – Parte dos resultados reportados por cada participante: EP de sorotipos de salmonela (dados reais).

	A	B	C	D
1	Participant code	Item 1	Item 2	Item 3
2	Expected result	Monophasic variant of <i>S. Typhimurium</i>	<i>S. Orion</i>	<i>S. Virchow</i>
3	Lab 1	Monophasic variant of <i>S. Typhimurium</i>	<i>S. Orion</i>	<i>S. Othmarschen</i>
4	Lab 2	Monophasic variant of <i>S. Typhimurium</i>	<i>S. Orion</i>	Not identify
5	Lab 3	Monophasic variant of <i>S. Typhimurium</i>	<i>S. Orion</i>	<i>S. Virchow</i>
6	Lab 4	Monophasic variant of <i>S. Typhimurium</i>	<i>S. Orion</i>	<i>S. Virchow</i>
7	Lab 5	Monophasic variant of <i>S. Typhimurium</i>	<i>S. Orion</i>	<i>S. Virchow</i>
8	Lab 6	Monophasic variant of <i>S. Typhimurium</i>	<i>S. Orion</i>	<i>S. Thompson</i>
9	Lab 7	Monophasic variant of <i>S. Typhimurium</i>	<i>S. Orion</i>	<i>S. Virchow</i>
10	Lab 8	Monophasic variant of <i>S. Typhimurium</i>	<i>S. Orion</i>	<i>S. Virchow</i>
11	Lab 9	Monophasic variant of <i>S. Typhimurium</i>	<i>S. Orion</i>	<i>S. Virchow</i>
12	Lab 10	Monophasic variant of <i>S. Typhimurium</i>	<i>S. Orion</i>	<i>S. Virchow</i>
13	Lab 11	Monophasic variant of <i>S. Typhimurium</i>	<i>S. Orion</i>	<i>S. Anatum</i>
14	Lab 12	Monophasic variant of <i>S. Typhimurium</i>	<i>S. Orion</i>	<i>S. Virchow</i>

Fonte: elaboração própria.

O provedor do ensaio detinha o resultado esperado de cada item de teste (cepa de salmonela) enviado aos participantes. O resultado reportado por cada participante era a classificação do sorotipo de salmonela (Figura 16). Nos resultados reportados pelos participantes (tabela 29) tem-se que “acerto” é quando o laboratório identificou corretamente o sorotipo de salmonela e “erro” é quando não identificou corretamente.

Houve casos em que o laboratório não conseguiu identificar o sorotipo enviado. Observa-se que o laboratório 6 apresentou o menor percentual de classificações concordantes com o resultado esperado (aproximadamente 79 %).

Tabela 29 – Resultados reportados pelos participantes: EP de sorotipos de salmonela (dados reais).

Laboratório	Erro	Acertos	Não identificou
Lab 1	5,3 %	89,5 %	5,3 %
Lab 2	5,3 %	84,2 %	10,5 %
Lab 3	0 %	100 %	0 %
Lab 4	5,3 %	94,7 %	0 %
Lab 5	0 %	100 %	0 %
Lab 6	15,8 %	78,9 %	5,3 %
Lab 7	0 %	100 %	0 %
Lab 8	10,5 %	89,5 %	0 %
Lab 9	5,3 %	94,7 %	0 %
Lab 10	0 %	100 %	0 %
Lab 11	5,3 %	89,5 %	5,3 %
Lab 12	0 %	100 %	0 %

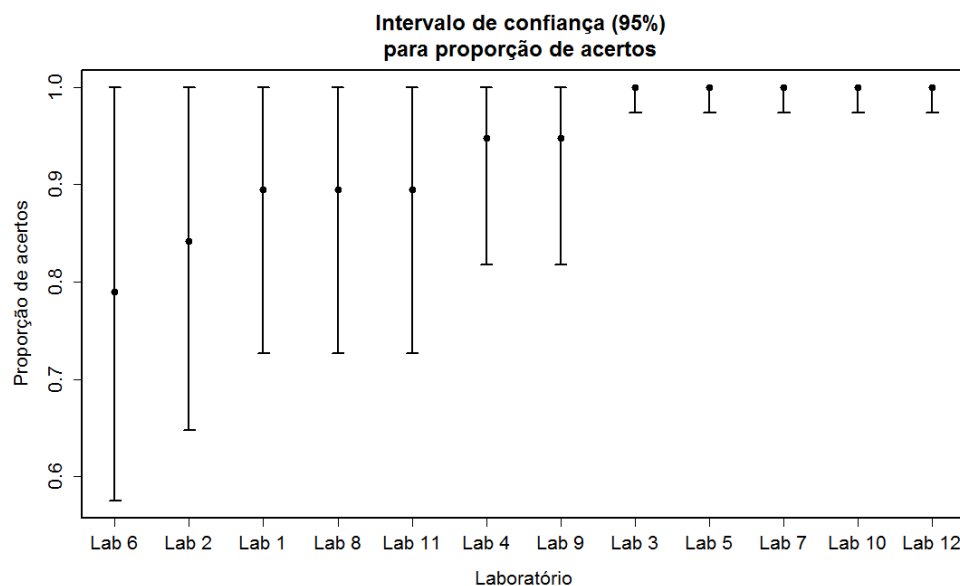
Fonte: elaboração própria.

No gráfico a seguir, foram construídos intervalos com 95 % de confiança para a proporção de classificações em concordância com o resultado esperado. Os intervalos de confiança foram construídos a partir da equação

$$p \pm z_{\alpha/2} \sqrt{pq/(n-1)} \pm 1/2n \quad \text{Equação 130}$$

Em que p é o percentual de acertos, $q = 1 - p$ é o percentual de “não acertos”, n é a quantidade de categorias e $1/2n$ é a correção de continuidade para tamanhos amostrais pequenos. Os intervalos foram construídos assumindo-se que o procedimento de seleção das cepas de salmonela foi por amostragem aleatória simples. É um procedimento de seleção que garante que todas as amostras de tamanho n têm a mesma probabilidade de serem escolhidas [103]. Pode-se observar que o laboratório 6 apresenta a maior dispersão.

Figura 17 – Intervalo de confiança para o percentual de acertos: EP de sorotipos de salmonela (dados reais).



Fonte: elaboração própria.

O coeficiente de Gower (S_{ij}) está acima de 0,7 para todos os participantes (tabela 30). Pela classificação sugerida nesta tese, todos os laboratórios são classificados como satisfatórios. Todos os laboratórios apresentam coeficiente kappa de Cohen estatisticamente diferente de zero considerando um nível de significância de 5 % (p-valor menor que 0,05). Pela classificação de Landis e Koch (1977), quase todos os laboratórios apresentam grau de concordância perfeito ou quase perfeito com o resultado esperado. A exceção é o laboratório 6, o qual apresentou grau de concordância substancial (tabela 30). Pela classificação sugerida nesta tese, todos os laboratórios têm resultados satisfatórios. Os coeficientes alfa de Krippendorff e Gwet (γ) estão acima de 0,7 para todos os participantes (tabela 30) e pela classificação sugerida apresentam resultados considerados satisfatórios.

Tabela 30 – Medidas de concordância para cada um dos participantes: EP de sorotipos de salmonela (dados reais).

Laboratório	Gower	kappa de Cohen	Krippendorff	Gwet
Lab 1	0,8947	0,8895	0,8921	0,8895
Lab 2	0,8421	0,8348	0,8382	0,8342
Lab 3	1	1	1	1
Lab 4	0,9474	0,9446	0,9460	0,9446
Lab 5	1	1	1	1
Lab 6	0,7895	0,7797	0,7846	0,7795
Lab 7	1	1	1	1
Lab 8	0,8947	0,8895	0,8921	0,8895
Lab 9	0,9474	0,9446	0,9460	0,9446
Lab 10	1	1	1	1
Lab 11	0,8947	0,8895	0,8921	0,8895
Lab 12	1	1	1	1

Fonte: elaboração própria.

No que se refere a abordagem proposta (Figura 3), optou-se pelo teste de Cohen, pois a tabela de contingência apresenta frequências esperadas menores que 5. O teste indica que não há diferença estatisticamente significativa (p-valor 0,2216) entre as proporções de classificações em consonância com o resultado esperado (proporções de acerto). Em outras palavras, há homogeneidade das proporções o que indica que os todos os laboratórios apresentam resultados satisfatórios segundo a metodologia proposta. Resultado satisfatório significa que o percentual de acerto não difere estatisticamente nem do valor designado (resultado esperado) nem entre os participantes do ensaio de proficiência.

6.2.2 Ensaio de Proficiência de Sorotipos de Salmonela (dados simulados)

Para fins de análise, foram simulados resultados a partir dos dados do ensaio de proficiência sobre a capacidade dos laboratórios em identificar sorotipos de salmonela.

Mais de 2500 diferentes sorotipos foram identificados [63] e simulou-se um ensaio em que 114 sorotipos foram enviados para os laboratórios participantes. Nesta simulação preservou-se os mesmos percentuais de acerto obtidos no ensaio conduzido pelo Istituto Zooprofilattico Sperimentale delle Venezie (dados reais).

A manutenção destes percentuais conduziu aos mesmos resultados obtidos pelos coeficientes de Gower, kappa de Cohen e Gwet (tabela 30). O coeficiente alfa de Krippendorff difere, em média, 0,2 % dos valores obtidos com os dados reais. Essa pequena diferença não influenciou na classificação sugerida e todos os laboratórios tiveram seus resultados considerados satisfatórios em todas as medidas de concordância consideradas nesta tese. O mesmo não ocorreu com o teste de Cohen que produziu conclusões diferentes das observadas nos dados reais. No teste de Cohen observa-se que não há homogeneidade entre as proporções, ou seja, há diferença estatisticamente significativa ($p\text{-valor} < 2,2 \cdot 10^{-16}$) entre as proporções de classificações em consonância com o resultado esperado (proporções de acerto).

Nas comparações múltiplas duas a duas (tabela 31) observa-se que a proporção de classificações corretas de sorotipos de salmonela do laboratório 6 difere estatisticamente, ao nível de significância de 5 %, de 5 dos 12 participantes (laboratórios 3, 5, 7, 10 e 12). Há indícios de que o laboratório 6 apresenta resultados discrepantes.

Tabela 31 – Teste de Cohen (comparações múltiplas): EP de sorotipos de salmonela (dados simulados).

Par	Diferença	Valor crítico	p-valor	Decisão
Expected result-Lab 6	0,859666	0,6073536	0,002038242	Rej. H0
Lab 3-Lab 6	0,859666	0,6073536	0,002038242	Rej. H0
Lab 5-Lab 6	0,859666	0,6073536	0,002038242	Rej. H0
Lab 6-Lab 7	0,859666	0,6073536	0,002038242	Rej. H0
Lab 6-Lab 10	0,859666	0,6073536	0,002038242	Rej. H0
Lab 6-Lab 12	0,859666	0,6073536	0,002038242	Rej. H0

Fonte: elaboração própria.

Na abordagem proposta (Figura 3), o resultado do laboratório 6 é classificado como insatisfatório. Na comparação entre os métodos (tabela 32), há divergência de classificação apenas para o laboratório 6. Há indícios de que a metodologia proposta é

mais propensa a captar diferenças. Os resultados evidenciam que o tamanho amostral pode impactar a análise quando a metodologia proposta é empregada.

Tabela 32 – Interpretação dos resultados obtidos pelas medidas de concordância e pelo teste de Cohen: EP de sorotipos de salmonela (dados simulados).

Laboratório	Gower	kappa de Cohen	Krippendorff	Gwet	Cohen
Lab 1	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
Lab 2	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
Lab 3	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
Lab 4	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
Lab 5	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
Lab 6	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Insatisfatório
Lab 7	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
Lab 8	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
Lab 9	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
Lab 10	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
Lab 11	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório
Lab 12	Satisfatório	Satisfatório	Satisfatório	Satisfatório	Satisfatório

Fonte: elaboração própria.

6.2.3 Tamanho amostral para dados categóricos

Medidas de concordância e testes de propoção

Dados categóricos podem ser expressos por meio de proporções [9, 12] e, nos ensaios de proficiência, essas proporções referem-se ao percentual de classificações em consonância com o resultado esperado. Nos ensaios de proficiência categóricos, há interesse em estimar a proporção de classificações corretas de cada um dos participantes. Para estimar essas proporções faz-se necessário definir o tamanho mínimo de amostra para obter esse percentual com algum grau de confiança.

Conforme já mencionado, o tamanho da amostral se refere à quantidade de itens de ensaio enviados aos laboratórios participantes. Supondo-se que um provedor do

ensaio de proficiência tenha uma população “virtualmente infinita” de itens de ensaio (teoricamente não há limite para a quantidade de itens que um provedor pode produzir) e considerando-se uma amostragem aleatórios sem reposição destes itens, tem-se a seguinte equação para definir tamanho amostral mínimo (n) para estimar a proporção de acerto de cada participante:

$$n = \frac{\left(z_{\alpha/2}\right)^2 \cdot \frac{NPQ}{N-1}}{\varepsilon^2 + \left(z_{\alpha/2}\right)^2 \cdot \frac{PQ}{N-1}} \quad \text{Equação 131}$$

Em que P é a proporção populacional ($Q = 1 - P$), $z_{\alpha/2}$ é o quantil da distribuição normal padrão e N é o tamanho populacional. Por fim, α é o nível de significância e ε representa o erro máximo admissível (ambos definidos pelo provedor do ensaio) [103].

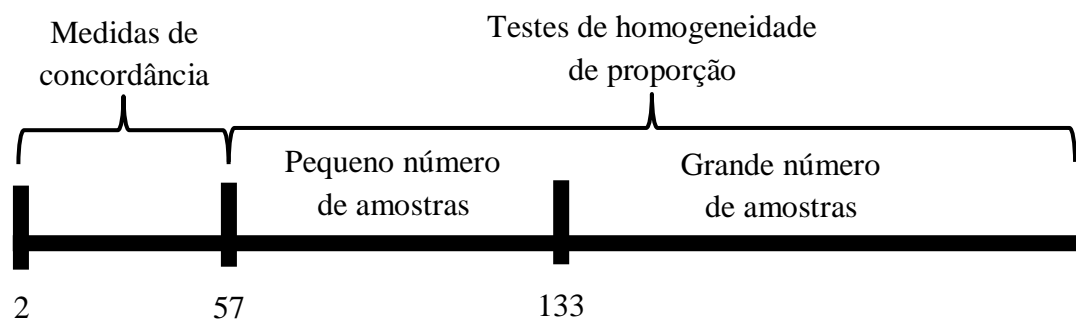
Como não há informação prévia sobre a proporção populacional considera-se $P = Q = 0,5$ [103]. Supondo-se que, com 95 % de confiança ($\alpha = 0,05$), o provedor do ensaio deseja que a proporção seja estimada com erro amostral máximo de 5 pontos percentuais para mais ou para menos ($\varepsilon = 0,05$), faz-se necessário que a quantidade mínima de itens de ensaio enviados para cada laboratório participante seja de 385.

Por questões técnicas inerentes a cada técnica analítica, e por causa de custos elevados, a produção desta quantidade de itens de ensaio não é viável. Por outro lado, embora haja menção, não há uma definição explícita do que é considerado um grande ou pequeno número de amostras (Krippendorff, 2004; Marascuilo, 1966; Cohen, 1967). Dentro deste contexto, foram realizadas simulações variando o percentual de acerto dos participantes para avaliar a capacidade dos testes de homogeneidade de proporção, assim como das medidas de concordância, em captar as diferenças entre os participantes (identificar proporções discrepantes).

Por se tratar de dados categóricos em que a variável qualitativa é nominal, os dados foram simulados do seguinte modo: (i) variou-se a quantidade de colunas (entre 2 e 133) que representam os itens de ensaio; (ii) nas linhas, que representam os resultados esperados e os resultados dos participantes, foram atribuídos os rótulos referentes as cepas de salmonela para cada item de ensaio. O layout dos dados simulados seguiu o layout apresentado na figura 16. Os resultados possibilitaram chegar as seguintes conclusões:

- (i) Nas amostras inferiores a 57 itens de ensaios ($2 \leq n < 57$) as medidas de concordância, em especial o coeficiente de Gwet, apresentaram melhor desempenho em captar as diferenças entre os participantes quando comparadas aos testes de homogeneidade de proporção.
- (ii) Nas amostras entre 57 inclusive e 133 itens exclusive ($57 \leq n < 133$) o teste de Cohen (*omnibus* e *post hoc*) mostrou maior capacidade em identificar percentuais atípicos quando comparado às medidas de concordância e aos testes de qui-quadrado e Marascuilo. Estes resultados fornecem indícios de que quantidades de itens dentro destes limites podem ser consideradas pequena, ou seja, pequeno número de amostras.
- (iii) Os testes de qui-quadrado e Marascuilo apresentaram melhor desempenho em identificar proporções atípicas quando a quantidade de itens de ensaios enviados aos participantes foi maior ou igual a 133 ($n \geq 133$). Os resultados fornecem evidências de que quantidade de itens acima deste valor podem ser consideradas como grande número de amostras.

Figura 18 – Quantidade de itens de ensaio enviados aos participantes de um ensaio de proficiência categórico.



Fonte: elaboração própria.

Na figura 18 observa-se os tamanhos amostrais em que as medidas de concordância ($2 \leq n < 57$) e os testes de homogeneidade de proporção ($n \geq 57$) são adequados. Os intervalos sobre o que pode ser considerado pequeno ($57 \leq n < 133$) e grande ($n \geq 133$) número de amostras também é considerado na figura 18.

É importante destacar que tamanhos amostrais considerados pequenos nos testes de homogeneidade de proporção ($57 \leq n < 133$) podem não ser viáveis para os laboratórios participantes do EP devido a questões de tempo e custo. Neste contexto,

quando a quantidade de itens de ensaio enviados aos participantes for inferior a 57 unidades recomenda-se utilizar a medida de concordância α de Krippendorff por fornecer grau de concordância mesmo quando há valores ausentes e/ou quando há um pequeno número de amostras [79].

Testes globais de homogeneidade de proporção

A metodologia proposta para avaliar o desempenho de laboratórios participantes de um ensaio de proficiência com dados categóricos baseia-se em testes globais de homogeneidade de proporção. Na presente tese foram sugeridos, para esta finalidade, o teste de Cohen (*omnibus*) quando há um pequeno número de amostras e o teste de qui-quadrado nos casos em que há uma grande quantidade de itens de ensaio. A opção por estes testes em detrimento ao teste exato de Fisher decorre dos seguintes motivos:

- (i) Segundo Modak (2008), o teste exato de Fisher é recomendado em tabelas de contingência com duas linhas ($r = 2$) e duas colunas ($c = 2$) [104]. Os ensaios de proficiência com dados categóricos geralmente têm mais do que dois laboratórios participantes ($r > 2$) e, eventualmente, podem ter mais do que 3 categorias ($c > 2$).
- (ii) Segundo Gregg (2002), o teste exato de Fisher é recomendado quando o tamanho amostral é inferior a 30 itens e/ou a frequência esperada é inferior a 5 em cada célula da tabela de contingência [105]. No que se refere aos ensaios de proficiência com dados categóricos, os dados simulados mencionados anteriormente forneceram indícios de que, para amostras inferiores a 57 itens, as medidas de concordância são mais adequadas para avaliar o desempenho dos laboratórios participantes.
- (iii) Segundo Gregg (2002), o teste de qui-quadrado fornece uma razoável aproximação para o teste exato de Fisher nos casos em que o tamanho amostral é de, pelo menos, 30 itens e a frequência esperada é superior a 5 em cada célula da tabela de contingência [105]. Considerando o contexto de ensaios de proficiência com dados categóricos, o teste de qui-quadrado é recomendado para amostras maiores ou iguais a 133 conforme retro citado.
- (iv) Embora o teste exato de Fisher seja recomendado para frequências esperadas inferiores a 5 itens e pequenas amostras ($n < 30$) os dados simulados no ensaio de proficiência categórico forneceram evidências de que o teste não é recomendado para

avaliar desempenho dos laboratórios participantes. Observou-se que, para tamanhos amostrais de pelo menos 8 itens $n \geq 8$, houve casos em que foi necessário calcular o p-valor do teste exato de Fisher utilizando simulação de Monte Carlo [53]. Esse procedimento demanda esforço computacional superior para obter resultados equivalentes aos obtidos pelos testes globais de Cohen ($57 \leq n < 133$) e qui-quadrado ($n \geq 133$). Cabe ainda observar que o teste de Cohen não apresenta restrição no que se refere a frequência esperada 5 em cada célula da tabela de contingência.

6.3 ENSAIO DE PROFICIÊNCIA MULTIVARIADO

6.3.1 Comparação interlaboratorial de RMN em peixe

Diversos requisitos devem ser atendidos para que a metabolômica baseada em ressonância magnética nuclear (RMN) e a técnica metabólica relacionada possam ser adotadas no monitoramento ambiental e na avaliação de risco químico [106].

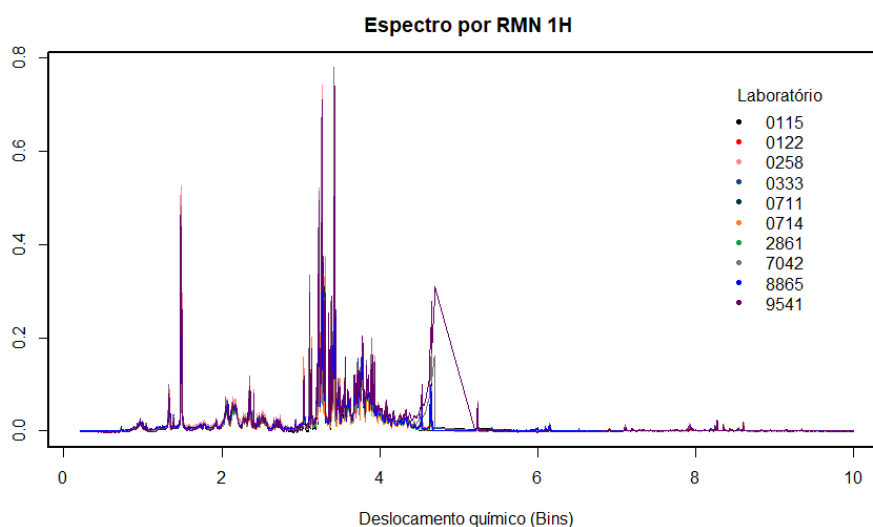
Realizou-se uma comparação interlaboratorial (CI) para avaliar a eficácia da metabolômica RMN ^1H em gerar conjuntos de dados comparáveis a partir de amostras derivadas do meio ambiente. Sete laboratórios dos Estados Unidos, um do Canadá, um do Reino Unido e um da Austrália participaram da intercomparação [106].

Nesta CI, analisaram-se misturas de metabólitos sintéticos e amostras de origem biológica de extratos de fígado de linguado europeu de locais limpos e contaminados. As misturas sintéticas foram preparadas com várias concentrações de glicose, citrato, fumurato, glutamina, alanina e nicotinato além de serem preparadas em tampão de fosfato de sódio 100 mmol/L (pH 7,0), transferidas quantitativamente para tubos de microcentrífuga, secas em um concentrador centrífugo e seladas com parafilme (106).

O linguado europeu adulto feminino (*Platichthys flesus*) foi coletado na foz dos rios Alde (local de controle não poluído) e Tyne (local poluído) no Reino Unido. As amostras dos tecidos do fígado foram imediatamente dissecadas, congelados em nitrogênio líquido e armazenados a $-80\text{ }^{\circ}\text{C}$ até a extração por um laboratório. As amostras foram extraídas, cada uma, usando um método metanol: clorofórmio: água e homogeneizador baseado em esferas Precellys-24 [106].

Cada participante forneceu como resultado um espectro por RMN ^1H . Os espectros foram reportados com deslocamento químico variando de 10 a 0,2 partes por milhão (ppm). A região de 4,7 a 5,2 ppm foi excluída devido artefatos de supressão de solvente de água e o espectro foi renormalizado. Os espectros foram segmentados em regiões (*bins*) de 0,005 ppm perfazendo um total de 1860 variáveis em cada espectro [93].

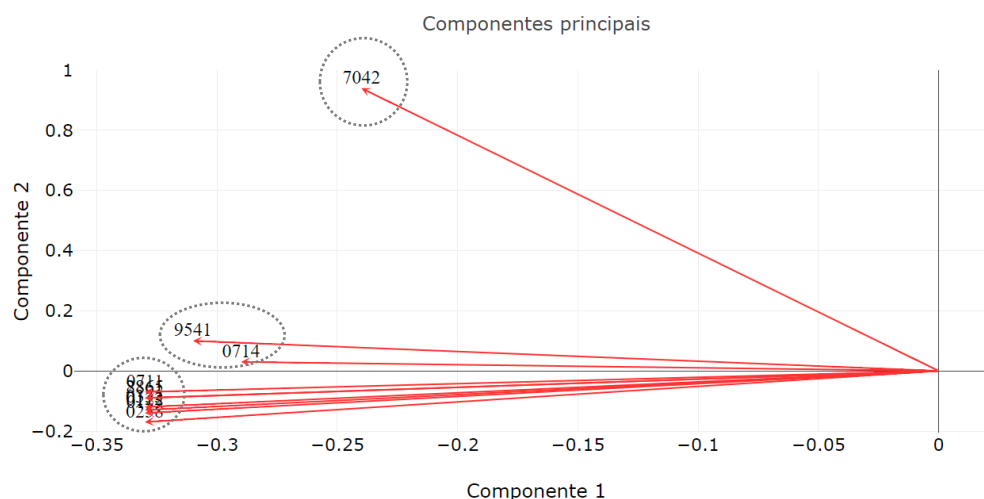
Figura 19 – Espectro RMN ^1H de cada laboratório participante.



Fonte: elaboração própria.

Na análise de componentes principais (PCA) as duas primeiras componentes explicam 94 % da variabilidade do sistema (a primeira componente, sozinha, explica 89 %). A interpretação das componentes principais pode ser feita por meio do gráfico das cargas (*loadings*). Observa-se que 7 laboratórios possuem espectros altamente correlacionados e agrupam-se no segundo quadrante da figura 20. Os laboratórios 0714 e 9541 agrupam-se no terceiro quadrante e, apesar da proximidade, distoam do primeiro grupo identificado. Por fim, o espectro reportado pelo laboratório 7042 distoa significativamente dos demais participantes.

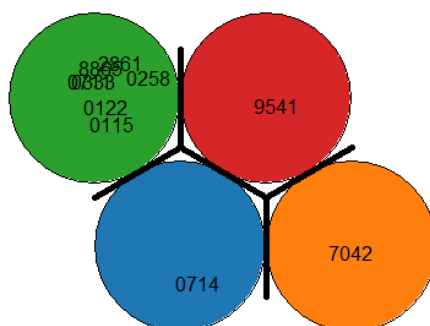
Figura 20 – Análise de componentes principais.



Fonte: elaboração própria.

Os mapas auto-organizáveis de Kohonen apresentam resultados similares aos encontrados pela técnica de PCA. Observa-se no gráfico abaixo que os laboratórios formam 4 grupos (*clusters*) em que um dos grupos agrega 7 laboratórios participantes. Os laboratórios 0714, 7042 e 9541 não se agrupam e formam *clusters* de tamanho 1.

Figura 21 – Mapas auto-organizáveis de Kohonen



Fonte: elaboração própria.

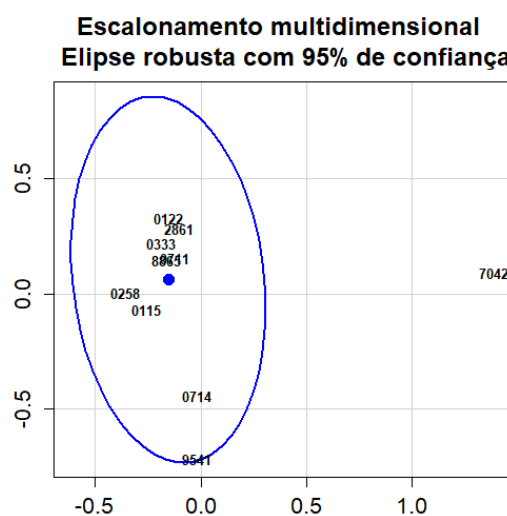
No método de zscore multivariado, a matriz de distância interespectral D_k foi obtida a partir das distâncias de Kullback-Leibler, Mahalanobis, Hellinger e Jensen-Shannon. Os valores $\hat{D}_{i,k}$ (distância média do espectro $s_{i,k}$ para outro espectro no conglomerado S_k) foram ajustados a uma distribuição lognormal. A escolha da distribuição de probabilidade baseou-se no Q-Q plot o qual mostrou a aderência dos valores $\hat{D}_{i,k}$ à distribuição lognormal [93].

Cada valor $Z_{i,k}$ é uma indicação de onde o espectro $s_{i,k}$ se encontra em relação aos demais no conglomerado S_k . No caso da distribuição lognormal, tem-se que $Z_{i,k}(1/2) = 1$ e $Z_{i,k}(0.95) \approx 5$. Neste contexto, $Z_{i,k} = 1$ indica que $s_{i,k}$ está mais próximo do centro de S_k ao passo que $Z_{i,k}$ maior que 5 indica que $s_{i,k}$ está fora do intervalo de confiança de 95% no conglomerado S_k [93].

Os valores $\|T_{i,L}\|$ foram ajustados a uma distribuição lognormal e o escore estatístico projetado \hat{Z}_i foi calculado para cada conjunto de dados. Os escores calculados a partir das distâncias de Kullback-Leibler, Mahalanobis, Hellinger e Jensen-Shannon apresentaram indícios de que os participantes 0714, 7042 e 9541 reportaram resultados (espectros) considerados atípicos (*outliers*) [93].

No escalonamento multidimensional, observa-se a distribuição espacial dos laboratórios a partir dos resultados (espectros) reportados. A elipse robusta com 95 % de confiança indica que o laboratório 7042 apresenta um espectro que difere estatisticamente dos demais participantes, pois está fora da elipse. Na abordagem proposta há indícios de que este laboratório apresenta um resultado considerado insatisfatório. Os demais laboratórios, que se encontram localizados dentro da elipse, apresentam um resultado satisfatório, ou seja, não diferem estatisticamente entre si (Figura 22).

Figura 22 – Escalonamento multidimensional e elipse de confiança robusta (MDS-RCE 2D).



Fonte: elaboração própria.

Tem-se a seguir um comparativo entre os resultados obtidos pela abordagem sugerida na presente tese e as conclusões obtidas a partir do zscore multivariado proposto por Sheen *et al.* (2017).

Tabela 33 – Classificação dos resultados pelos métodos de escalonamento multidimensional - elipse de confiança robusta (MDS-RCE 2D) e zscore multivariado.

Laboratório	MDS-RCE (2D)	Zscore multivariado
0115	Satisfatório	-
0122	Satisfatório	-
0258	Satisfatório	-
0333	Satisfatório	-
0711	Satisfatório	-
0714	Satisfatório	<i>Outlier</i>
2861	Satisfatório	-
7042	Insatisfatório	<i>Outlier</i>
8865	Satisfatório	-
9541	Satisfatório	<i>Outlier</i>

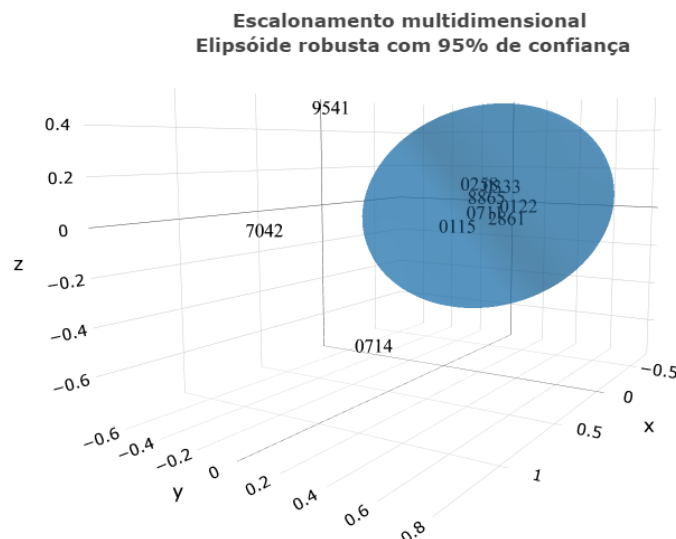
Fonte: elaboração própria.

No intuito de validar as conclusões obtidas pelo método proposto (MDS-RCE) realizou-se a análise tridimensional. Observa-se na figura 23 que mais laboratórios apresentam resultados que destoam dos demais participantes (ao nível de significância de 5 %). Ao todo, três laboratórios (0714, 7042 e 9541) encontram-se fora da elipsóide robusta com 95 % de confiança (Figura 23). Na metodologia proposta estes participantes apresentam resultados considerados insatisfatórios, ou seja, estatisticamente diferente dos demais (Figura 23).

E importante destacar que os laboratórios 0714 e 9541 encontram-se mais afastados do centro da elipse de confiança robusta conforme pode ser observado na Figura 22. O laboratório 9541 em particular, está situado no limite de confiança (Figura 22). Neste contexto, há indícios de que estes participantes podem ter reportado um resultado considerado insatisfatório pela metodologia sugerida na presente tese. Ao se construir a elipsoide robusta de confiança (Figura 23), estes laboratórios estão situados fora dos limites de confiança tendo seu resultado classificado como insatisfatório. Recomenda-se utilizar a análise 3D vista na figura 23 para investigar se

laboratórios situados em regiões limítrofes na análise 2D (como por exemplo na figura 22) são, ou não, atípicos em relação aos demais participantes. Adicionalmente, sugere-se utilizar o mesmo nível de confiança em ambas as análises.

Figura 23 – Escalonamento multidimensional e elipsóide de confiança robusta (MDS-RCE 3D).



Fonte: elaboração própria.

Na tabela 34, observa-se que os laboratórios considerados insatisfatórios pela abordagem tridimensional (Figura 23) são os mesmos que apresentaram indícios de terem reportados resultados atípicos pelo método de zscore multivariado.

Tabela 34 – Classificação dos resultados pelos métodos de escalonamento multidimensional - elisóide de confiança robusta (MDS-RCE 3D) e zscore multivariado.

Laboratório	MDS-RCE (3D)	Zscore multivariado
0115	Satisfatório	-
0122	Satisfatório	-
0258	Satisfatório	-
0333	Satisfatório	-
0711	Satisfatório	-
0714	Insatisfatório	<i>Outlier</i>
2861	Satisfatório	-
7042	Insatisfatório	<i>Outlier</i>
8865	Satisfatório	-
9541	Insatisfatório	<i>Outlier</i>

Fonte: elaboração própria.

7. CONCLUSÃO

Neste trabalho foi apresentada a aplicação web em Shiny/R *Web Application for Proficiency Testing Provider* (WAPT). O software permite, de forma intuitiva e amigável, realizar os cálculos estatísticos necessários para avaliar o desempenho de laboratórios participantes de ensaios de proficiência e comparações interlaboratoriais. A aplicação se encontra em versão pronta para ser lançada.

Cabe ainda destacar que, na presente tese, foram sugeridos métodos estatísticos para avaliação de desempenho de acordo com o tipo de resultado reportado: univariado, categórico e multivariado. Estes métodos foram comparados com os descritos ou mencionados na norma ISO 13528:2015 e sugeridos na literatura.

De modo geral, os resultados preliminares obtidos fornecem indícios de que as metodologias propostas nesta tese se mostraram promissoras como métodos de avaliação de desempenho de laboratórios participantes de ensaios de proficiência nos três tipos de resultados reportados.

No contexto em que o resultado reportado é numérico e univariado, comparou-se as conclusões obtidas com as estatísticas de desempenho (*scores*) com as obtidas pela análise de variância para modelos de efeitos fixos. Pode-se observar divergência de classificações para alguns laboratórios nas duas abordagens. Essas divergências decorrem do fato de o método de *scores* não considerar os pressupostos de normalidade e homocedasticidade em sua análise. Observou-se que os *scores* apresentam probabilidade significativa de classificar os resultados reportados pelo participante como satisfatórios quando há evidências, pela metodologia proposta, de que não o são.

Os resultados preliminares fornecem indícios de que o método estatístico sugerido aprimora a classificação de desempenho porque considerar em sua análise os pressupostos de normalidade e homocedasticidade. Além disso, a metodologia sugerida apresenta mecanismos para sanar a violação destes pressupostos. Neste contexto, os resultados obtidos fornecem elementos que permitem sugerir a análise de variância como método de avaliação de desempenho.

No que se refere a metodologia proposta para análise de dados univariados, cabe destacar que não foi utilizada a ANOVA com correção de Welch porque este teste depende de que as variâncias dos resultados reportados por cada um dos laboratórios

participantes do ensaio de proficiência sejam não nulas. No ensaio de proficiência referente à determinação do ácido benzóico em suco de laranja o laboratório 04 apresenta variância igual a 0 (zero).

No ensaio de proficiência cujos resultados reportados são variáveis categóricas comparou-se as classificações obtidas por meio das medidas de concordância com as obtidas nos testes de homogeneidade de proporção. Os resultados indicam que a escolha da metodologia depende da quantidade de itens de ensaio enviados aos laboratórios participantes. Outro aspecto observado é que as medidas de concordância, eventualmente, podem não captar a disparidade do resultado obtido pelo participante e classificar equivocadamente seu desempenho. Os testes de homogeneidade de proporção possibilitaram identificar o resultado atípico de um dos participantes nos dados simulados. Este resultado preliminar fornece indícios de que a metodologia proposta tem potencial para ser considerada como método de avaliação de desempenho. Por fim, recomenda-se que o número de itens de ensaio fornecidos aos participantes seja levado em consideração pelo provedor do ensaio.

Na comparação interlaboratorial com dados multivariados utilizou-se o método de escalonamento multidimensional (MDS) que permitiu identificar a similaridade/dissimilaridade entre os resultados reportados. Os laboratórios foram representados como pontos em um espaço bi/tri dimensional e a técnica de elipse/elipsóide de confiança robusta (RCE) permitiu identificar os participantes com medições (espectro) atípicas.

As conclusões obtidas pela técnica de MDS-RCE foram comparadas com os resultados fornecidos pelo método de zscore multivariado proposto por Sheen *et al.* (2017). Observou-se que, em ambas as abordagens, os mesmos laboratórios apresentam indícios de terem reportado resultados considerados discrepantes em relação aos demais participantes.

As técnicas multivariadas de análise de componentes principais (PCA) e mapas auto-organizáveis de Kohonen mencionadas na fundamentação teórica corroboraram as conclusões obtidas pela metodologia sugerida (MDS-RCE) para avaliação de desempenho. Os resultados obtidos sugerem que a metodologia MDS-RCE tem potencial para ser considerada como método de avaliação de desempenho no contexto multivariado.

No contexto da metrologia, as abordagens propostas em cada um dos cenários (univariado, categórico e multivariado) buscam aperfeiçoar os mecanismos de avaliação dos laboratórios. Participantes melhor avaliados podem aprimorar seus procedimentos, quando necessário, e melhorar o serviço prestado a indústria. Cabe ainda destacar que as metodologias propostas fornecem um grau de confiança associado aos testes.

Como sugestão para trabalhos futuros propõe-se, no contexto de ensaios de proficiência com dados categóricos, investigar (e/ou desenvolver) medidas de concordância específicas para dados binários (além de sugerir formas de avaliar o desempenho dos participantes) e compará-las com os métodos apresentados nesta tese.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **ABNT NBR ISO/IEC 17043** – Avaliação de conformidade – Requisitos gerais para ensaios de proficiência. Rio de Janeiro: ABNT, 2001.
- [2] DOCUMENTO ORIENTATIVO. **DOQ-CGCRE-020** – Definições de termos utilizados nos documentos relacionados à acreditação de laboratórios, produtores de materiais de referência e provedores de ensaios de proficiência. Rio de Janeiro: DOQ, 2018.
- [3] INTERNATIONAL ORGANIZATION OF STANDARDIZATION. **ISO/IEC 17043(E)** – Conformity assessment – General requirements for proficiency testing. Switzerland: ISO, 2010.
- [4] INTERNATIONAL ORGANIZATION OF STANDARDIZATION. **ISO 13528(E)** – Statistical methods for use in proficiency testing by interlaboratory comparison. Switzerland: ISO, 2015.
- [5] ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **ABNT NBR ISO/IEC 17025** – Requisitos gerais para a competência de laboratórios de ensaio e calibração. Rio de Janeiro: ABNT, 2017.
- [6] ANALYTICAL METHODS COMMITTEE. **AMC TB 11** – Understanding and acting on scores obtained in proficiency testing schemes. London: AMC, 2002. URL: http://www.rsc.org/images/proficiency-testing-technical-brief-11_tcm18-214875.pdf. Acesso em 04 NOV. 2019.
- [7] ANALYTICAL METHODS COMMITTEE. **AMC TB 78** – Proficiency testing of sampling. London: AMC, 2017. URL: http://www.rsc.org/images/TB_78_tcm18-249822.pdf. Acesso em 15NOV2019.
- [8] VOCABULÁRIO INTERNACIONAL DE METROLOGIA. Duque de Caxias: Inmetro, 2012.
- [9] RUMSEY, D. **Estatística Para Leigos**. Tradução da 2ª ed. Alta Books, Rio de Janeiro, 2019.
- [10] POWERS, D. A.; XIE , Y. **Statistical Methods for Categorical Data Analysis**. 2nd ed. Emerald Group Publishing, Bingley, UK, 2008.

- [11] STEWART, A. **Basic Statistics and Epidemiology: A Practical Guide**. 4th ed. Taylor & Fancis Group, London, UK, 2016.
- [12] AZEN, R.; WALKER, C. M. **Categorical Data Analysis for the Behavioral and Social Sciences**. Taylor & Fancis Group, London, UK, 2011.
- [13] ROUSSEEUW, P. J.; HUBERT, M. **Robust statistics for outlier detection**. WIREs Data Mining and Knowledge Discovery, v. 1, p. 73–79, jan. 2011.
- [14] INTERNATIONAL ORGANIZATION OF STANDARDIZATION. **ISO Guide 35(E)** – Reference materials – Guidance for characterization and assessment of homogeneity and stability. Switzerland: ISO, 2017.
- [15] ANALYTICAL METHODS COMMITTEE. **AMC TB 18A** – What is proficiency testing? Guide for end-users of chemical data. London: AMC, 2005. URL: http://www.rsc.org/images/proficiency-testing-technical-brief-18A_tcm18-214885.pdf. Acesso em 02DEZ2019.
- [16] HUBER, P. J. **Robust Estimation of a Location Parameter**. Annals of Mathematical Statistics, v. 35, n. 1, p. 73-101, mar. 1964.
- [17] THODE, H. C. **Testing for Normality**. Marcel Dekker, New York, 2002.
- [18] ANALYTICAL METHODS COMMITTEE. **Robust statistics – How not to reject outliers – Part 1 – Basics concepts**. Analyst, v. 114, p. 1693-1697, dez1989.
- [19] TRANTER, R. L. **Design and Analysis in Chemical Research**. Sheffield Academic Press, v. 15, jun 2001. URL: <https://doi.org/10.1002/cem.650>. Acesso em 11MAR2020.
- [20] VERHEIJ, B.; WIERING, M. 29th Benelux Conference on Artificial Intelligence. Springer, nov. 2017.
- [21] FURNO, M. and VISTOCCO, D. **Quantile Regression: Estimation and Simulation, Volume 2**. John Wiley & Sons, Oxford, UK, 2018.
- [22] HAMPEL, F. R. **The Influence Curve and Its Role in Robust Estimation**. Journal of the American Statistical Association, v. 69, n. 346, p. 383-393, jun 1974.
- [23] RIVERA, C.; RODRÍGUEZ, R. **Horwitz equation as quality benchmark in ISO/IEC 17025 testing laboratory**. Mexico, 2014. URL: <https://pdfs.semanticscholar.org/d6d6/a38d1a9e01e526ca4e2b5b8d804670e5414f.pdf?>

_ga=2.161059835.71560855.1544098209-1095917405.1544098209. Acesso em 05MAI2020.

[24] EURACHEM. Selection, use and interpretation of proficiency testing (PT) schemes by laboratories. London, UK, 2000.

[25] THOMPSON, M.; WOOD, R. **The international harmonized protocol for the proficiency testing of (chemical) analytical laboratories**. Pure and Applied Chemistry, v. 65, n. 9, p. 2123-2144, jul 1993.

[26] ELLISON, S. L. R; BARWICK, V; FARRANT, T. J. D. **Practical Statistics for the Analytical Scientist: A Bench Guide**. 2nd ed. Royal Society of Chemistry, London, UK, 2009.

[27] GASCA-ARAGON, H.; DUEWER, D. L. **The evaluation of the scoring systems: the fixed effects model under known variances**. Accreditation and Quality Assurance, v. 21, p. 255–263, jul 2016.

[28] CASTELAZO, I.; MITANI, Y. **On the use of the mean square error as a proficiency index**. Accreditation and Quality Assurance, v. 17, p. 95–97, jan 2012.

[29] MITANI, Y.; LARA-MANZANO, J. V.; RODRIGUES-LOPEZ, A. **Proficiency testing scheme for the harmonization and comparability of analytical measurements**. Accreditation and Quality Assurance, v. 13, p. 421–426, abr 2008.

[30] ARVIZU-TORRES, R.; PEREZ-CASTORENA, A.; SALAS-TELLEZ, J. A.; MITANI-NAKANISHI, Y. **Biological and environmental reference materials in CENAM**. Fresenius' Journal of Analytical Chemistry, v. 370, p. 156–159, jun 2001.

[31] Harmonized Protocol for the Proficiency Testing of Sampling of Environmental Matrices. Chemistry International, v. 33, p. 22, mai 2011. URL: <https://www.degruyter.com/view/j/ci.2011.33.issue-3/ci.2011.33.3.22a/ci.2011.33.3.22a.xml>. Acesso em 06AGO2020.

[32] RAMSEY, M. H. **Sampling proficiency testing as a means to improve both the quality of sampling, and estimates of measurement uncertainty from sampling**. The Sixth International Proficiency Testing Conference. Romania, 2017. URL: <http://sro.sussex.ac.uk/72818/1/Proceedings-pt-conf-2017.pdf>. Acesso em 15SET2020.

- [33] ARGYRAKI, A.; RAMSEY, M. H.; THOMPSON, M. **Proficiency testing in sampling: pilot study on contaminated land**. *Analyst*, v. 120, p. 2799–2804, dez 1995.
- [34] ALBANO, F. M.; RODRIGUES, M.; ALBANO, J. F. **Garantia da qualidade analítica através de programas de comparação interlaboratorial**. VII SEPROSUL – Semana de Engenharia de Produção Sul-Americana, Uruguay, nov 2007.
- [35] ROSARIO, P.; MARTÍNEZ, J. L.; SILVÁN; J. M. **Comparison of different statistical methods for evaluation of proficiency test data**. *Accreditation and Quality Assurance*, v. 13, n. 9, p. 493–499, jun 2008.
- [36] MONTGOMERY, D. C. **Design and Analysis of Experiments**, 10th ed. John Wiley & Sons, Oxford, UK, 2020.
- [37] ANALYTICAL METHODS COMMITTEE. **AMC TB 82 Are my data normal?** London: AMC, 2017. URL: http://www.rsc.org/images/TB-82_tcm18-250061.pdf. Acesso em 09OUT2019.
- [38] YAP, B. M.; SIM, C. H. **Comparisons of various types of normality tests**. *Journal of Statistical Computation and Simulation*, v. 81, p. 2141-2155, mai 2011.
- [39] GUJARATI, D. N.; PORTER, D. C. **Basic Econometrics**. 5th ed. McGraw–Hill, Atlanta, 2009.
- [40] ROBERTS, M. J.; RUSSO R. **A Student's Guide to Analysis of Variance**. Taylor & Fancis Group, London, UK, 1999.
- [41] RUTHERFORD, A. **ANOVA and ANCOVA: A GLM Approach**. John Wiley & Sons, Oxford, UK, 2011.
- [42] QUINN, G. P.; KEOUGH, M. J. **Experimental Design and Data Analysis for Biologists**. Cambridge University Press, Cambridge, 2002.
- [43] GLAZ, J.; POZDNYAKOV, V.; WALLENSTEIN, S. **Scan Statistics: Methods and Applications**. Springer, New York, 2009.
- [44] TAMHANE, A. C. **Statistical Analysis of Designed Experiments: Theory and Applications**. John Wiley & Sons, Oxford, UK, 2009.
- [45] BOTTLE, A.; AYLIN, P. **Statistical methods for healthcare performance monitoring**. Taylor & Fancis Group, London, UK, 2016.

- [46] SELVIN, S. **Statistical Tools for Epidemiologic Research**. Oxford University Press, Oxford, 2011.
- [47] Post-Hoc and Multiple Comparison Test: An Overview with SAS and R Statistical Package. *International Journal of Statistics and Medical Informatics*, v. 1, n. 1, 2016. URL: <https://ssrn.com/abstract=2944598>. Acesso em 04JAN2021.
- [48] ABDI, H. and WILLIAMS, L. J. **Tukey's honestly significant difference (HSD) test**. *Encyclopedia of Research Design*. Thousand Oaks, California: Sage, 1-5, 2010. URL: <https://personal.utdallas.edu/~herve/abdi-HSD2010-pretty.pdf>. Acesso em 10FEV2021.
- [49] WILLIAMS, L. J. and ABDI, H. Fisher's least significant difference (LSD) test. *Encyclopedia of research design*. Thousand Oaks, California: Sage, 1-5, 2010. URL: <https://personal.utdallas.edu/~herve/abdi-LSD2010-pretty.pdf>. Acesso em 04MAR2021.
- [50] KUTNER, M. H.; NACHTSHEIM, C. J.; NETER, J.; LI, W. **Applied linear statistical models**. 5th ed. McGraw-Hill, Atlanta, 2005.
- [51] DUNN, O. J. **Multiple comparisons using rank sums**. *Technometrics*, v. 6, n. 3, p. 241-252, ago 1964.
- [52] GREENE, W. H. **Econometric Analysis**. 5th ed. Prentice Hall, New Jersey, 2003.
- [53] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, 2022.
- [54] FLORENS, J.; MARIMOUTOU, V.; PÉGUIN-FEISSOLLE, A. **Econometric Modeling and Inference**. Cambridge University Press, Cambridge, 2007.
- [55] PINHEIRO, J. C.; BATES, D. M. **Mixed-Effects Models in S and S-PLUS**. Springer, New York, 2000.
- [56] BENJAMINI, Y.; HOCHBERG, Y. **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society Series B*, v. 57, n. 1, p. 289–300, nov 1995.
- [57] BENJAMINI, Y. e YEKUTIELI, D. **The control of the false discovery rate in multiple testing under dependency**. *Annals of Statistics*, v. 29, n. 4, p. 1165–1188, ago 2001.

- [58] Recent Developments in Multiple Comparison Procedures. Institute of Mathematical Statistics, Lecture Notes, Monograph Series, 2004.
- [59] GAVRILOV, Y.; BENJAMINI, Y.; SARKAR, S. K. **An adaptative step-down procedure with proven FDR control under independence.** The Annals of Statistics, v. 37, n. 2, p. 619–629, abr 2009.
- [60] ROYSTON, P. **Remark AS R94: A remark on Algorithm AS 181: The W test for normality.** Applied Statistics, V. 44, P. 547-551, mar 1995.
- [61] MONTGOMERY, D. C. **Introduction to Statistical Quality Control.** 6th ed. John Wiley & Sons, Oxford, UK, 2009.
- [62] GOWER, J. C. **A General Coefficient of Similarity and Some of Its Properties.** Biometrics, v. 27, n. 4, p. 857-871, dez 1971.
- [63] MANCIN, M.; BARCO, L.; SACCARDIN, C. RICCI, A. **Proposed statistical analysis to evaluate qualitative proficiency testing of Salmonella serotyping.** Accreditation and Quality Assurance, v. 20, p. 305–310, mai 2015.
- [64] LANDIS, J. R.; KOCH, G. G. **The Measurement of Observer Agreement for Categorical Data.** Biometrics, v. 33, n. 1, p. 159-174, mar 1977.
- [65] OSBORNE, J. W. **Best Practices in Quantitative Methods.** Sage Publications, California, 2008.
- [66] CANDELL-RIERA, J.; ORTEGA-ALCALDE, D. **Nuclear Cardiology in Everyday Practice.** Springer Science + Business Media Dordrecht, Barcelona, 1994.
- [67] DORNHEGE, G.; MILLÁN, J. R.; HINTERBERGER, T.; MCFARLAND, D. J.; MÜLLER, K. R. **Toward Brain-computer Interfacing.** MIT Press, Massachusetts, 2007.
- [68] COHEN, J. **A coefficient of agreement for nominal scales.** Educational and Psychological Measurement, v. 20, n. 1, p. 37-46, abr 1960.
- [69] ESTRELA, C. **Metodologia Científica: Ciência, Ensino, Pesquisa.** 3^a ed. Artes Médicas, Porto Alegre, 2018.
- [70] OLSON, K. A. **Manual Physical Therapy of the Spine.** Saunders Elsevier, Pensilvânia, 2009.

- [71] SALKIND, N. J. **Encyclopedia of Measurement and Statistics**. Sage Publications, California, 2007.
- [72] PETT, M. A. **Nonparametric Statistics for Health Care Research: Statistics for Small Samples and Unusual Distributions**. Sage Publications, California, 1997.
- [73] GWET, K. L. **Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters**. 4th ed. Advanced Analytics, Gaithersburg, 2014.
- [74] GWET, K. L. **Computing inter-rater reliability and its variance in the presence of high agreement**. The British journal of mathematical and statistical psychology, v. 61, p. 29-48, mai 2008.
- [75] XIE, Z.; Gadepalli, C.; CHEETHAM, B. **A Study of Chance-Corrected Agreement Coefficients for the Measurement of Multi-Rater Consistency**. International Journal of Simulation: Systems, Science & Technology, v. 19, n. 2, abr 2018.
- [76] BOVE, G.; CONTI, P. L.; MARELLA, D. **An ordinal measure of interrater absolute agreement**. Cornell University, 2019. URL: <https://doi.org/10.48550/arXiv.1907.09756>. Acesso em 28MAI2021.
- [77] LEIGHTON, J. P. **Using Think-Aloud Interviews and Cognitive Labs in Educational Research**. Oxford University Press, Oxford, 2017.
- [78] LATIFI, H.; VALBUENA, R. **3D Remote Sensing Applications in Forest Ecology: Composition, Structure and Function**. MDPI Books, Switzerland, 2019.
- [79] KRIPPENDORFF, K. **Reliability in Content Analysis: Some Common Misconceptions and Recommendations**. Human Communication Research, v. 30, p. 411–433, jul 2004.
- [80] TAYLOR, C.; MARCHI, A. **Corpus Approaches to Discourse: A Critical Review**. Taylor & Fancis Group, London, UK, 2018.
- [81] BREZINA, V. **Statistics in Corpus Linguistics: A Practical Guide**. Cambridge University Press, Cambridge, 2018.
- [82] CHANCE, B. L. and ROSSMAN, A. J. **Investigating Statistical Concepts, Applications, and Methods**. Duxbury Press, California, 2006.

- [83] COHEN, J. **An alternative to Marascuilo's "large-sample multiple comparisons" for proportions.** Psychological Bulletin, v. 3, n. 3, p. 199-201, mar 1967.
- [84] HAY-JAHANS, C. R. **Companion to Elementary Applied Statistics.** Taylor & Fancis Group, London, UK, 2019.
- [85] VIEIRA, S. **Bioestatística: Tópicos Avançados.** 3^a ed. Elsevier Brasil, Rio de Janeiro, 2010.
- [86] BERENSON, M.; LEVINE, D.; SZABAT, K. A.; KREHBIEL, T. C. **Basic Business Statistics: Concepts and Applications.** Pearson Prentice Hall, London, UK, 2013.
- [87] MARASCUILO, L. A. **Large-sample multiple comparisons.** Psychological Bulletin, v. 65, n. 5, p. 280-290, mai 1966.
- [88] JOHNSON, R. A. and WICHERN, D. W. **Applied Multivariate Statistical Analysis.** 6th ed. Pearson Prentice Hall, London, UK, 2007.
- [89] ASAN, U.; ERCAN, S. **An Introduction to Self-Organizing Maps.** Computational Intelligence Systems in Industrial Engineering, Atlantis Computational Intelligence Systems, volume 6. Atlantis Press, Paris, 2012. URL: https://link.springer.com/chapter/10.2991/978-94-91216-77-0_14. Acesso em 01JUL2021.
- [90] HUA, G.; SKALETISKY, M.; WESTERMANN, K. **Exploratory Analysis of CIA Factbook Data Using Kohonen Self-Organizing Maps.** Case Studies for Business, Industry and Government Statistics, v. 3, n. 1, p. 48-59, nov 2008.
- [91] ONG, J.; ABIDI, S. **Data Mining Using Self-Organizing Kohonen Maps: A Technique for Effective Data Clustering & Visualization.** Proceedings of the International Conference on Artificial Intelligence, Las Vegas, Nevada, USA, v. 1, p. 261-264, jul 1999.
- [92] OLIVEIRA, L. O. L. **Mapas Auto-organizáveis de Kohonen Aplicados ao Mapeamento de Ambientes de Robótica Móvel.** Dissertação de mestrado, Programa de Pós-Graduação em Ciência da Computação, Universidade Federal de Santa Catarina, Santa Catarina, 2001.

- [93] SHEEN, D. A.; ROCHA, W. F.C.; LIPPA, K. A.; BEARDEN, D. W. **A scoring metric for multivariate data for reproducibility analysis using chemometric methods**. *Chemometrics and Intelligent Laboratory Systems*, v. 162, p. 10-20, jan 2017.
- [94] LEEUW, J.; MAIR, P. **Multidimensional scaling using majorization: SMACOF in R**. *Journal of Statistical Software*, v. 31, p. 1-30, ago 2009. URL: <https://www.jstatsoft.org/article/view/v031i03>. Acesso em 12JUN2021.
- [95] GROENEN, P. J. F.; van de Velden, M. **Multidimensional scaling**. *Econometric Institute Report EI 2004-15*, Erasmus University Rotterdam, Rotterdam, Netherlands, abr 2004. URL: <https://repub.eur.nl/pub/1274>. Acesso em: 12JUN2021.
- [96] AGRAFIOTIS, D. K.; RASSOKHIN, D. N.; LOBANOV, V. S. **Multidimensional Scaling and Visualization of Large Molecular Similarity Tables**. *Journal of Computational Chemistry*, v. 22, n. 5, p. 488–500, mar 2001.
- [97] MORRISON, A.; ROSS, G.; CHALMERS, M. **Fast Multidimensional Scaling Through Sampling, Springs, and Interpolation. Information Visualization**. *Information Visualization*, v. 2, p. 68–77, mar 2003.
- [98] SCHUMACKER, R.; TOMEK, S. **Understanding Statistics Using R**. Springer, New York, 2013.
- [99] DALGAARD, P. **Introductory Statistics with R**. 2nd ed. Springer, New York, 2008.
- [100] BEELEY, C.; SUKHDEVE, S. R. **Web Application Development with R Using Shiny**. 3rd ed., Packt Publishing, Mumbai, 2018.
- [101] RESNIZKY, H. G. **Learning Shiny**. Packt Publishing, Mumbai, 2015.
- [102] CARVALHO, L. J.; REGO, E. C. P.; GARRIDO, B. C. **Quantification of benzoic acid in beverages: the evaluation and validation of direct measurement techniques using mass spectrometry**. *Analytical Methods*, n. 8, p. 2955–2960, abr 2016.
- [103] COCHRAN, W. G. **Sampling Techniques**. 3rd ed. John Wiley & Sons, Oxford, UK, 1977.

- [104] MODAK, R. K. **Anesthesiology Keywords Review**. 2nd ed. Lippincott Williams & Wilkins, Connecticut, 2013.
- [105] GREGG, M. B. **Field Epidemiology**. 2nd ed. Oxford University Press, Oxford, 2002.
- [106] VIANT, M. R.; BEARDEN, D. W.; BUNDY, J. G.; BURTON, I. W.; COLLETTE, T. W.; EKMAN, D. R.; EZERNIEKS, V.; KARAKACH, T. K.; LIN, C. Y.; ROCHFORD, S.; ROPP, J. S.; TENG, Q.; TJEERDEMA, R. S.; WALTER, J. A.; WU, H. **International NMR-Based Environmental Metabolomics Intercomparison Exercise**. Environmental Science & Technology, v. 43, p. 219-225, jan 2009.
- [107] DAS, N. **Non-parametric Control Chart for Controlling Variability Based on Rank Test**. Economic Quality Control, v. 23, n. 2, p. 227–242, jan 2008.
- [108] INTERNATIONAL ORGANIZATION OF STANDARDIZATION. **ISO 13528(E)** – Statistical methods for use in proficiency testing by interlaboratory comparison. Switzerland: ISO, 2005.
- [109] PODANI, J. **Extending Gower's General Coefficient of Similarity to Ordinal Characters**. Taxon, v. 48, n. 2, p. 331-340, mai, 1999.
- [110] LOESCH, C.; HOELTGEBAUM, M. **Métodos estatísticos multivariados**. Editora Saraiva, São Paulo, 2012.

APÊNDICE

A. DESVIO PADRÃO ROBUSTO

No que se refere à dispersão de um conjunto de dados, existem duas medidas de variabilidade robusta que, quando multiplicadas por uma constante, fornecem uma estimativa para o desvio padrão populacional que não é impactada pela presença de valores atípicos na amostra (*outliers*). Essas medidas são o desvio mediano absoluto (*median absolute deviation*) e o intervalo interquartilico (*interquartile range*) cujas respectivas estimativas robustas para o desvio padrão são

$$\hat{\sigma} = k \cdot MAD(\mathbf{x}) \quad \text{Equação 132}$$

$$\hat{\sigma} = k \cdot IQR(\mathbf{x}) \quad \text{Equação 133}$$

Em que

$$IQR(\mathbf{x}) = Q_3 - Q_1 \quad \text{Equação 134}$$

$$MAD(\mathbf{x}) = \text{mediana}(|x_i - \text{mediana}(\mathbf{x})|) \quad \text{Equação 135}$$

Sendo que Q_i é o i -ésimo quartil, k é um valor que depende da distribuição dos dados e $\mathbf{x} = (x_1, \dots, x_n)$.

Considerando-se, por exemplo, que os dados têm distribuição normal com média μ , variância σ^2 e função de distribuição acumulada $F(x)$ tem-se que:

$$F(\mu + MAD) = 0,75 \quad \text{Equação 136}$$

$$P(X \leq \mu + MAD) = 0,75 \quad \text{Equação 137}$$

$$P(X - \mu \leq MAD) = 0,75 \quad \text{Equação 138}$$

$$P\left(Z \leq \frac{MAD}{\sigma}\right) = 0,75 \quad \text{Equação 139}$$

Da função de distribuição acumulada da normal padrão tem-se que $P(Z \leq z) = 0,75$ implica em $z = 0,6744898$, logo:

$$\frac{MAD}{\sigma} = 0,6744898 \quad \text{Equação 140}$$

$$\sigma = \frac{MAD(x)}{0,6744898} \quad \text{Equação 141}$$

$$\sigma = 1,482602 \cdot MAD(x) \quad \text{Equação 142}$$

De modo análogo o mesmo resultado é obtido tomando-se $F(\mu - MAD) = 0,25$.
É importante destacar que

$$F(\mu + MAD) - F(\mu - MAD) = 0,5 \quad \text{Equação 143}$$

No caso do intervalo interquartil obtém-se, a partir do terceiro quartil,

$$P(X \leq Q_3) = 0,75 \quad \text{Equação 144}$$

$$P\left(Z \leq \frac{Q_3 - \mu}{\sigma}\right) = 0,75 \quad \text{Equação 145}$$

$$\frac{Q_3 - \mu}{\sigma} = z_{0,75} \quad \text{Equação 146}$$

$$Q_3 = \mu + z_{0,75}\sigma \quad \text{Equação 147}$$

E, a partir do primeiro quartil, obtém-se

$$P(X \leq Q_1) = 0,25 \quad \text{Equação 148}$$

$$P\left(Z \leq \frac{Q_1 - \mu}{\sigma}\right) = 0,25 \quad \text{Equação 149}$$

$$\frac{Q_1 - \mu}{\sigma} = z_{0,25} \quad \text{Equação 150}$$

$$Q_1 = \mu + z_{0,25}\sigma \quad \text{Equação 151}$$

Tomando-se por base os resultados supracitados, o desvio padrão pode ser obtido do seguinte modo:

$$\begin{aligned}
IQR(x) &= Q_3 - Q_1 \\
&= \mu + z_{0,75}\sigma - \mu - z_{0,25}\sigma \\
&= z_{0,75}\sigma - z_{0,25}\sigma \\
&= \sigma(z_{0,75} - z_{0,25})
\end{aligned}$$

Equação 152

Logo,

$$\begin{aligned}
\sigma &= \frac{IQR(x)}{(z_{0,75} - z_{0,25})} \\
&= \frac{IQR(x)}{1,34898}
\end{aligned}$$

Equação 153

$$\sigma = 0,7413011 \cdot IQR(x)$$

Equação 154

Por fim, tem-se que o desvio padrão σ de um conjunto de dados $Normal(\mu; \sigma^2)$ pode ser estimado por $\hat{\sigma} = 1,482602 \cdot MAD(x)$ ou $\hat{\sigma} = 0,7413011 \cdot IQR(x)$.

No caso em que os dados possuam distribuição uniforme com parâmetros a e b , ou seja, $X \sim Uniforme[a; b]$ cabe destacar que a função de distribuição acumulada é dada por

$$F(x) = \frac{x - a}{b - a}$$

Equação 155

Logo,

$$\hat{\sigma} = \frac{2}{\sqrt{3}} \cdot MAD(x)$$

Equação 156

Em que

$$\frac{2}{\sqrt{3}} \cong 1,154701$$

Equação 157

A tabela 35 contém alguns exemplos de desvios-padrões robustos obtidos a partir de distribuições de probabilidade.

Tabela 35 – Desvio-padrão robusto.

Distribuição	MAD	IQR
$X \sim \text{Normal}(\mu; \sigma^2)$	$\hat{\sigma} = 1,482602 \cdot \text{MAD}(\mathbf{x})$	$\hat{\sigma} = 0,7413011 \cdot \text{IQR}(\mathbf{x})$
$X \sim \text{exponencial}(\lambda)$	$\hat{\sigma} = 2,078087 \cdot \text{MAD}(\mathbf{x})$	$\hat{\sigma} = 0,9102392 \cdot \text{IQR}(\mathbf{x})$
$X \sim \text{Uniforme}[a; b]$	$\hat{\sigma} = 1,154701 \cdot \text{MAD}(\mathbf{x})$	$\hat{\sigma} = 0,5773503 \cdot \text{IQR}(\mathbf{x})$
$X \sim \text{Logística}(\mu; \theta^2)$	$\hat{\sigma} = 1,650991 \cdot \text{MAD}(\mathbf{x})$	$\hat{\sigma} = 0,8254957 \cdot \text{IQR}(\mathbf{x})$

Fonte: elaboração própria.

B. ESTIMATIVA ROSUSTA MULTIVARIADA

O vetor de médias robustas $\hat{\mu}_{rob}$ e a matriz de variância-covariância robusta S_{rob} , que compõem a construção da elipse e da elipsóide de confiança robustas, são obtidas a partir do processo iterativo [53] abaixo descrito. O algoritmo a seguir é descrito para o caso bivariado, mas pode ser estendido para o caso p-variado de maneira análoga.

Passo1: $\hat{\mu} = [\bar{x} \quad \bar{y}]$ e $w_i = 1 + p/v \quad \forall i = 1, \dots, n$ em que p é o número de variáveis ($p = 2$ considerando um conjunto de dados bivariado $(x; y)$) e v é o número de graus de liberdade da distribuição t multivariada.

$$(x; y) = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix} \quad \text{Equação 158}$$

Passo2: Calcular a matriz

$$X = \begin{bmatrix} (x_1 - \bar{x}) & (y_1 - \bar{y}) \\ \vdots & \vdots \\ (x_n - \bar{x}) & (y_n - \bar{y}) \end{bmatrix} \quad \text{Equação 159}$$

Passo 3: Calcular $svd(A)$ em que

$$A = \begin{bmatrix} \sqrt{\frac{w_1}{\sum_{i=1}^n w_i}} (x_1 - \bar{x}) & \sqrt{\frac{w_1}{\sum_{i=1}^n w_i}} (y_1 - \bar{y}) \\ \vdots & \vdots \\ \sqrt{\frac{w_n}{\sum_{i=1}^n w_i}} (x_n - \bar{x}) & \sqrt{\frac{w_n}{\sum_{i=1}^n w_i}} (y_n - \bar{y}) \end{bmatrix} \quad \text{Equação 160}$$

Tem-se que $svd(A)$ é a decomposição em valor singular da matriz A , ou seja, $svd(A) = USV^T$.

Passo 4: Calcular a matriz

$$w' = XVS' = \begin{bmatrix} w'_{11} & w'_{12} \\ \vdots & \vdots \\ w'_{n1} & w'_{n2} \end{bmatrix} \quad \text{Equação 161}$$

Em que

$$S' = \begin{bmatrix} 1/s_1 & 0 \\ 0 & 1/s_2 \end{bmatrix} \quad \text{Equação 162}$$

Passo 5: Calcular o vetor

$$Q = [Q_1 \quad \cdots \quad Q_n] = [(w'_{11})^2 + (w'_{12})^2 \quad \cdots \quad (w'_{n1})^2 + (w'_{n2})^2] \quad \text{Equação 163}$$

Passo 6: Calcular os novos pesos

$$w_i^* = \frac{(v + p)}{(v + Q_i)} \quad \forall i = 1, \dots, n \quad \text{Equação 164}$$

Passo 7: Estimar $\hat{\mu}_{rob}$

$$\hat{\mu}_{rob} = [\bar{x}_{rob} \quad \bar{y}_{rob}] = \left[\frac{\sum_{i=1}^n w_i^* x_i}{\sum_{i=1}^n w_i^*} \quad \frac{\sum_{i=1}^n w_i^* y_i}{\sum_{i=1}^n w_i^*} \right] \quad \text{Equação 165}$$

Passo 8: A nova estimativa $\hat{\mu}_{rob}$ e os pesos w_i^* são utilizados para obter novos valores no passo 1 ($\hat{\mu} = \hat{\mu}_{rob}$ e $w_i = w_i^*$). Este procedimento é repetido até que os valores de w_i^* converjam, ou seja, $|w_i - w_i^*| < \varepsilon$.

Passo 9: Calcular a matriz de variância-covariância robusta S_{rob} .

$$S_{rob} = \begin{bmatrix} \frac{\sum_{i=1}^n w_i^* (x_i - \bar{x}_{rob})^2}{n} & \frac{\sum_{i=1}^n w_i^* (x_i - \bar{x}_{rob})(y_i - \bar{y}_{rob})}{n} \\ \frac{\sum_{i=1}^n w_i^* (x_i - \bar{x}_{rob})(y_i - \bar{y}_{rob})}{n} & \frac{\sum_{i=1}^n w_i^* (y_i - \bar{y}_{rob})^2}{n} \end{bmatrix} \quad \text{Equação 166}$$

C. HOMOGENEIDADE/ESTABILIDADE

Estudos de homogeneidade e estabilidade podem constituir uma etapa preliminar na condução de ensaio de proficiência para garantir a integridade do item de ensaio a ser fornecido aos participantes. No estudo de homogeneidade busca-se a garantia da manutenção das propriedades físico-químicas do lote do material estudado [14]. Estabilidade é a capacidade do material de referência em manter o valor de uma determinada propriedade dentro de limites especificados por um período de tempo pré-estabelecido, quando estocado nas condições especificadas [14].

C.1 Homogeneidade

Estudo de homogeneidade constitui um estudo de análise de variância para modelos de efeitos aleatórios. Tem-se a seguir uma breve descrição da abordagem paramétrica, não paramétrica (Kruskal-Wallis) e do modelo FGLS.

Na análise de variância com um fator para modelos com efeitos aleatórios, os p tratamentos (amostras) representam uma amostra aleatória de uma grande população de tratamentos para a qual se deseja estender as conclusões obtidas a partir da amostra [36]. Neste experimento testam-se hipóteses em relação à variabilidade σ_τ^2 (variância entre os tratamentos), ou seja, $H_0: \sigma_\tau^2 = 0$ versus $H_1: \sigma_\tau^2 > 0$. A hipótese nula é rejeitada se $F_{obs} > F_{p-1; N-p; \alpha}$ em que $F_{p-1; N-p; \alpha}$ é o quantil da distribuição Fisher-Snedecor e α é o nível de significância [36]. Nos modelos com efeitos aleatórios, se a hipótese nula é aceita significa que não existe variabilidade entre os tratamentos (homogeneidade). Este modelo é recomendado quando os resíduos são normais e homocedásticos.

Neste tipo de planejamento de experimentos existem dois componentes de variabilidade a serem estimados: (i) variância entre os tratamentos σ_τ^2 ; (ii) variância do erro (dentro dos tratamentos) σ^2 . Estes componentes são estimados por

$$\hat{\sigma}^2 = MS_E \quad \text{Equação 167}$$

$$\hat{\sigma}_\tau^2 = \frac{MS_{trat} - MS_E}{n_o} \quad \text{Equação 168}$$

Em que

$$n_o = \frac{1}{p-1} \cdot \left[\sum_{i=1}^p n_i - \frac{\sum_{i=1}^p n_i^2}{\sum_{i=1}^p n_i} \right] \quad \text{Equação 169}$$

A variabilidade do modelo de efeitos aleatórios é estimada a partir da soma dos componentes da variância: $V(y_{ij}) = \hat{\sigma}_\tau^2 + \hat{\sigma}^2$. Em algumas situações $\hat{\sigma}_\tau^2 < 0$ e, como não existe variância negativa, sugere-se considerar $\hat{\sigma}_\tau^2 = 0$. Por fim $\hat{\sigma}_\tau^2 / (\hat{\sigma}_\tau^2 + \hat{\sigma}^2)$ reflete a proporção da variabilidade das observações que é resultado da variabilidade entre os tratamentos [36].

Nos casos em que os resíduos forem não normais, mas homocedásticos sugere-se avaliar a homogeneidade por meio do teste de Kruskal-Wallis. Se $K > \chi_{k-1}^2$ rejeita-se a hipótese nula, ou seja, as amostras não são homogêneas. As componentes da variância são estimadas do mesmo modo que no modelo linear normal.

Nos casos em que os resíduos forem não normais e heteroscedástico, sugere-se avaliar a homogeneidade por meio do modelo FGLS. Se $F_1 > F_{(p-1);(n-p);\alpha}$ rejeita-se a hipótese nula, ou seja, as amostras não são homogêneas. As componentes da variância são estimadas do mesmo modo que no modelo linear normal.

C.2 Estabilidade

No estudo de estabilidade emprega-se modelos de regressão para verificar eventuais tendências ao longo do tempo [14] dos itens de ensaio. Tem-se a seguir uma descrição dos modelos OLS e FGLS.

No método de mínimos quadrados ordinários (*ordinary least squares* – OLS) os parâmetros são estimados a partir da equação matricial

$$b = (X^T X)^{-1} X^T Y \quad \text{Equação 170}$$

A significância de cada parâmetro do modelo ($H_0: \beta_k = 0$ versus $H_1: \beta_k \neq 0$) é avaliada a partir da estatística de teste

$$t_{obs} = b_k / s(b_k) \quad \text{Equação 171}$$

Em que

$$s^2(b) = MSE(X^T X)^{-1} \quad \text{Equação 172}$$

Sendo que $k = 0$ representa o intercepto, $k = 1$ representa o coeficiente angular e MSE é o erro quadrático médio obtido da análise de variância. Se $|t_{obs}| > t_{n-p}$ rejeita-se a hipótese nula em que t_{n-p} é o quantil da distribuição t -student com $n - p$ graus de liberdade ($p = 2$ no modelo linear normal) [14, 50].

No método dos mínimos quadrados generalizados “factíveis” (*feasible generalized least squares* – FGLS) a equação estimada é

$$b = (X^T \hat{\Omega}^{-1} X)^{-1} X^T \hat{\Omega}^{-1} Y \quad \text{Equação 173}$$

Em que $\hat{\Omega}$ é uma matriz diagonal de pesos. O teste de significância dos parâmetros do modelo é análogo ao modelo OLS [52], entretanto tem-se que

$$s^2(b) = (SQG/(n - p)) \cdot (X^T \hat{\Omega}^{-1} X)^{-1} \quad \text{Equação 174}$$

Em que

$$SQG = (Y - Xb)^T \hat{\Omega}^{-1} (Y - Xb) \quad \text{Equação 175}$$

C.2.1 Gráficos de controle

O aplicativo disponibiliza gráficos de controle como ferramenta adicional de análise no estudo de estabilidade. Pode-se construir gráficos para dados normais (\bar{x} , R e S) e não normais. No que tange gráficos de controle para dados não normais, o aplicativo contém: (i) média móvel exponencialmente ponderada; (ii) método proposto por Nandini Das em 2008 [107].

O gráfico de controle é uma técnica de monitoramento muito eficaz quando fontes não usuais de variabilidade estão presentes em um determinado processo [61]. O gráfico de controle determina, estatisticamente, uma faixa denominada limites de controle com a finalidade de verificar se o processo está sob controle, isto é, isento de causas especiais [61]. Supondo-se que m amostras de tamanho n estejam disponíveis os dados necessários para a construção dos gráficos de controle estão dispostos na tabela 36.

Tabela 36 – Gráficos de controle: base de dados.

Amostras	Observações				Médias (\bar{x}_i)	Amplitudes (R_i)	Desvio- padrão (S_i)
1	x_{11}	x_{12}	\cdots	x_{1n_1}	\bar{x}_1	R_1	S_1
2	x_{21}	x_{22}	\cdots	x_{2n_2}	\bar{x}_2	R_2	S_2
\vdots	\vdots	\vdots		\vdots	\vdots	\vdots	\vdots
m	x_{m1}	x_{m2}	\cdots	x_{mn_m}	\bar{x}_m	R_m	S_m
-	-	-		-	$\bar{\bar{x}}$	\bar{R}	\bar{S}

Fonte: Montgomery, 2009 [61].

Em que

$$R_i = \max(x_{i1}, x_{i2}, \dots, x_{in_i}) - \min(x_{i1}, x_{i2}, \dots, x_{in_i}) \quad \text{Equação 176}$$

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \quad \text{Equação 177}$$

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \quad \text{Equação 178}$$

$$\bar{\bar{x}} = \frac{1}{m} \sum_{i=1}^m \bar{x}_i \quad \text{Equação 179}$$

$$\bar{R} = \frac{1}{m} \sum_{i=1}^m R_i \quad \text{Equação 180}$$

$$\bar{S}^2 = \frac{1}{m} \sum_{i=1}^m S_i^2 \quad \text{Equação 181}$$

$$\bar{S} = \frac{1}{m} \sum_{i=1}^m S_i \quad \text{Equação 182}$$

Recomenda-se que m seja, pelo menos, igual 20 ou 25 e n é usualmente igual a 4, 5 ou 6 [61].

Existem 4 tipos de gráficos de controle para variáveis: \bar{x} , R , S^2 e S . O primeiro monitora o valor médio do processo e os demais monitoram a variabilidade. O gráfico mais comumente utilizado para monitorar a variabilidade é o gráfico para amplitude R . O gráfico de controle \bar{x} monitora o nível médio da qualidade de um processo, sendo assim monitora a variabilidade entre amostras, ou seja, a variabilidade no processo ao longo do tempo. O gráfico de controle R mede a variabilidade dentro da amostra, ou seja, a variabilidade instantânea do processo em um dado instante de tempo. O gráfico de controle para S é recomendável quando os tamanhos das amostras são variáveis ou moderadamente grandes [61]. Os limites de controle para cada gráfico estão descritos na tabela 37.

Tabela 37 – Gráficos de controle: Limites de controle.

Limites $k - \sigma$	Limites $\alpha - fixado$
Gráfico de controle \bar{x}	
$LSC = \bar{\bar{x}} + k \cdot \frac{\bar{R}/d_2}{\sqrt{n}}$	$LSC = \bar{\bar{x}} + z_{1-\alpha/2} \cdot \frac{\bar{R}/d_2}{\sqrt{n}}$
$LM = \bar{\bar{x}}$	$LM = \bar{\bar{x}}$
$LIC = \bar{\bar{x}} - k \cdot \frac{\bar{R}/d_2}{\sqrt{n}}$	$LIC = \bar{\bar{x}} - z_{1-\alpha/2} \cdot \frac{\bar{R}/d_2}{\sqrt{n}}$
Gráfico de controle R	
$LSC = \bar{R} + k \cdot d_3 \cdot \frac{\bar{R}}{d_2}$	$LSC = W_{(1-\alpha/2)} \cdot \frac{\bar{R}}{d_2}$
$LM = \bar{R}$	$LM = \bar{R}$
$LIC = \bar{R} - k \cdot d_3 \cdot \frac{\bar{R}}{d_2}$	$LIC = W_{\alpha/2} \cdot \frac{\bar{R}}{d_2}$
Gráfico de controle S	
$LSC = \bar{S} + k \cdot \bar{S} \cdot \sqrt{\frac{1}{(c_4)^2} - 1}$	$LSC = \sqrt{\frac{\bar{S}^2 \cdot \chi_{(n-1); 1-\alpha/2}^2}{n-1}}$
$LM = \bar{S}$	$LM = \bar{S}$
$LIC = \bar{S} - k \cdot \bar{S} \cdot \sqrt{\frac{1}{(c_4)^2} - 1}$	$LIC = \sqrt{\frac{\bar{S}^2 \cdot \chi_{(n-1); \alpha/2}^2}{n-1}}$

Fonte: Montgomery, 2009 [61].

Os limites de controle podem ser subdivididos em *limites $k - \sigma$* (k vezes o desvio-padrão em que k é normalmente igual a 3) e *limites $\alpha - fixado$* (em que α é usualmente igual a 0,05). Nos gráficos de controle tem-se que $z_{1-\alpha/2}$ é o quantil da

distribuição normal padrão, $\chi^2_{(n-1); 1-\alpha/2}$ e $\chi^2_{(n-1); \alpha/2}$ são os quantis da distribuição qui-quadrado com $n - 1$ graus de liberdade, LSC é o limite superior de controle, LM é a linha média e LIC é o limite inferior de controle. As quantidades d_2 , d_3 , c_4 , $W_{(1-\alpha/2)}$ e $W_{\alpha/2}$ são valores tabelados que dependem de n . Por fim, nos gráficos de controle de variabilidade (R , S^2 e S) se $LIC < 0$ assume-se que $LIC = 0$ [61].

O gráfico de controle da média móvel exponencialmente ponderada constitui uma alternativa ao gráfico de controle \bar{x} de Shewhart quando as observações não apresentam distribuição normal ou nos casos em que é necessário detectar pequenas mudanças na média do processo.

A média móvel exponencialmente ponderada (*Exponentially weighted moving average* – EWMA) é definida por $z_i = \lambda \bar{x}_i + (1 - \lambda)z_{i-1}$ em que $0 \leq \lambda \leq 1$ (em geral valores de λ no intervalo $0,05 \leq \lambda \leq 0,25$ geram bons resultados com $\lambda = 0,05$, $\lambda = 0,10$ e $\lambda = 0,20$, sendo escolhas bem populares) e $z_0 = \bar{\bar{x}}$. É importante observar que no gráfico de controle \bar{x} os valores plotados no gráfico são \bar{x}_i ao passo que no gráfico de controle da média móvel exponencialmente ponderada os valores plotados são z_i [61]. Este gráfico é utilizado para monitorar a média do processo e os respectivos limites de controle são dados por:

$$LSC = \bar{\bar{x}} + L \cdot \frac{\bar{R}}{d_2} \cdot \sqrt{\frac{\lambda}{(2 - \lambda)} [1 - (1 - \lambda)^{2i}]} \quad \text{Equação 183}$$

$$LM = \bar{\bar{x}} \quad \text{Equação 184}$$

$$LIC = \bar{\bar{x}} - L \cdot \frac{\bar{R}}{d_2} \cdot \sqrt{\frac{\lambda}{(2 - \lambda)} [1 - (1 - \lambda)^{2i}]} \quad \text{Equação 185}$$

Nandini Das (2008) propôs um gráfico de controle não paramétrico para monitorar a variabilidade do processo baseado no teste de Mood. Duas amostras consecutivas independentes x_{11}, \dots, x_{1n} e x_{21}, \dots, x_{2m} são combinadas em uma única amostra e os postos (*ranks*) são obtidos. A estatística de teste é definida por:

$$M = \sum_{j=1}^n \left(r_{ij} - \frac{N+1}{2} \right)^2 \quad \text{Equação 186}$$

Em que r_{ij} são os postos da i -ésima amostra e $N = n + m$. É importante observar que $j = 1, \dots, n$ quando $i = 1$ e $j = 1, \dots, m$ quando $i = 2$. A estatística M tem distribuição aproximadamente normal com média $E\{M\}$ e variância $Var\{M\}$ e, por conseguinte,

$$Z = \frac{M - E\{M\}}{\sqrt{Var\{M\}}} \quad \text{Equação 187}$$

Tem distribuição normal padrão em que

$$E\{M\} = \frac{n}{12} (N - 1)(N + 1) \quad \text{Equação 188}$$

$$Var\{M\} = \frac{nm}{180} (N + 1)(N - 2)(N + 2) \quad \text{Equação 189}$$

Considerando-se k amostras de tamanho n_{ij} ($i = 1, \dots, k$), calcula-se a estatística Z acima definida para cada duas amostras consecutivas obtendo-se assim $k - 1$ valores de Z . Os valores de Z são plotados no gráfico e se algum ponto estiver fora dos limites de controle há um indicativo de que o processo está fora de controle com respeito à variabilidade. Os limites de controle são: $LSC = 3$, $LM = 0$ e $LIC = -3$ [107].