UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE
LABORATÓRIO DE ENGENHARIA E EXPLORAÇÃO DE PETRÓLEO

TRATAMENTO ESTATÍSTICO DE DADOS GEOQUÍMICOS

Márcio Luís Carvalho Araújo

Tamires dos Santos Soares

Thiago Rocha Gomes

Dâmaris Machado de Araújo

Laysla Nicoli Pereira da Rocha

MACAÉ - RJ
NOVEMBRO - 2023

Sumário

Capítulo 1

Introdução

Os métodos estatísticos são utilizados no processamento de informações, visando explorar conjuntos numéricos e derivar conclusões valiosas a partir deles. Essas técnicas envolvem a análise e interpretação de dados para oferecer insights significativos e embasar decisões informadas.

1.1 Escopo do Problema

Em experimentos de geoquímica de laboratório, é comum notar que certas amostras se destacam com valores diferentes das demais. Para ilustrar esse fenômeno, podemos considerar as percentagens das frações NSO[1], saturados[2] e aromáticos [3]

No laboratório tem n pesquisadores que utilizam a amostra 09, por exemplo, para calcular suas percentagens de saturados, aromáticos e NSO. É de se saber que cada pesquisador pode encontrar valores diferentes destas percentagens para essa mesma amostra. Esses valores normalmente variam pouco de experimento para experimento, mas existem casos em que a percentagem sofre uma considerável variação (Tabela 1.1).

Tabela 1.1: Percentagens de SAT, ARO e NSO para a amostra 09.

Pesquisador	% SAT	% ARO	% NSO
Fernanda	50	20	30
Hilda	47	23	30
Laercio	52	25	23
Mateus	**10**	**70**	20
Natieli	55	25	20

Conforme indicado na tabela 1.1, as percentagens calculadas pelo pesquisador Mateus diferem significativamente das encontradas pelos outros pesquisadores para as frações

[1]Compostos NSO são compostos do petróleo que contém átomos de nitrogênio, enxofre e oxigênio em sua estrutura molecular.

[2]Hidrocarbonetos que possuem apenas ligações simples entre os átomos de carbono de sua estrutura.

[3]Hidrocarbonetos que possuem o anel benzênico em sua estrutura molecular.

SAT e ARO. Esses valores são considerados atípicos, e para confirmar essa suspeita, são realizados testes como o Teste do escore z modificado, Teste de Grubbs, Teste de Dixon, Teste de Cochran e Teste de Doerffel.

Em estatística, um outlier, ou valor atípico, refere-se a uma observação que se distancia consideravelmente das demais na série, sendo inconsistente ou "fora do padrão". A presença de outliers geralmente prejudica a interpretação dos resultados dos testes estatísticos aplicados às amostras. Em resumo, um outlier é uma observação que se desvia notavelmente das demais, levantando suspeitas de que pode ter sido influenciado por um mecanismo diferente.

A determinação de *outliers*, por exemplo, pode indicar processos geoquímicos raros (como mineralizações), que pode ser usado para exploração de hidrocarbonetos. De um modo geral eliminam-se os *outliers* quando eles representam erros claros, mas frequentemente eles representam irregularidades interessantes que merecem um estudo mais detalhado. Na verdade para alguns conjuntos de dados, os *outliers* são a característica mais importante. A identificação de valores pertencentes a um conjunto de dados que possam ser caracterizados como *outliers*, bem como o tratamento que se deve dar a eles é tema importante no tratamento estatístico de dados. A regressão linear pode ser usada por exemplo para a determinação do gradiente térmico (variação da temperatura com a profundidade) e seria de grande importância para a determinação de temperaturas em rochas geradoras e reservatórios.

1.2 Objetivo

O objetivo geral do *software* é propor uma série de tratamentos estatísticos como a regressão linear e os testes de hipóteses, bem como alguns testes que identificam valores irregulares (*outliers*).Este trabalho implementa alguns testes que identificam, que são: Teste do escore z modificado, Teste de Grubbs, Teste de Dixon, Teste de Cochran e Teste de Doerffel , além de testes de hipóteses e regressão linear que são aplicados em dados geoquímicos.

1.3 Organização do Documento

O presente documento está organizado como uma apostila, contendo a Especificação, a Elaboração onde também uma teoria básica, para melhor compreensão do programa, e o passo a passo da utilização do mesmo. , a Análise Orientada a Objeto (AOO), a Implementação, os Testes e a Documentação.

Capítulo 2

Especificação

Apresenta-se neste capítulo a especificação do sistema a ser modelado e desenvolvido.

2.1 Especificação do programa - descrição dos requisitos

A elaboração deste modelo foi concebida com o objetivo de apresentar uma abordagem estatística para o tratamento de dados geoquímicos, utilizando técnicas como regressão linear, testes de hipóteses e detecção de valores atípicos (outliers).

2.1.1 Definições da interface

Devido à sua simplicidade, o software será dotado de uma interface de linha de comando. Adicionalmente, é imprescindível que seja multiplataforma, sendo capaz de executar em sistemas como GNU/Linux, UNIX, Mac OS e Windows.

2.1.2 Entrada e saída de dados

A entrada e a saída de dados irá depender do tipo de teste selecionado. Para melhor observá-los indica-se a análise dos diagramas. A entrada de dados poderá ser feita via arquivo de disco ou via teclado.

2.2 Casos de uso do Programa

A Tabela 2.1 mostra a descrição do caso de uso do programa. Estão descritas todas as etapas realizadas no funcionamento do programa. Observe que se o usuário escolher uma opção não correspondente a nenhuma das alternativas para saída do programa (1-Selecionar o teste estatístico, 2-Fornecer os dados geoquímicos, 3-Aplicar o teste, e 4-Analisar resultados), um exemplo de cenário alternativo, a frase “Opção inválida” irá aparecer na tela, e o usuário deverá selecionar outra opção.

Tabela 2.1: Casos de uso do programa

Nome do caso de uso:	Tratamento estatístico de dados geoquímicos.
Resumo/descrição:	Escolher um método estátistico para tratar dados geoquímicos.
Etapas:	1. Selecionar o tratamento estatístico. 2. Fornecer dados geoquímicos. 3. Aplicar o tratamento. 4. Analisar resultados.
Cenário Alternativo	Um cenário alternativo envolve uma entrada errada do usuário, escolhendo um número acima de 4 (que não representa nenhuma das opções de saída do programa). Neste caso, a frase "Opção inválida" aparece na tela.

2.3 Diagrama de caso de uso geral do programa

O diagrama de caso de uso geral do programa é apresentado na Figura 2.1. Observe que o ator (o usuário) deve escolher um teste estatístico, e em seguida fornecer os dados que deseja analisar, que será a entrada do programa, contendo os informações geoquímicas, como o exemplo das percentagens de SAT, ARO e NSO. Posteriormente, o usuário deverá então selecionar a opção de aplicar o teste. De posse dos resultados, os mesmos serão analisados e interpretados.

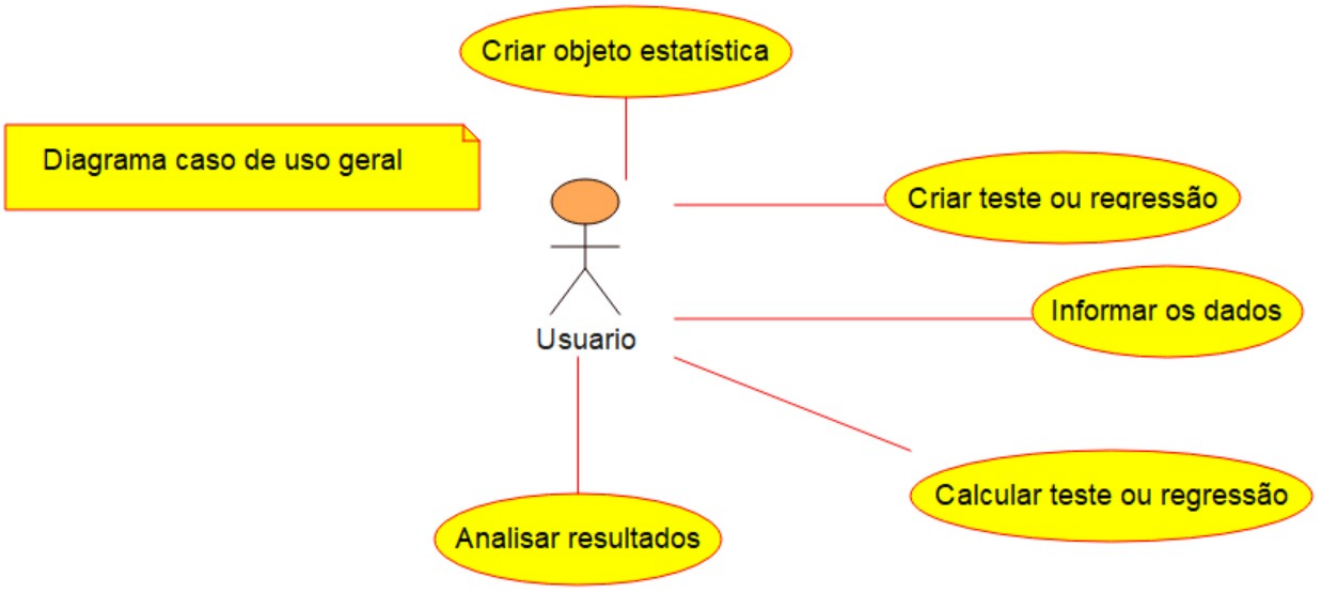


Figura 2.1: Diagrama de caso de uso – Caso de uso geral

2.4 Diagrama de caso de uso específico do programa

Apresenta-se na Figura **??** um diagrama de caso de uso específico do programa. Nele está detalhado o caso de uso específico que é "Testes" onde um teste deverá ser escolhido para tratar os dados geoquímicos. O usuáro cria um objeto "Teste" e informa ao sistema os dados geoquímicos e o nível de significância, posteriormente é solicitada a realização do teste através do caso de uso "Realizar Teste". Os resultados são então gerados e informados ao usuário para que o mesmo possa analisá-lo.

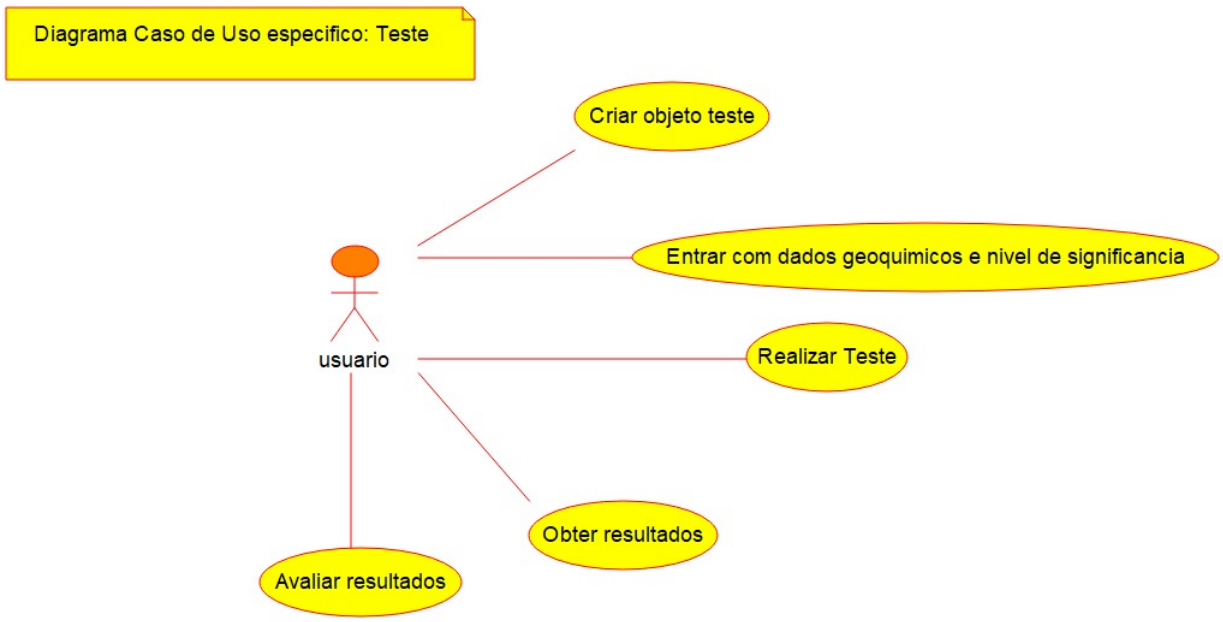


Figura 2.2: Diagrama de caso de uso específico – Detalhamento dos testes a serem utilizados

Já para o caso de uso específico "Regressão Linear", o objeto é criado e o usuário deve fornercer os valores de x e y. O sistema então retorna o coeficiente de correlação e a regressão linear ao usuário, o que pode ser melhor visualizado através de um gráfico plotado no `gnuplot`. Com isso, o usuário pode então analisar os resultados. O diagrama de caso de uso específico "Regressão Linear" é apresentado na figura **??**.

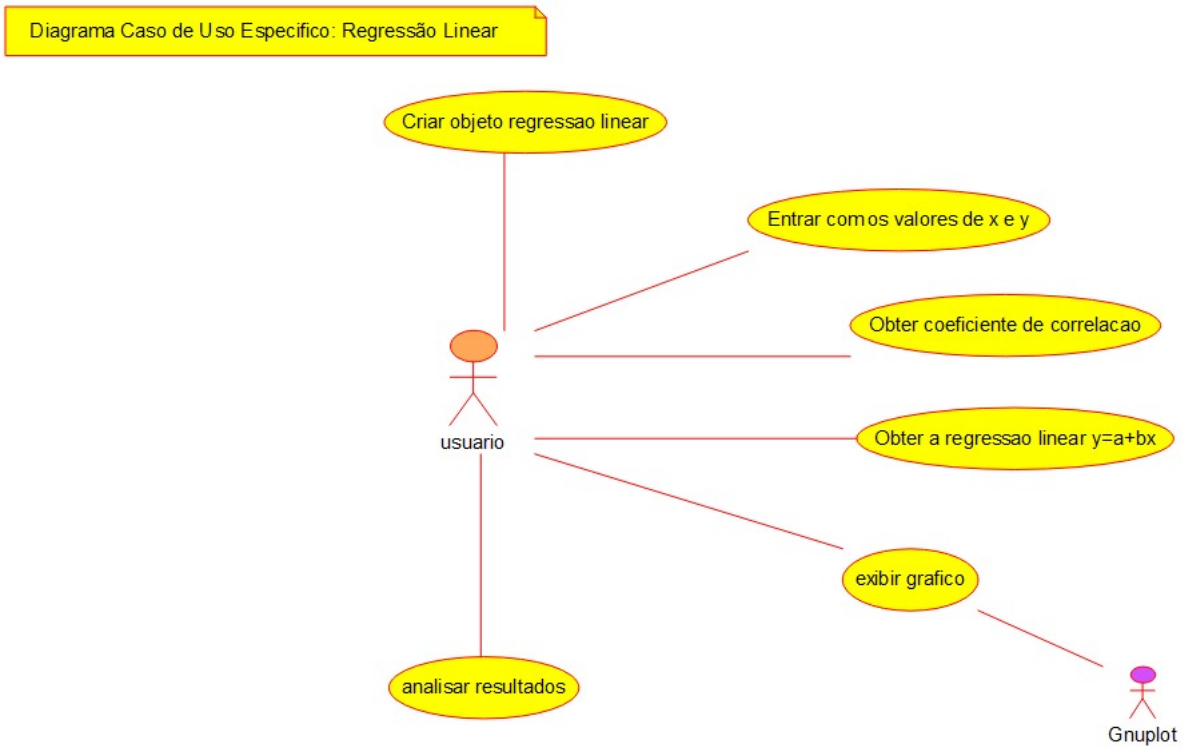


Figura 2.3: Diagrama de caso de uso específico – Detalhamento da regressão linear

Capítulo 3

Elaboração

Neste capítulo, abordaremos o processo de elaboração, que inclui a exploração de conceitos associados ao programa em desenvolvimento, a análise do domínio relevante e a identificação de pacotes.

3.1 Revisão de conceitos básicos de estatística

Apresenta-se a seguir uma revisão de estatística.

3.1.1 Conceitos estatísticos básicos

- População: Refere-se a qualquer conjunto de indivíduos ou valores, podendo ser finita ou infinita.
- Amostra: Constitui uma parcela da população, geralmente selecionada com o propósito de realizar inferências sobre a totalidade da população.
- Amostra representativa: Caracteriza-se por apresentar as propriedades relevantes da população na mesma proporção em que ocorrem na própria população.
- Amostra aleatória: Trata-se de uma amostra de N valores ou indivíduos obtida de maneira que todos os conjuntos possíveis de N valores na população tenham igual probabilidade de serem escolhidos.
- Frequência: Corresponde ao número de elementos em um intervalo de interesse dividido pelo número total de elementos.
- Média amostral:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3.1}$$

onde :

x_i = i-ésimo valor

N = Número total de valores na amostra

- Desvio:

$$d_i = x_i - \bar{x} \tag{3.2}$$

- Variância amostral:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N} {d_i}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{3.3}$$

- Desvio padrão amostral

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} {d_i}^2} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{3.4}$$

3.1.2 Regressão linear

O emprego da regressão linear é apropriado apenas quando a correlação entre as variáveis é estatisticamente notável. A correlação quantifica a intensidade ou extensão do vínculo entre duas variáveis, ao passo que a regressão oferece uma equação que expressa esse relacionamento em termos matemáticos. Em uma perspectiva ideal, desejaríamos antecipar os valores precisos de uma variável com base na relação existente; no entanto, na prática, somos capazes de prever apenas valores médios ou esperados.

O método dos mínimos quadrados é uma estratégia para adequar pontos a uma reta, baseando-se na minimização da soma dos quadrados das distâncias verticais dos pontos até a reta; essa reta resultante, conhecida como reta dos mínimos quadrados, reta de regressão ou reta de regressão estimada, tem seus valores de a e b da equação da reta $y = a + bx$ estimados com base em dados amostrais.

O método dos mínimos quadrados busca minimizar a soma dos quadrados dos resíduos, ou seja, minimizar $\sum_{i=1}^{n} e_i^2$.

A premissa é que, ao minimizar essa soma, encontraremos os valores de a e b que proporcionarão a menor diferença entre a previsão de y e o y realmente observado.

Assim, dado um conjunto de pares ordenados x_i, y_i, é possivel relacionar x_i e y_i utilizando-se um modelo de regressão linear bastante simples, a regressão linear. A mesma pode ser escrita da seguinte forma:

$$y_i - a - bx_i + \varepsilon_i \tag{3.5}$$

sendo a e b os coeficientes da equação linear e ε_i o erro. Note que

$$E\{\varepsilon_i\} = \sigma^2 \tag{3.6}$$

Pode-se determinar os coeficientes a e b utilizando-se as equações:

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x^2)} \tag{3.7}$$

$$a = \frac{(\sum x^2)(\sum y) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x^2)} = \frac{\sum y}{n} - b\frac{\sum x}{n} \tag{3.8}$$

Um coeficiente r^2, conhecido como coeficiente de ajuste da regressão, pode ser determinado utilizando-se a equação abaixo:

$$r^2 = \frac{\sum \bar{y}}{\sum y^2} = b^2\frac{\sum x^2}{\sum y^2} = \frac{(\sum xy)^2}{(\sum x^2)(\sum y^2)} \tag{3.9}$$

Para visualizar o processo da regressão linear, consulte a 4.14 que apresenta o diagrama de atividades correspondente.

3.1.3 Distribuição normal

Uma distribuição estatística é uma função que descreve como uma variável aleatória se comporta. Essa variável pode assumir valores em um conjunto predefinido relacionado ao sistema em questão, cada um com uma probabilidade específica determinada pela distribuição de probabilidades. Identificando ou estimando essa distribuição, podemos calcular a probabilidade de ocorrência de qualquer valor de interesse.

A distribuição normal é contínua, significando que a variável pode ter qualquer valor dentro de um intervalo previamente determinado. para uma variável com distribuição normal, o intervalo é $(-\infty, +\infty)$, indicando que ela tem a capacidade de adotar qualquer valor real, pelo menos teoricamente. Uma distribuição contínua da variável x é definida pela sua densidade de probabilidade $f(x)$.

$$f(x)dx = \frac{1}{\sigma\sqrt{2\pi}}\exp\frac{-(x-\mu)^2}{2\sigma^2}dx \tag{3.10}$$

onde:

$f(x)dx$ = Densidade de probabilidade da variável aleatória x

μ= Média populacional

σ^2= Variância populacional

Empregaremos a notação $x \approx N(\mu, \sigma^2)$ para indicarmos que x se distribui normalmente com média μ e variância σ^2. Para a distribuição normal padrão teremos:

$x \approx N(0, 1)$

$$f(x)dx = \frac{1}{\sigma\sqrt{2\pi}}\exp\frac{x^2}{2} \tag{3.11}$$

.

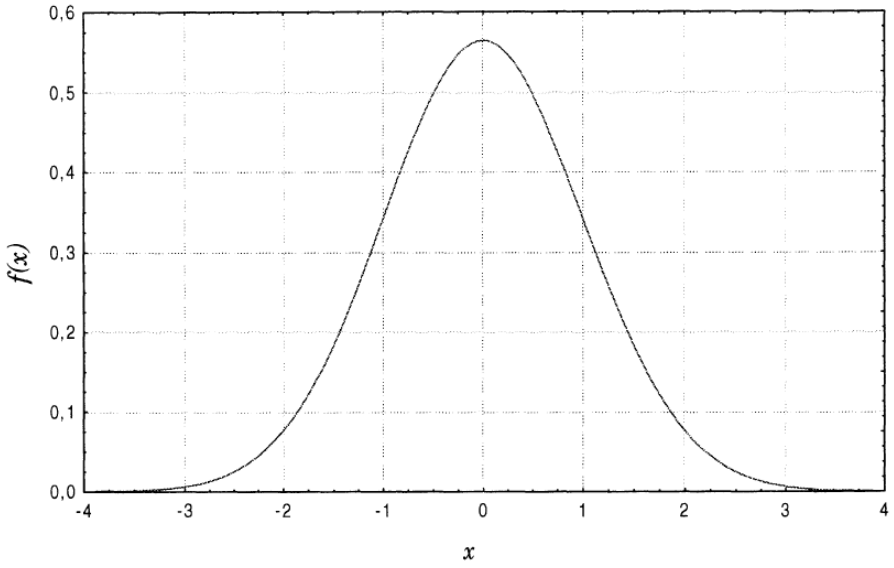


Figura 3.1: Distribuição de frequências de uma variável aleatória x. Note que x é o afastamento em relação à média (que é zero), em número de desvios padrão.

- Variável Normal Padronizada

  $\mathbf{x}$ = Variável aleatória com distribuição $N(\mu, \sigma^2)$

  $\mathbf{z}$ = Variável aleatória com distribuição $N(0, 1)$

$$z = \frac{x - \mu}{\sigma} \tag{3.12}$$

3.1.4 Distribuição t de student

A distribuição t student é uma distribuição de probabilidade teórica. Ela é simétrica, tem formato de sino, e assemelha-se à curva normal padrão, contudo com caudas mais amplas. Em outras palavras, uma simulação usando a distribuição t de Student pode resultar em valores mais extremos do que uma simulação utilizando a distribuição normal. O único parâmetro que a distingue da normal e define sua forma é o número de graus de liberdade v. Quanto maior for esse parâmetro, mais próxima da normal ela será.

Suponha que Z tenha a distribuição normal com média 0 e variância 1, que V tenha a distribuição Chi-quadrado com v graus de liberdade, e que Z e V sejam independentes. Então:

$$t = \frac{Z}{\sqrt{\frac{V}{v}}} \tag{3.13}$$

tem a distribuição t de Student com v graus de liberdade.

A função densidade de probabilidade é:

$$f(t) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\Gamma(\frac{v}{2})}(1 + \frac{t^2}{v})^{-(\frac{v+1}{2})} \tag{3.14}$$

em que Γ[1] é a função gama. Usando-se a função beta β[2], a função densidade de probabilidade pode ser escrita como:

$$f(t) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\beta(\frac{1}{2}, \frac{v}{2})}(1 + \frac{t^2}{v})^{-(\frac{v+1}{2})} \tag{3.15}$$

A distribuição t de Student surge de forma orgânica no contexto de determinar a média de uma população (que segue uma distribuição normal) a partir de uma amostra. Nesse cenário, a média e o desvio padrão da população são desconhecidos, embora se pressuponha que a distribuição seja normal.

Supondo que o tamanho da amostra n seja muito menor que o tamanho da população, temos que a amostra é dada por n variáveis aleatórias normais independentes $X_1, ..., X_n$, cuja média $\bar{X}_n = (X_1 + ... + X_n)/n$ é o melhor estimador para a média da população.

Considerando $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ como a variância amostral, temos o seguinte resultado:

A variável aleatória t dada por:

$$t = \frac{X_n - \mu}{S_n/\sqrt{n}} \tag{3.16}$$

ou

$$t = \sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \tag{3.17}$$

segue uma distribuição t de Student com $v = n - 1$ graus de liberdade.

[1]Em Matemática, a funçao Γ é uma função definida por $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$

[2]Em Matemática, a funçao β, também chamada de integral de Euler de primeiro tipo, é a função definida por: $\beta(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$

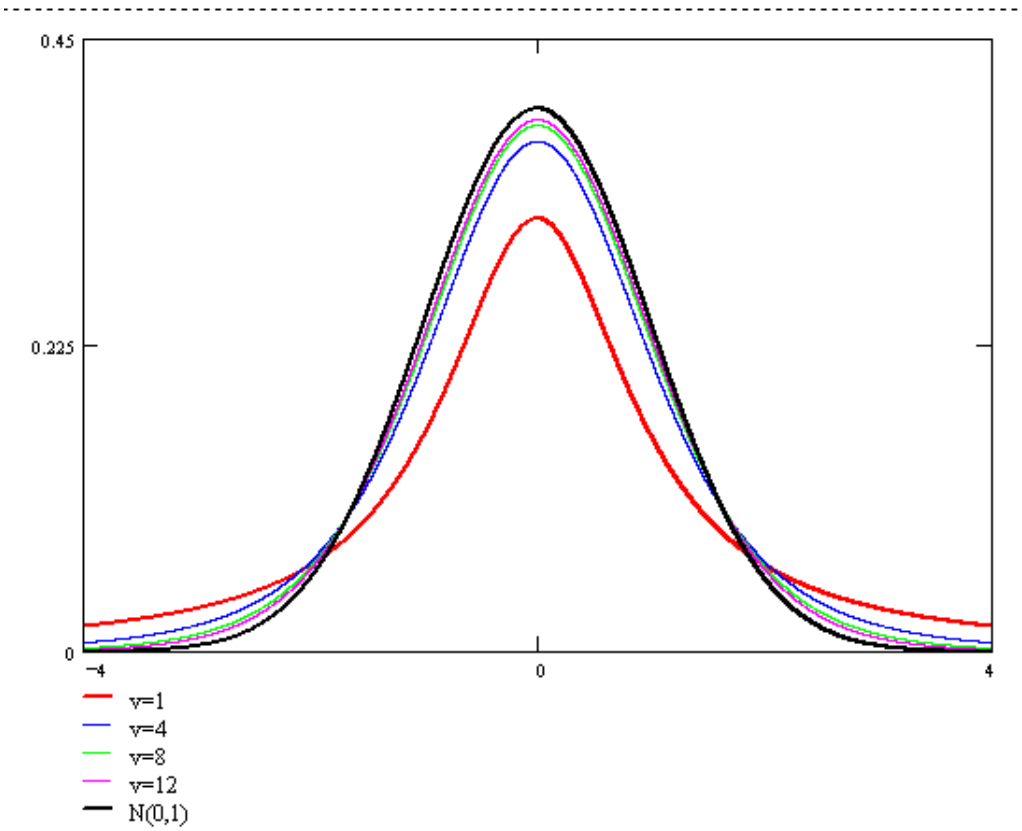


Figura 3.2: A função densidade da distribuição de Student para alguns valores de v e da distribuição normal

3.1.5 Teste de hipóteses

Em estatística, um Teste de Hipóteses[3] é um método usado para verificar se os dados são compatíveis com alguma hipótese, podendo, muitas vezes, sugerir a não-validade de uma hipótese. O teste de hipóteses é uma abordagem estatística que utiliza a análise de uma amostra para avaliar parâmetros desconhecidos em uma população, ou seja, para verificar a validade de afirmações sobre esses parâmetros.

Um Teste de Hipóteses pode ser paramétrico ou não-paramétrico. Testes paramétricos dependem de parâmetros da amostra, como média e desvio padrão. A escolha entre testes paramétricos e não paramétricos é influenciada pelo tamanho da amostra e pela distribuição da variável em estudo.

Os testes de hipóteses são sempre constituídos por duas hipóteses, a hipótese nula H_0 e a hipótese alternativa H_1.

• Hipótese nula (H_o) : afirmação estatística que sugere a ausência de efeitos ou diferenças notáveis em um experimento ou estudo. Essencialmente, é a posição inicial que os pesquisadores buscam verificar ou questionar.

[3] A expressão teste de significância foi criada por Ronald Fisher: "*Critical tests of this kind may be called tests of significance, and when such tests are available we may discover whether a second sample is or is not significantly different from the first.*"

• Hipótese alternativa (H_1) : procura evidenciar a presença de efeitos estatisticamente relevantes.

• Nível de significância: a probabilidade de rejeitar a hipótese nula quando ela é efetivamente verdadeira (ERRO).

Finalidade: avaliar afirmações sobre os valores de parâmetros.

O valor-p é uma estatística muito utilizada para sintetizar o resultado de um teste de hipóteses. Formalmente, o valor-p é definido como a probabilidade de se obter uma estatística de teste igual ou mais extrema quanto aquela observada em uma amostra, assumindo verdadeira a hipótese nula.

3.1.6 Teste de hipóteses para médias

O teste consiste em verificar, através de uma amostra, se a média da população atende o caso em teste (conforme desejemos testar diferença, valor inferior ou valor superior a uma referência para a média), para um certo nível de significância desejado.

Inicialmente devemos calcular:

$$Z_{calc} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{3.18}$$

onde:

μ= média esperada da população

$\bar{x}$= média da amostra

σ= desvio padrão da população

n=tamanho da amostra

Em seguida consultamos na tabela da curva normal o Z correspondente a cada caso.

Finalmente verificamos se Z_{calc}se encontra na área de rejeição conforme o caso em teste.

- Caso 1 - Unilateral ou unicaudal à esquerda

  $H_0 : \mu = \mu_0$

  $H_1 : \mu < \mu_0$

  Rejeitar se:

  Z_{calc}<-Z_α

- Caso 2 - Unilateral ou unicaudal à direita

  $H_0 : \mu = \mu_0$

  $H_1 : \mu > \mu_0$

  Rejeitar se:

  Z_{calc}>Z_α

- Caso 3 - Bilateral

  $H_0 : \mu = \mu_0$

  $H_1 : \mu \neq \mu_0$

  Rejeitar se:

  $Z_{calc} <$ -$Z_{\alpha/2}$

  $Z_{calc} > Z_{\alpha/2}$

onde Z_α é o valor crítico tabelado para um nível de significância α e μ_0 é o valor da hipótese a ser testada.

Para o caso onde o desvio padrão da população é desconhecido, devemos utilizar a fórmula 3.19.

$$t_{calc} = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{3.19}$$

onde:

s=desvio padrão da amostra

Em seguida, consultamos a tabela da distribuição t de Student para encontrar o t correspondente.

Finalmente verificamos se t_{calc}se encontra na área de rejeição conforme o caso em teste.

- Caso 1 - Unilateral ou unicaudal à esquerda

  $H_0 : \mu = \mu_0$

  $H_1 : \mu < \mu_0$

  Rejeitar se:

  $t_{calc} <$-t_α

- Caso 2 - Unilateral ou unicaudal à direita

  $H_0 : \mu = \mu_0$

  $H_1 : \mu > \mu_0$

  Rejeitar se:

  $t_{calc} > t_\alpha$

- Caso 3 - Bilateral

  $H_0 : \mu = \mu_0$

  $H_1 : \mu \neq \mu_0$

  Rejeitar se:

  $t_{calc} <$ -$t_{\alpha/2}$

  $t_{calc} > t_{\alpha/2}$

onde t_α é o valor crítico tabelado para um nível de significância α e μ_0 é o valor da hipótese a ser testada.

Veja nas Figuras 4.20 e **??** os diagramas de atividades dos testes de hipótese.

3.1.7 *Outliers*

O software também utiliza os testes estatísticos para identificar os *outliers* (irregularidades) que estão relacionados a alguns dados experimentais do laboratório de geoquímica.

Observações que se destacam das demais ou não seguem o padrão geral são chamadas de outliers. Esses pontos também são chamados de "anômalos", contaminantes, incomuns, extremos, anormais ou aberrantes.

Desde os primeiros esforços para analisar dados, a questão das observações outliers surgiu. Inicialmente, pensava-se em removê-las da análise. As opiniões divergiam: alguns recomendavam excluir observações "inconsistentes", enquanto outros argumentavam que todas as observações deveriam contribuir igualmente, sem exclusões, para o resultado final.

Antes de decidir o que deverá ser feito às observações *outliers* é conveniente ter conhecimento das causas que levam ao seu aparecimento. Em muitos casos as razões da sua existência determinam as formas como devem ser tratadas. Assim, as principais causas que levam ao aparecimento de *outliers* são:

- Erros de medição;
- Erros de execução;
- Variabilidade inerente dos elementos da população.

A análise de outliers, independentemente de suas origens, pode ser dividida em várias etapas. Na fase inicial, focamos na identificação de observações que podem ser consideradas diferentes. Essa identificação é feita subjetivamente, usando métodos como análise gráfica ou, em conjuntos de dados pequenos, pela observação direta dos valores. Assim, identificamos as observações que têm o potencial de serem outliers.

Na segunda etapa, busca-se eliminar a subjetividade da fase anterior. O objetivo é confirmar se as observações identificadas como potenciais outliers realmente o são. Testes são conduzidos nessas observações consideradas "preocupantes". Os testes escolhidos devem ser os mais adequados para a situação em análise. As observações suspeitas são avaliadas quanto à sua discordância. Se a hipótese de algumas observações serem outliers for aceita, elas são chamadas de discordantes. Uma observação é considerada discordante se parecer inconsistente com os outros valores após a aplicação de um critério estatístico objetivo. O termo "discordante" é frequentemente usado como sinônimo de outlier.

Na última etapa, é essencial decidir como lidar com as observações discrepantes. Embora a exclusão direta seja uma alternativa, como previamente indicado, isso não é aconselhável, a menos que os pontos fora da curva sejam causados por erros irreversíveis. Em outros cenários, é vital tratar meticulosamente as observações identificadas como pontos

fora da curva, pois elas encerram informações valiosas sobre características fundamentais dos dados e podem ser cruciais para entender a população à qual a amostra pertence. Abaixo, serão delineados testes comuns utilizados na geoquímica para detectar pontos fora da curva.

3.1.8 Teste do escore z modificado

Este teste é mais abrangente do que o teste que considera como outlier os valores que ultrapassam a soma da média aritmética com três desvios padrão, ou a média menos três desvios padrão. Isso acontece porque tanto a média quanto o desvio padrão são afetados pela presença do outlier. O teste do escore z modificado usa estimadores robustos, como a mediana, assegurando que os valores usados para definir um outlier não tenham sido impactados por ele.

Exemplo: Montaremos uma verificação da presença de um *outlier* com este teste (Tabela 3.1).

Tabela 3.1: Exemplo da verificação da presença de um *outlier*

Dado Original (Xi)	| Xi- Xm|	Zi
3.2	0.1	-0.34
3.3	0.0	0.00
8.1	4.8	16.19
3.2	0.1	-0.34
2.9	0.4	-1.35
3.7	0.4	1.35
3.1	0.2	-0.67
3.5	0.2	0.67
3.3	0.0	0.00
9.2	5.9	19.90

1. Calcular a mediana dos dados brutos, que vale $3,3$.

2. Determinar a coluna com os valores dos desvios absolutos, definida por $|Xi - Xm|$.

3. Determinar a média aritmética dos desvios absolutos (MAD), valores que constam da coluna criada no passo anterior, que vale $0,2$ neste caso.

4. Calcular os valores de z modificado para cada observação, gerando a coluna três da tabela anterior; este valor é representado por $z * i$, que vale $z * i = 0,6745(Xi - Xm)/MAD$. Para a terceira observação se tem $z*i = 0,6745(8,1–3,3)/0,2 = 16,19$; Para a quarta observação se tem $z * i = 0,6745(3,2–3,3)/0,2 = -0,34$; Para a décima observação se tem $z * i = 0,6745(9,2–3,3)/0,2 = 19,90$;

5. Considerar como *outliers* valores de$z * i > 3,5$, ou seja, no caso estudado são considerados outliers os valores relativos a $16,19$ e $19,90$da terceira coluna.

Veja na Figura 3.1.8 o diagrama de atividades para o teste Z modificado.

3.1.9 Teste de Grubbs

Este teste é usado para dados que têm uma distribuição lognormal. O teste de Grubbs é explicado com um exemplo prático. Ele é especialmente útil para avaliar a variabilidade entre laboratórios.

Exemplo:

Tabela 3.2: Exemplificando o teste de Grubbs

Dado Original (Xi)	Ln (Xi)	Com rank
2.15	0.77	0.77
11.76	2.46	1.14
5.08	1.63	1.63
3.12	1.14	2.19
12.87	2.55	2.46
32.13	3.47	2.55
219	5.39	2.98
19.69	2.98	3.47
179	5.19	3.87
9609	9.17	4.31
327	5.79	4.62
74.2	4.31	5.19
102	4.62	5.39
47.8	3.87	5.79
8.97	2.19	9.17

1. Calcular a média e o desvio padrão dos dados já transformados em logaritmos naturais, respectivamente $3,70$ e $2,17$.

2. Colocar os dados logtransformados em ordem crescente, com *rank* .

3. Se houver suspeita de *outlier* para o menor valor se faz.

4. $$\tau = \frac{Media - X_1}{S} \tag{3.20}$$

5. se houver suspeita de *outlier* para o maior valor se faz

6. $$\tau = \frac{X_n - Media}{S} \tag{3.21}$$

7. No presente exemplo, suspeitando-se do maior valor se tem

8. $$\tau_{15} = \frac{9,17 - 3,70}{2,17} = 2,52 \tag{3.22}$$

9. Para um $\alpha = 0,05$ se determina o τ crítico para$n = 15$, no caso $2,409$.

10. Se o valor calculado for maior que o crítico se rejeita a hipótese nula e se conclui que o dado testado é um outlier; no caso presente, se rejeita a hipótese nula e o valor testado é um outlier.

11. A Tabela 3.3 contém os valores críticos para o teste de Grubbs, com α valendo $0,10$,$0,05$,$0,025$, $0,01$ e $0,005$, unicaudais, ao se usar teste bicaudal se deve adotar a mesma tabela com o dobro das probabilidades α; esta tabela tem incrementos unitários para tamanhos de amostra entre 3 e 40 observações e incrementos de 10 unidades entre amostras com 40 a 140 observações.

Tabela 3.3: Valores críticos de Grubbs

n \ α	0.10	0.05	0.025	0.01	0.005
3	1.148	1.153	1.155	1.155	1.155
4	1.425	1.463	1.481	1.492	1.496
5	1.602	1.672	1.715	1.749	1.764
6	1.729	1.822	1.887	1.944	1.973
7	1.828	1.938	2.020	2.097	2.139
8	1.909	2.032	2.126	2.221	2.274
9	1.977	2.110	2.215	2.323	2.387
10	2.036	2.176	2.290	2.410	2.482
11	2.088	2.234	2.355	2.485	2.564
12	2.134	2.285	2.412	2.550	2.636
13	2.175	2.331	2.462	2.607	2.699
14	2.213	2.371	2.507	2.659	2.755
15	2.247	2.409	2.549	2.705	2.806
16	2.279	2.443	2.585	2.747	2.852
17	2.309	2.475	2.620	2.785	2.894
18	2.335	2.504	2.651	2.821	2.932
19	2.361	2.532	2.681	2.854	2.968
20	2.385	2.557	2.709	2.884	3.001
21	2.408	2.580	2.733	2.912	3.031
22	2.429	2.603	2.758	2.939	3.060
23	2.448	2.624	2.781	2.963	3.087
24	2.467	2.644	2.802	2.987	3.112
25	2.486	2.663	2.822	3.009	3.135
26	2.502	2.681	2.841	3.029	3.157
27	2.519	2.698	2.859	3.049	3.178
28	2.534	2.714	2.876	3.068	3.199
29	2.549	2.730	2.893	3.085	3.218
30	2.563	2.745	2.908	3.103	3.236
31	2.577	2.759	2.924	3.119	3.253
32	2.591	2.773	2.938	3.135	3.270
33	2.604	2.786	2.952	3.150	3.286
34	2.616	2.799	2.965	3.164	3.301
35	2.628	2.811	2.979	3.178	3.316
36	2.639	2.823	2.991	3.191	3.330
37	2.650	2.835	3.003	3.204	3.343
38	2.661	2.846	3.014	3.216	3.356
39	2.671	2.857	3.025	3.228	3.369
40	2.682	2.866	3.036	3.240	3.381
50	2.768	2.956	3.128	3.336	3.483
60	2.837	3.025	3.199	3.411	3.560
70	2.893	3.082	3.257	3.471	3.622
80	2.940	3.130	3.305	3.521	3.673
90	2.981	3.171	3.347	3.563	3.716
100	3.017	3.207	3.383	3.600	3.754
110	3.049	3.239	3.415	3.632	3.787
120	3.078	3.267	3.444	3.662	3.817
130	3.104	3.294	3.470	3.688	3.843
140	3.129	3.318	3.493	3.712	3.867

Veja na Figura 4.17 o diagrama de atividades do teste de Grubbs.

3.1.10 Teste de Dixon

Avaliando a diferença entre os valores máximo e mínimo e seus vizinhos, o teste de Dixon para valores extremos gera uma razão "r" que é associada a uma distribuição específica. O teste de Dixon é frequentemente utilizado para identificar poucos pontos fora da curva, especialmente quando o tamanho da amostra varia de 3 a 25 observações. Os dados são organizados, e uma estatística é calculada para o maior ou menor valor suspeito de ser um ponto fora da curva. Após estabelecer um nível de significância, esse valor é comparado com um valor crítico na tabela. Se for menor que o valor crítico, a hipótese nula, que indica a ausência de pontos fora da curva, não é rejeitada. Se a hipótese nula for rejeitada (quando o valor calculado é maior que o valor crítico), conclui-se que o valor testado é um ponto fora da curva. Para verificar a presença de outros pontos fora da curva, o teste é repetido, mas seu poder diminui à medida que o número de repetições aumenta. Alguns autores argumentam que o teste de Dixon não é mais a escolha mais adequada, devido à disponibilidade de métodos mais eficazes.

Exemplo:

1. Os dados devem ser ordenados do menor para o maior, sendo o menor valor o de ordem 1, e o maior valor o de ordem N

2. Chama-se$Z$ ao valor numérico do dado de ordem N, ou seja, $Z(1)$ é o valor numérico do menor resultado e $Z(N)$ é o valor numérico do resultado de maior valor numérico, $Z(N–1)$ é o valor do penúltimo dado em ordem crescente de valor numérico; ao se proceder ao teste Q se chama QM ao valor mais elevado (suspeito de ser *outlier*) e Qm ao valor menor (suspeito de ser *outlier*).

3. Procede-se ao teste de Dixon, de acordo com três situações:

4. havendo entre 3 e 7 observações

5. $$QM = [Z(N)–Z(N–1)]/[Z(N)–Z(1)] \tag{3.23}$$

6. $$Qm = [Z(2)–Z(1)]/[Z(N)–Z(1)] \tag{3.24}$$

7. havendo entre 8 e 12 observações

8. $$QM = [Z(N)–Z(N–1)]/[Z(N)–Z(2)] \tag{3.25}$$

9. $$Qm = [Z(2)–Z(1)]/[Z(N-1)–Z(1)] \tag{3.26}$$

10. havendo entre 13 e 25 observações

11.

$$QM = [Z(N)–Z(N–2)]/[Z(N)–Z(3)] \tag{3.27}$$

12.

$$Qm = [Z(3)–Z(1)]/[Z(N–2)–Z(1)] \tag{3.28}$$

Se existirem mais de 25 observações, o teste não é aplicável, sendo preciso procurar uma abordagem alternativa. A Tabela 3.4 apresenta os valores críticos do teste de Dixon para valores de α iguais a $0,10$, $0,05$ e $0,01$ unicaudais, para o caso bicaudal deve-se usar os mesmos valores críticos mas duplicando as probabilidades nos cabeçalhos das colunas. Esta tabela é válida ao se aplicar o teste de Dixon para conjuntos de dados que se ajustem à distribuição normal.

Tabela 3.4: Valores críticos de Dixon

n	$\alpha = 0,10$	$\alpha = 0,05$	$\alpha = 0,01$
3	0,886	0,941	0,988
4	0,679	0,765	0,889
5	0,557	0,642	0,780
6	0,482	0,560	0,698
7	0,434	0,507	0,637
8	0,479	0,554	0,683
9	0,441	0,512	0,635
10	0,409	0,477	0,597
11	0,517	0,576	0,679
12	0,490	0,546	0,642
13	0,467	0,521	0,615
14	0,492	0,546	0,641
15	0,472	0,525	0,616
16	0,454	0,507	0,595
17	0,438	0,490	0,577
18	0,424	0,475	0,561
19	0,412	0,462	0,547
20	0,401	0,450	0,535
21	0,391	0,440	0,524
22	0,382	0,430	0,514
23	0,374	0,421	0,505
24	0,367	0,413	0,497
25	0,360	0,406	0,489

Veja na Figura 4.18 o diagrama de atividades do teste de Dixon.

3.1.11 Teste de Cochran

O teste de Cochran é utilizado para analisar a consistência interna de um laboratório. O teste de Cochran é definido pela estatística C, que vale

$$C = S^2max/\sum_{i=1}^{p} S_i^2 \tag{3.29}$$

onde $Smax$ é o desvio padrão máximo no conjunto; a hipótese nula parte do princípio que a estatística C tem uma distribuição aproximada à de *Qui quadrado* com (m–1) graus de liberdade, onde m representa o número de variáveis.

A eficácia do teste de Cochran é impactada pela não normalidade dos dados e requer o uso de uma tabela específica, denominada tabela de Cochran.

O teste de Cochran é uma variante do teste t (de Student, que compara conjuntos cujas variabilidades não sejam muito diferentes entre si), quando as amostras apresentam diferenças de variabilidade, verificada por um teste F.

Em resumo, o teste de Cochran requer a ordenação crescente para cada par de repetições, seguido pela aplicação da equação 3.29 e avaliar o resultado obtido em relação ao valor de referência tabelado para este teste. Se o valor for menor que o tabelado, sugere a inexistência de dispersão. Se for maior, aponta para a presença de dispersão em relação à amplitude.

Exemplo:

Tabela 3.5: Exemplificando o teste de Cochran

	A	B	C	D
Média	2.75	3.50	6.25	9.00
Variância	2.214	0.857	1.071	1.714
Desvio Padrão	1.488	0.926	1.035	1.309

O valor de C será$2,214/5,856 = 0,378$com 4 grupos e 7 graus de liberdade, que são respectivamente$k$ e(n–1), quatro conjuntos e cada conjunto com oito valores. O valor crítico para C, considerado um α igual a $0,05$, é de$0,5365$. Em conclusão, se rejeita a hipótese nula de que as variâncias sejam iguais. Os valores críticos de Cochran estão nas Tabelas 3.6 e 3.7.

Tabela 3.6: Valores críticos do teste de Cochran para $\alpha = 0.05$

α = 0,05

N° grupos	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
GL																	
1	9985	9669	9065	8412	7808	7271	6798	6385	6020	5410	4709	3894	3434	2929	2370	1737	0998
2	9750	8709	7679	6838	6161	5612	5157	4775	4450	3924	3346	2705	2354	1980	1567	1131	0632
3	9392	7977	6841	5981	5321	4800	4377	4027	3733	3264	2758	2205	1907	1593	1259	0895	0495
4	9057	7457	6287	5441	4803	4307	3910	3584	3311	2880	2419	1921	1656	1377	1082	0765	0419
5	8772	7071	5895	5065	4447	3974	3595	3286	3029	2624	2195	1735	1493	1237	0968	0682	0371
6	8534	6771	5598	4783	4184	3726	3362	3067	2823	2439	2034	1602	1374	1137	0887	0623	0337
7	8332	6530	5365	4564	3980	3535	3185	2901	2666	2299	1911	1501	1286	1061	0827	0583	0312
8	8159	6333	5175	4387	3817	3384	3043	2768	2541	2187	1815	1422	1216	1002	0780	0552	0292
9	8010	6167	5017	4241	3682	3259	2926	2659	2439	2098	1736	1357	1160	0958	0745	0520	0279
10	7880	6025	4884	4118	3568	3154	2829	2568	2353	2020	1671	1303	1113	0921	0713	0497	0266
16	7341	5466	4366	3645	3135	2756	2462	2226	2032	1737	1429	1108	0942	0771	0595	0411	0218
36	6602	4748	3720	3066	2612	2278	2022	1820	1655	1403	1144	0879	0743	0604	0462	0316	0165
144	5813	4031	3093	2513	2119	1833	1616	1446	1308	1100	0889	0675	0567	0457	0347	0234	0120
∞	5000	3333	2500	2000	1667	1429	1250	1111	1000	0833	0667	0500	0417	0333	0250	0167	0083

Tabela 3.7: Valores críticos do teste de Cochran para $\alpha = 0.01$

α = 0,01

N° grupos	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
GL																	
1	9999	9933	9676	9279	8828	8376	7945	7544	7175	6528	5747	4799	4247	3632	2940	2151	1225
2	9950	9423	8643	7885	7218	6644	6152	5727	5358	4751	4069	3297	2821	2412	1915	1371	0759
3	9794	8831	7814	6957	6258	5685	5209	4810	4469	3919	3317	2654	2295	1913	1508	1069	0585
4	9586	8335	7212	6329	5635	5080	4627	4251	3934	3428	2882	2288	1970	1635	1281	0902	0489
5	9373	7933	6761	5875	5195	4659	4226	3870	3572	3099	2593	2048	1759	1454	1135	0796	0429
6	9172	7606	6410	5531	4866	4347	3932	3592	3308	2861	2386	1877	1608	1327	1033	0722	0387
7	8988	7335	6129	5259	4608	4105	3704	3378	3106	2680	2228	1748	1495	1232	0957	0668	0357
8	8823	7107	5897	5037	4401	3911	3522	3207	2945	2535	2104	1646	1406	1157	0898	0625	0334
9	8674	6912	5702	4854	4229	3751	3373	3067	2813	2419	2002	1567	1388	1100	0853	0594	0316
10	8539	6743	5536	4697	4084	3616	3248	2950	2704	2320	1918	1501	1283	1054	0816	0567	0302
16	7949	6059	4884	4094	3529	3105	2779	2514	2297	1961	1612	1248	1060	0867	0668	0461	0242
36	7067	5153	4057	3351	2858	2494	2214	1992	1811	1535	1251	0960	0810	0658	0503	0344	0178
144	6062	4230	3251	2644	2229	1929	1700	1521	1376	1157	0934	0709	0595	0480	0363	0245	0125
∞	5000	3333	2500	2000	1667	1429	1250	1111	1000	0833	0667	0500	0417	0333	0250	0167	0083

Estas tabelas contêm os valores críticos (C) do teste de Cochran para uniformidade

de variâncias em amostras de tamanho idêntico. Todos os valores das duas tabelas devem ser divididos por 10.000, ou seja, elas contêm apenas a parte decimal, a parte inteira vale sempre zero. Assim, na tabela relativa a $\alpha = 0,05$, para graus de liberdade (GL) valendo 5 e dois grupos o valor tabelado vale $0,8772$.

Veja o diagrama de atividades do teste de Cochran na Figura 4.19.

3.1.12 Teste de Doerffel

É um teste de fácil aplicação proposto por Doerffel em 1967, corroborado por Dean e Dixon em 1981, conforme citado por Wellmer (1998). Esse teste é adequado para conjuntos de dados pequenos e é essencialmente caracterizado por

$$Q = (X_a - X_r)/R \tag{3.30}$$

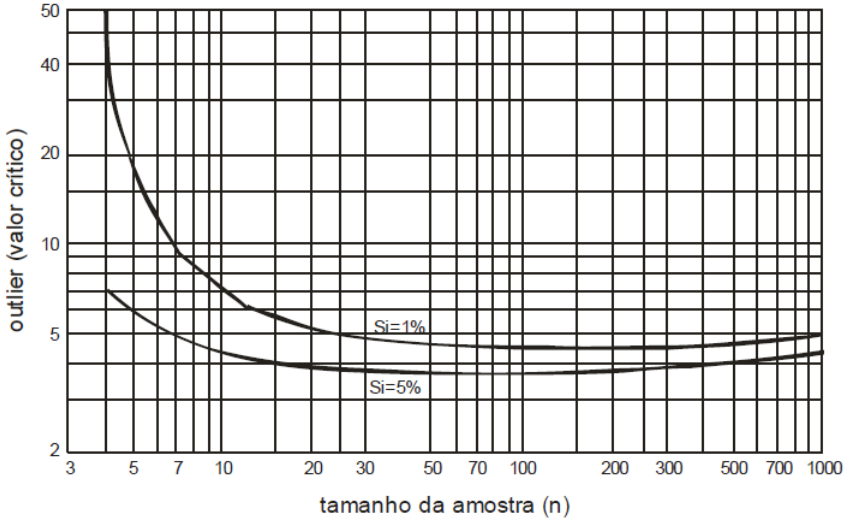
em que X_a é o valor que se suspeita seja um *outlier*, X_r é o valor adjacente (mais próximo) dele, R representa a amplitude dos dados (valor máximo – valor mínimo), e Q é o valor do teste. O valor testado será aceito se Q calculado for inferior ao valor tabelado por Doerffel (1967) e por Dean & Dixon (1981), na Tabela 3.8, reproduzida a seguir.

Tabela 3.8: Tabela de Doerffel e Dean & Dixon

n (Tamanho)	Q de Doerffel ($\alpha = 0.05$)	Q de Dean & Dixon ($\alpha = 0.05$)
3	0.97	0.94
4	0.84	0.76
5	0.73	0.64
6	0.64	0.56
7	0.59	0.51
8	0.54	0.47
9	0.51	0.44
10	0.49	0.41

Anteriormente, em 1962, Doerffel havia sugerido um método para identificar outliers com base em um diagrama, o valor é considerado aberrante por este método se superar a soma $(\mu + S * g)$, sendo que a média aritmética e o desvio padrão devem ser calculados sem o valor suspeito (pois ele afeta estes valores), e o valor de g pode ser obtido a partir de um diagrama específico para isto, este valor representa o *threshold* de um valor aberrante; este diagrama está representado na Figura 3.3.

Figura 3.3: Valores críticos de Doerffel



Exemplo:

Wellmer (1998) apresenta um exemplo de aplicação com os valores expressos em WO(%) iguais a $0,8$; $1,4$; $0,7$; $2,4$; $4,6$; $2,1$ e $1,5$; sendo o valor $4,6\%$ suspeito de ser um *outlier*. O valor adjacente a ele vale $2,4$. A aplicação do teste:

$$Q = (X_a - X_r)/R \tag{3.31}$$

$$(4,6 - 2,4)/(4,6 - 0,7) = 2,2/3,9 = 0,56 \tag{3.32}$$

Escolhendo-se o nível de significância ($Si = 5\%$) se verifica que o valor de$Q$ é menor que o valor de Si correspondente, logo o valor $4,6$ é aceitável, ou seja, não deve ser classificado como um *outlier*.

Veja na Figura 4.16 o diagrama de atividades do teste de Doerffel.

3.1.13 Análise de Variância

A Análise de Variância, ANOVA, é uma técnica estatística para comparar médias de três ou mais grupos e verificar se existem diferenças significativas entre eles. Ela avalia se a variação entre as médias é maior do que a variação dentro dos grupos, considerando a hipótese nula de que não há diferença significativa. É comumente usada em experimentos para analisar o efeito de diferentes variáveis independentes.

Os pincípios básicos para essa análise são:

1. Amostras aleatórias e independentes.

2. Teste paramétrico, ou seja, populações com distribuição normal.

3. Variâncias populacionais são iguais, então se refere a variabilidade dos dados em uma

população estatística completa. A fórmula para calcular a variância populacional é dada por:

$$\sigma^2 = \frac{1}{N}\Sigma_{i=1}^{N}(Xi - \mu)^2 (33)$$

onde, N é relativo ao tamanho da população, Xi são os valores individuais na população e μseria a média da população.

A variância populacional desempenha um papel crucial como indicador da dispersão dos dados em relação à média da população. Uma variância maior sugere uma dispersão mais ampla dos valores em relação à média. No entanto, na prática, é comum usar a estimativa da variância baseada em uma amostra da população, aplicando a fórmula da variância amostral. Isso é frequentemente realizado para compreender a distribuição dos dados de maneira mais precisa.

3.1.14 Teste Kolmogorov-Smirnov

O teste de Kolmogorov-Smirnov é usado para verificar se uma amostra segue uma distribuição específica. Ele compara a distribuição acumulada observada com a esperada, geralmente uma distribuição contínua como a normal. É essencial destacar que este teste é não paramétrico, o que significa que não faz suposições específicas sobre a forma da distribuição da amostra.

A estatística de teste D é a maior discrepância vertical absoluta entre a distribuição acumulada observada e a distribuição teórica esperada. A comparação com o valor crítico de D, definido para um certo nível de significância, determina se a hipótese nula (H_0- afirma que a amostra segue a distribuição teórica especificada) é rejeitada. Em resumo, no teste de Kolmogorov-Smirnov unidimensional para a distribuição contínua acumulada, a estatística de teste D é calculada como a maior discrepância vertical absoluta entre a distribuição empírica acumulada ($F_n(x)$) e a distribuição teórica acumulada ($F(x)$). Então a fórmula geral seria:

$$D = max|F_n(x) - F(x)| \ (34)$$

Em resumo, o teste de Kolmogorov-Smirnov é usado para verificar se uma amostra segue a mesma distribuição de probabilidade que uma distribuição teórica específica. É frequentemente utilizado em análises estatísticas para avaliar a adequação de um conjunto de dados a uma distribuição específica.

3.1.15 Teste U (Mann-Whitney)

O Teste U de Mann-Whitney é uma análise estatística que avalia se existem diferenças significativas entre duas amostras independentes. É empregado quando os dados não atendem aos requisitos necessários para conduzir um teste t-paramétrico, como o teste t de Student.

O procedimento do teste seria da seguinte forma:

1. Hipóteses (H_0e H_1).
2. Rankings: as amostras devem ser combinadas e classificadas do menor para o maior valor e então deverá ser atribuido a cada amostra um rank.
3. Soma dos ranks
4. Estatística U: menor valor entre a soma dos ranks dos dois grupos (pode ser necessário ajustar para o tamanho da amostra). Podemos usar também:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 (35)$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 (36)$$

5. Aproximação da normal Z, através de:

$$z = \frac{(U - m_U)}{\sigma_U} \ (37)$$

sendo m_Ua média e $\sigma_U$seria o desvio padrão.

Se a estatística U for menor do que o valor crítico, pode-se rejeitar a hipótese nula, indicando que há diferença estatisticamente significativa entre as duas amostras.

3.1.16 Teste Kruskal-Wallis

O Teste de Kruskal-Wallis é uma técnica estatística não paramétrica usada para avaliar se existem diferenças significativas entre três ou mais grupos independentes. Este teste oferece uma alternativa à análise de variância (ANOVA) quando os dados não atendem aos pressupostos necessários para realizar a ANOVA paramétrica. Em vez de focar nas médias, o Teste de Kruskal-Wallis compara as classificações médias entre os grupos. Ele é aplicado quando o objetivo é verificar se há diferenças estatisticamente significativas nas medianas entre esses grupos.

Etapas para realização do teste Kruskal-Wallis:

1. Formulação de hipóteses: H_0e H_1.
2. Organização e classificação dos dados do menor ao maior valor.
3. Obter ranks para cada grupo, usando:$R_i = \sum_{j=1}^{ni} r_{ij}$ (38)
4. Calculando H: É importante esclarecer que esse cálculo de H, conforme a equação (39), só será utilizado quando não houver empates dos valores analisados das amostras ou quando esse valor for muito pequeno. $H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{Ri^2}{n_i} - 3(N+1)$ (39)

Existem algumas observações sobre o teste H, são elas:

1. se k=3 e $n_i \leq 5$, deve ser utilizada a tabela padrão da estatística H.
2. se K $>$ 3 e $n_i \geq 5$, H tem aproximadamente a distribuição Qui-quadrado.

3.2 Análise de domínio

- Área relacionada:

A disciplina fundamental associada ao desenvolvimento de software é a Estatística, que se baseia nas teorias probabilísticas para descrever a frequência de eventos, tanto em estudos observacionais quanto em experimentos, modelando a aleatoriedade e a incerteza para estimar ou permitir a previsão de fenômenos futuros, conforme a situação.

- Sub-área relacionada:

A área correlata é a geoquímica do petróleo. Esta disciplina é responsável pela identificação de rochas geradoras em bacias sedimentares, análise de maturação, biodegradação e avaliação do potencial gerador dessas rochas. A geoquímica pode ser definida como a aplicação de métodos químicos no estudo de fenômenos geológicos. No contexto específico da geoquímica do petróleo, trata-se da utilização de métodos da química orgânica na geologia do petróleo. A geoquímica faz uso da estatística para analisar os resultados dos experimentos.

3.3 Identificação de pacotes – assuntos

- Geoquímica: Utilizada para fornecer os dados experimentais.

- Estatística: Utilizada para identificar as anomalias dos dados experimentais.

3.4 Diagrama de pacotes – assuntos

A Figura 3.5 representa um diagrama de pacotes simplificado, para o programa a ser desenvolvido. Nele estão contidos os pacotes (assuntos) descritos anteriormente.

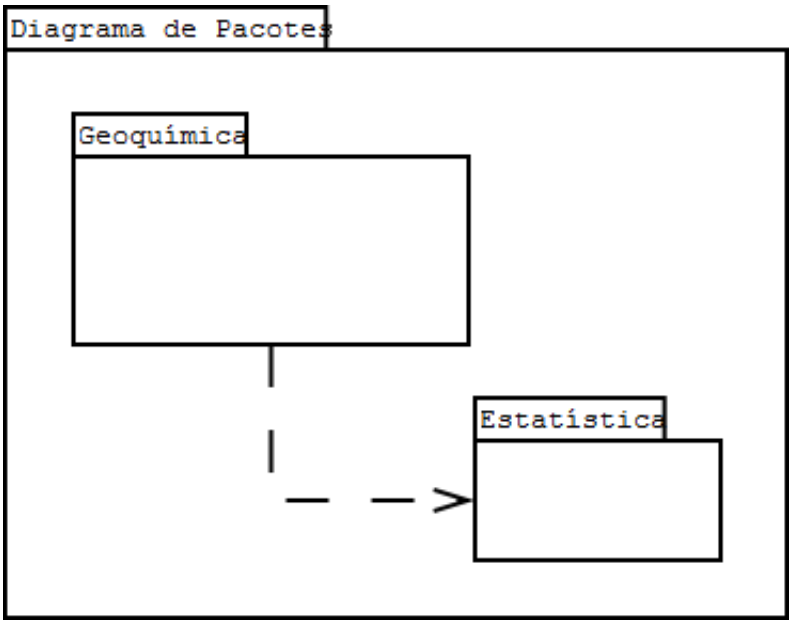


Figura 3.5: Diagrama de Pacotes

Figura 3.4: Tabela de Mann-Whitney

Critical Values of the Mann-Whitney U
(Two-Tailed Testing)

n_2	α	n_1																	
		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	.05	--	0	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
	.01	--	0	0	0	0	0	0	0	0	1	1	1	2	2	2	2	3	3
4	.05	--	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
	.01	--	--	0	0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	.05	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
	.01	--	--	0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	.05	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
	.01	--	0	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	.05	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
	.01	--	0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	.05	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
	.01	--	1	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9	.05	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
	.01	0	1	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10	.05	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
	.01	0	2	4	6	9	11	13	16	18	21	24	26	29	31	34	37	39	42
11	.05	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
	.01	0	2	5	7	10	13	16	18	21	24	27	30	33	36	39	42	45	48
12	.05	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
	.01	1	3	6	9	12	15	18	21	24	27	31	34	37	41	44	47	51	54
13	.05	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
	.01	1	3	7	10	13	17	20	24	27	31	34	38	42	45	49	53	56	60
14	.05	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
	.01	1	4	7	11	15	18	22	26	30	34	38	42	46	50	54	58	63	67
15	.05	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
	.01	2	5	8	12	16	20	24	29	33	37	42	46	51	55	60	64	69	73
16	.05	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
	.01	2	5	9	13	18	22	27	31	36	41	45	50	55	60	65	70	74	79
17	.05	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
	.01	2	6	10	15	19	24	29	34	39	44	49	54	60	65	70	75	81	86
18	.05	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
	.01	2	6	11	16	21	26	31	37	42	47	53	58	64	70	75	81	87	92
19	.05	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
	.01	3	7	12	17	22	28	33	39	45	51	56	63	69	74	81	87	93	99
20	.05	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
	.01	3	8	13	18	24	30	36	42	48	54	60	67	73	79	86	92	99	105

Capítulo 4

AOO – Análise Orientada a Objeto

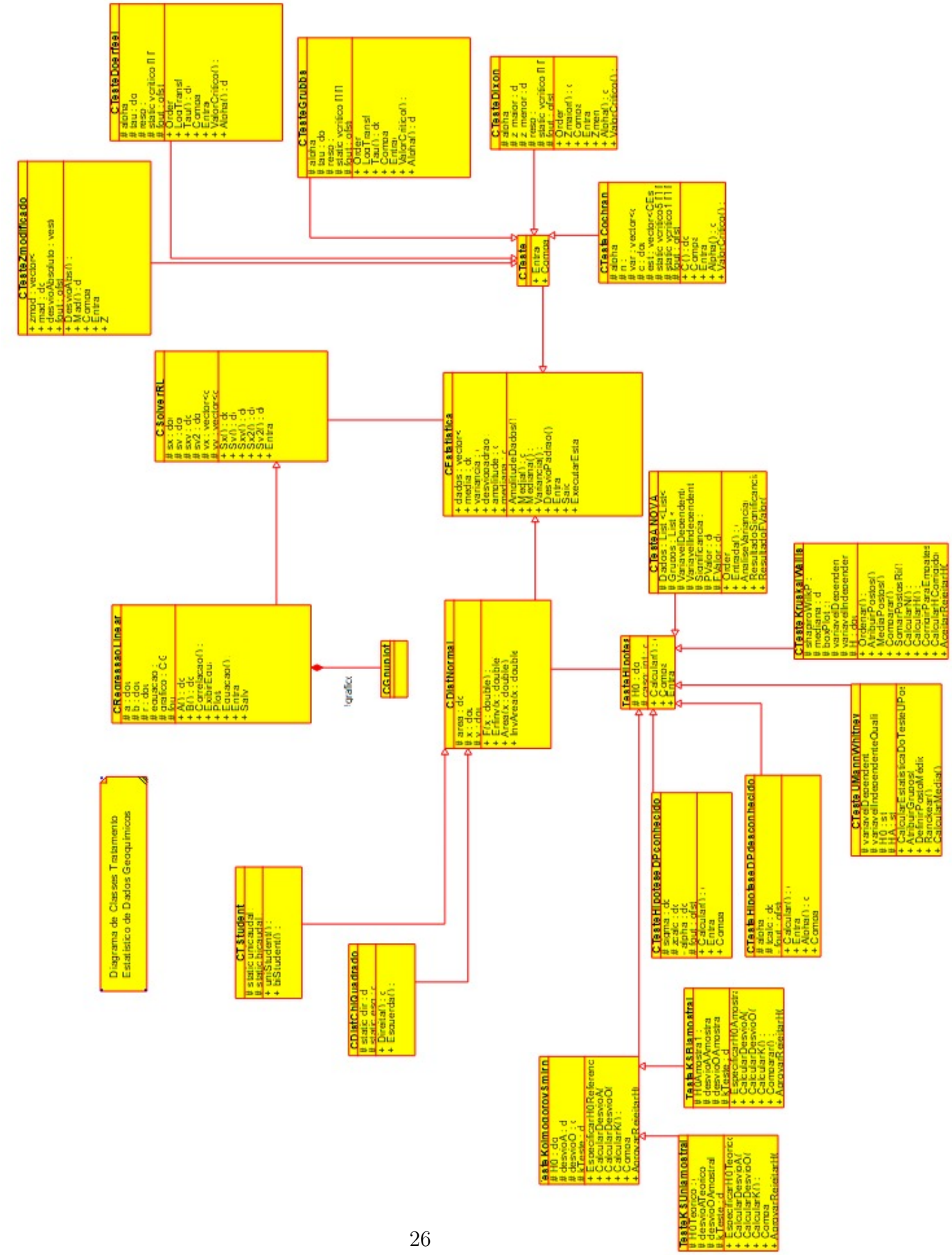
A terceira etapa no desenvolvimento de um sistema é a Análise Orientada a Objeto (AOO). Nessa etapa, são aplicadas diversas regras para identificar objetos de interesse, estabelecer relações entre pacotes, classes, atributos, métodos, heranças, associações, agregações, composições e dependências.

O modelo de análise deve ser conciso e simplificado, concentrando-se no que precisa ser feito, sem se preocupar com o processo de implementação.

O resultado da análise é um conjunto de diagramas que identificam os objetos e seus relacionamentos.

4.1 Diagramas de classes

O diagrama de classes é apresentado na Figura 4.1.


CTesteDoerfel
CTesteGrubbs
CTesteDixon
CTesteZmodificado
CTesteCochran
CTeste
CSolveRL
CEstatistica
CTesteANOVA
CTesteKruskalWallis
CRegressaoLinear
CGnuplot
CDistNormal
CTesteHipoteseDPconhecido
CTesteHipoteseDPdesconhecido
CTesteUMannWhitney
CT Student
CDistChiQuadrado
TesteKSUnamostral
TesteKSKolmogorovSmirn
TesteKSBiamostral
Diagrama de Classes Tratamento
Estatistico de Dados Geoquimicos

Os métodos Get() e Set() e os métodos construtores e destrutores foram omitidos por serem padrão em linguagens de programação orientada a objeto. Entretanto, todos eles foram implementados no código do programa.

A classe CTesteHipotese é uma classe abstrata e possui o método virtual puro calcularZ(). Esta classe é uma classe base que possui como classes derivadas CTesteHipoteseDPconhecido, CTesteHipoteseDPdesconhecido, CTesteANOVA, CTesteUMannWhitney, CTesteKruskalWallis, CTesteKolmogorovSmirnov. Para o teste de hipótese será implementado o polimorfismo, onde o usuário poderá escolher o teste de hipótese desejado. Nessa implementação será criado um ponteiro do tipo CTesteHipótese através do qual o programa irá executar o método calcularZ() do teste escolhido em tempo de execução.

A classe CTeste é uma classe abstrata e possui dois métodos virtuais puros Entrada() e Compara(). Esta classe é que possui como classes derivadas CTesteZmodificado , CTesteGrubbs, CTesteDoerffel, CTesteDixon, CTesteCochran. Para os testes para detecção de anomalia será implementado o polimorfismo, onde o usuário poderá escolher o teste a ser realiazado. Nessa implementação será criado um ponteiro para um objeto da classe base através do qual o programa irá executar os métodos Entrada() e Compara().

4.1.1 Dicionário de classes

- Classe CSolverRL: Classe que possui um método que calcula o somatório de$x$, somatório de y, somatório dexy, somatório de x^2, somatório de y^2 que serão utilizados pela classe CRegressaoLinear para o cálculo dos coeficientes a e b da equação da reta$y = a + bx$ e o coeficiente de correlação linear.

- Classe CRegressaoLinear: Classe representativa de uma regressão linear que descreve o relacionamento entre duas variáveis em termos matemáticos. Veja seção 3.1.2.

- CTeste : Classe de interface que permite o uso do polimorfismo nas classes de teste de identificação de anomalias (*outliers*).

- Classe CTesteZmodificado: Classe representativa de um teste realizado pra identificar anomalias. Veja seção 3.1.8.

- Classe CTesteDoerffel: Classe representativa de um teste realizado pra identificar anomalias. Veja seção 3.1.12.

- Classe CTesteGrubbs: Classe representativa de um teste realizado pra identificar anomalias (muito útil para testar variabilidade entre laboratórios). Veja seção 3.1.9.

- Classe CTesteDixon: Classe representativa de um teste realizado pra identificar anomalias (recomendado quando o número de observações está entre 3 e 25). Veja seção 3.1.10.

- Classe CTesteCochran: Classe representativa de um teste realizado pra identificar anomalias (recomendado para estudar a variabilidade interna de um laboratório). Veja seção 3.1.11.
- Classe CEstatística: Classe representativa dos cálculos estatísticos básicos. Veja seção 3.1.1.
- Classe CTesteHipotese: Classe representativa de um teste cujo objetivo é verificar se os dados são compatíveis com alguma hipótese. Veja seção 3.1.5.
- Classe CTesteHipoteseDPdesconhecido: Classe representativa de um teste cujo objetivo é verificar se os dados são compatíveis com alguma hipótese. Veja seção 3.1.6.
- Classe CTesteANOVA é a classe representativa de um teste cujo objetivo é conduzir a Análise de Variância (ANOVA), armazenar os resultados, e verificar se os dados são compatíveis com alguma hipótese.
- Classe CTesteHipoteseDPconhecido: Classe representativa de um teste cujo objetivo é verificar se os dados são compatíveis com alguma hipótese. Veja seção 3.1.6.
- Classe CDistNormal: Classe representativa de distribuição normal padrão de probabilidade. Veja seção **??**.
- Classe CDistTStudent: Classe representativa da distribuiçao t de Student de probabilidade. Veja seção **??**.
- Classe CDistChiQuadrado: Classe representativa da distribuição Chi-Quadrado de probabilidade.
- Classe CTesteANOVA: Classe representativa de um teste (ANOVA) cujo objetivo é verificar se os dados são compatíveis com alguma hipótese.
- Classe CTesteKolmogorovSmirnov: Classe representativa de um teste (Kolmogorov Smirnov) cujo objetivo é verificar se os dados são compatíveis com alguma hipótese.
- Classe CTesteUMannWhitney: Classe representativa de um teste (U Mann de Whitney) cujo objetivo é verificar se os dados são compatíveis com alguma hipótese.
- ClasseCTesteKruskalWallis: Classe representativa de um teste () cujo obKruskall Wallis cujo objetivo é verificar se os dados são compatíveis com alguma hipótese.

4.2 Dicionário de classes com atributos/métodos

- Classe CSolverRL:

  Atributos:

  - double sx - somatório dos valores de x.
  - double sy - somatório dos valores de y.
  - double sxy - somatório dos valores de $x * y$.
  - double sx2 - somatório dos valores de x^2.
  - double sy2 - somatório dos valores de y^2.
  - vector<double> vx - vetor de ordenada x do ponto (x, y).
  - vector<double> vy - vetor de ordenada y do ponto (x, y).

  Métodos:

  - double Sx() - soma todos os valores contidos no vetor vx da classe CRegressaoLinear.
  - double Sy() - soma todos os valores contidos no vetor vy da classe CRegressaoLinear.
  - double Sxy() - soma dos produtos $x_i * y_j$, com $i = j$, percorrendo todos os vetores.
  - double Sx2() - somatório do quadrado de cada elemento do vetor vx.
  - double Sy2() - somatório do quadrado de cada elemento do vetor vy.
  - void Entrada() - solicita os dados ao usuário.

- Classe CRegressaoLinear:

  Atributos:

  - double a - coeficiente linear.
  - double b - coeficiente angular.
  - double r - coeficiente de correlação.
  - string equacao - armazena o nome da reta encontrada pela regressão.
  - CGnuplot grafico - objeto da classe CGnuplot utilizado para acessar o programa *Gnuplot*.
  - ofstream fout - Objeto associado à arquivo de disco para escrita.

  Métodos:

  - double A() - calcula o valor do coeficiente linear da reta.
  - double B() - calcula o valor do coeficiente angular da reta.
  - double Correlacao() - calcula a associação numérica entre duas variáveis.
  - void ExibirEquacao() - após o cálculo, exibe a equação da reta.

  - void Plotar() - plota um gráfico da reta $y = a + bx$.
  - string Equacao() - retorna a equação como uma string para utilização no método Plotar().
  - void Salvar() - salva o resultado em arquivo de disco.

- Classe CTesteZmodificado:

  Atributos:

  - vector <double> desvioAbsoluto - armazena o módulo dos desvios absolutos dos dados.
  - vector <double> desvio - armazena os desvios absolutos dos dados.
  - double mad - mediana dos módulos dos desvios absolutos.
  - vector<double> zmod - armazena os valores da estatística calculada do teste.
  - ofstream fout - Objeto associado à arquivo de disco para escrita.

  Métodos:

  - double DesvioAbs() - calcula os desvios absolutos e seu módulo e preenche os vetores desvioAbsoluto e desvio.
  - double Mad() - calcula a mediana do módulo dos desvios absolutos.
  - double Z() - calcula os valores de z modificado.
  - void Compara() - compara a estatística calculada.
  - void Entrada() - solicita os dados ao usuário.

- Classe CTesteDoerffel:

  Atributos:

  - double valorSuspeito - valor suspeito de ser uma anomalia (*outlier*).
  - double valorAdjacente - valor mais próximo do *outlier.*
  - double q - estatística do teste.
  - static double valorCritico[][] - matriz que armazena os valores críticos do teste.
  - vector <double> dif - guarda os valores das diferenças entre o valor suspeito e os demais.
  - ofstream fout - Objeto associado à arquivo de disco para escrita.

  Métodos:

  - double ValorAdjacente() - calcula o valor adjacente, mais próximo do *outlier.*
  - double Q() - calcula a estatística do teste.

  - void Compara() - compara o valor calculado ao da tabela para saber se será aceito ou rejeitado.
  - void Entrada() - solicita os dados ao usuário.
  - double ValorCritico() - retorna o valor crítico do teste.

- Classe CTesteGrubbs:

  Atributos:

  - int alpha - nível de significância.
  - double tau - estatística do teste.
  - int resp - identifica a opção do usuário por calcular o maior ou menor valor.
  - static double valorCritico[][] - matriz que armazena os valores críticos.
  - ofstream fout - Objeto associado à arquivo de disco para escrita.

  Métodos:

  - double LogTransform() - método que calcula o logaritmo natural.
  - double Ordenar() - métodos que ordena os dados.
  - double Tau() - cálcula a estatística do teste de Grubbs.
  - void Compara() - compara o valor calculado ao da tabela para saber se será aceito ou rejeitado.
  - void Entrada() - solicita os dados ao usuário.
  - double Alpha() - método que retorna o valor da significância.
  - double ValorCritico() - retorna o valor crítico tabelado.

- Classe CTesteDixon:

  Atributos:

  - int alpha - nível de significância.
  - int resp - identifica a opção do usuário se descartará o maior ou menor elemento.
  - double z_maior - maior valor de z.
  - double z_menor - menor valor de z.
  - static double valorCritico[][] - armazena os valores críticos do teste.
  - ofstream fout - Objeto associado à arquivo de disco para escrita.

  Métodos:

  - void Ordenar() - método que coloca os dados obtidos em ordem crescente.

  - double Zmaior() - calcula a estatística do teste de Dixon para o maior elemento.
  - double Zmenor()- calcula a estatística do teste de Dixon para o menor elemento.
  - void Compara() - compara o valor calculado ao da tabela para saber se será aceito ou rejeitado.
  - void Entrada() - solicita os dados ao usuário.
  - double Alpha() - retorna o nível de significância do teste.
  - double ValorCritico() - retorna o valor crítico tabelado.

- Classe CTesteCochran:

  Atributos:

  - int alpha - nível de significância.
  - int n - quantidade de grupos de amostras.
  - vector<double> var - armazena a variância de cada grupo de amostra.
  - double c - estatística do teste Cochran.
  - std:: vector<CEstatistica> est - vetor que armazena os dados dos grupos.
  - static double vcritico5[][] - tabela do valor crítico para alpha igual 0.05.
  - static double vcritico1[][] - tabela do valor crítico para alpha igual 0.01.
  - ofstream fout - Objeto associado à arquivo de disco para escrita.

  Métodos:

  - double C() - calcula a estatística do teste Cochran.
  - void Compara() - compara o valor calculado ao da tabela para saber se será aceito ou rejeitado.
  - void Entrada() - solicita os dados ao usuário.
  - double Alpha() - retorna o valor do nível de significância utilizada.
  - double ValorCritico() - retorna o valor crítico tabelado.

- Classe CEstatística:

  Atributos:

  - int numeroDados - número de dados utilizados.
  - vector<double> dados - armazena os dados da amostra.
  - double media - média dos dados.
  - double variancia - variância dos dados.
  - double desviopadrao- desvio padrão dos dados.

  - double amplitude - amplitude dos dados.
  - double mediana - mediana dos dados.

  Métodos:

  - double AmplitudeDados() - calcula e retorna a amplitude dos dados (valor máximo - valor mínimo).
  - double Media() - calcula e retorna a media.
  - double Mediana() - calcula e retorna a mediana.
  - double Variancia() - calcula e retorna a variância.
  - double DesvioPadrao() - calcula e retorna o desvio padrão.
  - void Entrada() - solicita os dados ao usuário.

- Classe CTesteHipotese:

  Atributos:

  - double H0 - valor da hipótese nula.
  - int caso - identifica o caso escolhido pelo usuário (unicaudal à esquerda, unicaudal à direita ou bicaudal).

  Métodos:

  - double Calcular() - calcula a estatística do teste de hipótese.
  - double Entrada() - solicita dados ao usuário.
  - double Compara() - compara a estatística do teste calculada com o valor tabelado.

- Classe CTesteHipoteseDPdesconhecido:

  Atributos:

  - int alpha - valor que identifica o nível de significância.
  - double tcalc - estatística do teste.
  - static double valoresTabelados[][] - armazena os valores tabelados.
  - ofstream fout - Objeto associado à arquivo de disco para escrita.

  Métodos:

  - double Calcular() - calcula a estatística do teste de hipótese para um desvio padrão desconhecido.
  - void Entrada() - solicita os dados ao usuário.

  - void Compara() - método que compara a estatística calculada e o valor crítico tabelado.
  - double Alpha() - método que retorna o valor da significância.
  - double ValorCritico() - retorna o valor crítico do teste.

- Classe CTesteHipoteseDPconhecido:

  Atributos:

  - double sigma - valor do desvio padrão populacional.
  - double zcalc - estatística calculada.
  - double alpha - identifica o nível de significância.
  - ofstream fout - Objeto associado à arquivo de disco para escrita.

  Métodos:

  - double Calcular() - calcula a estatística do teste de hipótese para um desvio padrão desconhecido.
  - void Entrada() - solicita os dados ao usuário.
  - void Compara () - Compara o valor calculado e o valor tabelado.

- Classe CDistNormal:

  Atributos:

  - double area - área que representa a probabilidade acumulada.
  - double x - abcissa da distribuição normal.
  - double y - ordenada da distribuição normal.

  Métodos:

  - double F (double) - calcula a ordenada da distribuição normal.
  - double Erfinv (double) - aproximação da função erro inversa por séries de Taylor.
  - double Area (double) - calcula a área a esqueda do parâmetro _x utilizando a função erro.
  - double InvArea (double) - o usuário entra com o valor da probabilidade e método retorna a abcissa correspondente.

- Classe CDistTStudent:

  Atributos:

  - static double unicaudal[][] - tabela de valores para testes unicaudais.

  - static double bicaudal[][] - tabela de valores para testes bicaudais.

  Métodos:

  - double uniStudent(int _x, int _y) - retorna o valor crítico para testes unicaudais.
  - double biStudent(int x, int y) - retorna os valores críticos para testes bicaudais.

- Classe CDistChiQuadrado:

  Atributos:

  - static double dir[][] - tabela de valores para testes unicaudais à direita.
  - static double esq[][] - tabela de valores para testes unicaudais à esquerda.

  Métodos:

  - double Direita(int x, int y) - retorna o valor crítico para testes unicaudais à direita.
  - double Esquerda(int x, int y) - retorna os valores críticos para testes unicaudais à esquerda.

- Classe CTesteANOVA

  Atributos:

  - double Dados: List - dados dos estudos geoquímicos.
  - double Grupos: List - dados de cada grupo
  - double VariavelDependente- variável intervalar/razão e contínua.
  - double VariavelIndependente- variável qualitatica com dois ou mais níveis
  - double FValor- razão F(variância entre as médias dos grupos / variância dentro dos grupos)

  Métodos:

  - string Ordenar()- organiza os dados
  - double Entrada()- solicita os dados ao usuário.
  - string AnaliseVariancia()- analisa a variância.
  - double ResultadosSignificancia()- resultado de significância.
  - double FValor()- calcula a razão F(variância entre as médias dos grupos / variância dentro dos grupos).

- Classe CTesteKruskalWallis

  Atributos:

  - double shapiroWilkP- dados shapiroWilkp.
  - double mediana- dados medianas.
  - string boxPlot- analise boxPlot.
  - double VariavelDependente- variável intervalar/razão e contínua.
  - double VariavelIndependente- variável qualitatica com dois ou mais níveis.
  - double H- valor da hipótese.

  Métodos:

  - string Ordenar()- organizar de forma crescente do menor para o maior.
  - double AtribuirPostos()- posicionar cada valor.
  - double MediaPostos()- media para os postos que ocupam as mesmas colocações.
  - double Comparar()- analisar postos de cada grupo.
  - double SomarPostosRi()- somar os postos para cada grupo.
  - douvle CalcularN()- soma de n por grupos.
  - double CalcularH()- calcular H através da fórmula (H= 12/N(N+1).[Somatório (Ri^2/ni)-(3(N+1))]).
  - double CorrigirParaEmpates()- aplicar a corração de empates (CH- 1.[(Somatório t^3-t)/N^3-N).
  - string AceitarRejeitarH0()- aceitar ou rejeitar hipótese.

- Classe CTesteUMannWhitney

  Atributos:

  - double variavelDependente- intervalar/razão e continua.
  - double variavelIndependenteQualitativa- qualitativa com dois níveis.
  - string H0- média dos dados que foram amostrados que são iguais.
  - string HA- média dos dados que foram amostrados que não são iguais.

  Métodos:

  - double CalcularEstatisticaDoTesteUPosicoes()-
  - string AtribuirGrupos()- definir grupos.
  - string DefinirPostoMedio()- definir o posto médio de cada grupo.
  - double Ranckear()- ranckear os dados em ordem crescente de dois grupos comparados.

  - double CalcularMedia()- calcular media de ambos grupos apos ranckear e analisar.

- Classe CTesteKolmogorovSmirnov

  Atributos:

  - double H0- distribuição acumulada.
  - double desvioA- frequência relativa acumulada.
  - double desvioO- frequência relatica acumulada.
  - doubler kTeste- máximo Ki

  Métodos:

  - double EspecificarH0Referencial()- especificar a distribuição acumulada H0
  - double CalcularDesvioA()- determinar os desvios Ki= Ai-Oi
  - double CalcularDesvioO()- determinar os desvios Ki= Ai-Oi
  - double Calculark-kteste= máximo (ki)
  - string AprovarRejeitarH0()-aceitar ou rejeitar hipótese.

- Classe CTesteKSUniamostral

  Atributos:

  - double H0Teorico- especificar a distribuição acumulada teórica H0
  - double desvioATeorico- frequência relativa acumulada teórica.
  - double desvioOAmostral- frequência relatica acumulada amostral.
  - doubler kTeste- máximo Ki.

  Métodos:

  - double EspecificarH0Teorico()- especificar a distribuição acumulada teórica H0.
  - double CalcularDesvioA()- determinar os desvios Ki= Ai teórica-Oi amostral.
  - double CalcularDesvioO()- determinar os desvios Ki= Ai teórica-Oi amostral.
  - double Calculark--kteste= máximo (ki)
  - string Comparar ()- analisar
  - string AprovarRejeitarH0()-aceitar ou rejeitar hipótese.

- Classe CTesteBiamostral

  Atributos:

  - double H0Amostra1- especificar a distribuição acumulada amostra1 H0

- double desvioAAmostra- frequência relativa acumulada amostra1.
- double desvioOAmostra-frequência relatica acumulada amostral2.
- doubler kTeste-máximo Ki.

Métodos:

- double EspecificarH0Amostra()-especificar a distribuição acumulada amostra1 H0.
- double CalcularDesvioA()- determinar os desvios Ki= Ai amostra1-Oi amostra2.
- double CalcularDesvioO()-determinar os desvios Ki= Ai amostra1-Oi amostra2.
- double Calculark-kteste= máximo (ki).
- string Comparar ()-analisar.
- string AprovarRejeitarH0()-aceitar ou rejeitar hipótese.

4.3 Diagrama de sequência – eventos e mensagens

O diagrama de seqüência enfatiza a troca de eventos e mensagens e sua ordem temporal. Contém informações sobre o fluxo de controle do programa. Costuma ser montado a partir de um diagrama de caso de uso e estabelece o relacionamento dos atores (usuários e sistemas externos) com alguns objetos do sistema.

4.3.1 Diagramas de sequência

O diagrama de seqüência da Figura 4.2 mostra o comportamento típico de um objeto da classe CRegressaoLinear.

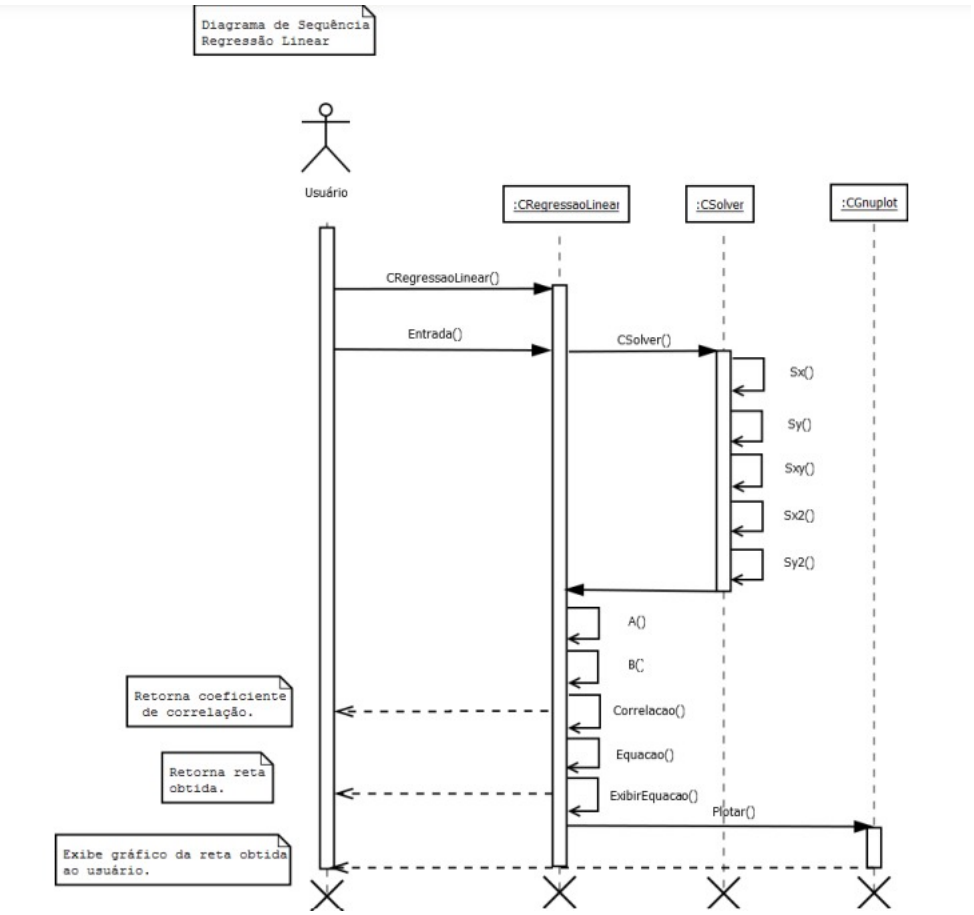


Figura 4.2: Diagrama de Sequência típico da classe CRegressaoLinear

O diagrama de sequência da Figura 4.3 mostra o comportamento típico de um objeto da classe CTesteZmodificado.

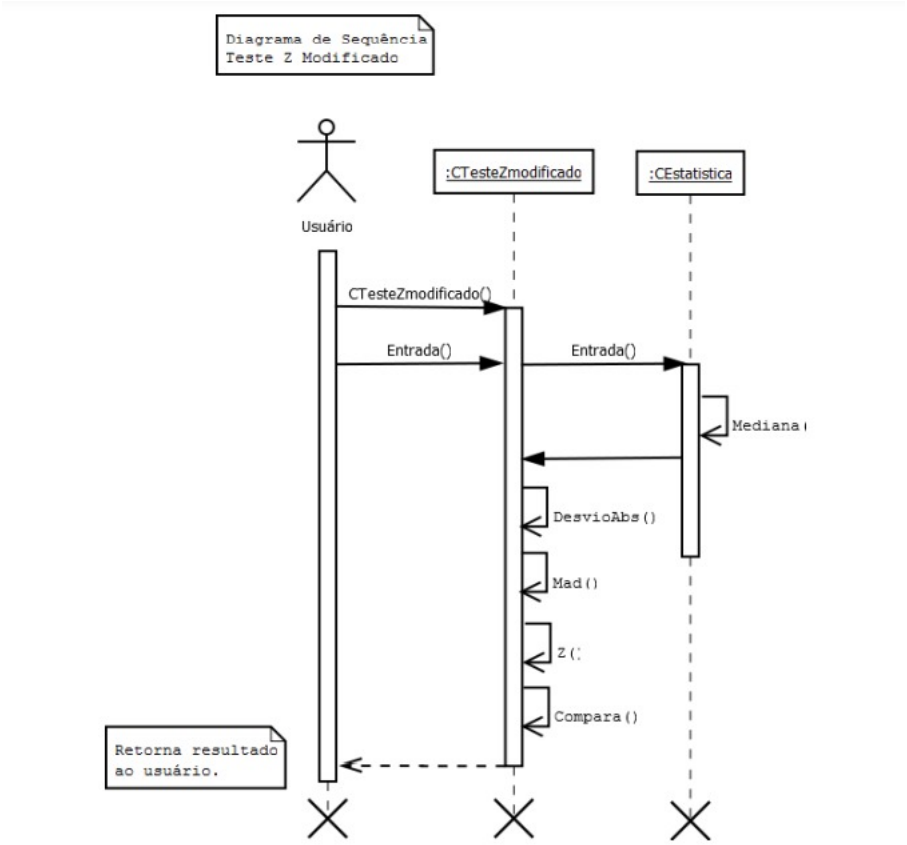


Figura 4.3: Diagrama de Seqüência típico da classe CTesteZmodificado

A Figura 4.4 mostra o diagrama de sequência típico da classe CTesteDoerffel.

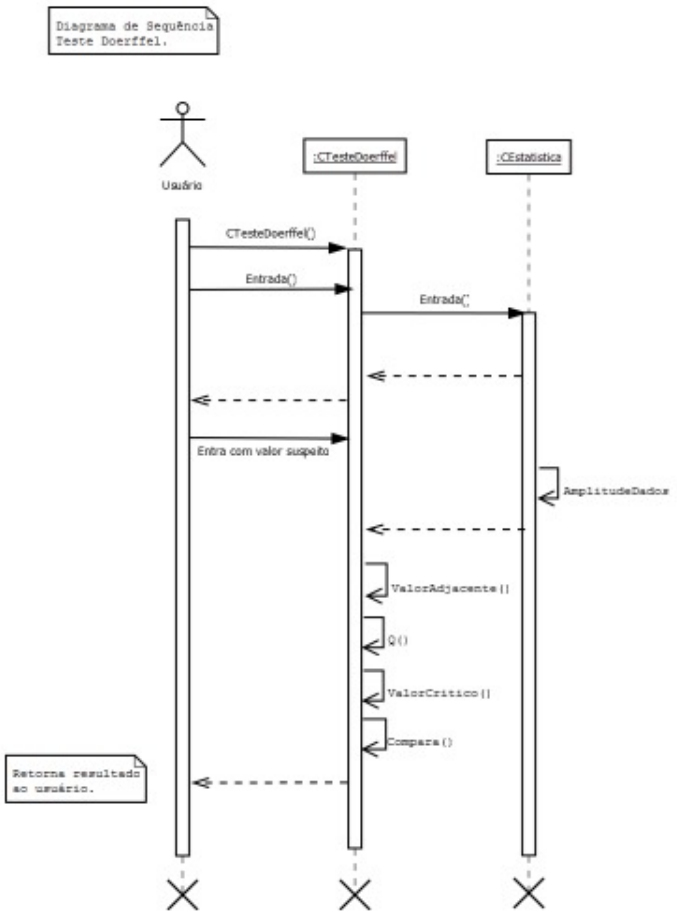


Figura 4.4: Diagrama de Sequência típico da classe CTesteDoerffel

A Figura 4.5 mostra o diagrama de sequência típico da classe CTesteGrubbs.

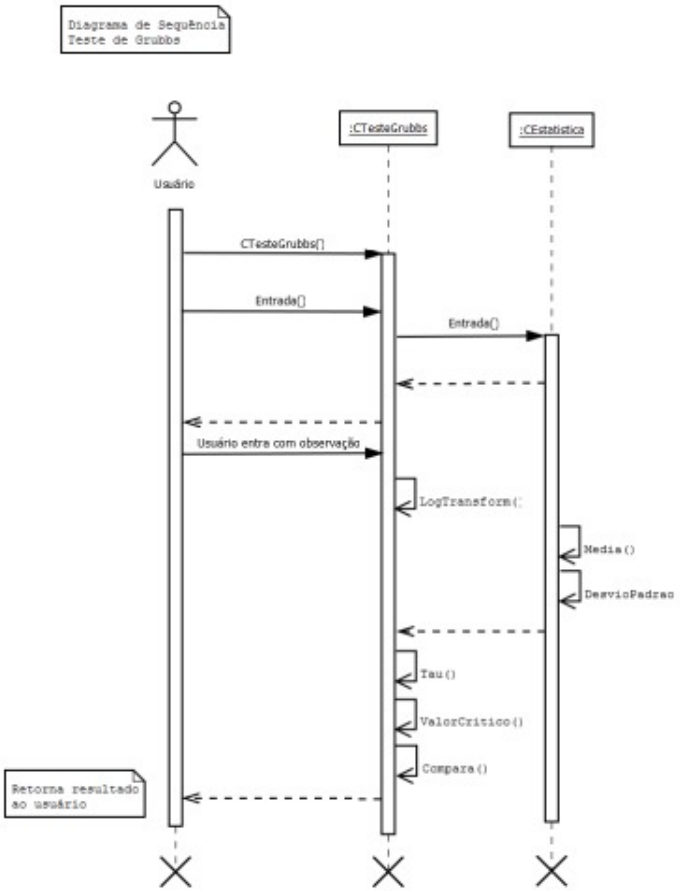


Figura 4.5: Diagrama de Sequência típico da classe CTesteGrubbs

A Figura 4.6 mostra o diagrama de sequência típico da classe CTesteDixon.

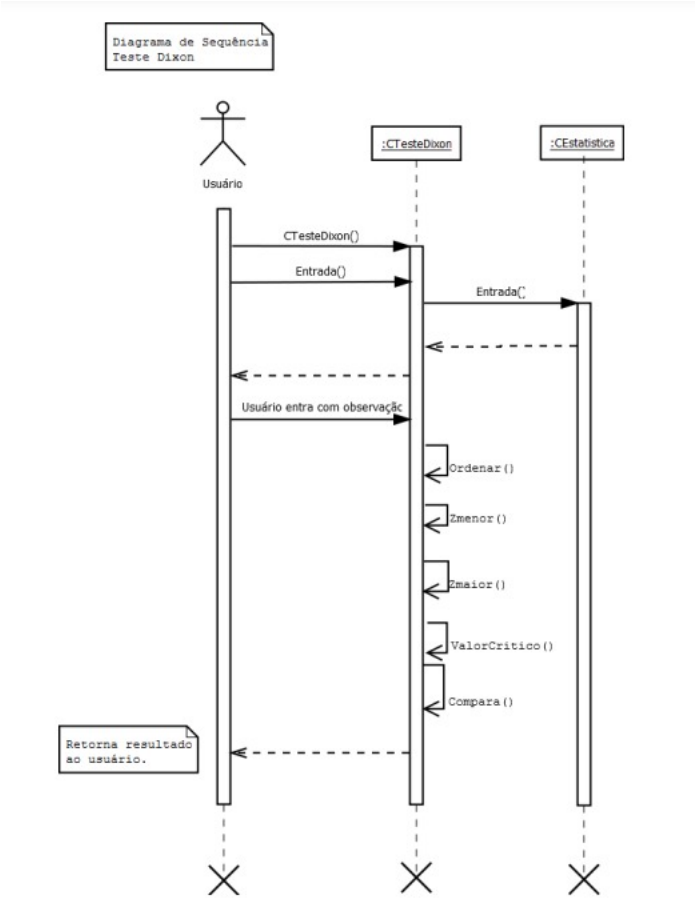


Figura 4.6: Diagrama de Sequência típico da classe CTesteDixon

A Figura 4.7 mostra o diagrama de sequência típico da classe CTesteCochran.

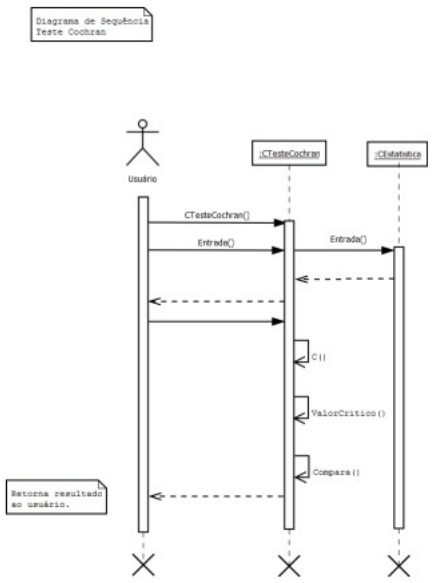


Figura 4.7: Diagrama de Sequência típico da classe CTesteCochran

A Figura 4.8 mostra o diagrama de sequência típico da classe CTesteHipoteseDPconhecido.

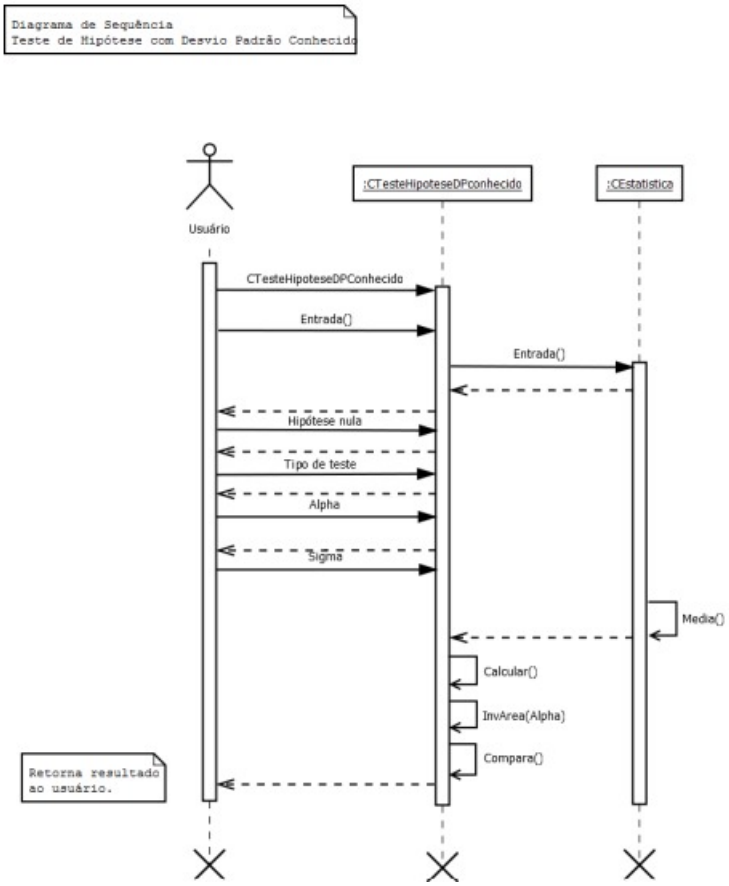


Figura 4.8: Diagrama de Sequência típico da classe CTesteHipoteseDPconhecido

A Figura 4.9 mostra o diagrama de sequência típico da classe CTesteHipoteseDPdesconhecido.

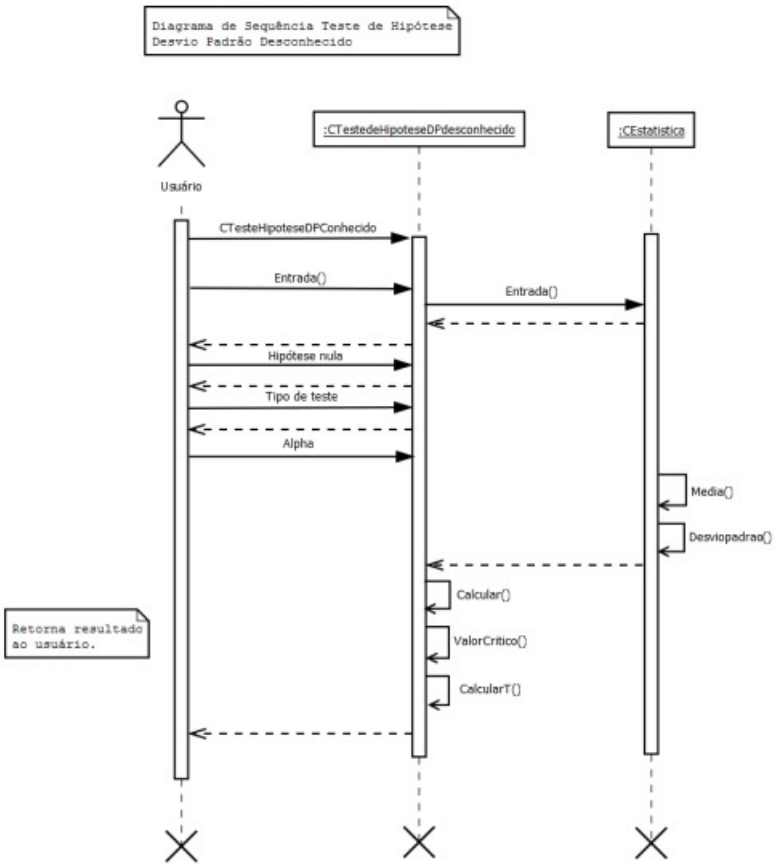


Figura 4.9: Diagrama de Sequência típico da classe CTesteHipoteseDPdesconhecido

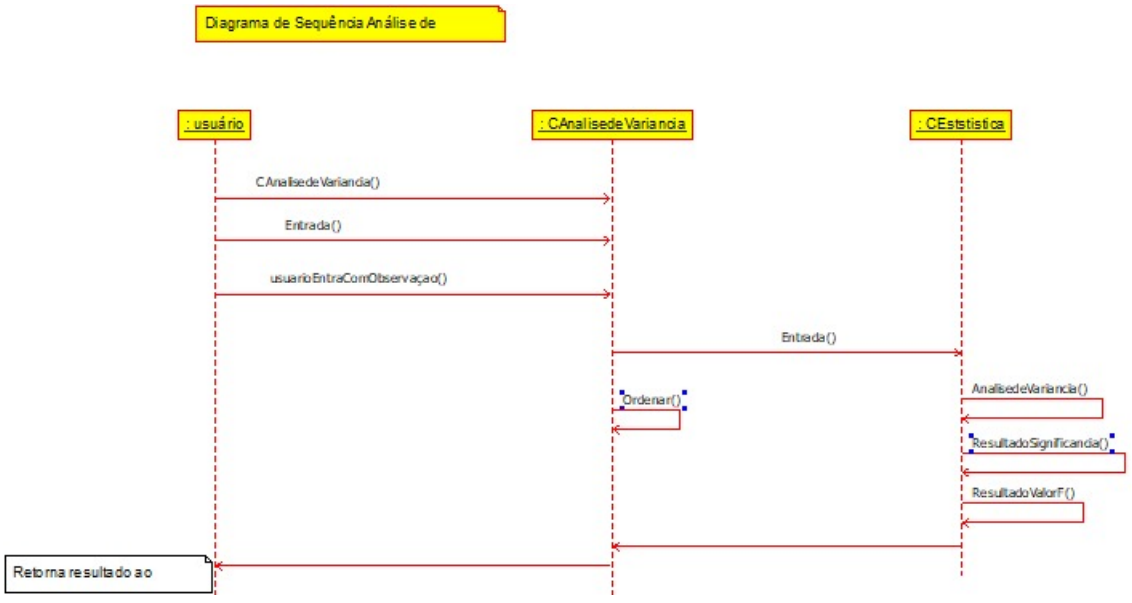


Figura 4.10: Digrama de Sequência tipo de classe CTesteANOVA

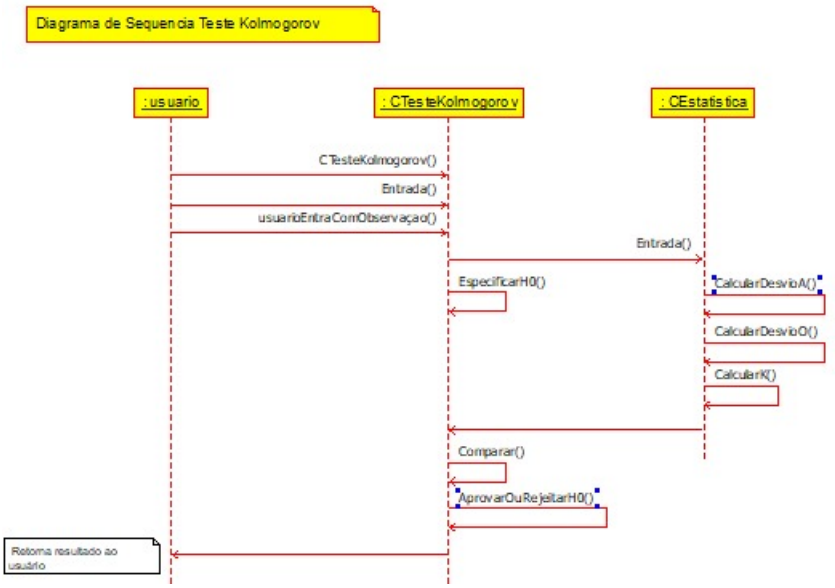


Figura 4.11: Diagrama de Sequência classe CTesteKolmogorovSmirnov

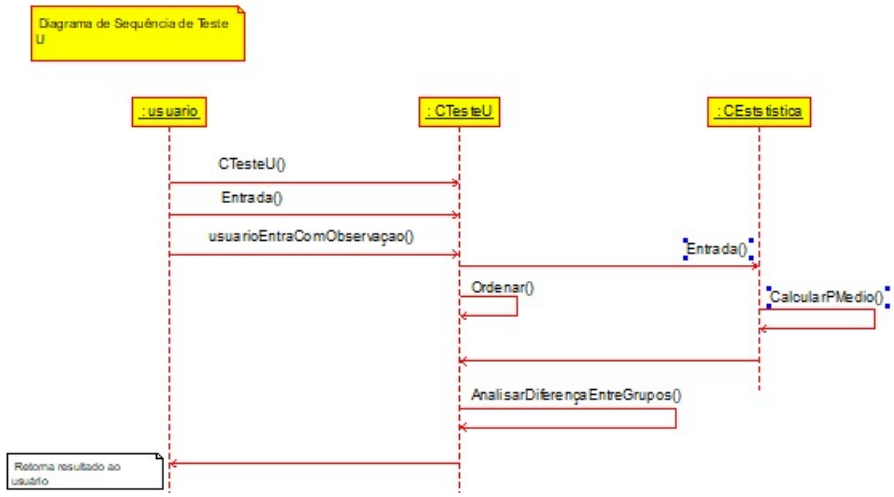


Figura 4.12: Diagrama de Sequência Classe CTesteUMannWhitney

4.4 Diagrama de atividades

Veja na Figura 4.14 o diagrama de atividades correspondente a uma atividade específica do diagrama de máquina de estado.

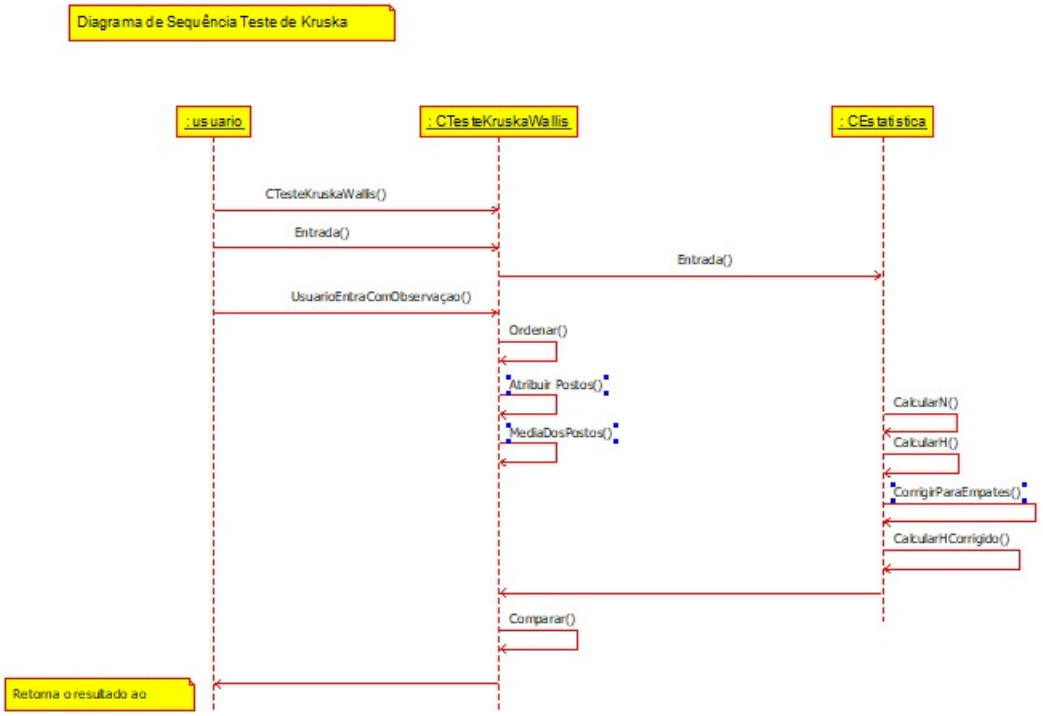


Figura 4.13: Diagram de Sequência classe CTesteKruskalWallis

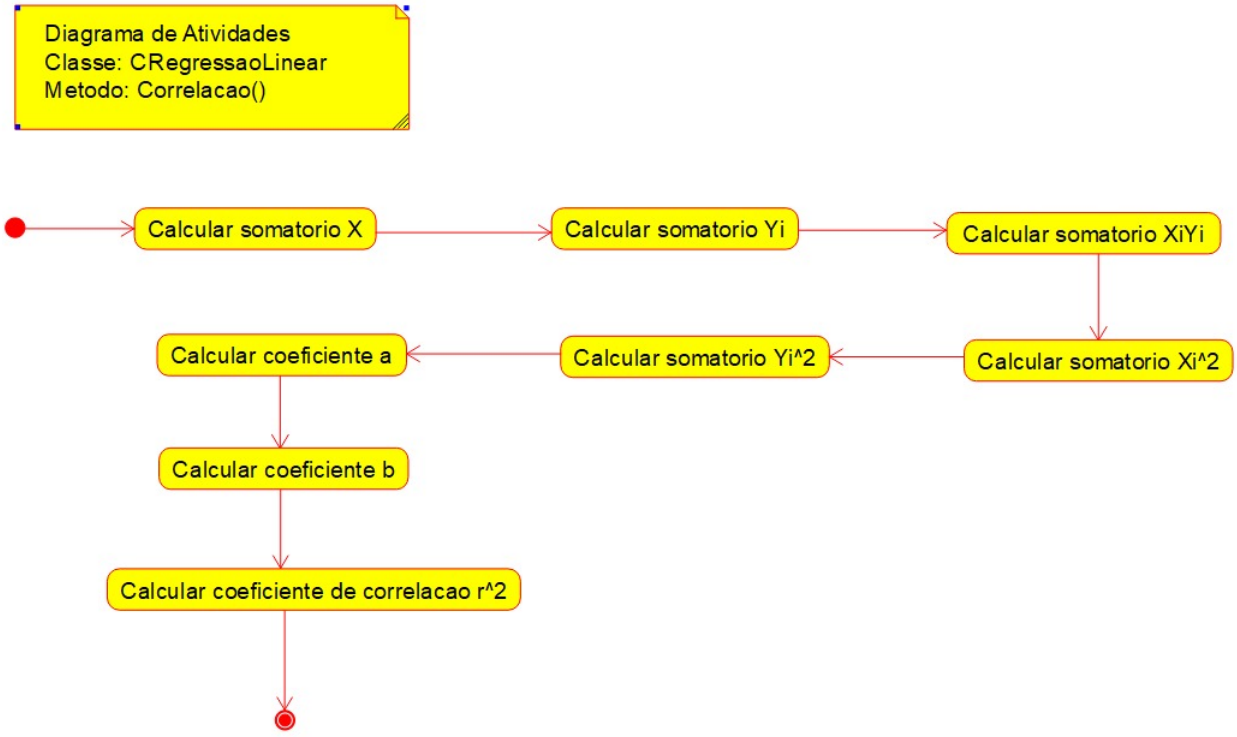


Figura 4.14: Diagrama de Atividades da classe CRegressãoLinear::Correlacao()

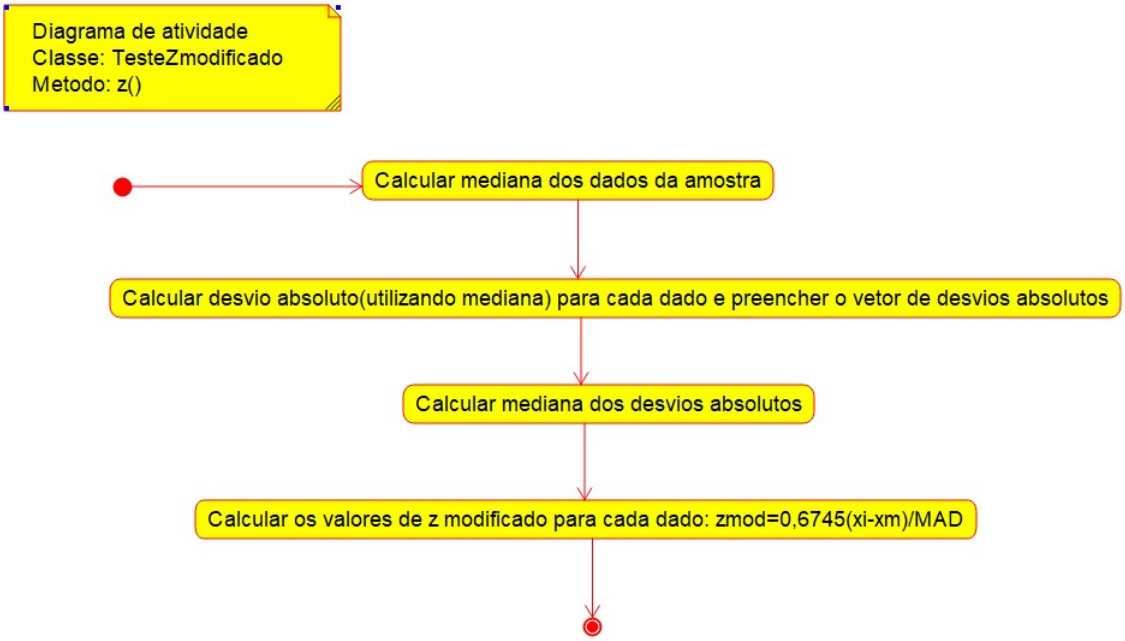


Figura 4.15: Diagrama de Atividades da classe CTesteZmodificado::Z()

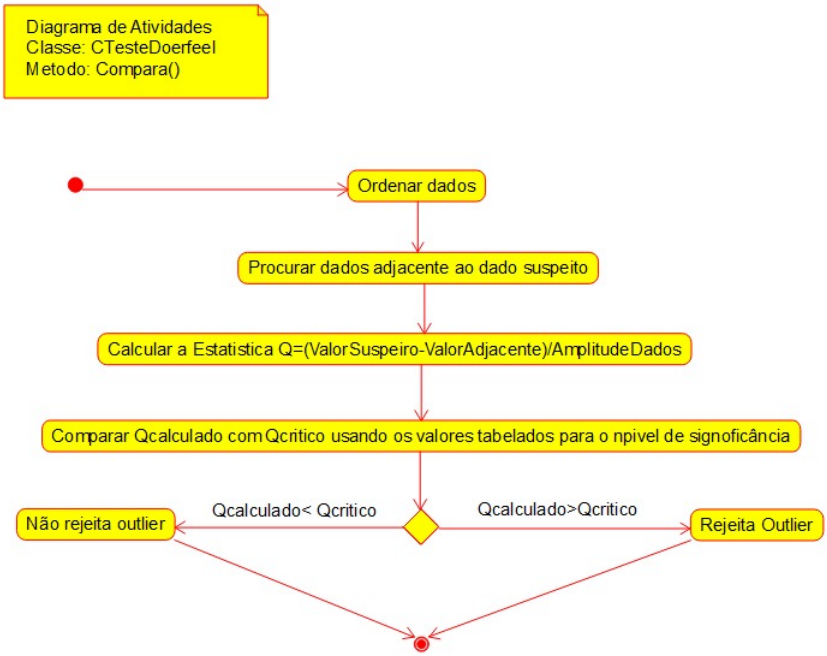


Figura 4.16: Diagrama de Atividades da classe CTesteDoerfeel::Compara()

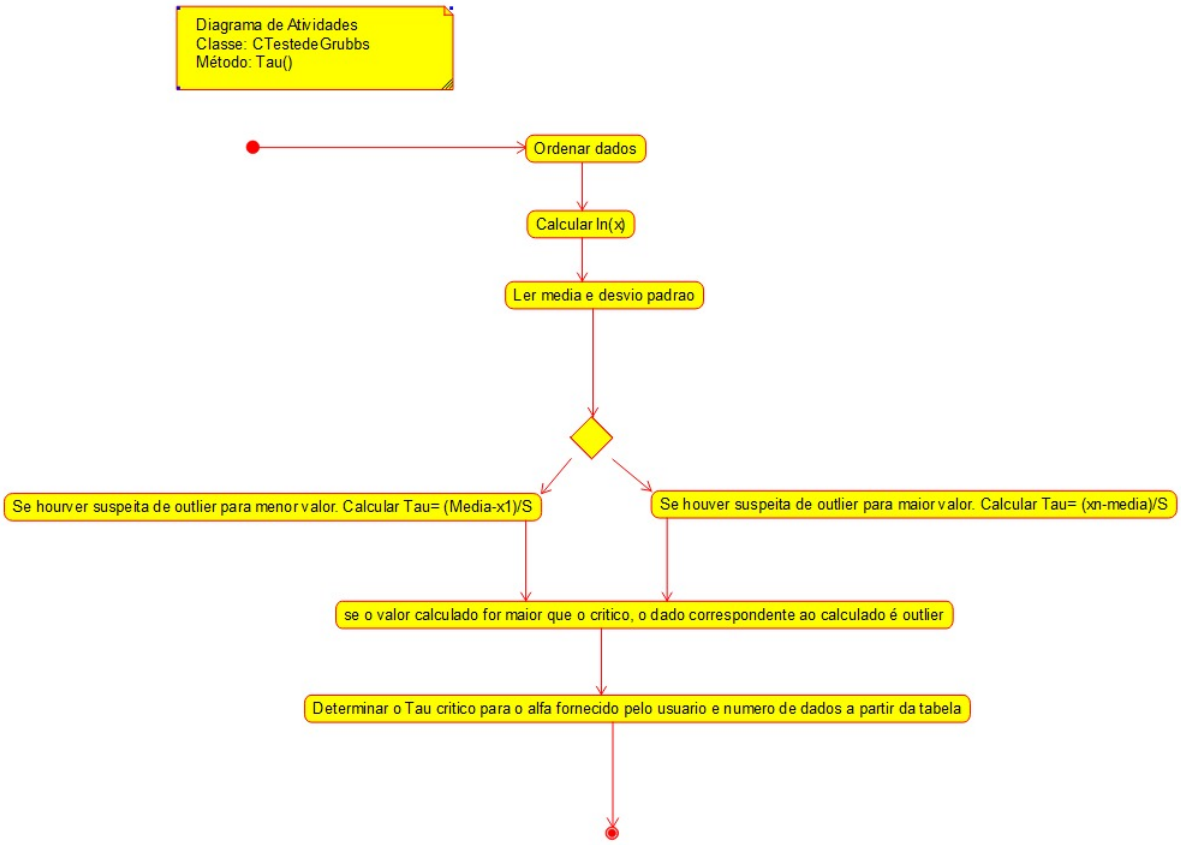


Figura 4.17: Diagrama de Atividades da classe CTesteGrubbs::Tau()

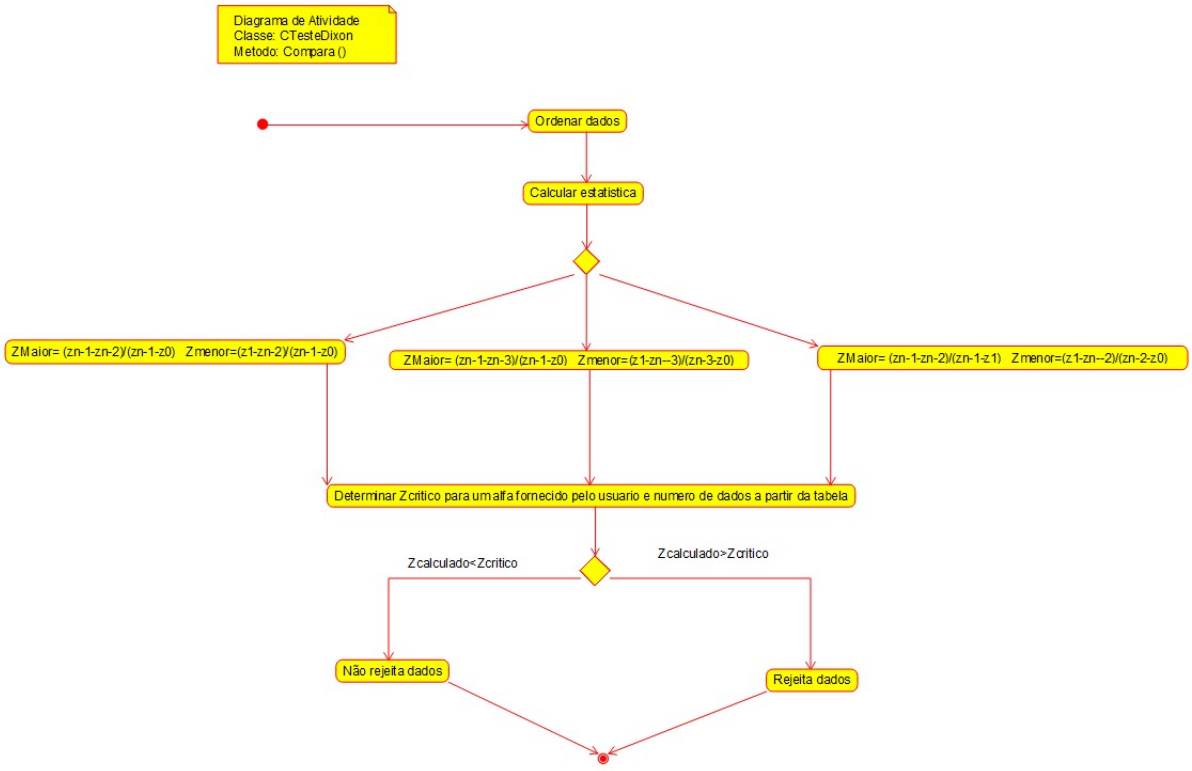


Figura 4.18: Diagrama de Atividades da classe CTesteDixon::Compara()

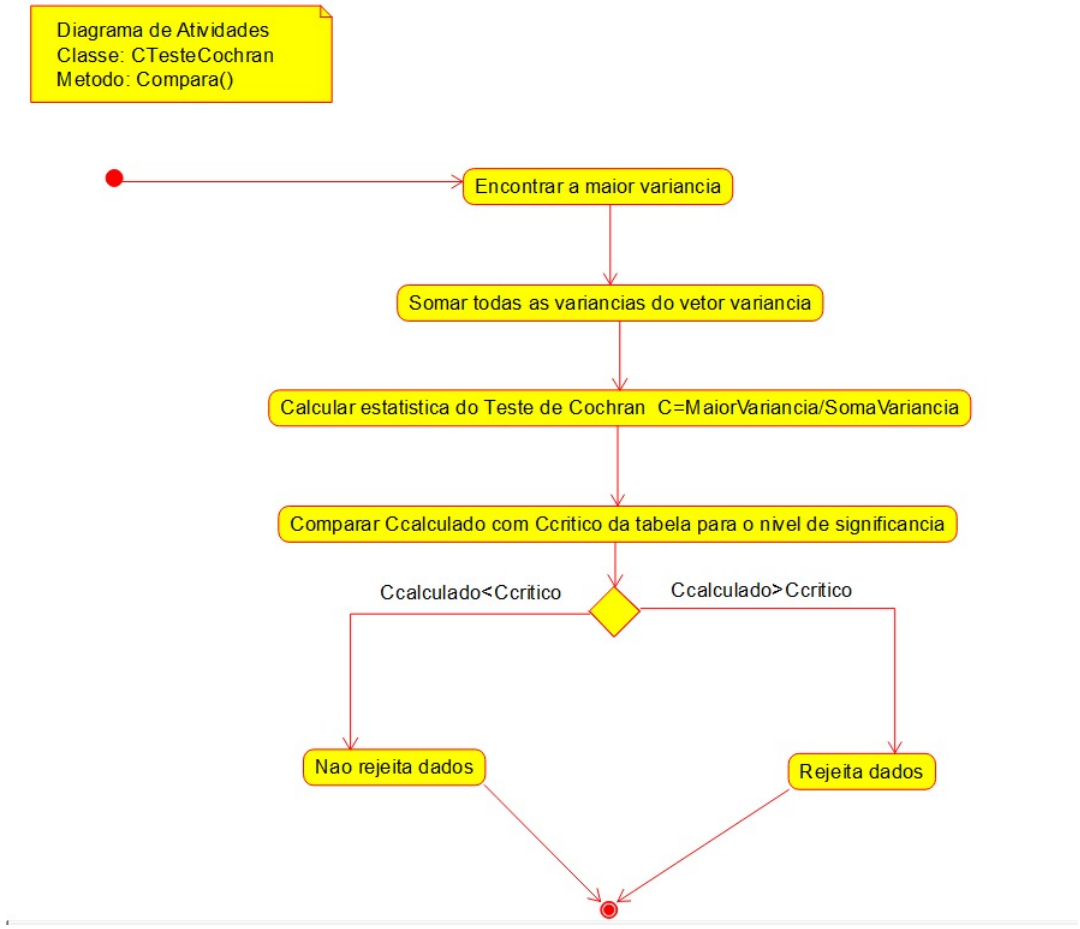


Figura 4.19: Diagrama de Atividades da classe CTesteCochran::Compara()

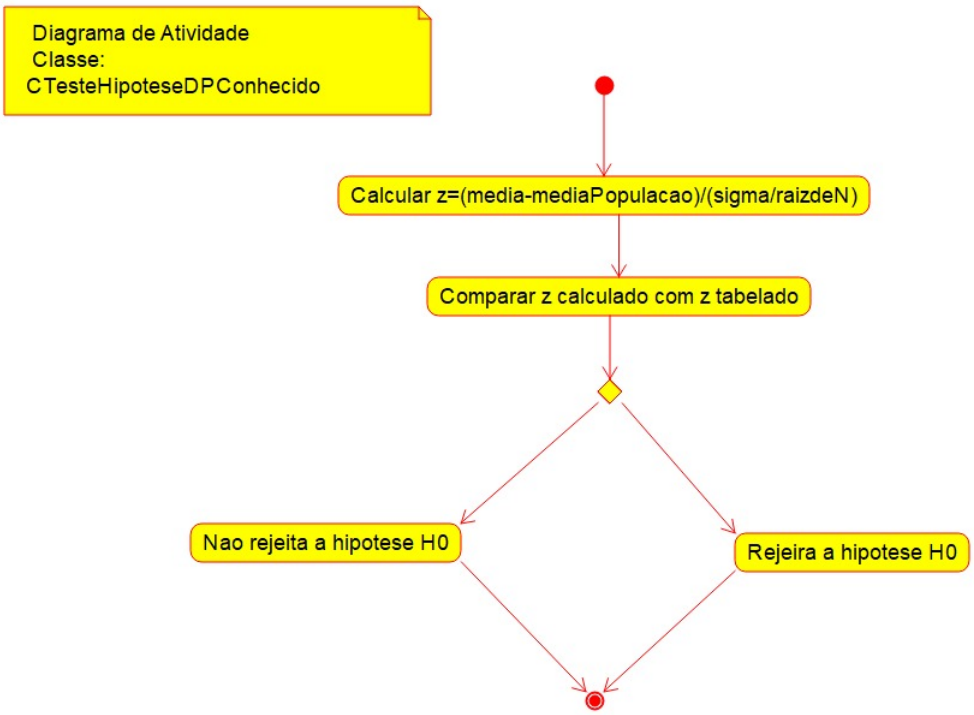


Figura 4.20: Diagrama de Atividades da classe CTesteHipoteseDPconhecido::Compara()

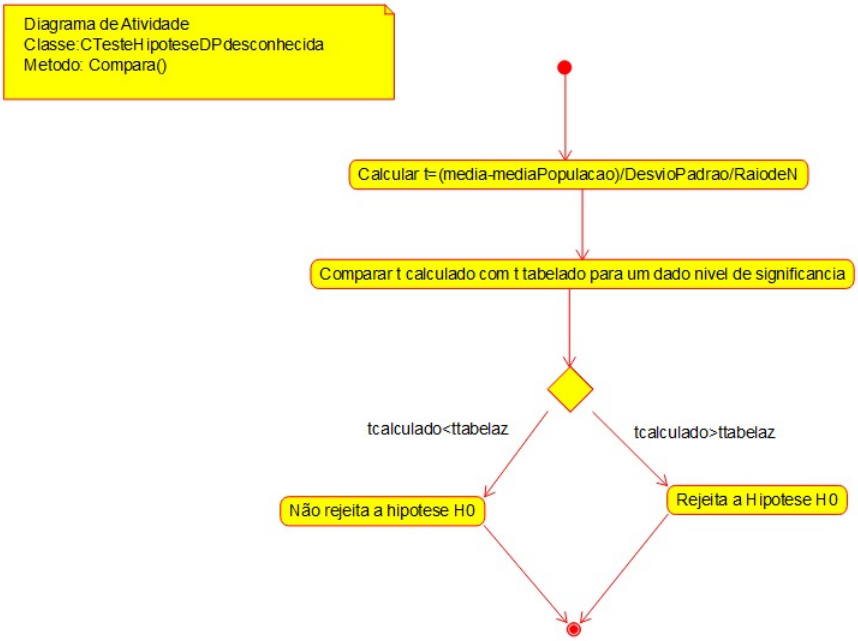


Figura 4.21: Diagrama de Atividades da clasee CTesteHipoteseDPdesconhecido::Compara()

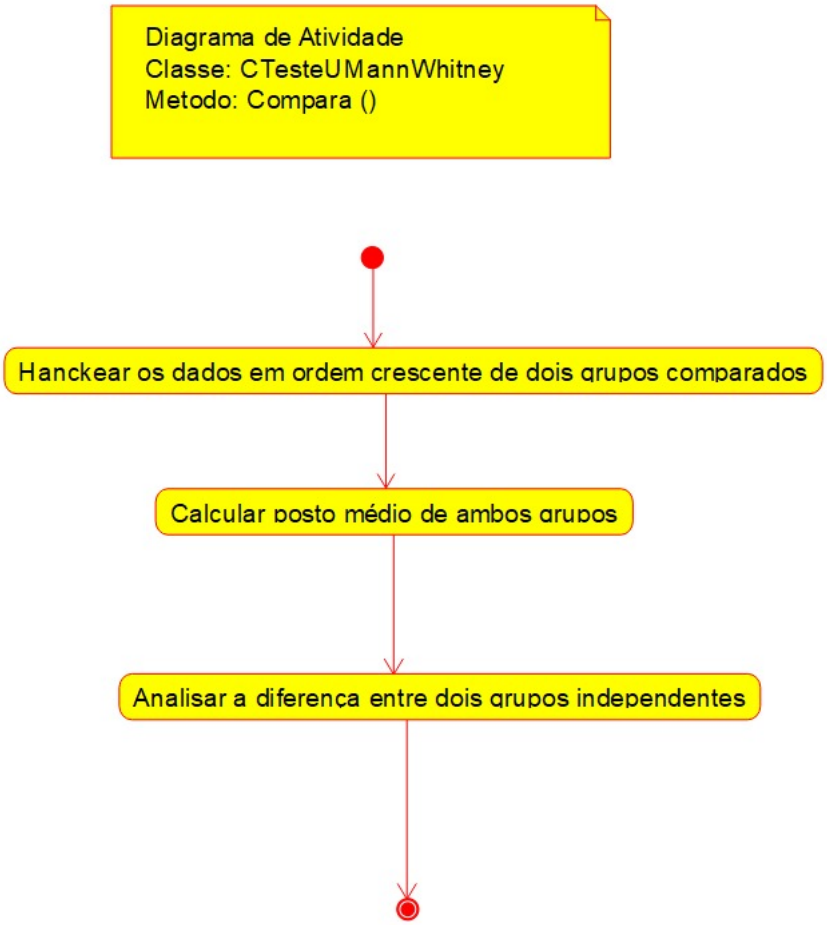


Figura 4.22: Diagrama de Atividades da classe CTesteUMannWhitney:Comparar()

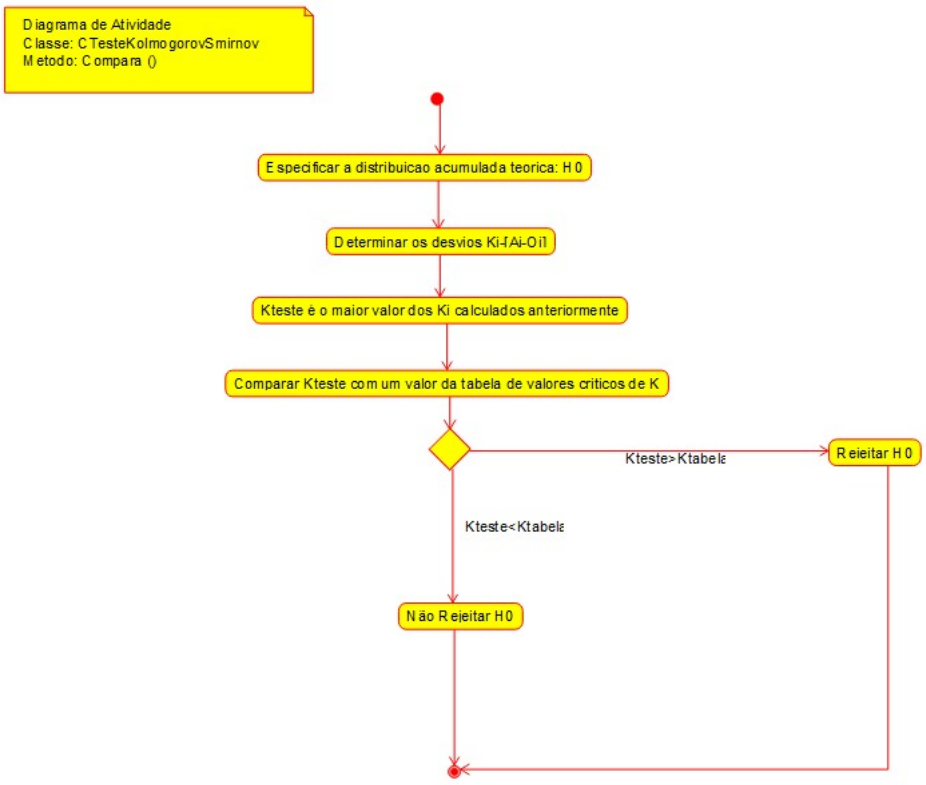


Figura 4.23: Diagrama de atividades classe CTesteKolmogorovSmirnov

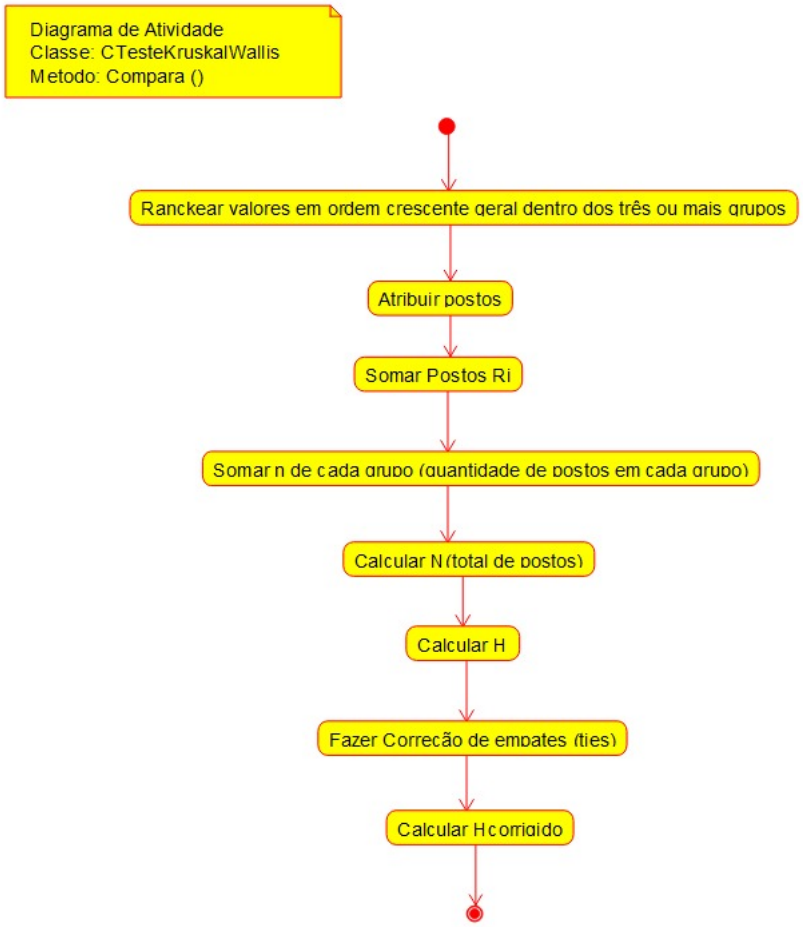


Figura 4.24: Diagrama de Atividades ckasse CTesteKruskalWallis

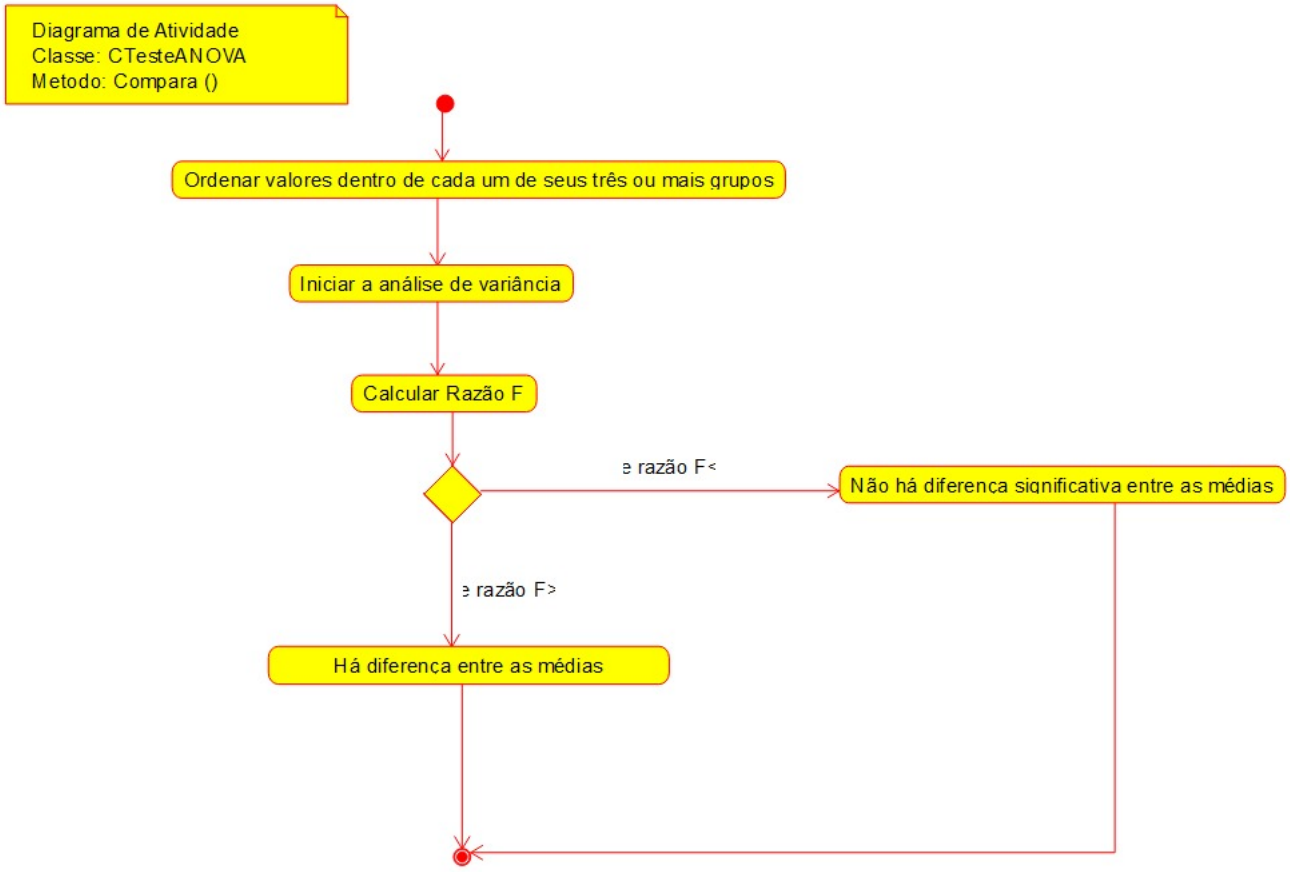


Figura 4.25: Diagrama de atividade classe CTesteANOVA

Capítulo 5

Projeto

5.1 Projeto do sistema

Após a fase de análise orientada a objetos, inicia-se a concepção do projeto do sistema. Isso engloba etapas essenciais, como estabelecer protocolos, desenvolver a interface API, gerenciar recursos, subdividir o sistema em subsistemas, alocar esses subsistemas ao hardware, escolher estruturas de controle, selecionar plataformas e bibliotecas externas, e adotar padrões de projeto. Além disso, implica tomar decisões conceituais e políticas que compõem a base do projeto. É vital estabelecer normas para documentação, nomenclatura de classes, retorno e parâmetros, assim como as características da interface do usuário e desempenho. Esse processo estabelece diretrizes para assegurar a consistência e eficiência ao longo do desenvolvimento do projeto.

5.2 Projeto Orientado a Objeto – POO

Na etapa de projeto orientado a objetos, que segue a criação do projeto do sistema, consideramos as decisões tomadas nessa fase. Incorporamos a análise realizada e as características da plataforma escolhida. Detalhamos o funcionamento do programa, adicionando atributos e métodos para resolver problemas específicos não identificados durante a análise. Buscamos otimizar a estrutura de dados e os algoritmos para reduzir tempo de execução, uso de memória e custos.

- Recursos: O programa utiliza o HD, o processador e o teclado do computador e o software livre `Gnuplot` para gerar gráficos.
- Plataformas: O programa é multiplataforma, funcionando tanto no Windows quanto no GNU/ Linux. Foram utilizadas bibliotecas padrão como iostream, cmath, vector, string, entre outras.
- Controle: Caso o usuário entre com algum dado errado será enviada uma mensagem de erro.

- Ambiente de desenvolvimento integrado: O programa foi compilado no GNU/Linux, utilizando ocompilador g++, software livre de simples utilização.