Joel Nuno Rodrigues Lopes
Licenciado em Engenharia e Gestão Industrial

Estudos quimiométricos em amostras de arroz nacional: caracterização do perfil de aminoácidos e sua correlação com o teor de arsénio

Dissertação para obtenção do Grau de Mestre em
Engenharia e Gestão Industrial

Orientadora: Doutora Ana Sofia Leonardo Vilela de Matos,
Professora Auxiliar, FCT - UNL

Júri:

Presidente:	Profª. Doutora Isabel Maria Lopes Nunes
Arguente:	Profª. Doutora Ayana Maria Xavier Furtado Mateus
Vogais:	Doutora Isabel Palmira Joaquim Castanheira
	Profª. Doutora Ana Sofia Leonardo Vilela de Matos

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Julho 2014

Estudos quimiométricos em amostras de arroz nacional: caracterização do perfil de aminoácidos e sua correlação com o teor de arsénio

AGRADECIMENTOS

Há 5 anos atrás, quando entrei neste projeto da minha vida que estou agora prestes a terminar, o pensamento e as palavras dos meus pais eram mais ou menos estas: “Um ladrão rouba um tesouro, mas não furta a inteligência. Uma crise destrói uma herança, mas não uma profissão. Não importa se você não tem dinheiro, você é uma pessoa rica, pois possui o maior de todos os capitais: a sua inteligência. Invista nela. Estude!” (Augusto Cury - médico, psiquiatra, psicoterapeuta e escritor). E é aos meus pais, à minha irmã, ao meu primo, aos meus tios e avós que quero agradecer de forma majestosa todo o apoio, não só nesta etapa, mas ao longo de toda a minha vida.

Em segundo lugar, um grande obrigado à minha orientadora, Ana Sofia Matos, que foi incansável na ajuda e no apoio, e teve uma enorme paciência para mim ao longo destes quase 5 meses da dissertação. Encontrei uma frase que descreve o discurso de apoio e motivação dados pela professora: "Aplica-te a todo o instante com toda a atenção...para terminar o trabalho que tens nas tuas mãos...e liberta-te de todas as outras preocupações. Delas ficarás livre se executares cada ação da tua vida como se fosse a última.” (Marco Aurélio – Imperador Romano). Também agradecer à Dra. Isabel Castanheira em nome do Instituto Nacional de Saúde Dr. Ricardo Jorge por fazer com que fosse possível este trabalho, e um obrigado especial à pessoa que me ajudou dentro do laboratório, Carla Mota, pela ajuda e prontidão nos assuntos mais incógnitos para mim.

Após estes anos de faculdade levo mais que um curso, levo uma aprendizagem para a vida: “Posso ter defeitos, viver ansioso e ficar irritado algumas vezes, mas não esqueço de que minha vida é a maior empresa do mundo. E que posso evitar que ela vá à falência. Ser feliz é reconhecer que vale a pena viver, apesar de todos os desafios, incompreensões e períodos de crise. Ser feliz é deixar de ser vítima dos problemas e se tornar autor da própria história. (…) Ser feliz é não ter medo dos próprios sentimentos. É saber falar de si mesmo. É ter coragem para ouvir um “não”. É ter segurança para receber uma crítica, mesmo que injusta.” (Augusto Cury). Esta frase é dedicada a todos os meus amigos, a quem eu agradeço imenso.

Para finalizar os agradecimentos, um obrigado à Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa por me ter dado os meios e as condições durante esta minha passagem.

A frase que resume, em boa matemática, os meus 5 anos no curso de engenharia que estou agora prestes a terminar, onde as variáveis orientam grande parte da vida: “Se A é o sucesso, então A é igual a X mais Y mais Z. O trabalho é X; Y é o lazer; e Z é manter a boca fechada.” (Albert Einstein - Físico). Foram 5 anos em que me esforcei, soube aproveitar quando podia e estar calado (talvez nem sempre) quando assim o era exigido.

Não consegui numa página individualizar mas sei quem me apoia e contribui para a minha vida de forma positiva. Mais uma vez, obrigado a todos vocês.

Muito obrigado!

RESUMO

O arroz é um dos alimentos básicos mais importantes para a população mundial, sendo um dos cereais mais consumidos em todo o mundo. Possui um alto teor em hidratos de carbono devido à alta concentração de amido, contém ainda proteínas, vitaminas, minerais e poucas gorduras. A quantidade de proteína a ingerir é requisito para uma dieta adequada (0,75g/kg/dia), devido ao desempenho vital que esta tem na saúde humana. O arroz pelo seu papel determinante na alimentação mundial faz com que os aminoácidos, constituintes das proteínas, mereçam o foco deste estudo. Por outro lado, o arroz pelo seu tipo de cultivo é uma das maiores fontes de ingestão de arsénio para o Homem, um importante agente cancerígeno e contaminante da cadeia alimentar. Isto faz com que este elemento seja igualmente merecedor de análise no presente estudo.

Neste estudo foram analisadas, ao nível dos diferentes aminoácidos e do arsénio, 39 amostras de diferentes tipos e regiões de arroz nacional que foram remetidas para uma análise multivariada. Foi feita uma caracterização e posterior comparação entre tipos/variedades/região de arroz, que demonstra para ambos os tipos de estatística (ANOVA e Kruskal-Wallis), diferenças entre variedades, arroz integral e arroz branco. Verifica-se que ao analisar pelas várias características do arroz, não existem diferenças ao nível do arsénio e que, através da correlação de Spearman, este se correlaciona positivamente com arroz integral e negativamente com arroz branco. Na análise de *clusters*, os aminoácidos (variáveis) foram 3 conjuntos: baixa, média e alta concentração. Por sua vez, as amostras dividem-se pela variedade, formando ainda um *cluster* em que existe uma fusão de variedades. Para classificação de arroz no futuro, com base no perfil de aminoácidos, foi possível a criação de um modelo *k-NN* cujo erro de classificação fosse nulo.

Palavras-chave: Arroz, Aminoácidos, Arsénio, Estatística Multivariada, *Clusters*, Correlação de Spearman, *k-Nearest Neighbors*

ABSTRACT

Rice is a major staple food for the world population, being one of the most consumed cereals worldwide. It has a high content of carbohydrates due to the high starch concentration, it also contains proteins, vitamins, minerals, and low fat. The amount of protein to ingest is a requirement of a proper diet (0,75g/kg/day), due to its vital performance in the human health. The rice, by its decisive role in the global feeding makes the amino acids, constituents of proteins, deserving of the focus of this study. On the other hand, the rice by its cultivation type, is a major source of arsenic intake for the humans, an important carcinogenic contaminant in the food chain. This makes this element also worthy of analysis in this study.

In this study were analyzed, in terms of amino acids and arsenic levels, 39 national rice samples of different types and regions that were referred to a multivariate analysis. It was made a characterization and subsequent comparison between types/varieties/region of rice, which demonstrates for both types of statistics (ANOVA and Kruskal-Wallis), differences between varieties, brown rice and white rice. It appears that when analyzing the different characteristics of rice, there are no differences in the arsenic level, and by Spearman correlation, this is positively correlated with brown rice, and negatively correlated with white rice. In cluster analysis, the amino acids (variables) were 3 groups: low, medium and high concentration. In turn, the samples are divided by the range still forming a cluster in which there is a fusion of varieties. For classification of the rice in the future, based on the amino acid profile, it was possible to create a k-NN model with null error classification.

Keywords: Rice, Amino Acids, Arsenic, Multivariate Statistics, Clusters, Spearman Correlation, k-Nearest Neighbors

ABREVIATURAS

ANOVA	*Analysis of Variance* (Análise de variância)
APARROZ	Agrupamento de produtores de arroz do vale do sado
Asi	Arsénio Inorgânico
ATSDR	*Agency for toxic substances and disease registry* (Agência de substâncias tóxicas e registo de doenças)
CA	*Cluster Analysis* (Análise de *clusters*)
DAN	Departamento de Alimentação e Nutrição
EFSA	*European Food Safety Authority* (Autoridade Europeia para a Segurança dos Alimentos)
EPA	*Environmental Protection Agency* (Agência de proteção ambiental)
FAO	*Food and Agriculture Organization* (Organização para a Agricultura e Alimentação)
FAPAS	*Food Analysis Performance Assessment Scheme* (Regime de avaliação no desempenho da análise de alimentos)
HCA	*Hierarchical Cluster Analysis* (Análise hierárquica de clusters)
HPLC	*High-performance liquid chromatography* (Cromatografia líquida de alta performance)
ICP-MS	*Inductively Coupled Plasma – Mass Spectrometry*
INSA	Instituto Nacional de Saúde Doutor Ricardo Jorge
IPAC	Instituto Português de Acreditação
k-NN	*k-Nearest Neighbors* (k-Vizinhos mais próximo)
LDA	*Linear Discriminant Analysis* (Análise discriminante linear)
PCA	*Principal Components Analysis* (Análise de Componentes Principais)
PDA	*Photodiode array detector* (Detetor de fotodíodos)
UE	União Europeia
UPLC	*Ultra-performance liquid chromatography* (Cromatografia líquida de ultra performance)
WHO (OMS)	*World Health Organization* (Organização Mundial da Saúde)

SIMBOLOGIA

AD	Estatística de teste do teste de Anderson-Darling
CV	Coeficiente de variação
d	Distância entre objetos
D_n	Diferença máxima entre as funções distribuição acumulada
e	Resíduo
g.l.	Graus de liberdade
H	Estatística de teste de Kruskal-Wallis
H_0	Hipótese Nula
H_1	Hipótese Alternativa
k	Número de amostras
MS_B	*Mean square between* (Desvio quadrático médio entre níveis ou tratamentos)
MS_W	*Mean square whitin* (Desvio quadrático médio dentro dos níveis ou tratamento, ou variância)
n	Tamanho da amostra
N	Tamanho total de todas as amostras
R	Rank do valor observado
r	Correlação de Pearson
r_s (ρ)	Correlação de Spearman
s	Desvio-padrão da amostra
s^2	Variância da amostra
S_p	Desvio-padrão agrupado
SS_B	*Between Sum of Squares* (Soma das variações entre os níveis ou tratamentos)
SS_T	*Total sum of squares* (Soma total dos desvios quadráticos)
SS_W	*Whithin Sum of Squares* (Soma das variações dentro dos níveis ou tratamentos, ou Erro)
W	Estatística de teste do teste de Shapiro-Wilk
W_0	Estatística de teste do teste de Levene
W_{50}	Estatística de teste do teste de Brown-Forsythe
α	Nível de significância e Erro do tipo I
β	Erro do tipo II
μ	Média da população
x	Valor observado
$\bar{x}$	Média da amostra
$\bar{\bar{x}}$	Média amostral global
σ^2	Variância da população

ÍNDICE GERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

CAPÍTULO 1 – INTRODUÇÃO

1.1. Enquadramento e Motivação

O arroz é um dos alimentos básicos mais importantes para a população humana mundial, sendo responsável por 20% da energia proveniente da alimentação ao nível global, fornecendo 536 kcal/capita/dia (Almeida & Marques, 2013; Marques, 2009). Este cereal é um amigo inseparável das cozinhas portuguesas, já que é o acompanhamento habitual dos seus pratos. Isto faz com que Portugal seja o maior consumidor de arroz da Europa (cerca de 17 kg/capita/ano), cujo consumo ronda as 180 mil toneladas por ano (Almeida & Marques, 2013).

O arroz possui um alto teor em hidratos de carbono devido à alta concentração de amido, contém ainda proteínas, vitaminas e minerais, e igualmente importante, poucas gorduras (Walter, Marchezan, & Avila, 2008). Uma boa nutrição é a base de uma boa saúde, e o arroz, pelo seu consumo e pelas suas características, tem todo o interesse em ser analisado. É aqui que entra a química alimentar, que trata a composição dos alimentos, e as características físico-químicas e suas mudanças durante o processamento, armazenamento e manuseamento a que os alimentos são submetidos (Fennema, 1996).

A quantidade de proteína (constituída por aminoácidos) que deve ser consumida é um requisito de uma dieta adequada, pois os aminoácidos têm um papel vital na saúde humana. Os aminoácidos estão divididos entre aminoácidos essenciais (os que o organismo não produz), não essenciais (os que organismo produz) e os condicionalmente essenciais (que são essenciais em condições fisiológicas especiais). E, se por um lado, a produção de aminoácidos não essenciais está assegurada pelo organismo, por outro, a variedade, quantidade e qualidade de aminoácidos essenciais depende da alimentação de cada um, fazendo então com que o arroz tenha um papel decisivo na alimentação do Homem (Balch, 2006).

O tipo de cultivo a que o arroz é submetido em Portugal, condições de inundação quase permanente, faz dele uma importante fonte de exposição ao arsénio inorgânico, que ocorre principalmente através do consumo de água subterrânea, assim como da ingestão de alimentos preparados ou irrigados durante a sua produção com essa água (Dwivedi et al., 2012; WHO, 2010). O arsénio é um elemento

químico pertencente à tabela periódica, sendo o arsénio inorgânico altamente tóxico (conhecido agente cancerígeno e contaminante da cadeia alimentar) (Dwivedi et al., 2012; WHO, 2010).

Com tudo isto, e em colaboração com o Instituto Nacional de Saúde Dr. Ricardo Jorge (INSA), laboratório nacional de referência, nasce o interesse em realizar a presente dissertação com a intenção de investigar estatisticamente as análises químicas feitas a arroz nacional. Pela quantidade de variáveis provenientes das análises químicas, é necessário o recurso a um segmento particular da estatística – a estatística multivariada. Em traços gerais, com o intuito de enquadrar o projeto, na Figura 1.1 estão as áreas e etapas da dissertação.

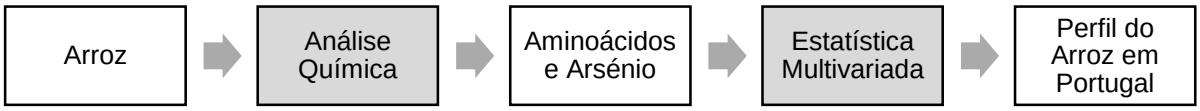

Figura 1.1 - Áreas e etapas ordenadas da presente dissertação

O interesse que o autor da dissertação tem pelo tema da agricultura, a capacidade em usar a estatística como ferramenta de trabalho, e a curiosidade em fazer uma investigação deste carácter, são as razões pelas quais se tornou cativante e motivante a escolha da mesma.

1.2. Objetivos

A presente dissertação, tem como principal objetivo, a caracterização do arroz produzido e/ou comercializado em Portugal através de uma análise estatística multivariada. Pretende-se, através dos dados recolhidos e, com base nos diferentes tipos de arroz, traçar um perfil para cada um destes.

Os dados foram recolhidos e analisados pelo Departamento de Alimentação e Nutrição (DAN) do INSA, tendo posteriormente sido disponibilizados os valores referentes às concentrações dos vários aminoácidos presentes no arroz proveniente de vários produtores ou estabelecimentos comerciais, bem como as concentrações de arsénio.

O objetivo inicial passa por caracterizar os vários tipos de arroz, e investigar sobre diferenças que possam ou não existir entre estes. Esta caracterização é feita quer ao nível dos aminoácidos quer no arsénio. Após esta etapa, pretende-se correlacionar o arsénio com os demais aminoácidos com o objetivo de tentar perceber se o tipo de arroz ou outra característica influenciam a concentração do mesmo. Segue então uma caracterização do arroz ao nível dos aminoácidos recorrendo à análise de *clusters*. Por fim, e se porventura existirem diferenças significativas nas concentrações dos aminoácidos pelas características do arroz, pretende-se criar um modelo matemático válido para posterior identificação do arroz com base nas leituras cromatográficas.

1.3. Estrutura da dissertação

Estruturalmente a dissertação encontra-se dividida em 6 capítulos. O presente capítulo (Capitulo 1), visa enquadrar o problema a ser estudado e os objetivos a atingir no final do estudo.

O segundo e terceiro capítulo (fundamentos teóricos do arroz e estatística multivariada, respetivamente), também geralmente designados de revisão bibliográfica, expõem e fundamentam os conceitos teóricos presentes ao longo de toda a dissertação, que nomeadamente assentam sobre as áreas do arroz, da química alimentar e da estatística multivariada.

O capítulo quarto apresenta a metodologia seguida, quer pelo INSA na análise química, quer pela análise aos dados, levada a cabo pelo autor desta dissertação. A metodologia foi feita, tendo por base o alinhamento dos conceitos apresentados no capítulo precedente.

No quinto capítulo, são apresentados os resultados obtidos através da sequência de processos descritos na metodologia, e respetiva análise/conclusão aos mesmos.

Finalmente, no sexto e último capítulo, são apresentadas conclusões relativamente ao estudo/trabalho realizado. São ainda, apresentadas algumas melhorias e recomendações para trabalhos futuros no mesmo âmbito.

CAPÍTULO 2 – FUNDAMENTOS TEÓRICOS ARROZ

2.1. O arroz

2.1.1. Características e importância do arroz

O arroz é uma planta pertencente à família das gramíneas (*Poeceae*), sendo esta monocotiledónea, isto é, produz sementes cujo embrião apresenta um único cotilédone (pode ser visto na Figura 2.1). Para além disto, é uma planta monocárpica anual, gerando apenas flor e fruto uma única vez (Cheajesadagul, Arnaudguilhem, Shiowatana, Siripinyanond, & Szpunar, 2013; Gonzálvez, Armenta, & Guardia, 2011).

Figura 2.1 - Planta de arroz com os estolões e embriões que contêm os cotilédones

A origem do arroz é motivo de discórdia, ainda assim, existem evidências arqueológicas que apontam para a China como origem, à 8000 anos atrás (Drumond, 2012).

Atualmente existem duas espécies cultivadas de arroz: o africano (*Oryza Glaberrima*) que se estima ser cultivado á cerca de 3500 anos, e o asiático (*Oryza Sativa*), que se divide em duas variedades com base na origem, sendo que o cultivo deste iniciou à aproximadamente 7000 anos atrás. As variedades do arroz de origem asiática são a Indica (do lado indiano), designada de *Oryza sativa* variedade Indica e, a Japónica (do lado chinês), nomeada por *Oryza sativa* variedade Japónica. Estima-se então, que existam alguns milhares de tipos de arroz, como consequência do amplo histórico no cultivo, no entanto, as variedades mais produzidas são as de origem asiática (Drumond, 2012; Marques, 2009).

O arroz mantém praticamente a sua qualidade nutricional original mesmo após a sua laboração, e é considerado um produto natural, rico em hidratos de carbono e proteínas, que para além da secagem e branqueamento (processos físicos) não envolve qualquer outro tipo de processamento, nem requer a utilização de aditivos ou conservantes. O arroz é superior ao trigo em hidratos de carbono disponíveis, todavia detém um teor de proteína inferior ao do trigo. O amido é o componente mais importante deste cereal revelando-se fundamental na determinação do seu comportamento na cozedura e da sua qualidade alimentar. O arroz, ainda que inferior ao trigo, é rico em proteínas e uma importante fonte de micronutrientes (ferro, potássio, fósforo, magnésio, vitaminas B1, B2 e B6), e hidratos de carbonos (fibras). A principal fonte destas vitaminas e minerais é a película - razão pela qual o arroz integral tem uma qualidade nutricional superior à do arroz branqueado. O arroz não possui gordura, colesterol nem glúten, o que o torna num alimento adequado a qualquer dieta alimentar. Alimenta sem engordar nem causar problemas de alergias ou intolerâncias alimentares. Na verdade, 20% da energia proveniente da alimentação ao nível global é fornecida pelo arroz, já que este (branco) fornece 536 kcal/capita/dia (Almeida & Marques, 2013; Marques, 2009).

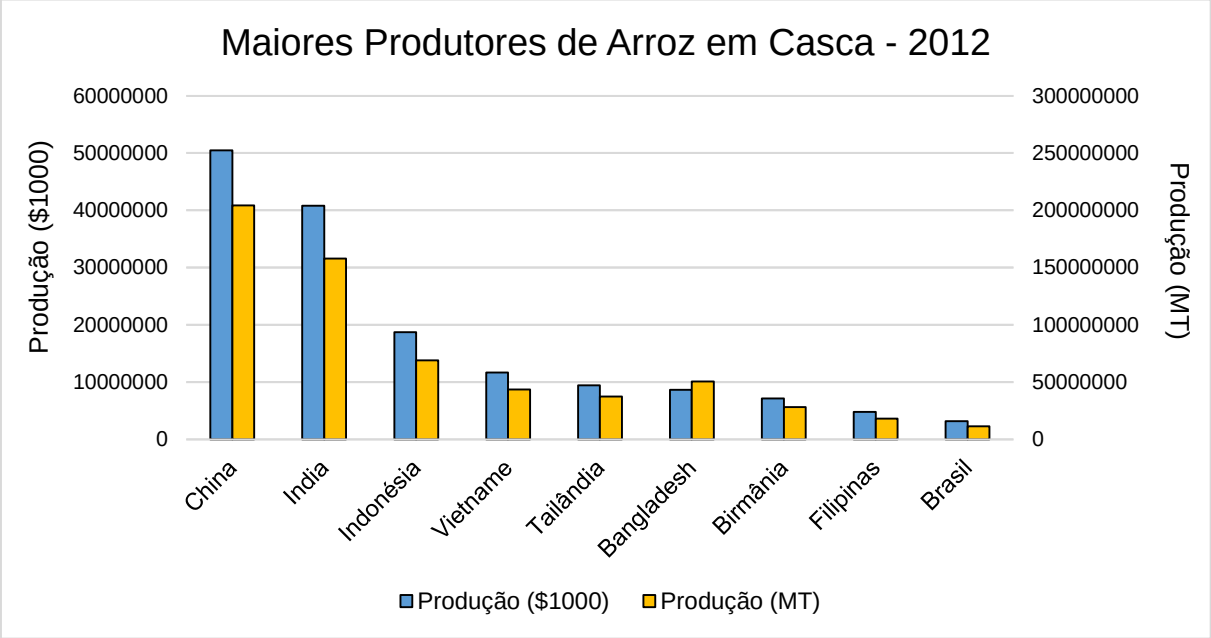

Figura 2.2 - Gráfico dos países com maior produção de arroz (em casca) (FAOSTAT, 2012)

O arroz é um dos alimentos básicos mais importantes para a população humana mundial, especialmente nos continentes asiático e americano. Sendo este, o segundo cereal mais produzido no mundo, estando entre o milho (o mais produzido) e o trigo (o terceiro cereal mais produzido em todo o mundo), cobrindo cerca de 9% da terra arável. Contudo, é o cereal que está em primeiro lugar quando a produção é medida em valor monetário, ou seja, por outras palavras, é o cereal que a nível monetário mais foi produzido. Entre os países líderes na produção de arroz, como se pode visualizar na Figura 2.2, destacam-se a China, a índia, a Indonésia e o Vietname. No entanto, o país que mais exportou arroz na última década foi a Tailândia (Almeida & Marques, 2013; Gonzálvez et al., 2011).

Em 2012, segundo dados da Organização para a Agricultura e Alimentação (FAO), foram produzidos cerca de 720 milhões de toneladas de arroz em casca (*paddy*), divididos por 164 milhões de hectares em 112 países. A China como líder, produziu cerca de 204 milhões de toneladas, estando à frente da Índia que ocupou a maior área com o cultivo de arroz (44 milhões de hectares) (Almeida & Marques, 2013; FAOSTAT, 2012).

2.1.2. Tipos de arroz

Como referido no ponto anterior, existem inúmeros tipos de arroz, com isto a legislação portuguesa em vigor classifica o arroz em categorias, com base nas diferentes características do mesmo. As características pelas quais o decreto-lei faz a diferenciação e, as respetivas categorias em que cada uma se divide, estão presentes na Tabela 2.1 (Decreto-Lei n. 62/2000 de 19 de Abril, 2000).

De todas elas, destacam-se as categorias de arroz quanto ao comprimento dos grãos de arroz, do ponto vista comercial, sendo estas definidas como: arroz de grãos redondos – arroz cujos grãos tenham um comprimento inferior ou igual a 5,2 mm e cuja relação comprimento/ largura seja inferior a 2; arroz de grãos médios – arroz cujos grãos tenham um comprimento superior a 5,2 mm e inferior ou igual a 6,0 mm e cuja relação comprimento/ largura seja inferior a 3; e arroz de grãos longos – arroz de grãos com um comprimento superior a 6,0 mm e cuja relação comprimento/ largura seja superior a 2 e inferior a 3, ou superior ou igual a 3 (Cotarroz, sem data-b; Decreto-Lei n. 62/2000 de 19 de Abril, 2000; Drumond, 2012).

Em Portugal as variedades mais produzidas e/ou consumidas de arroz, são conhecidas de outra forma, a variedade *Oryza sativa* variedade Indica é normalmente designada por Arroz Agulha, que é um arroz de grãos longos – pois apresenta um grão com um comprimento superior a 6,0 mm e uma relação comprimento/largura superior a 3. Já a variedade *Oryza sativa* variedade Japónica, intitula-se por Arroz Carolino, que é igualmente um arroz de grãos longos – apresenta, geralmente, um grão com um comprimento superior a 6,0 mm e uma relação comprimento/largura de aproximadamente 2,5 (Almeida & Marques, 2013; Drumond, 2012).

Tabela 2.1 - Tipos de arroz em função da característica pela legislação em vigor

Característica	**Categorias em que se divide**
Estado físico do arroz	Arroz em casca (*paddy*)
	Arroz descascado, em película ou meio preparo
	Arroz semibranqueado
	Arroz branqueado
Comprimento dos grãos de arroz	Arroz de grãos redondos
	Arroz de grãos médios
	Arroz de grãos longos
Tratamento a que o arroz é sujeito	Arroz estufado ou vaporizado (*parboiled*)
	Arroz pré-cozido
	Arroz glaciado
	Arroz matizado
Comercialização	Classe comercial
	Tipo comercial
Características dos grãos de arroz, trincas e defeitos	Grão inteiro
	Grão despontado
	Grão partido ou trinca
	Grão verde
	Grão deformado
	Grão danificado
	Grão fendido
	Grão gessado
	Grão estriado de vermelho
	Grão vermelho
	Grão manchado (grão *taché*)
	Grão amarelo
	Grão ambarino
	Grão escuro (*peck*)
	Casca
	Farelo de casca
	Sêmea
	Gérmen
	Farinha
	Impurezas

Para além destas variedades, existem outras no mercado, como é o caso do arroz vaporizado (ou estufado), que é submetido a um tratamento industrial com vapor de água, cuja apresentação final é firme e de cor dourada (deve-se ao amido ficar gelatinizado durante a vaporização). Este arroz é muito rico em fibras (hidratos de carbono) e sais minerais (micronutrientes). Um tipo de arroz que está a ganhar fama devido às suas propriedades é o arroz integral, que é descascado e limpo sem sofrer branqueamento, e ficando com isso, rico em fibras, minerais e vitaminas. Este arroz pode ser de grão longo ou curto. Existem ainda, os tipos de arroz aromáticos: o arroz *Basmati*, que pode ser de origem indiana ou paquistanesa, e o arroz *Jasmine*, de origem tailandesa, apresentando ambos um grão longo. Dentro das variedades de arroz de grão médio, destaca-se o arroz *Carnaroli*, sendo este aconselhado para a preparação de *risotto,* Por fim, destacam-se o arroz glutinoso, tipicamente usado em práticos asiáticos como o caso do *sushi* e, o arroz *Arborio* que também pode ser usado na confecção de *risotto,* dentro dos tipos de arroz que apresentam um grão redondo (ou curto) (Almeida & Marques, 2013; Arrozeiras Mundiarroz, sem data; Cotarroz, sem data-b; Drumond, 2012; Marques, 2009; Novarroz, sem data-d).

Na Figura 2.3 são apresentados exemplos de arroz integral e branco, onde são visíveis as diferentes de cor entre eles.

Figura 2.3 - Imagem representativa de arroz integral e branco

2.1.3. O arroz em Portugal

Foi no reinado de D. Dinis que surgem as primeiras referências escritas sobre a cultura do arroz, nas quais se indicava que este se destinava somente às classes mais nobres. A cultura do arroz em Portugal julga-se ter sido introduzida na zona do Baixo Mondego, com sementes vindas da região de Sevilha. No início do século XVIII, este cereal já era documentado, considerando registos da sua presença em zonas do estuário do Tejo, onde inclusivamente eram facultados incentivos à sua produção. Mais tarde, por volta de 1909 estas culturas foram crescendo para outras regiões do país, após se ter elaborado um conjunto de regras para a preparação dos terrenos e da gestão da água, com o intuito de expandir o cultivo a outras variedades de arroz (Drumond, 2012; Novarroz, sem data-a, sem data-c).

Atualmente, em Portugal o cultivo do arroz é feito em três regiões: Vale do Sado, Vale do Tejo e Sorraia e Vale do Mondego (Baixo Mondego).

No que toca a quantidades, em Portugal consomem-se cerca de 180 mil toneladas por ano, fazendo com que se torne no maior consumidor de arroz da Europa (cerca de 17 kg/capita/ano). Deste consumo anual, 44% é arroz carolino, 45% é arroz agulha e os restantes 11% são de outros tipos. Produzem-se anualmente 120 mil toneladas, das quais 72% é de arroz carolino, 27% é de arroz agulha e apenas 1% de outros tipos de arroz. Por conseguinte, conclui-se que Portugal é autossuficiente em arroz carolino, fazendo com que, alguma da produção seja exportada. No entanto, é necessário por ano, importar cerca de 80 mil toneladas, das quais 90% é de arroz agulha (Almeida & Marques, 2013).

2.1.4. O cultivo do arroz

O arroz é cultivado como uma planta anual monocárpica (dá flor apenas uma vez), apesar de algumas variedades em áreas tropicais poderem crescer como perenes, isto é, deixar os estolões no

campo para a produção da próxima cultura. Em Portugal cultiva-se em condições de inundação quase permanente, porém noutras regiões, como em algumas zonas do norte do Brasil, pode ser cultivado em condições de sequeiro ou submersão mais ou menos profunda.

O arroz pode atingir alturas entre 1 e 1,8 metros com folhas finas e compridas. Estas podem variar entre 50 e 100 centímetros e, entre 2 e 2,5 centímetros de comprimento e largura, respetivamente. Já o grão de arroz pode ter entre 5 e 12 milímetro de comprimento e, entre 2 e 3 milímetros de espessura. O ciclo de crescimento deste cereal é de 3 a 6 meses (90 a 180 dias), dependendo da variedade e do ambiente em que for cultivado.

Especificamente em Portugal, os campos no Inverno encontram-se em repouso após mais uma época de colheitas, e à medida que as chuvas vão caindo, os campos vão-se transformando em autênticos lagos, até que, em meados da Primavera tem início um novo ciclo de cultivo.

Durante os meses de Fevereiro e Março preparam-se os terrenos, sendo que a limpeza das valas é feita com a finalidade de irrigar e drenar os campos. Após a limpeza, os campos são gradados, o que faz com os terrenos fiquem nivelados e permite a criação de lama, ideal para receber o arroz previamente germinado (ou “chumbado” – as sementes são colocadas em água para incharem e ganhar peso). Isto facilita o seu enraizamento, evitando que seja arrancado e deslocado pelas pequenas ondas formadas pelo vento. Além das operações de preparação do solo, a incorporação de adubo de fundo e controlo inicial da água de rega é fundamental a realização de uma correta sementeira, para que durante as primeiras cinco a seis semanas da campanha seja possível obter um bom estabelecimento do arrozal. A sementeira decorre já no final de Abril e, pode ser realizada a seco, neste caso sempre por via terrestre, ou com os canteiros inundados podendo esta ser por via terrestre ou via aérea – avião.

Tal como qualquer outra cultura agrícola, também o arroz precisa de cuidados para o seu crescimento e nutrição, de quantidades adequadas e oportunas de nutrientes que extrai do solo ou dos fertilizantes. Em Junho, quando o arroz já tem alguma altura acima da água é adubado. Nos adubos encontram-se nutrientes como o azoto, fósforo, potássio, cálcio, magnésio, enxofre, ferro e silício. A sua adição em determinada quantidade aumenta a velocidade de crescimento, a matéria seca e o rendimento do grão. Também este processo pode ser feito por via terrestre ou aérea.

Em meados de Junho voltam os trabalhos ao campo. Nesta altura procede-se à monda do arroz – eliminação de ervas daninhas, que atualmente é feita com recurso a produtos químicos (herbicidas), e para que estes surtam efeito, os campos devem ter pouca água. Os herbicidas controlam não só os infestantes da planta, como também doenças e pragas que possam surgir. Nos climas temperados a cultura do arroz depende fundamentalmente da disponibilidade de água, sendo que os níveis desta são, normalmente, controlados através de comportas. A temperatura, qualidade, variações de nível e até os organismos animais e vegetais que a água contém, conferem-lhe particular valor e, estão diretamente relacionados com o rendimento da cultura.

Em Setembro chega a altura de ceifar o arroz. A colheita deve ser realizada quando a humidade do grão alcança determinados valores. A colheita com baixas humidades origina aumento da percentagem de grãos partidos durante os processos de laboração com o arroz, tal como humidades elevadas conduzem a menores produções e ao aumento de grãos verdes e gessados. Em tempos o arroz era cortado à mão, mas atualmente essa prática apenas acontece para cortar alguns cantos que a ceifeira mecânica não consegue alcançar, para isso os campos são drenados previamente para que as ceifeiras entrem no campo. Posto isto, o arroz é levado para secar nas eiras ou em secadores mecânicos e posteriormente ser descascado e branqueado. O descascar do arroz é feito no moinho e, só após isso, pode finalmente ser consumido. (Campo, sem data; Cotarroz, sem data-a; Marques, 2009; Reaño, Sackville, & Romero, 2008)

2.1.5. Agricultura e certificação biológica

Segundo a legislação em vigor no país, a agricultura biológica deverá utilizar sobretudo recursos renováveis em que os desperdícios e subprodutos deverão ser reciclados, a fim de restituir os nutrientes à terra. A produção biológica deverá contribuir para manter e aumentar a fertilidade dos solos e impedir a sua erosão, dando preferência à nutrição dada pelos ecossistemas dos solos e não por fertilizantes solúveis espalhados nas terras. No entanto, os fertilizantes, os corretivos do solo e os produtos fitofarmacêuticos só deverão ser utilizados se forem compatíveis com os objetivos e princípios da produção biológica.

Com o propósito de clareza para os consumidores em todo o mercado comunitário, é conveniente tomar medidas ao nível da União Europeia. Como tal, tornou-se obrigatória a aplicação do logotipo da UE em todos os produtos alimentares biológicos pré-embalados produzidos na UE, bem como a nomeação das entidades que possam conceder a permissão de tal aplicação às empresas produtoras deste tipo de produtos.

A agricultura biológica tem presente uma série de regras que visam a obtenção da certificação biológica presente nos rótulos, das quais se destacam (Regulamento (CE) Nº 834/2007 do Conselho de 28 de Junho, 2007):

- A fertilidade e a atividade biológica dos solos são mantidas e aumentadas pela rotação plurianual das culturas, incluindo leguminosas e outras culturas para a adubação verde, e pela aplicação de estrume ou de matérias orgânicas, de preferência ambos compostados, provenientes da produção biológica;
- Só podem ser utilizados fertilizantes e corretivos dos solos autorizados para utilização na produção biológica;
- Não podem ser utilizados fertilizantes minerais azotados;
- Todas as técnicas de produção vegetal utilizadas devem impedir ou reduzir ao mínimo eventuais contribuições para a contaminação do ambiente;

- A prevenção dos danos causados por parasitas, doenças e infestantes deve assentar principalmente na proteção dos predadores naturais, na escolha das espécies e variedades, na rotação das culturas, nas técnicas de cultivo e em processos térmicos;
- Em caso de ameaça comprovada para uma cultura, só podem ser utilizados produtos fitofarmacêuticos autorizados para utilização na produção biológica;
- No caso das sementes, e respetivas plantas-mãe devem ter sido produzidas segundo as regras supracitadas durante pelo menos uma geração ou, no caso de culturas perenes, dois ciclos vegetativos;
- Só podem ser utilizados na produção vegetal produtos de limpeza e desinfeção autorizados para utilização na produção biológica.

A Itália é um exemplo no que diz respeito à produção de arroz biológico, cultivava já em 2007 grandes áreas com este tipo de agricultura, a rondar os 14 mil hectares. Tomando isso por base, o agrupamento de produtores de arroz do vale do Sado (APARROZ), testou culturas de arroz pelo meio da agricultura biológica na zona de Montemor-o-Velho. O objetivo desta experiência era mostrar que seria possível explorar este nicho de mercado em Portugal (APARROZ, 2007). Um dos animais que é usado no controlo biológico de pragas em campos de arroz irrigado é o Marreco-de-Pequim, uma ave que consegue ter um bom rendimento aquando da infestação de determinada praga (João & Azambuja, 2005).

No que toca à certificação de produtos biológicos em Portugal existem, segundo o Instituto Português de Acreditação (IPAC), diversas entidades competentes para fazer a certificação biológica que foram aprovadas para tal pela legislação nacional.

2.1.6. O arroz na cozinha portuguesa

O arroz é o acompanhamento habitual dos pratos portugueses. As imensas possibilidades que proporciona na cozinha foram aproveitadas desde sempre na cozinha portuguesa, passando a fazer parte da cultura culinária do país.

O carácter polivalente do arroz, a sua suavidade, e a capacidade de absorver aromas, sabores e texturas, permitem que este faça parte de receitas típicas de Norte a Sul de Portugal. Dos pratos mais simples à gastronomia mais avançada, a cozinha portuguesa encontra no arroz o parceiro perfeito para potenciar o protagonismo do alimento que o acompanha, seja carne, peixe, marisco ou legumes.

Pratos como o arroz de cabidela, o arroz de pato ou o arroz de sarrabulho são exemplos bem conhecidos dos pratos confecionados com a combinação da carne com o arroz. Sem esquecer os numerosos exemplos que associam o peixe e marisco ao arroz, de como são exemplo, o arroz de polvo, ou o arroz de marisco

O arroz carolino tem um maior poder de absorção dos sabores e, depois de cozido, o grão fica solto, e envolvido num molho cremoso e aveludado (empapado). Tradicionalmente, o arroz carolino é

empregue em receitas típicas de carne, de peixe e eventualmente, de marisco ou de legumes. Pelo contrário, o arroz agulha é um arroz mais solto, que não empapa com tanta facilidade, usado normalmente para fazer o tradicional arroz branco.

As mil e uma maneiras de cozinhar arroz em Portugal fazem do país um dos maiores consumidores de arroz da União Europeia, e do arroz, um amigo inseparável das cozinhas portuguesas (Martins, 2012; Novarroz, sem data-b)

2.2. Química alimentar

2.2.1. Enquadramento

Uma boa nutrição é a base de uma boa saúde, e como tal, todo o ser humano necessita de quatro nutrientes básicos: água, hidratos de carbono, proteínas e lípidos, sendo ingeridos maioritariamente através de comida. A quantidade de proteína a ser ingerida diariamente, como parte de uma dieta adequada nutricionalmente, é identificada como requisito. Este é definido pela quantidade alimentar a ingerir conseguindo satisfazer as necessidades do organismo, e tem o valor mínimo diário de 0,75g por cada kg do individuo (Balch, 2006; WHO, FAO, & UNU, 2007).

A ciência alimentar (*food science*) é um assunto interdisciplinar que envolve bacteriologia, química, biologia e engenharia. A química, o ponto fundamental desta ciência, trata a composição dos alimentos, e as características físico-químicas e suas mudanças durante o processamento, armazenamento e manuseamento a que os alimentos são submetidos. A cor, o sabor, a textura, o valor nutritivo e a segurança são importantes atributos da qualidade dos alimentos, e nestas áreas, grandes avanços têm sido feitos nos últimos anos para alterações que podem ocorrer nestes durante o processamento, armazenamento ou manuseamento. A química alimentar está intimamente relacionada com química, bioquímica, química fisiológica, botânica, zoologia e biologia molecular (Fennema, 1996).

Quanto à constituição do arroz, e para cimentar o que foi mencionado num ponto anterior (onde se falam das características do arroz), tem-se que, o amido é o seu principal constituinte (cerca de 90% dos hidratos de carbono presentes no arroz; os restantes são açúcares), e possui quantidades menores de proteínas e lípidos. No entanto, existem diferenças entre o arroz integral e o arroz branco, pois este é sujeito a um processo de branqueamento onde a quantidade de alguns nutrientes diminui. Para um arroz branco tem-se em valores aproximados 87,58% de amido, 8,94% de proteínas e 0,36% de lípidos; por outro lado para um arroz integral tem-se 74,12% de amido, 10,46% de proteínas e 2,52% de lípidos. Em suma, o arroz é, uma excelente fonte de energia, devido à alta concentração de amido, contendo ainda proteínas, vitaminas e minerais, e possui baixo teor de lípidos (gorduras) (Walter et al., 2008).

2.2.2. Aminoácidos

Os aminoácidos são as unidades estruturais básicas das proteínas formados por um grupo amina, um grupo carboxilo e uma cadeia lateral que difere de aminoácido para aminoácido. Os aminoácidos constituintes das proteínas são vinte, que podem ser agrupados tendo em conta o seu grupo R ou propriedade nutricional. Dentro destes vinte, a glutamina e a asparagina contêm também o grupo amida. Para além destes já foram descobertos outros cerca de 600 aminoácidos que não são incorporados nas proteínas, denominados de aminoácidos livres.

Relativamente à composição nutricional do arroz, o que interessa são os aminoácidos proteicos e estes, podem ser divididos com base na polaridade da sua cadeia lateral (Campos, 2009; Nelson, Lehninger, & Cox, 2008).

Para além dessa divisão, os aminoácidos podem ser divididos entre essenciais – aqueles que o organismo não consegue produzir mas são necessários para o seu funcionamento; e não essenciais – aqueles que o organismo igualmente necessita, no entanto, consegue produzi-los. Para além destas duas categorias existe uma terceira, designada por aminoácidos condicionalmente essenciais, que são alguns dos aminoácidos não essenciais que por vezes (em condições fisiológicas especiais) se tornam essenciais para o organismo. Dos vinte aminoácidos proteicos, nove são aminoácidos essenciais, cinco não essenciais, e seis são condicionalmente essenciais. Na Figura 2.4, são apresentadas as categorias e os aminoácidos pertencentes a cada uma delas (Insel, Turner, & Ross, 2004; Trumbo, Schlicker, Yates, & Poos, 2002; WHO et al., 2007). Dentro dos 9 aminoácidos essenciais, existem 2 grupos que se podem formar com aminoácidos que se agrupam. A fenilalanina e a tirosina formam um grupo designado de aminoácidos aromáticos; e a metionina, juntamente com a cisteína, constituem os aminoácidos sulfurados. Por vezes, estes aminoácidos são quantificados juntos, bem como os seus requisitos que são fornecidos como um total dos dois que formam os diferentes grupos (WHO et al., 2007).

Os aminoácidos assinalados de outra cor não aparecem nos dados das análises químicas, devido ao processo de análise, no qual é usada uma hidrólise ácida que destrói ou modifica quimicamente a asparagina, a glutamina e triptofano. A asparagina e glutamina são convertidas nos respetivos ácidos, o ácido aspártico e o ácido glutâmico e são analisados e quantificados como tal. O triptofano é completamente destruído e a sua mensuração terá de ser feita recorrendo a um tipo de hidrólise diferente (Wilson & Walker, 2010).

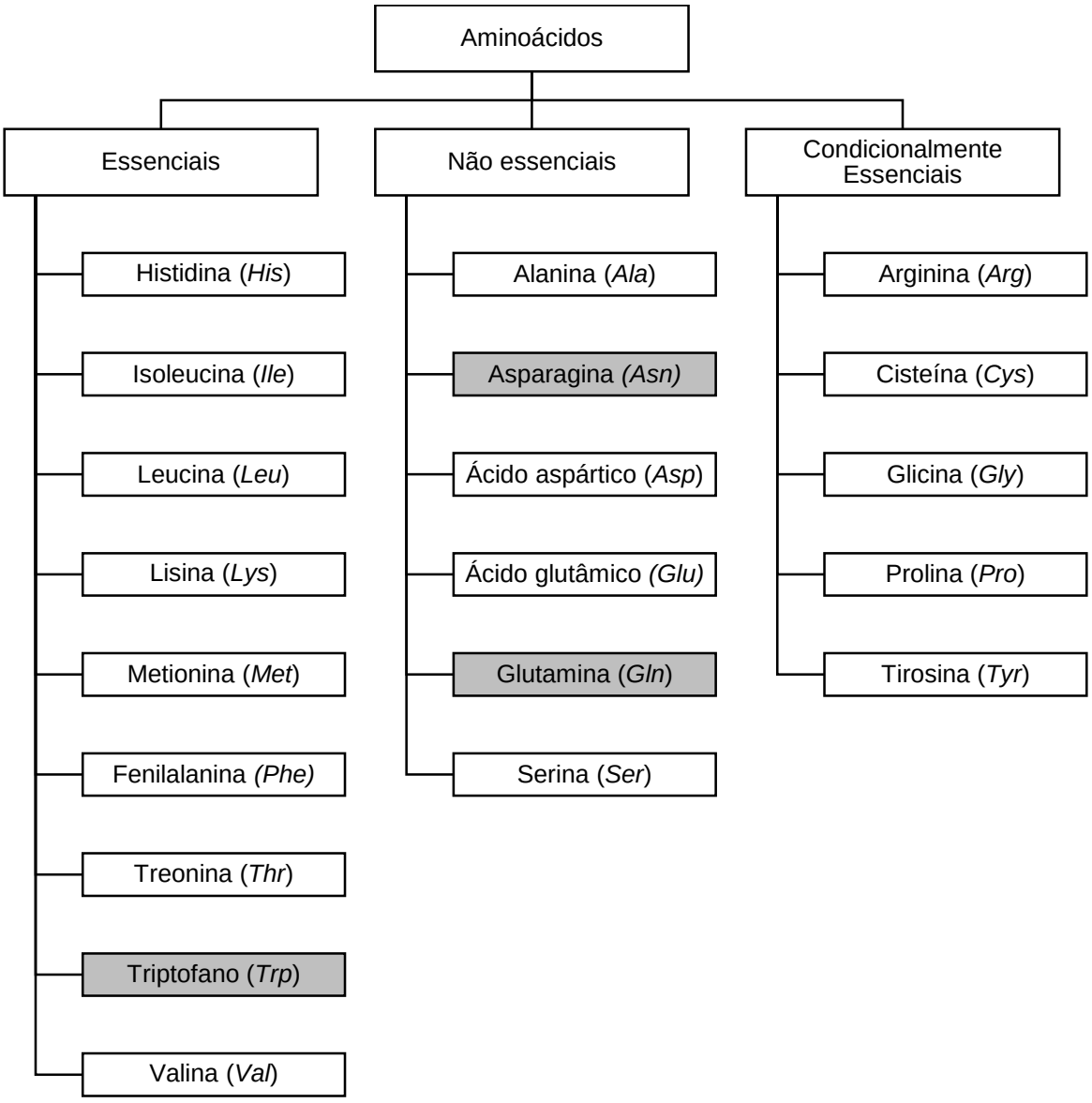

Figura 2.4 - Aminoácidos organizados pela sua dispensabilidade

Para um melhor entendimento da importância dos aminoácidos para o ser humano é descrito de forma sucinta a importância de alguns dos aminoácidos (Balch, 2006):

- A alanina desempenha um papel importante na transferência de nitrogénio dos tecidos periféricos para o fígado, protegendo também a acumulação de substâncias tóxicas, que são libertados nas células musculares;
- A arginina retarda o crescimento de tumores e cancro devido à sua atuação de melhoraria no sistema imunitário[1]. É também benéfica para distúrbios ao nível do fígado;
- A asparagina é necessária para manter o equilíbrio no sistema nervoso central, isto é, evita excessos de nervos ou calma extrema;

[1] Sistema imunitário – sistema de defesas naturais do organismo contra as diversas patologias (Shakir, Hussain, Javeed, Ashraf, & Riaz, 2011).

- O ácido aspártico é favorável para a fadiga e depressão, pois aumenta os níveis de energia no ser humano. Por vezes, é encontrado em níveis elevados em pessoas com epilepsia e níveis baixos em pessoas com depressões ou com fadiga crónica;
- O ácido glutâmico é importante no metabolismo de açúcares e ajuda em perturbações de personalidade. É usado no tratamento de epilepsia, úlceras e atrasos mentais;
- A glutamina é o aminoácido mais abundante nos músculos do corpo humano porque este ajuda na construção e manutenção dos mesmos;
- A glicina retarda a degeneração dos músculos fornecendo creatina[2] adicional;
- A histidina é importante no crescimento e reparação de tecidos, protege o corpo de danos provocados por radiação e, auxilia na remoção dos metais pesados presentes no sistema;
- A isoleucina, a leucina e a valina atuam juntos para proteger os músculos e atuam como combustível. O arroz integral é uma fonte natural de leucina, já os ovos são uma fonte alimentar de isoleucina. De acrescentar ainda, que a valina tem um efeito estimulante;
- A lisina ajuda na absorção de cálcio e é importante em recuperações de cirurgias e lesões desportivas pois ajuda a contruir proteína muscular;
- A metionina auxilia o fortalecimento do cabelo e das unhas, melhorando, simultaneamente, a saúde da pele e é igualmente benéfico para a degradação das gorduras, evitando sua acumulação no fígado e nas artérias que podem obstruir a corrente sanguínea;
- A fenilalanina pode melhorar o humor, diminuir a dor, reduzir o apetite ou até mesmo ajudar na aprendizagem e memorização; já a prolina melhora a textura da pele e ajuda na cicatrização de cartilagens e, também fortalece tendões e articulações;
- A serina é útil para o metabolismo adequado de gorduras, para o crescimento dos músculos e para a conservação de um sistema imunitário saudável;
- A treonina melhora o funcionamento do fígado e encontra-se presente no coração, sistema nervoso central e músculo-esquelético;
- O triptofano é conhecido pelas suas propriedades calmantes ajudando a controlar a hiperatividade em crianças, alivia o stress e é bom para o coração.

Pelo referido anteriormente, pode-se atestar o papel vital na saúde humana dos aminoácidos com os efeitos benéficos a vários níveis potenciados pelos mesmos e, se por um lado, a produção de aminoácidos não essenciais está assegurada pelo seu organismo, por outro, a variedade, quantidade e qualidade de aminoácidos essenciais depende da alimentação de cada um, originando assim o papel preponderante do arroz na alimentação do Homem.

[2] Creatina - derivado de aminoácido que pode ser encontrado em maior concentração nos músculos. É usada como suplemento alimentar devido aos seus benefícios: melhora o desempenho físico, reduz a fadiga, acelera a recuperação e o crescimento dos músculos (Moret, Prevarin, & Tubaro, 2011).

2.2.3. Análise de aminoácidos - cromatografia

As proteínas e peptídeos são macromoléculas formadas por aminoácidos interligados e organizados em sequência. Os peptídeos são moléculas menores que as proteínas e constituídos por poucos aminoácidos. Com isto, as análises de aminoácidos podem ser utilizadas para quantificar ou identificar proteínas e/ou peptídeos (Campos, 2009; Nelson et al., 2008; Wilson & Walker, 2010).

Para a análise de aminoácidos, é usada uma técnica de separação, mais especificamente a cromatografia – método de separação físico-química. Dentro da cromatografia existem diversas técnicas que podem ser agrupadas de várias formas. No entanto, para o presente estudo só irá ser abordada a técnica UPLC (*ultra-performance liquid chromatography*), por ter sido a usada na determinação dos aminoácidos no arroz. Esta técnica é um tipo de cromatografia em que se encontra dentro das cromatografias de coluna líquidas (Heftmann, 2004; Nollet & Toldra, 2012; Wilson & Walker, 2010).

O princípio de funcionamento da cromatografia em coluna consiste em colocar uma amostra (solução composta de solvente e solutos) numa coluna. Na cromatografia existem duas fases: a fase móvel (que faz com que a amostra progrida na coluna) e a fase estacionária – coluna cromatográfica (com o objetivo das partículas que vão juntamente com a fase móvel se ligarem a esta), sendo que esta depende do objetivo da análise. A fase móvel é um solvente (é o seu estado que determina o tipo de cromatografia: líquido, gás ou gás pressurizado), já a fase estacionária é sólida. Com a ajuda da fase móvel a amostra irá percorrer a coluna e, uma vez que cada componente na amostra irá interagir de forma diferente com a coluna, originar-se-ão diferentes taxas de fluxo para os diferentes componentes, conduzindo à construção do cromatograma em função do "tempo de ligação" entre as partículas e a fase estacionária (Atkins & Jones, 1999; Heftmann, 2004; Nollet & Toldra, 2012). O cromatograma é o resultado da análise cromatográfica (Heftmann, 2004), que por sua vez será extrapolado para concentrações. Na Figura 2.5 está representado um exemplo de um cromatograma.

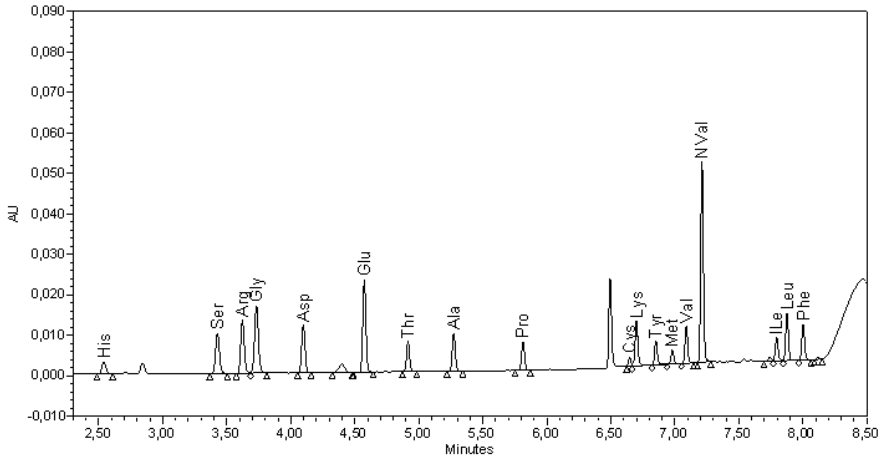

Figura 2.5 - Exemplo de um cromatograma

O HPLC (*high-performance liquid chromatography*) é uma técnica preferencial e bastante usada, quando comparada a outras técnicas de cromatografia, uma vez que apresenta a vantagem de utilizar

pressões elevadas. Desta forma, como a fase móvel é forçada sobre pressão a atravessar a coluna, a separação dos componentes não irá depender apenas da força gravítica e a separação dos componentes será mais eficiente (Atkins & Jones, 1999; Heftmann, 2004; Nollet & Toldra, 2012). No entanto, o UPLC é bastante melhor que o HPLC. É um processo simples, reproduzível e preciso, que consegue reduzir o tempo de análise. O UPLC necessita de apenas 40% do tempo necessário pelo HPLC para obter a separação dos dezassete aminoácidos com uma resolução clara (Boogers, Plugge, Stokkermans, & Duchateau, 2008).

2.2.4. Arsénio

O arsénio (As) é um elemento químico pertencente à tabela periódica, que pode ser encontrado naturalmente no seu estado orgânico ou inorgânico (Asi) (Heikens, 2006). Pode ser encontrado no ar, na água, no solo, em águas subterrâneas, pedras, e em metais como chumbo e cobre. Isto faz do arsénio o vigésimo elemento mais abundante na crosta terrestre. O As encontra-se combinado com o oxigénio, cloro e enxofre na sua forma inorgânica, e com carbono e hidrogénio na forma orgânica (ATSDR & EPA, 2007). As espécies inorgânicas mais importantes são o arsenato (AsV) e o arsenito (AsIII), já os ácidos monometilarsónico (MMA) e dimetilarsínico (DMA) são as espécies orgânicas mais comuns (Simões, 2014).

O arsénio inorgânico é altamente tóxico, sendo um conhecido agente cancerígeno (faz parte do grupo 1 dos agentes cancerígenas – "cancerígenas para humanos") e contaminante da cadeia alimentar (Dwivedi et al., 2012). A ingestão de arsénio inorgânico pode levar a intoxicação se for durante um longo período ou a efeitos que podem levar anos para se desenvolver, dependendo do nível de exposição. Efeitos esses que incluem lesões de pele, sintomas gastrointestinais, diabetes, efeitos sobre o sistema renal, doenças cardiovasculares e cancro (tais como: de pele, bexiga e pulmão) devido às propriedades carcinogénicas deste elemento. De destacar que compostos de arsénio orgânicos, são menos prejudiciais à saúde devido à sua menor toxicidade, e são rapidamente eliminados pelo corpo (Dwivedi et al., 2012; WHO, 2010, 2011).

A exposição humana a níveis elevados de arsénio inorgânico está altamente interligada à água subterrânea que contém naturalmente altos níveis de arsênico inorgânico. Ocorre principalmente através do consumo dessa água ou da ingestão de alimentos preparados ou irrigados durante a sua produção com essa água (Dwivedi et al., 2012; WHO, 2010). O arroz é, então, uma importante fonte de exposição ao arsénio inorgânico devido à sua produção ser muitas vezes em condições inundadas, em especial para as populações dependentes de uma dieta básica de arroz (Dwivedi et al., 2012).

A parte comestível do arroz é o grão, porém os caules e folhas da planta são normalmente secos e, a palha formada destina-se à alimentação animal. Com isto, é gerada outra via de exposição humana a este elemento. No grão, o As distribui-se de forma desigual nas diferentes zonas: a casca é a zona do grão onde se acumula maior quantidade de As, seguida do farelo. Portanto, o tipo de processamento

a que o arroz é submetido pode reduzir a concentração de As no grão, nomeadamente o branqueamento do arroz branco em que é retirado o farelo (Simões, 2014).

Em águas para consumo humano e para rega a legislação portuguesa refere como limite de As 10 µg/L e 10 mg/L, respetivamente. O limite para as águas de rega é variável consoante as culturas, especificamente no caso da rega para plantações de arroz, o limite passa a ser de 0,05 mg/L. Porém a legislação nacional não contempla ainda limites para os elementos contaminantes, entre os quais o arsénio, nos solos. São utilizados então os valores de referência de países como a Holanda e o Canadá (Simões, 2014).

Recentemente foi proposto para a União Europeia (WHO/FAO) um limite máximo de arsénio para arroz cru de 0,3 mg/kg (300 ppb). No entanto, ainda não existe limite máximo para o arsénio no arroz na Europa e nos Estados Unidos (FAO/WHO, 2012).

2.2.5. Análise do arsénio - espectrometria de massa

Os elementos que são necessários para os seres vivos em quantidades muito pequenas têm sido variadamente designados como micronutrientes, elementos-traço ou microelementos. No entanto, existe uma distinção destes significados entre as plantas e os animais/Homem: para as plantas micronutrientes são os elementos-traço essenciais e elementos-traço os não essenciais; já para os animais e humanos os elementos-traço são apenas aqueles que são encontrados em pequenas concentrações. Apenas de realçar que os elementos-traço mesmo que essenciais (quando ingeridos de menos podem causar distúrbios nutricionais), quando ingeridos em demasia podem ser tóxicos (Adriano, 2001; Pais & Jones, 1997). O arsénio faz, então, parte do grupo dos elementos-traço.

Para análise dos elementos-traço é usada a espectrometria de massa, mais especificamente o ICP-MS (*Inductively Coupled Plasma – Mass Spectrometry* ou em português, espectrometria de massa com plasma acoplada indutivamente). Existem diversos tipos de espectrometria, no entanto esta é dedicada quase exclusivamente a elementos-traço devido às concentrações muito pequenas que consegue medir. Esta técnica consiste em que os iões de carga única, formados num plasma à pressão atmosférica, sejam extraídos para um analisador de massa quadrupólo para deteção. O árgon tem sido indicado como uma boa fonte de emissão e tem sido o plasma de preferência para a espectrometria de massa (Krull, 1991).

CAPÍTULO 3 – ESTATÍSTICA MULTIVARIADA

3.1. A estatística e os diferentes tipos de análise

A estatística é a ciência que permite recolher, organizar, apresentar, analisar e interpretar dados quantitativos com o objetivo de tomar melhores decisões. A estatística é uma parte da Matemática Aplicada, podendo ser dividida em 3 áreas, a estatística descritiva – que tem por objetivo descrever e resumir os dados, a fim de se poder tirar conclusões acerca das características dos mesmos (através da análise exploratória); a probabilidade – que se traduz na ferramenta matemática que deduz a partir de um modelo as propriedades de um fenómeno aleatório; e, a inferência estatística (ou estatística indutiva) – conjunto de métodos que permite inferir o comportamento de uma população a partir do conhecimento da amostra. Por população entende-se como o conjunto de todos os elementos em estudo que tem pelo menos uma característica em comum e, por amostra, um subconjunto da população (Neto, 2004).

Dentro da estatística descritiva existem diferentes tipos de análises estatísticas que diferem pelo número de variáveis que se estão a estudar simultaneamente. Se o estudo se restringir apenas a uma variável então é utilizada uma análise univariada. No entanto, se o estudo for de duas variáveis então é uma análise bivariada e, o que a distingue da análise univariada é que estuda ainda a relação entre as variáveis. Em último lugar, tem-se a análise multivariada, que por sua vez analisa três ou mais variáveis. Por outras palavras, a análise multivariada refere-se a todos os métodos estatísticos que ao mesmo tempo analisam várias medidas sobre cada caso (objeto ou indivíduo) da amostra em estudo (Corrar, Paulo, & Filho, 2007; Rencher, 2005).

No presente estudo, devido ao elevado número de variáveis que se pretende analisar e estudar para além da estatística descritiva, ir-se-á dar especial atenção a algumas técnicas de análise multivariada.

3.2. Objetivo e aplicação da análise multivariada

O objetivo subjacente à utilização de análise multivariada a um conjunto complexo de dados é retirar o máximo de informação possível dos mesmos. Este tipo de análise poderá servir para controlo de qualidade, otimização e controlo de processos ou pesquisa e desenvolvimento.

A análise multivariada, por usufruir da capacidade de conseguir analisar um conjunto significativo de variáveis, tem diversos campos de aplicação, podendo estes ir desde a área da educação, psicologia, à área da química, física, geologia, entre muitas outras (Rencher, 2005).

Existem inúmeras técnicas de análise multivariada, dependendo da finalidade (ou da resposta que se quer encontrar após a análise dos dados). Estas respostas podem ter a ver com o grau de relação entre variáveis, as diferenças entre grupos de amostras diferentes, a previsão de aderência de dados em grupos, a estrutura dos dados e/ou o tempo de decurso de determinados eventos (Mardia, Kent, & Bibby, 1980; Rencher, 2005).

Sendo o objetivo do presente estudo analisar o perfil de aminoácidos de diferentes tipos de arroz, analisar afinidades entre aminoácidos e possíveis correlações com o arsénio, a análise estatística multivariada irá constituir uma ferramenta importante e indispensável. Por se tratar de dados obtidos por análises químicas, tem uma designação própria – quimiometria. A quimiometria pode ser definida pela interação de métodos estatísticos e matemáticos com dados de origem química. Em suma, a quimiometria analisa, através da análise multivariada, diversas variáveis químicas medidas de várias amostras (Kumar, Bansal, Sarma, & Rawal, 2014).

3.3. Variáveis

Antes de introduzir técnicas de quimiometria, é fundamental abarcar o tema das variáveis. As variáveis, com base na natureza dos dados, podem ser classificadas como qualitativas (não-métricas ou categóricas) e/ou quantitativas (métricas). As variáveis qualitativas podem ser definidas por várias categorias, podendo ser nominais ou ordinais; já as variáveis quantitativas, expressas por números, podem ser discretas (quanto se usam apenas valores inteiros) ou contínuas (com toda a escala real à sua disposição).

Para os valores das variáveis existem diferentes escalas, podendo estas ser nominais, ordinais, intervalares e de razão, como pode ser visto na Figura 3.1, apresentada de seguida.

Para o estudo feito, por se tratar de dados químicos, as variáveis medidas são quantitativas contínuas, sendo usada uma escala de razão devido aos atributos do arroz que estão a ser medidos (Hair, Black, Babin, Anderson, & Tatham, 2006; Morais, 2005).

Outra forma de classificar as variáveis é pela sua manipulação, podendo ser variáveis independentes ou dependentes (Lino, 2009).

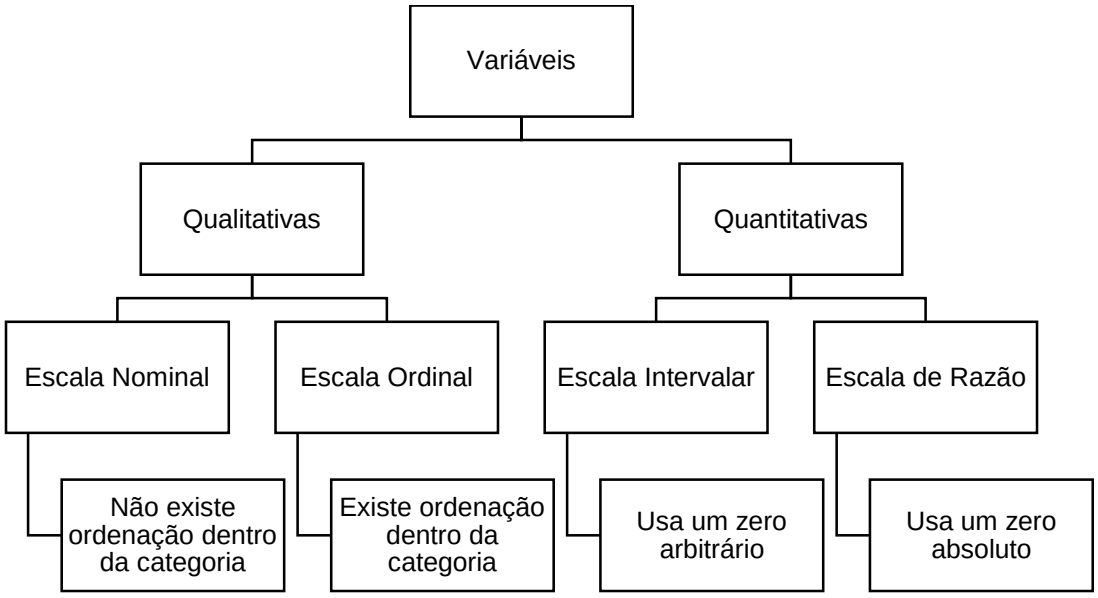

Figura 3.1 - Tipos de variáveis

Após o entendimento da escala de medida, e antes de aplicar qualquer técnica de análise multivariada aos dados, é necessária uma análise exploratória dos mesmos.

3.4. Conceitos básicos

É fundamental fazer uma explicação sucinta de conteúdos básicos de estatística que se encaixam, para já, na análise exploratória e, posteriormente noutras análises.

O teste de hipóteses é algo que, de certa forma, está associado à estatística, pois é um método de inferência estatística, que a partir de uma ou várias amostras permite averiguar se determinada hipótese sobre a ou as populações deve ou não ser rejeitada. A hipótese que se deseja testar é definida por Hipótese Nula (H_0) e, até prova estatística em contrário, é assumida como verdadeira. De salientar que, a hipótese nula contém sempre a igualdade. A prova estatística que possa levar à rejeição ou não da hipótese nula é fundamentada numa estatística de teste apropriada ao caso em estudo. A outra hipótese, designada por Hipótese Alternativa (H_1) contém o "oposto" da hipótese nula, e poderá ser bilateral ou unilateral dependendo da circunstância. Para decidir a rejeição ou não da hipótese nula, define-se o nível de significância (α) e compara-se com o valor da estatística de teste calculado. No entanto, podem cometer-se dois tipos de erros quando se faz um teste de hipóteses: o Erro Tipo I – rejeitar a hipótese nula sendo ela verdadeira; e, o Erro Tipo II – não rejeitar a hipótese nula quando esta é falsa. A probabilidade do Erro Tipo II é representada por β e, a do Erro Tipo I por α, ou seja, o nível de significância que é definido pelo avaliador (Pereira & Requeijo, 2012).

Outro conceito que está muito presente no dia-a-dia é o valor-p, um valor que é dado pelos *softwares* de estatística (e não só) quando é executado um teste estatístico nos mesmos. O valor-p é, geralmente, a ferramenta mais usada para medir evidências contra a hipótese nula (Sellke, Bayarri, & Berger, 2001). Por outras palavras, é a probabilidade de encontrar um valor da estatística de teste melhor ou igual ao observado, sabendo que a hipótese nula é verdadeira (Bayarri & Berger, 2000;

Devore, 2011). De uma forma mais leve, pode afirmar-se que esta ferramenta, também denominada nível descritivo do teste, representa o menor nível de significância ao qual a hipótese nula pode ser rejeitada.

3.5. Análise exploratória

Previamente a uma análise multivariada é essencial e necessário uma análise exploratória que envolve uma análise univariada com estatística descritiva às diversas variáveis. Durante esta análise é importante verificar a presença de *outliers*, pois estes podem enviesar os parâmetros e fazer com que sejam retiradas conclusões erradas no final. Após a obtenção dos dados em bruto, é necessário organizá-los, habitualmente feito através da construção de tabelas.

Os valores das variáveis medidas podem ser de diferentes unidades, de diferente ordem de grandeza ou medidos por diferentes instrumentos, o que faz com que as variáveis possam ter diferentes pesos. No caso de existir esse problema, pode ser necessário fazer uma nova ponderação – que consiste em multiplicar as variáveis por uma constante diferente entre elas, para que sejam alcançadas condições de igualdade; e/ou um novo dimensionamento – que consiste em colocar as variáveis numa nova escala igual para todas, através de um critério interno ou externo. Dentro dos critérios internos, os mais usados são o *mean centering*: que consiste em subtrair a média da variável a cada observação; a padronização: que para além da subtração da média, as observações são divididas pelo desvio-padrão; e a normalização: que consiste em subtrair cada observação pelo mínimo da variável, e posterior divisão pela diferença entre o máximo e mínimo dessa mesma variável. No que toca aos critérios externos tem-se, por exemplo, a transformação logarítmica, em que para todas as observações são calculadas o valor do seu logaritmo.

Calcular a média, a variância (e/ou o desvio padrão), testar a normalidade da distribuição ou a homogeneidade da variância, verificar a assimetria, testar hipóteses sobre a(s) média(s) ou variância(s) são ações que podem ser feitas com o intuito de conhecer melhor as variáveis individualmente e obter conclusões particulares antes de partir para a análise multivariada (Berrueta, Alonso-Salces, & Héberger, 2007).

3.5.1. Comparação de médias

De seguida apresentam-se o teste *t* de *Student* e a análise de variância (*ANalysis Of VAriance* - ANOVA), técnicas pertencentes à estatística paramétrica, que permitem testar hipóteses sobre as médias de uma ou várias populações (Huang & Paes, 2009).

3.5.1.1. Teste *t* de *Student*

O teste *t* de *Student* é um teste estatístico que visa testar a diferença das médias de duas amostras independentes.

Para este teste (teste bilateral) as hipóteses são:

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$

em que μ_1 e μ_2 são os valores das médias das populações.

Contudo, a execução do teste depende do facto das variâncias das duas amostras serem diferentes ou não. Como consequência de tais opções, é necessário verificar se existem diferenças significativas nas variâncias, sendo por isso usado o teste *F* de Fisher. Este teste está geralmente associado ao teste *t* de *Student* neste contexto.

O teste *F* de Fisher é um teste estatístico que testa a semelhança das variâncias, sendo as hipóteses do teste (teste bilateral) as seguintes:

H_0: $\sigma_1 = \sigma_2$
H_1: $\sigma_1 \neq \sigma_2$

onde σ_1 e σ_2 são os valores dos desvios padrão das populações. Já a estatística de teste é definida por

$$F_0 = \frac{{s_1}^2}{{s_2}^2} \tag{3.1}$$

onde s_1 e s_2 são os valores das variâncias das amostras 1 e 2, respetivamente.

Rejeita-se a hipótese nula se $F_{\frac{\alpha}{2};(n_1-1),(n_2-1)} < F_0 < F_{1-\frac{\alpha}{2};(n_1-1),(n_2-1)}$, valores que são tabelados e dependem, como se pôde ver, do nível de significância escolhido e do número de graus de liberdade (sendo n_1 e n_2 as dimensões das amostras 1 e 2, respectivamente).

Após a aplicação do teste *F* de Fisher, continua-se a apresentação do teste *t* de *Student* para as duas situações.

- **Se as variâncias não forem significativamente diferentes**

Para o caso em que $\sigma_1^2 \cong \sigma_2^2$ a estatística de teste é dada por (passando a seguir uma distribuição t de Student):

$$t_0 = \frac{\overline{x_1} - \overline{x_2}}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{3.2}$$

onde

n_1 e n_2 é o tamanho da amostra 1 e 2, respetivamente
$\bar{x}_1$ e $\bar{x}_2$ é a média da amostra 1 e 2, respetivamente
S_p é o desvio padrão agrupado

$$S_p = \sqrt{\frac{(n_1-1){s_1}^2+(n_2-1){s_2}^2}{(n_1+n_2-2)}} \tag{3.3}$$

Rejeita-se a hipótese nula quando $|t_0| > t_{\frac{\alpha}{2};(n_1+n_2-2)}$.

- **Se as variâncias forem significativamente diferentes**

No caso em que as variâncias, pelo teste *F* de Fisher, forem significativamente diferentes, a estatística de teste é dada por

$$t_0 = \frac{\overline{x_1}-\overline{x_2}}{\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}} \tag{3.4}$$

e o número de graus de liberdade é dado por

$$\vartheta = \frac{(\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2})^2}{\frac{(\frac{{s_1}^2}{n_1})^2}{n_1-1}+\frac{(\frac{{s_2}^2}{n_2})^2}{n_2-1}} \tag{3.5}$$

com base em ϑ (se não for número inteiro, deve adoptar-se o número inteiro imediatamente inferior), determina-se o valor crítico da estatística (valor tabelado na tabela da distribuição *t*), e rejeita-se a hipótese nula se $|t_0| > t_{\frac{\alpha}{2};\vartheta}$ (Agarwal, 2006; Huang & Paes, 2009; Park, 2009; Pereira & Requeijo, 2012).

3.5.1.2. Análise de variância a um fator

A análise de variância (ANOVA) é, também, uma técnica de teste para verificar a existência de diferenças significativas entre as médias das populações. No entanto, há vários modelos para a análise de variância, estando a escolha diretamente relacionada com o número de variáveis independentes de que esta análise testa. No estudo em questão, apenas irá ser desenvolvida a ANOVA a um fator (*one-way* ANOVA) atendendo à sua aplicação prática no capítulo seguinte. A principal diferença entre o teste *t* de *Student* e a ANOVA, é que o primeiro está limitado apenas a duas amostras na sua comparação, enquanto a segunda consegue comparar médias de mais que duas amostras. Pode-se então afirmar, que o teste *t* de *Student* é um caso especial da ANOVA a um fator.

A análise de variância é descrita por um modelo matemático de efeitos fixos:

$$x_{ij} = \mu + \tau_i + \varepsilon_{ij} \tag{3.6}$$

onde

x_{ij} são as observações independentes e normalmente distribuídas com variância homogénea

μ é a média global

τ_i é o efeito do nível *i*

ε_{ij} é uma variável aleatória normalmente distribuída com valor esperado nulo e variância constante

Como o objetivo da análise de variância passa por averiguar se os efeitos dos vários níveis (τ_i's) são ou não significativamente diferentes de zero, a hipótese nula na ANOVA pode ser formulada da seguinte forma:

H_0: $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$

H_1: $pelo\ menos\ duas\ \mu_{i's}\ serem\ diferentes,\ onde\ i = 1,2,3,\ldots,k$

com a seguinte notação:

μ_i é a média da população *i*

k é o numero de amostras (níveis ou tratamentos)

Tamanho das amostras: $n_1, n_2, \ldots, n_k$

Tamanho total de amostras: $N = n_1 + n_2 + \cdots + n_k$

Média das amostras: $\overline{x_1}, \overline{x_2}, \ldots, \overline{x_k}$

Desvio padrão das amostras: $s_{1,} s_2, \ldots, s_k$

Média amostral global: $\bar{\bar{x}}$

A ANOVA usa a estatística *F* para testar se todos os grupos possuem a mesma média da seguinte forma:

$$F = \frac{MS_B}{MS_W} = \frac{\frac{SS_B}{(k-1)}}{\frac{SS_W}{(N-k)}} \tag{3.7}$$

$$SS_B = \sum_{i=1}^{k} n_i(\overline{x_\iota} - \bar{\bar{x}})^2, \ \text{com k} - 1 \text{ graus de liberdade} \tag{3.8}$$

$$SS_W = \sum_{i=1}^{k}(n_i - 1){s_i}^2, \ \text{com N} - \text{k graus de liberdade} \tag{3.9}$$

$$SS_T = SS_B + SS_W, \quad \text{com N} - 1 \text{ graus de liberdade} \tag{3.10}$$

Considerando a variação total (ou soma total dos desvios quadráticos – SS_T), pode dizer-se que esta corresponde à soma dos quadrados dos desvios de todas as observações em relação à média global, sendo igual à soma das variações entre os níveis ou tratamentos (SS_B – *Between Sum of Squares*) e dentro dos níveis ou tratamentos (também designada por variação interior aos tratamentos ou, erro – SS_W *Whithin Sum of Squares*). A variação entre os níveis (SS_B) define-se como a soma ponderada dos quadrados das diferenças entre as médias dos níveis e a média global. Já a variação dentro dos níveis (SS_W) é definida como a soma dos quadrados dos desvios das observações em relação às médias dos respetivos níveis. Após ter as duas variações e, fazendo o quociente entre estas e o respetivo número de graus de liberdade, obtêm-se as variâncias – MS_B e MS_W. O quociente destas obtém o valor de F_0, que será posteriormente comparado com o valor crítico da distribuição *F* (tem em conta o nível de significância e o número de graus de liberdade do numerador e do denominador). Usualmente, é utilizada uma tabela (tabela da ANOVA) para dispor os valores e fazer os cálculos, com o intuito de facilitar, quer os cálculos quer a perceção destes. Um exemplo de uma tabela da ANOVA é apresentado de seguida na Tabela 3.1 (Agarwal, 2006; Montgomery & Runger, 2003; Pereira & Requeijo, 2012; Reddy, 2011).

Tabela 3.1 - Tabela geralmente usada na *one-way* ANOVA

Fonte de variação	Graus de Liberdade	SS	MS	F_0
Entre os níveis	k-1	SS_B	SS_B / (k-1)	MS_B / MS_W
Dentro dos níveis (Erro)	N-k	SS_W	SS_W / (N-k)	
Total	N-1	SS_T	SS_T / (N-1)	

3.5.1.3. Pressupostos

Após a breve explicação das duas técnicas usadas para verificar a existência ou não de diferenças significativas nas médias de duas ou mais populações é importante expor os pressupostos que são assumidos ao utilizar estas técnicas.

1 - Amostras sejam escolhidas independente e aleatoriamente;
2 - Amostram que sejam descritas por distribuições normais;
3 - Variâncias semelhantes (ou homogéneas) entre populações.

Ambas as técnicas pressupõem os três princípios atrás descritos. A normalidade e homogeneidade da variância podem ser testados nas respetivas amostras com recurso a vários testes, ou pela análise de resíduos, sendo este último um requisito obrigatório para a ANOVA. Os resíduos são calculados pela diferença entre o valor observado e o valor esperado (média da população).

$$e_{ij} = x_{ij} - \overline{x_i} \tag{3.11}$$

com:

$i = 1,2, \dots, k$

$j = 1,2, \dots, n$

Estes, quando representados graficamente devem dispor-se numa reta no gráfico da normalidade (valores esperados normalizados vs. resíduos), e de forma aleatória no gráfico da independência (resíduos vs. ordem das amostras) e da homogeneidade da variância (resíduos vs. valores previstos) (Agarwal, 2006; Bekiro, 2001; Bradley, 2007; Pereira & Requeijo, 2012). Isto deve-se ao facto de que as técnicas apresentadas pertencem à estatística paramétrica – ou seja, os parâmetros da população são conhecidos. Para verificar a normalidade e homogeneidade da variância das amostras, existem diversos testes que podem ser aplicados para verificação dos pressupostos supracitados, pois a não verificação dos mesmos pode invalidar a aplicação das respetivas técnicas. Estes testes são apresentados nos próximos pontos do presente documento.

3.5.2. Testes de normalidade

Existem vários testes para verificar a normalidade de uma amostra, no entanto foram escolhidos para explicação e demonstração, segundo a literatura, os testes mais potentes: o teste de Shapiro-Wilk e o teste de Anderson-Darling (Islam, 2011; Razali, Wah, & Sciences, 2011). Contudo, e como ao longo da vida académica, o teste de Kolmogorov-Smirnov foi sempre falado como o teste para testar a normalidade, foi também escolhido, sabendo à partida que é o mais fraco dos três testes (Razali et al., 2011).

3.5.2.1. Teste de Shapiro-Wilk

Entre os testes de normalidade existentes, o teste de Shapiro-Wilk, tornou-se o mais comum em amostras de reduzida dimensão (n <50) devido ao seu poder: é o teste de normalidade mais potente.

O teste de Shapiro-Wilk é realizado para testar a seguinte hipótese:

H_0: A população segue uma distribuição normal;
H_1: A população não segue uma distribuição normal.

E usa a seguinte estatística de teste:

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (3.12)$$

$y_{(i)}$ é o termo de ordem *i* dos *n* valores da amostra x ordenados por ordem crescente
a_i*'s* dependem dos valores esperados da estatística de uma distribuição normal padrão

Quando $W \leq W_\alpha$ (valor crítico tabelado) rejeita-se a hipótese nula, no entanto é mais comum fazer a rejeição ou não através do valor-p dado pelo *software*.

No entanto, pode fazer-se uma aproximação (que pode, inclusive, ser feita manualmente):

1º - Ordenar os *n* valores da amostra *x* crescentemente ($y_1 \leq y_2 \leq y_3 \leq \ldots \leq y_n$)
2º - Calcular a estatística de teste (coeficientes a_i's são valores tabelados):

$$W = \frac{\left(\sum_{i=1}^{k} a_{n-i+1}(y_{n-i+1}-y_i)\right)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}, k = \frac{n}{2} \quad (3.13)$$

3º - Apurar W_α na tabela e comparar com o resultado dado para rejeitar ou não a hipótese nula.

De destacar, que esta aproximação só pode ser feita para amostras com dimensão até 50, pois as tabelas dos coeficientes só permitem que tal aconteça (Abelquist, 2001; Matsushita, Puri, & Hayakawa, 1993; Sen & Srivastava, 1990; Shapiro & Wilk, 1965).

3.5.2.2. Teste de Anderson-Darling

O teste de Anderson-Darling é outro teste que, geralmente, se obtém bons resultados quando se testa a normalidade de uma amostra (Razali et al., 2011; Yap & Sim, 2011).

O teste de hipóteses é o mesmo que no outro teste atrás exposto (pois o objetivo é testar a normalidade) e, a estatística de teste é:

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)(\ln(y_i) + \ln(1-y_{n+1-i})) \quad (3.14)$$

$y_{(i)}$ é o termo de ordem *i* dos *n* valores da observação x ordenados por ordem crescente

Mais uma vez o resultado da estatística de teste é comparada com o valor crítico AD_α (valores tabelados dependendo do nível de significância) em que por sua vez é rejeitada ou não a hipótese nula se o valor da estatística for superior ou inferior, respetivamente (Engmann & Cousineau, 2011).

3.5.2.3. Teste de Kolmogorov-Smirnov

O teste de Kolmogorov-Smirnov, como já referido no início desta secção, aquando da justificação da escolha dos testes, é o mais fraco de todos eles (Razali et al., 2011), no entanto considerou-se importante introduzi-lo como termo de comparação.

Mais uma vez, destina-se a averiguar se uma amostra pode ser considerada como proveniente de uma população com uma determinada distribuição, normal neste caso específico. Este teste consiste em fazer uma comparação entre as funções distribuição de probabilidade da hipótese e da referência, ou seja, entre as funções distribuição da amostra e da distribuição normal (Feldman & Valdez-Flores, 2009).

Hipóteses em teste:

H_0: a população tem uma determinada distribuição (segue uma distribuição normal);
H_1: a população não segue essa determinada distribuição.

$$D_n = sup_x|F_0(x) - \widehat{F_n}(x)| \tag{3.15}$$

A função distribuição de probabilidade da distribuição normal é dada por

$$F_n(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \tag{3.16}$$

Esta função não é analiticamente integrável, pode apenas ser calculada numericamente.

Por outras palavras, D_n é a diferença máxima entre a função distribuição acumulada da hipótese e da amostra. Com base no tamanho da amostra (n) e no nível de significância (geralmente é 0,05), tem-se os valores tabelados de D_α (valor crítico), que se compara com D_n (valor do teste) e se rejeita ou não a hipótese nula (Feldman & Valdez-Flores, 2009; Pereira & Requeijo, 2012). Quando executado informaticamente o valor dado é o valor-p que é comparado diretamente com α.

3.5.3. Testes de homogeneidade da variância

Com o intuito de verificar outro dos pressupostos (a homogeneidade da variância) das técnicas que permitem comparar médias de populações, existem também variadas técnicas. Como já apresentado, tem-se o teste *F* de Fisher usado para o teste *t* de *Student*. Contudo, existem testes mais robustos, como os testes de Bartlett, de Levene, de Brown & Forsythe e de Cochran. O teste de Cochran permite a obtenção de bons resultados, sendo até muitas vezes um dos métodos prediletos nesta área, sendo o mais robusto para amostras que não seguem uma distribuição normal e de tamanhos muito diferentes. O teste de Bartlett é uma boa escolha mas apenas quando se verifica a normalidade nas amostras. Contudo, os testes de Levene e de Brown-Forsythe foram escolhidos, para além do

teste *F* de Fisher já apresentado, pois permitem resultados igualmente bons. Estes dois testes são os mais potentes em amostras de reduzida dimensão e cuja normalidade possa não se verificar, permitindo assim a validação dos valores obtidos. Estes testes são apresentados de seguida, de notar que o teste de Brown-Forsythe é mais robusto que o de Levene (Vorapongsathorn, Taejaroenkul, & Viwatwongkasem, 2004).

3.5.3.1. Teste de Levene

O teste de Levene compara a homogeneidade (semelhança) entre variâncias de várias populações. Se existirem *k* amostras aleatórias independentes entre si, então o teste de hipóteses será:

H_0: $\sigma_1^2=\sigma_2^2=\sigma_3^2=\ldots=\sigma_k^2$

H_1: apenas duas σ_i^2's serem diferentes

De notar que, cada amostra *i* das *k* existentes, tem n_i elementos, e o desvio entre a observação e a média é dado por $z_{ij} = |x_{ij} - \bar{x_i}|$ A estatística de teste (W_0) será dada por:

$$W_0 = \left(\frac{N-k}{k-1}\right)\frac{\sum_{i=1}^{k} n_i(\bar{z_i}-\bar{\bar{z}})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_j-\bar{z_i})^2} \tag{3.17}$$

A hipótese nula é rejeitada se $W_0 > F_{(k-1),(N-k),(1-\alpha)}$ – valor da tabela pertencente à distribuição *F* (Almeida, Elian, & Nobre, 2008; Brown & Forsythe, 1974).

3.5.3.2. Teste de Brown & Forsythe

Este teste foi uma modificação ao teste de Levene apresentado anteriormente, pois estes autores notaram que em algumas distribuições o teste era liberal. A alteração consistiu em substituir o estimador clássico do parâmetro de localização. Ao invés da média da amostra passou a utilizar-se a mediana e, o desvio *z* passou a ser $z_{ij}{}^{(m)} = |x_{ij} - \bar{x_i}|$.

$$W_{50} = \left(\frac{N-k}{k-1}\right)\frac{\sum_{i=1}^{k} n_i(\bar{z_i}^{(m)}-\bar{\bar{z}}^{(m)})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_j{}^{(m)}-\bar{z_i}^{(m)})^2} \tag{3.18}$$

O resultado dado pela estatística de teste é comparado com o mesmo valor crítico usado no teste de Levene, já que a estatística de teste é a mesma, mudando apenas o parâmetro de localização (Almeida et al., 2008).

3.5.4. Violação dos pressupostos

Quando os pressupostos das técnicas de estatística paramétrica não são verificados, existem várias alternativas que se podem considerar para o passo seguinte: usar as técnicas paramétricas de igual forma sabendo que os pressupostos foram violados e, conhecendo os efeitos associados a essa violação; transformar (dimensionar) os dados para algo que vise aceitar os pressupostos; e/ou usar técnicas de estatística não paramétrica.

Quando se escolhe a primeira alternativa, mais concretamente usar a análise de variância (ANOVA) com os pressupostos violados, existem efeitos dependentes do pressuposto em questão que devem ser conhecidos, sendo que estes estão presentes na Tabela 3.2 (Hecke, 2012; Lomax & Hahs-Vaughn, 2013).

Tabela 3.2 - Pressupostos e respetivos efeitos da sua violação

Pressuposto	Efeito da violação do pressuposto
Independência das amostras	Aumento da probabilidade de erro tipo I e/ou tipo II na estatística *F*
	Afeta os desvios padrão e tem interferência nas médias
Normalidade da amostra	Distorções na variação dentro dos níveis (SS_w); Aumento da probabilidade de erro tipo I e/ou tipo II
	Pequenos efeitos com tamanhos de amostras iguais ou aproximadamente iguais; Por outro lado, os efeitos decrescem quando o tamanho da amostra aumenta
Homogeneidade da variância	Efeitos mínimos com tamanhos de amostras iguais ou aproximadamente iguais

Para a segunda opção, existem diversas transformações que podem ser feitas. Existe, no entanto, investigadores que não aprovam esta alternativa, devido à transformação dos dados para algo que é novo e não os dados recolhidos. Estas transformações são técnicas de dimensionamento (como já exposto anteriormente), ainda que aqui são de destacar as técnicas que obedecem a determinado critério externo e não interno (Berrueta et al., 2007; Howell, 2012).

As transformações mais conhecidas, ou mais usadas, são a transformação logarítmica, a transformação por meio da raiz quadrada e a transformação recíproca (ou inversa), existindo ainda a transformação do arcsen. De uma forma sintetizada pode-se usar a transformação da raiz quadrada para distribuições com assimetrias moderadamente positivas ou negativas usando-se uma constante positiva ou negativa para o efeito ou, para distribuições com assimetrias substancialmente mais desviadas a transformação logarítmica (De Muth, 2006; Howell, 2012). Existe ainda uma transformação bem conhecida, a transformação de Box-Cox que permite assegurar a homogeneidade e normalidade dos dados (Pereira & Requeijo, 2012).

Todavia, existe ainda a alternativa de usar a estatística não-paramétrica, onde se mantêm os dados recolhidos e não tem interesse a distribuição dos mesmos (Montgomery & Runger, 2003).

3.5.5. Estatística não-paramétrica

Na estatística não-paramétrica existem técnicas cuja finalidade é a mesma que na estatística paramétrica. São apresentadas de seguida técnicas para comparação de amostras, e para pesquisa de correlações entre variáveis.

3.5.5.1. Teste de Kruskal-Wallis e teste de Mann-Whitney

Os testes de Kruskal-Wallis e Mann-Whitney são métodos para comparar as médias, e tal como acontece na estatística paramétrica (com a ANOVA e o teste *t* de *Student*), para o caso de existirem mais de duas amostras ou apenas duas amostras, respetivamente. Então, e por ser mais abrangente, segue-se uma explicação mais detalhada do teste de Kruskal-Wallis, que é basicamente uma extensão do teste de Mann-Whitney.

O teste de *Kruskal-Wallis*, também chamado teste H, é baseado em "*ranks*" (números de ordem). Com *k* amostras para testar se são idênticas, ordenam-se as observações de todas as amostras e atribui-se o número de ordem. Posto isso, é calculada a estatística de teste dada por:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{{R_i}^2}{n_i} - 3(N+1) \tag{3.19}$$

Em caso de existirem observações iguais a estatística muda um pouco, pois é necessário um fator de correção. Os números de ordem passam a ser iguais para todos os valores repetidos, sendo a média entre os números de ordem das observações repetidas. A estatística de teste H aproxima-se a uma distribuição χ^2 (qui-quadrado) com k −1 graus de liberdade, pelo que através desta se tira o valor crítico (Hecke, 2012; Kruskal & Wallis, 1952).

3.5.5.2. Correlação de Spearman e correlação de Kendall

A correlação de Pearson é a correlação paramétrica que mede a linearidade entre duas variáveis. O coeficiente de correlação de Pearson (r) mede o grau da correlação linear entre duas variáveis quantitativas, é um índice adimensional que se situa entre -1 (correlação negativa perfeita) e 1 (correlação positiva perfeita), sendo 0 o valor que demonstra que não existe dependência linear entre variáveis (Filho & Júnior, 2009). Para distribuições normais a correlação de Pearson é a mais eficiente, no entanto para amostras que não sigam uma distribuição normal, existem outras correlações que fazem parte da estatística não paramétrica – correlação de Spearman e correlação de Kendall. Mais uma vez as alternativas passam por usar ranks (Croux & Dehon, 2010). No entanto, estas duas correlações não apresentam diferenças significativas e, como o uso da correlação de Spearman é a mais comum, mais simples de fazer e interpretar, considerou-se apenas a aplicação desta correlação em termos práticos (Hauke & Kossowski, 2011; Taylor, 1987).

A correlação de Spearman, tratando-se de uma medida de correlação não-paramétrica, é muito semelhante à correlação de Pearson, usando ranks ao invés dos valores observados. A razão pela qual a correlação de Spearman mede consistência, e não forma (linearidade como Pearson), é apenas porque quando duas variáveis estão consistentemente relacionadas, os seus ranks estão relacionados. Tal como no teste de Kruskal-Wallis, os valores são ordenados por variável e são lhe atribuídos um número de ordem ("rank"). Após esta atribuição é calculada a diferença *d* entre os *ranks* e calculado o coeficiente de correlação pela expressão apresentada de seguida.

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} \quad (3.20)$$

A hipótese nula formulada para este caso é que "não existe correlação entre os conjuntos de dados", sendo a hipótese alternativa o inverso. O coeficiente r_s também é vulgarmente denominado por ρ (letra grega rho) e, varia igualmente entre -1, que significa monotonamente decrescente, enquanto um conjunto de dados aumenta o outro diminui, e 1 que expressa uma monotonia crescente (Borradaile, 2003; Gravetter & Wallnau, 2010; Sheskin, 2003). Em suma, a correlação de Spearman não afirma sobre a forma como se se relacionam mas sim com a consistência, isto é se ambos crescem ou se têm comportamentos opostos, não interessando se é linear, quadrático ou ainda outro.

3.6. Técnicas de reconhecimento de padrões

Hoje em dia, facilmente se dispõe uma grande quantidade de dados para analisar devido à rapidez com que estes se conseguem extrair, seja em que área for. Uma vez na posse de um conjunto formado por dados de várias variáveis, a identificação de padrões poderá constituir uma forma eficaz de extrair informação desses mesmos dados. Para tal, existem dois tipos de técnicas de reconhecimento de padrões nos dados: as técnicas supervisionadas, em que estas usam informações dos membros de classes já conhecidas; e as não supervisionadas, que tentam detetar padrões sem quaisquer informações, só apenas com os dados (Berrueta et al., 2007; Tan, Steinbach, & Kumar, 2005).

Na quimiometria, estas técnicas são muito usadas devido à sua versatilidade. As técnicas supervisionadas também são geralmente designadas de técnicas de classificação, pois a regra de classificação usada pelas mesmas é previamente conhecida. É apresentado na Figura 3.2 um organograma que expõe estas técnicas bem como as subdivisões que existem dentro de cada uma (Gredilla, Fdez-Ortiz de Vallejuelo, Diego, Madariaga, & Amigo, 2013).

Pela análise da figura e pela literatura, nas técnicas não supervisionadas, destacam-se a análise de *clusters* (CA – *clusters analysis*), e o grupo dos métodos baseados em modelos de fator, que se resume basicamente à análise de componentes principais (PCA – *principal components analysis*). Dentro das técnicas de classificação existem três distinções possíveis que podem ser feitas (ao todo 6 classes), ou seja, as técnicas são classificadas sempre de 3 formas. As duas técnicas que mais se destacam na literatura são a análise discriminante linear (LDA – *linear discriminant analysis*) que é uma técnica paramétrica, discriminante e probabilística; e a k-NN (*k-Nearest Neighbors*) que é uma técnica não-paramétrica e determinística, contudo é igualmente discriminante (Berrueta et al., 2007; Gredilla et al., 2013). Com base nos dados em estudo as técnicas serão escolhidas apropriadamente.

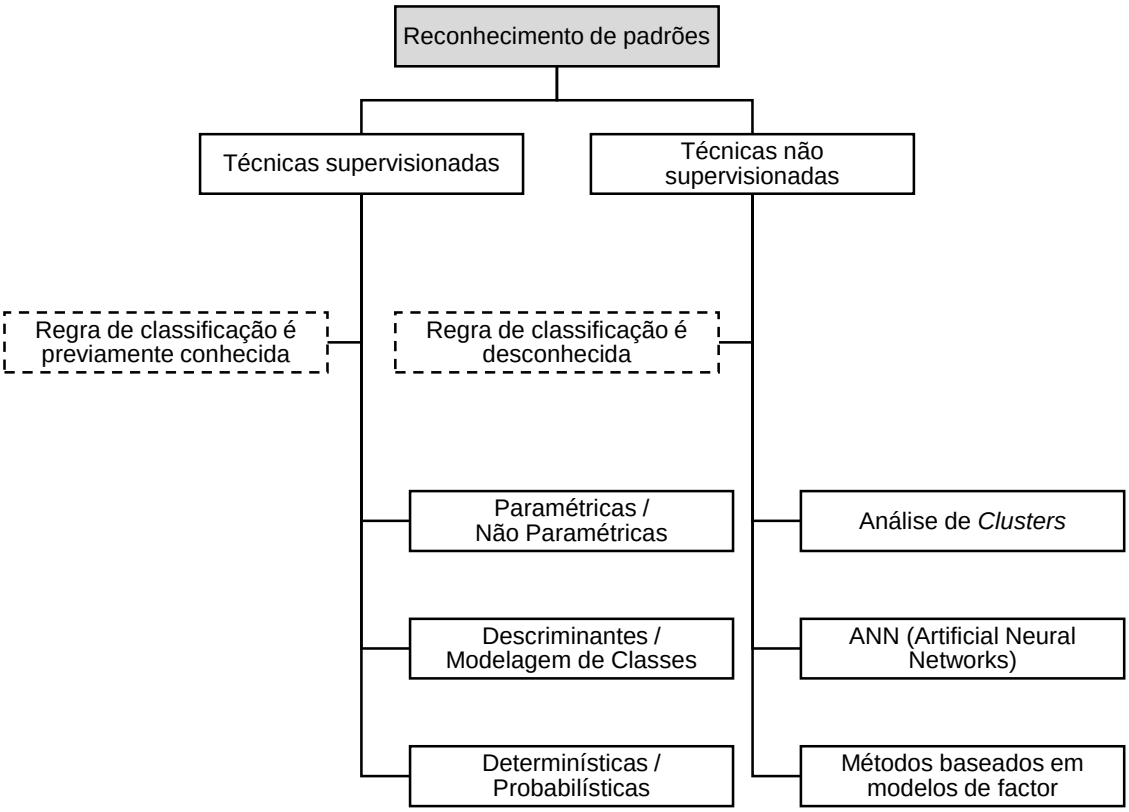

Figura 3.2 - Classificação das técnicas de reconhecimento de padrões

3.6.1. Análise de *Clusters* - HCA

A análise de *clusters* agrupa os dados, com algum tipo de semelhança de forma natural em grupos (*clusters*) que tenham significado do ponto de vista dessa semelhança. Basicamente, se os dados pertencem a um determinado *cluster*, de alguma forma, estarão relacionados. É importante perceber estas relações uma vez que estas podem revelar informações sobre os dados que até aqui não tinham sido entendidas. Em suma, a análise de *clusters* faz parte das técnicas de reconhecimento de padrões não supervisionadas, já que é autónomo na sua análise/agrupamento. Por vezes a análise de *clusters* é apenas um ponto de partida útil para outras finalidades, em que o único objetivo é aglomerar os dados de alguma forma. A análise de *clusters* tem revelado um papel importante nas diversas áreas, como a psicologia e outras ciências sociais, biologia, estatística, reconhecimento de padrões. Na análise de *clusters* muitas vezes não há especificação prévia sobre o número ou a natureza de *clusters* aos quais os objetos serão atribuídos. Obviamente, quanto melhor a similaridade entre os dados e maior a diferença entre os grupos, mais correto será o agrupamento (Berrueta et al., 2007; Ferreira & Hitchcock, 2009; Tan et al., 2005).

De entre todos os tipos de métodos de *clustering* (aglomeração), os dois mais relevantes são os métodos particionais (*partiotining methods*) e os métodos hierárquicos (*hierarchical methods*). Os particionais (ou não hierárquicos) que com base num número de grupos previamente especificado, agrupa os dados nesse mesmo número de grupos através de iterações que vão sendo feitas. Por outro lado, os métodos hierárquicos criam uma estrutura hierárquica que vai sendo construída. Esta

construção pode ser feita de duas formas: por aglomeração, em que se constrói o dendrograma (nome do gráfico que apresenta a estrutura hierárquica) de baixo para cima com a fusão de elementos/grupos; ou então, feita por divisão (divisivos), começando de cima para baixo, em que se tem um *cluster* e vai sendo divido e subdivido (Maimon & Rokach, 2006; Mooi & Sarstedt, 2011).

No estudo em causa, pelo desconhecimento do número de *clusters* (é exatamente o que se pretende descobrir bem como as suas características), os métodos que vão ser aplicados/estudados encontram-se dentro dos métodos hierárquicos por aglomeração (HCA – *hierarchical clusters analysis*). No entanto, dentro destes métodos é necessário escolher a medida de semelhança/similaridade e o algoritmo de formação dos *clusters* (Mooi & Sarstedt, 2011). De assinalar que são escolhidos métodos de aglomeração também pela limitação dos s*oftwares* escolhidos, pois estes não possuem métodos hierárquicos divisivos.

Existem três tipos de medidas de similaridade: as medidas de correlação, as medidas de distância e as medidas de associação. As medidas de associação são utilizadas para comparar objetos cujas características são medidas em escalas nominais ou ordinais. Nas medidas de correlação, a correlação entre colunas representa a correlação (ou semelhança) entre dois objetos, sendo que não são medidas muito utilizadas. Por fim, as medidas de distância medem a semelhança como a proximidade entre observações, podendo ser usadas várias distâncias para a medição (Hair et al., 2006). Isto é, para medir tal proximidade existem diferentes medidas que podem ser usadas. Uma maneira simples é a ligação entre 2 objetos através de uma linha, sendo designada de distância euclidiana e é a mais vulgarmente utilizada. Por outro lado, tem-se o quadrado da distância euclidiana, que oferece progressivamente mais peso aos objetos que estão mais distantes (Mooi & Sarstedt, 2011). Desta forma foi escolhido o quadrado da distância euclidiana, pois no dendrograma as diferenças são mais facilmente visualizáveis, isto, é os mais distantes apresentam maiores diferenças do que pela distância euclidiana. A expressão para o seu cálculo é apresentado de seguida.

$$d_{ij} = \sum_{k=1}^{n} (x_{ik} - x_{jk})^2 \tag{3.21}$$

Esta medida calcula a distância para todos os pares de objetos e apresenta todas estas distâncias em forma de matriz, denominada de matriz de proximidade ou matriz de distâncias. Basicamente é uma matriz quadrada com todos os objetos e as respetivas distâncias entre eles (seja qual for a medida escolhida). De realçar que a distância entre A e B é a mesma que entre B e A, e que entre os mesmos objetos é zero, isto para dizer que é uma matriz simétrica com a diagonal preenchida de zeros (Hair et al., 2006; Mooi & Sarstedt, 2011).

Após a escolha da medida de semelhança, segue-se a escolha do algoritmo de aglomeração (*clustering*), sendo que neste capítulo também a escolha é diversificada. Os 5 algoritmos mais populares entre todos são apresentados de seguida com uma breve explicação para cada um.

- Ligação simples (*single linkage*) ou critério do vizinho mais próximo:

Descobre os dois objetos para os quais a distância entre eles é a mínima e coloca-os no primeiro *cluster*, depois descobre a próxima distância mais curta e, ou esse novo objeto se junta ao *cluster* formado ou forma-se um novo *cluster* com dois objetos. Este processo continua até existir apenas um *cluster* que contenha todos os objetos. A distância entre dois *clusters* é a distância mais pequena de qualquer ponto de um *cluster* para qualquer ponto do outro.

- Ligação completa (*complete linkage*) ou critério do vizinho mais afastado:

Semelhante à ligação simples, mas tem por base a distância máxima. A distância máxima entre objetos do mesmo *cluster* representa a mais pequena esfera (mínimo diâmetro) que pode englobar todos os indivíduos. É chamada de ligação completa porque todos os objetos num dado *cluster* estão ligados uns aos outros por uma distância máxima (semelhança mínima).

- Ligação média (*average linkage*):

Aqui o procedimento é diferente da ligação simples e da ligação completa. A distância entre *clusters* é dada pela distância média de todos os objetos de um *cluster* com todos os objetos do outro.

- Método do centróide:

No método do centróide a distância entre *clusters* é a distância (Euclidiana quadrada ou simples) entre os seus centróides, sendo estes definidos pela média dos valores das observações nas variáveis em processo de aglomeração.

- Método de Ward:

A distância entre *clusters* no método de Ward corresponde à soma dos quadrados entre dois *clusters* para todo o conjunto de variáveis (Hair et al., 2006; Maimon & Rokach, 2006).

Por causa da sua versatilidade, o *clustering* tem emergido como um dos principais métodos de análise multivariada, sendo originalmente desenvolvida para classificação biológica (Saraçli, Dogan, & Dogan, 2013). Com isto, os métodos tem vindo a crescer devido a não ser algo exato. O *clustering* está no olho do investigador, e como tal, existem muitos métodos com formalizações matemáticas de diferentes autores (Estivill-Castro, 2003). Por exemplo, Ward definiu um método que deveria ser usado em amostras superiores a 100 (Ward, 1963), mas é muitas vezes usado em amostras pequenas pois o resultado revela-se igualmente consistente.

O resultado gráfico da análise de *clusters* é geralmente retratado na forma de árvores hierárquicas, designadas por dendrograma ou gráfico de árvore, que consistem em gráficos com ramificações que mostra o nível de distância onde houve uma combinação de objetos e por conseguinte formação de *clusters* (Hair et al., 2006; Mooi & Sarstedt, 2011; Saraçli et al., 2013). Pode ser visto um exemplo na Figura 3.3.

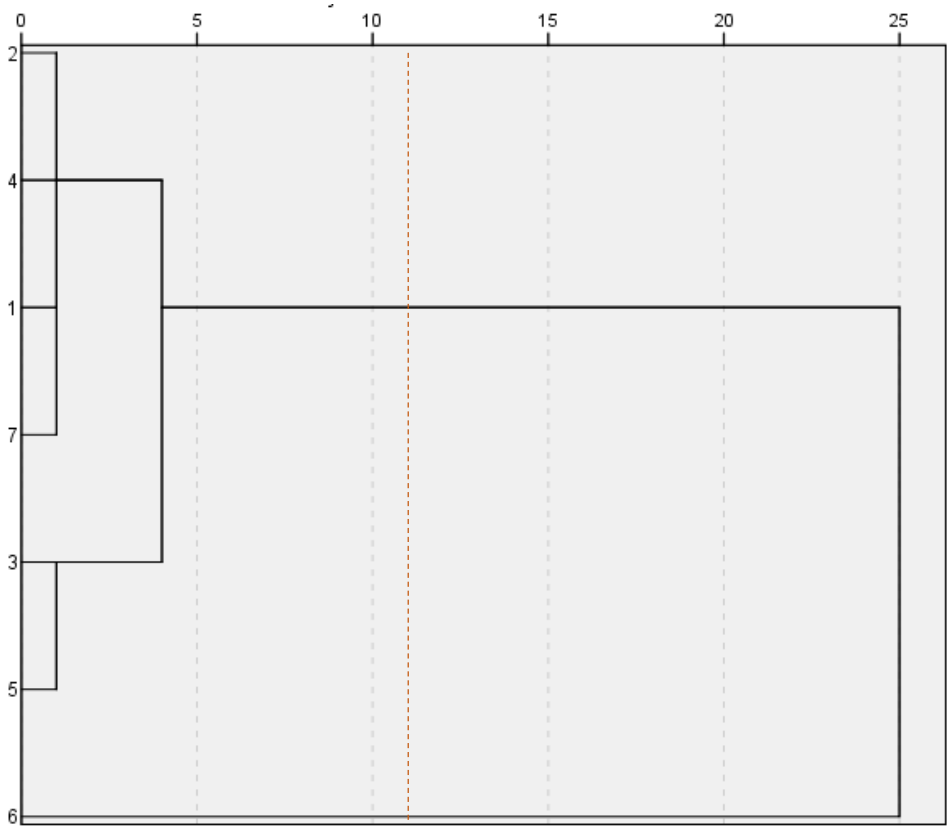

Figura 3.3 - Exemplo de um dendrograma

A análise de *clusters* é subjetiva, pelo que o principal problema se cinge na determinação do número de *clusters* ótimo. Pelo dendrograma, os *clusters* são dados pelo corte de ramos, a questão passa então a ser por onde cortar esses ramos. Os métodos matemáticos analisam o dendrograma e fazem o corte por um valor de altura constante, mas em dendrogramas complexos o desempenho destes é fraco. A maneira mais usual é cortar os ramos onde estes apresentarem uma altura maior, para além de ser uma *thumb rule* (regra de ouro) é um método bastante fiável, podendo ser feito manualmente (a olho) pelo investigador ou com métodos matemáticos mais avançados (Langfelder, Zhang, & Horvath, 2008). Existem avaliações que podem ser feitas internamente (sem recurso a conhecimento externo) ou externamente (comparando os resultados com agrupamentos conhecidos) para avaliar os diversos algoritmos (Saraçli et al., 2013), mas a escolha do método passa pela análise do dendrograma com base na facilidade de entender os *clusters* resultantes. Pela Figura 3.3, o resultado são 2 *clusters* em que um deles apenas contem um objeto (a tracejado está assinalado onde se cortam os ramos).

3.6.2. *k-NN* (*k-Nearest Neighbors*)

O *k-NN* (*k-Nearest Neighbors*) é um método não-paramétrico de identificação/classificação (Berrueta et al., 2007), sendo um dos melhores algoritmos de classificação conhecidos (Maimon & Rokach, 2006). Tal como acontece na análise de *clusters*, aqui a medida de similaridade tem várias opções mas a mais usada é a distância Euclidiana (Liu, 2011).

Dentro desde algoritmo existem três grupos: o grupo de treino, a classificação respetiva do grupo de treino, e o grupo de teste. Seja *d*, o elemento do grupo de teste (aquele que se pretende classificar), *D* o grupo de treino, e *k* o número de elementos "mais parecidos" (elementos próximos), então o algoritmo funciona da seguinte forma (Cios, Pedrycz, Swiniarski, & Kurgan, 2007; Liu, 2011):

1º - Calcula-se a distância entre d e todos os elementos de *D;*

2º - Escolhem-se os *k* exemplos de *D* que estão próximos de *d*, denominados por *P*;

3º - Classifica-se *d*, com a classe mais frequente (a classe maioritária) em *P*.

A versão mais simples deste algoritmo é quando o k é 1 (Cios et al., 2007), no entanto, normalmente não é suficiente para determinar a classe de *d*, devido à presença de ruido e *outliers* no grupo *D* (de treino). A importância de ter um *k* ótimo pode ser visto através do exemplo da Figura 3.4, em que a classificação varia para *k*'s diferentes (Liu, 2011).

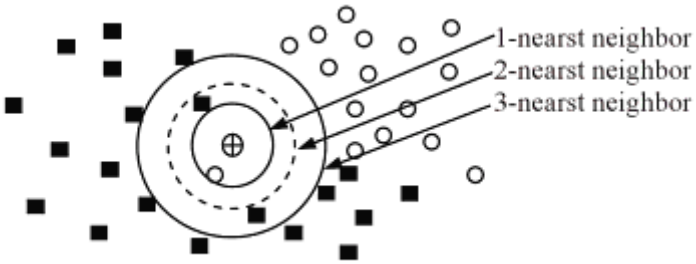

Figura 3.4 - Exemplo de uma classificação no modelo *k-NN* com diferentes *k*'s

É necessário um *k* superior a 1 para classificar mais fidedignamente *d*. Para a escolha de *k* é usualmente feita com recurso a um grupo de validação, ou a uma validação cruzada no grupo de treino. (Liu, 2011) O melhor *k* é determinado pela validação cruzada pelo método *leave-one-ou*t (Cios et al., 2007), que consiste em deixar apenas um elemento de teste e todos os restantes pertencem a *D*. Este método faz tantas classificações quantos elementos existirem, e no final dá o erro de classificação para o k definido (Witten, Frank, & Hall, 2011). Para escolher o melhor k é ir variando até que se tenha um erro de zero ou próximo deste.

3.7. Estatística multivariada aplicada a casos reais (ramo alimentar)

As técnicas de reconhecimento de padrões atrás apresentadas fazem parte das técnicas de quimiometria, e como tal, são regularmente aplicadas a produtos alimentícios. Leite, queijos, mel, bebidas alcoólicas, cereais, frutas, carnes e peixes são exemplos encontrados na bibliografia onde foram aplicadas estas técnicas. São vários os objetivos subjacentes à aplicação deste tipo de técnicas nesta área, como sejam: classificação por diversos critérios das amostras alimentares, caracterização de macro ou micro nutrientes, averiguação de adulterações no processamento,produção ou etiquetagem destas, entre outros (Berrueta et al., 2007).

Na universidade de Valência, ainda que com uma técnica estatística diferente (análise discrimante linear), estudaram-se as relações entre as origens geográficas do arroz e sua composição mineral (Gonzálvez et al., 2011).

No Brasil um estudo feito por (Diniz, Filho, Müller, Fernandes, & Palheta, 2013), onde foi aplicada uma análise de *clusters* às ervas medicinais e suas infusões provenientes da região amazónica. Este estudo usou igualmente a composição mineral dos chás.

Ainda aplicada às ervas medicinais, (Tokalıoğlu, 2012) na Turquia aplicou a análise de clusters e análise de componentes principais aos elementos-traço das mesmas. Obteve-se resultados semelhantes e corroborantes pelas duas técnicas usadas.

Na Ásia existe uma bebida chamada "*rice wine*", um género de *saké*, feita através da fermentação de arroz. Aplicada a estas bebidas, (Shen, Ying, Li, Zheng, & Zhuge, 2011) usaram a estatística multivariada para as classificar com base no tempo de envelhecimento das diferentes marcas. Este estudo teve como intenção criar uma estratégia efetiva para verificar a veracidade do tempo de envelhecimento apresentado nos rótulos, já que este tempo influencia o preço.

Para concluir, um exemplo de outro estudo que agrupou através da análise de *clusters* variedades cultivadas de arroz pelo poder de absorção e de ficar empapado do amido presente no arroz (Lee et al., 2012).

CAPÍTULO 4 – METODOLOGIA

Neste capítulo é apresentada e discutida a metodologia seguida ao longo da presente investigação. Na Figura 4.1 são apresentadas as quatro grandes etapas que permitiram cumprir os objectivos inicialmente delineados.

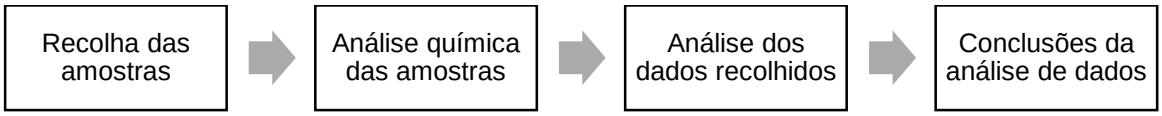

Figura 4.1 - Etapas da investigação subjacente à dissertação

As duas primeiras etapas foram desempenhadas por um laboratório – o Instituto Nacional de Saúde Doutor Ricardo Jorge (INSA) e as duas últimas pelo autor da dissertação, sendo todas elas explicadas nos subcapítulos posteriores.

4.1. Análises químicas

Este estudo consiste em analisar estatisticamente dados extraídos de um alimento, mais especificamente, as concentrações de aminoácidos proteicos e arsénio presentes no arroz, provenientes do laboratório. No total foram analisadas 39 amostras de arroz, que podem ser divididas entre branco ou integral, por variedade/tipo e por região, como se pode verificar na Figura 4.2, que é apresentada de seguida

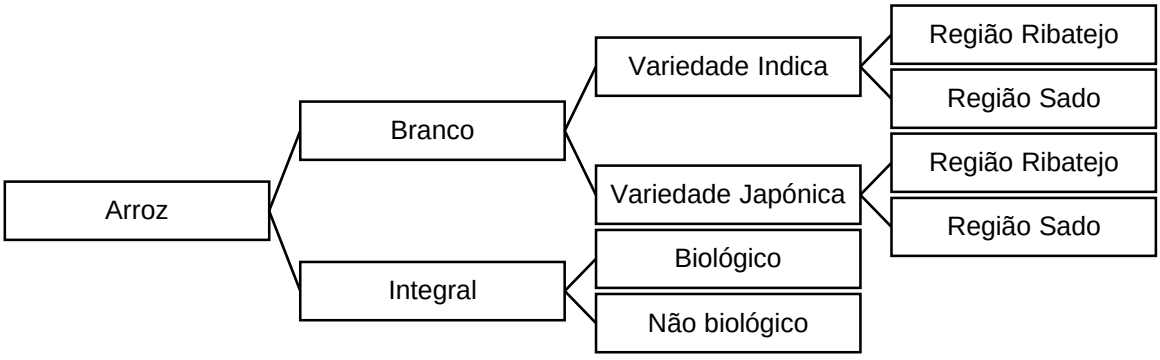

Figura 4.2 - Organigrama dos tipos de arroz presentes no estudo

.

De arroz branco foram recolhidas 22 amostras sendo as outras restantes 17 de arroz integral; já dentro do arroz integral existem 9 amostras de arroz integral não biológico e 8 de arroz integral de produção biológica. Por sua vez, 7 amostras de arroz branco são provenientes da região do Sado e as restantes 15 da região do Ribatejo; pela variedade tem-se que 12 amostras são arroz branco de variedade Indica (das quais 4 são da região do Sado e 8 do Ribatejo) e as remanescentes 10 são de variedade Japónica (3 da região do Sado e 7 do Ribatejo) . De destacar, que o arroz branco é oriundo diretamente de produtores nacionais, de anos de colheita entre 2009 e 2012, e o arroz integral é proveniente de pacotes comercializados em estabelecimentos comerciais, comprados nos anos de 2012 e 2013. Na Tabela 4.1 encontra-se a caracterização de cada amostra estudada.

Tabela 4.1 - Caracterização das amostras de arroz do estudo

Arroz Branco			
Amostra	Código	Tipo	Região
1.1	CD1	Índico	Ribatejo
1.2	CD4	Índico	Ribatejo
1.3	ML2	Índico	Ribatejo
1.4	ML4	Índico	Ribatejo
1.5	QF1	Índico	Ribatejo
1.6	QF2	Índico	Ribatejo
1.7	3	Índico	Ribatejo
1.8	HDL1	Índico	Sado
1.9	HDL2	Índico	Sado
1.10	HDL4	Índico	Sado
1.11	31	Índico	Sado
1.12	25	Índico	Ribatejo
2.1	2	Japónico	Ribatejo
2.2	5	Japónico	Ribatejo
2.3	6	Japónico	Ribatejo
2.4	8	Japónico	Ribatejo
2.5	9	Japónico	Ribatejo
2.6	1	Japónico	Sado
2.7	30	Japónico	Sado
2.8	32	Japónico	Sado
2.9	24	Japónico	Ribatejo
2.10	26	Japónico	Ribatejo

Arroz Integral	
Amostra	Tipo
3	Não Biológico
4	Biológico
5	Biológico
6	Biológico
7	Não Biológico
8	Não Biológico
9	Não Biológico
10	Não Biológico
11	Biológico
12	Não Biológico
13	Biológico
14	Biológico
15	Biológico
16	Biológico
19	Não Biológico
20	Não Biológico
21	Não Biológico

4.1.1. Instituto Nacional de Saúde Doutor Ricardo Jorge (INSA)

O Instituto Nacional de Saúde Doutor Ricardo Jorge (INSA) é uma instituição pública que está ao abrigo do Ministério da Saúde, fundada em 1899. É autónomo e, para além de laboratório nacional de referência, é também laboratório do Estado no sector da saúde e, observatório nacional de saúde. O INSA possui várias unidades onde opera: em Lisboa, onde se encontra a sede, no Porto e em Águas de Moura. O INSA está organizado, em termos técnico-científicos, em seis grandes departamentos,

onde se destaca o Departamento de Alimentação e Nutrição (DAN), de onde são provenientes os dados para o presente estudo (INSA, sem data-b).

O Departamento de Alimentação e Nutrição (DAN) desenvolve atividades nas áreas da segurança alimentar e nutrição: prevenindo doenças de origem alimentar e, melhorando o estado nutricional da população. Estas atividades são feitas através de investigação e desenvolvimento, vigilância, formação e consultoria. Este departamento é parceiro de organismos como a Organização Mundial de Saúde (OMS), a Organização para a Agricultura e Alimentação das Nações Unidas (FAO) e a Autoridade Europeia para a Segurança dos Alimentos (EFSA) (INSA, sem data-a).

4.1.2. Análises

Por se tratar de um laboratório de referência a nível nacional, os dados provenientes do mesmo têm que ser fidedignos, pelo que o laboratório tem um controlo interno para que não sejam extraídos resultados deturpados. No caso especifico do arroz, as amostras recolhidas são provenientes de vários pacotes ou vários lotes (dependendo se vem do pacote tradicional que se vende nas secções de mercearia ou, do produtor antes de ser embalado), são trituradas criando uma *pool*. Estas, são armazenadas em vácuo para que o arroz não perca as suas características até ao momento da sua análise, seja ele qual for.

- **Análise dos aminoácidos**

Os aminoácidos analisados, como referido anteriormente (Subcapítulo 2.2.2), no estudo em questão são dezassete devido ao tipo de hidrólise. São eles: a histidina (His), a serina (Ser), a arginina (Arg), a glicina (Gly), o ácido aspártico (Asp), o ácido glutâmico (Glu), a treonina (Thr), a alanina (Ala), a prolina (Pro), a cisteína (Cys), a lisina (Lys), a tirosina (Tyr), a metionina (Met), a valina (Val), a isoleucina (Ile), a leucina (Leu) e a fenilalanina (Phe).

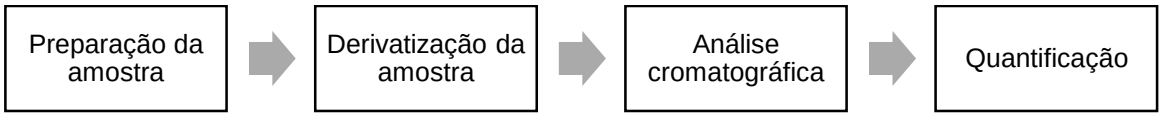

Figura 4.3 - Etapas da análise de aminoácidos

A análise aos aminoácidos (Figura 4.3) começa com a preparação da amostra, em que é feita a pesagem da amostra sucedida de uma hidrólise ácida em atmosfera inerte num micro-ondas *Milestone ETHOS 1 Series* equipado com sistema de digestão de vasos fechados. Após a preparação da amostra, é feita a derivatização que consiste na transformação dos aminoácidos em compostos fluorescentes altamente estáveis. De seguida, vem a fase principal, a análise cromatográfica feita através do equipamento *Acquity UPLC system* (Figura 4.4) proveniente da marca *Waters* equipado com um detetor de foto díodos (*PDA - photodiode array detector*). A análise cromatográfica é desempenhada numa coluna *BEH C18* em gradiente durante 10 minutos, ou seja, a fase móvel vai variando ao longo do tempo para que seja possível quantificar todos os aminoácidos, ácidos e básicos. Por fim, é feita a quantificação dos aminoácidos presentes no cromatograma com recurso a

curvas de calibração. As leituras são feitas pelo menos duas vezes (repetição), e cada amostra é replicada também no mínimo 2 vezes.

Figura 4.4 - Equipamento de análise cromatográfica (*Acquity UPLC system – Waters)*

- **Análise do arsénio**

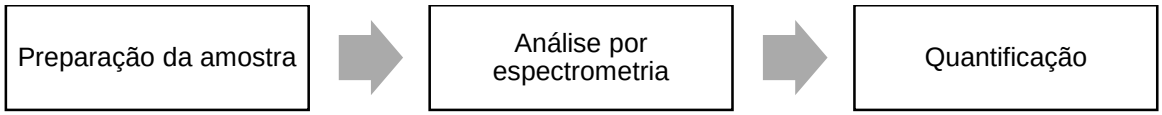

Figura 4.5 - Etapas da análise do arsénio

Por sua vez, na análise do arsénio (Figura 4.5), pela espectrometria de massa, alguns dos procedimentos são semelhantes aos da análise de aminoácidos. A amostra é pesada e preparada de igual forma para a hidrólise ácida num micro-ondas *Milestone ETHOS 1 Series,* equipado com sistema de digestão de vasos fechados. Após a hidrólise as amostras são analisadas com recurso a um espectrómetro *ICP-MS XSERIES II* da *Thermounicam* e para concluir é feita a quantificação que, mais uma vez, é feita recorrendo a extrapolação das curvas de calibração. Todo este procedimento é feito em salas limpas onde apenas se pode entrar com o equipamento adequado, sendo o próprio processo de filtração do ar feito de forma diferente das outras divisões do laboratório. Estes cuidados extremos devem-se ao facto da medição dos elementos-traço corresponderem a concentrações extremamente baixas, evitando dessa forma qualquer tipo de contaminação que possa surgir, e por sua vez adulterar os resultados da medição.

4.1.3. Controlo interno

No laboratório todas as medições são controladas, isto é, quer sejam os elementos-traço ou os aminoácidos, cada tubo é medido pelo menos 2 vezes e é calculada a sua média, sendo que por amostra de arroz existem pelo menos 2 tubos. Se as médias entre os tubos tiverem um coeficiente de variação (é dado pelo quociente entre o desvio padrão e a média; neste caso é a média das médias de leituras) superior a 10%, é repetida nova análise com uma nova amostra da *pool*, pois alguma das amostras estava díspar.

No que diz respeito ao controlo interno do laboratório, para além do controlo na medição, que é feito em cada amostra, existem quatro parâmetros fundamentais: as curvas de calibração, a adição de padrão interno, os ensaios interlaboratoriais e o uso de materiais de referência certificados. No INSA, as curvas de calibração são sempre construídas com um coeficiente de correlação (r) superior a 0,9950 (atingindo quase a linearidade), com o objetivo da leitura da amostra ser o mais correta possível. A adição de padrão interno no início da preparação da amostra, serve de verificação, aquando da leitura pelo equipamento e das perdas que este teve durante a hidrólise. Obviamente a escolha do padrão interno é feita para que não haja perda alguma e, se esta existir de alguma forma é sinal que algo aconteceu na hidrólise, não só com o padrão interno mas também com a amostra, pelo que as leituras serão rejeitadas para este caso (quando as perdas são superiores a 20%). Os ensaios interlaboratoriais servem para apurar se os equipamentos do laboratório estão a medir corretamente. Estes são feitos através da medição de amostras cegas, isto é, de amostras cujos valores reais são desconhecidos, enviados por outros laboratórios de referência. O INSA faz parte do FAPAS (*Food Analysis Performance Assessment Scheme)* que tem como objetivo reconhecer os laboratórios da legitimidade das suas análises através de *scores*. Por fim, no INSA são usados materiais de referência certificados, isto é semelhante aos ensaios interlaboratoriais, no entanto aqui são usadas amostras cujos valores reais são conhecidos, que serve igualmente para verificar a autenticidade das análises mas apenas internamente.

4.2. Análise dos dados

Antes de entrar na metodologia de análise dos dados propriamente dita, há que expor os *softwares* usados para tal. O *Microsoft Office Excel 2013* (ou semelhante) é a ferramenta informática obrigatória em qualquer estudo estatístico, quer para pré-tratamento dos dados, quer se necessário para cálculo de estatísticas descritivas. Para além deste *software* básico, outras ferramentas estatísticas mais importantes foram escolhidas, nomeadamente o *IBM SPSS Statistics 22* e o *Statsoft Statistica Software10*. Normalmente, no estudo em questão, ambos são usados servindo sobretudo de validação um do outro e, também, com o intuito de anular o erro humano associado ao uso destas ferramentas. Por fim, ainda é usado o *Mathworks Matlab R2013a*, para desempenhar funções que os outros *softwares* referidos anteriormente são incapazes. Em suma, e com as designações que serão usadas daqui em diante, usou-se o *Excel* para funções mais simples, o *SPSS* e o *Statistica* para utilidades mais estatísticas e, por fim, o *Matlab* para aplicações mais complexas.

Na Figura 4.6, estão esquematizados os principais passos que irão permitir atingir os 4 grandes objetivos do presente estudo estatístico, nomeadamente: comparação de médias entre populações caracterizadas previamente, correlação entre os diversos aminoácidos e o arsénio, análise de *clusters* aos aminoácidos e por fim, a criação de um modelo *k-NN* e avaliação da capacidade do mesmo.

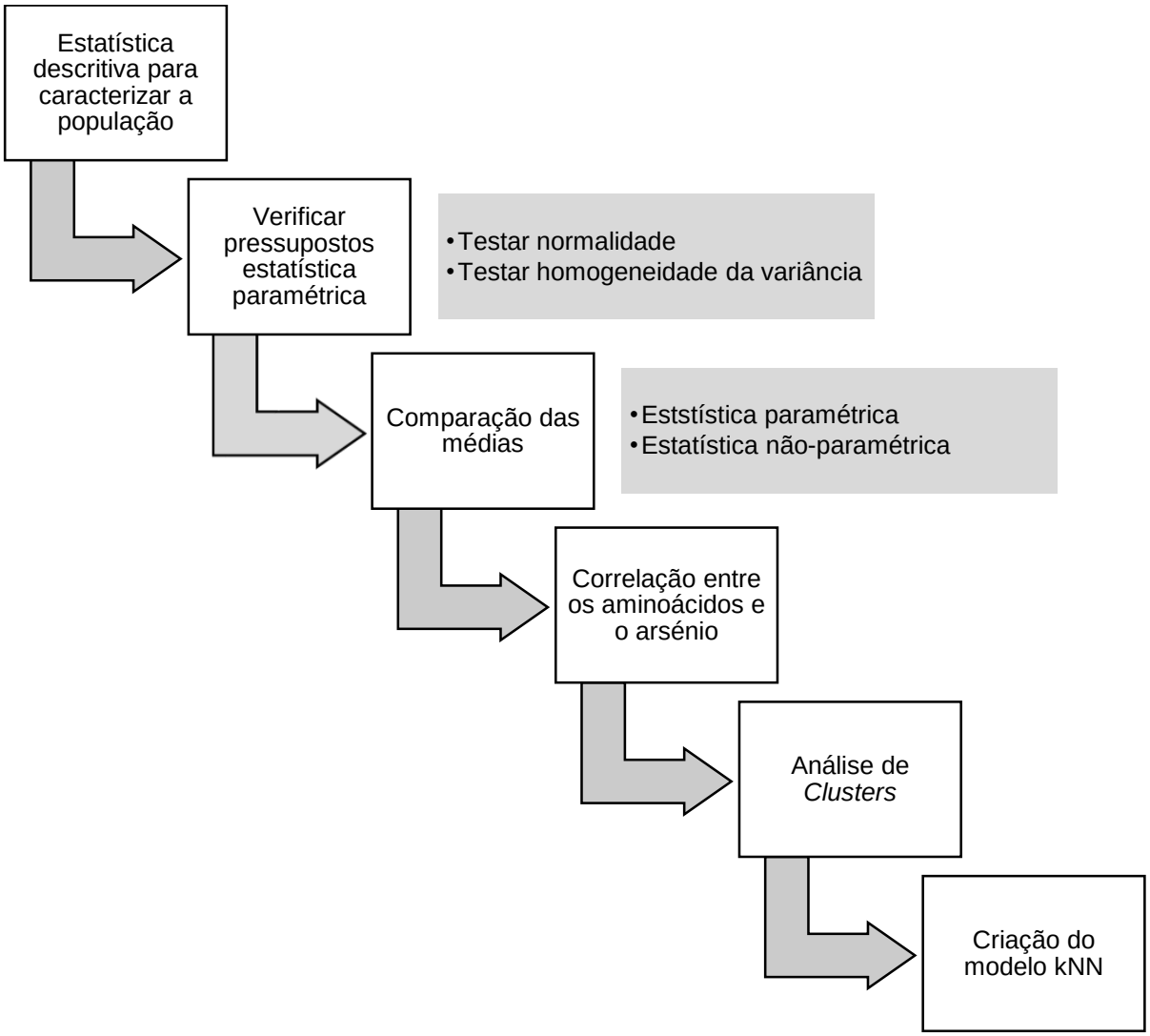

Figura 4.6 - Etapas da análise estatística dos dados, seguidas ao longo do presente estudo

O primeiro passo do estudo consiste em caracterizar a população através da estatística descritiva, isto é, com as diferentes médias das leituras de cada "tubo" foram calculadas a média e o desvio padrão de cada amostra (cada amostra foi analisada em 2 ou mais tubos), para os 17 aminoácidos recolhidos. Na caracterização, e daí em diante, usaram-se sempre "conjuntos" no estudo, ou seja, foram analisados os dados do arroz branco e os do arroz integral como se de duas populações se tratassem; para além deste caso, o mesmo aconteceu para o arroz integral biológico e não biológico, e, ainda, para o arroz branco de variedade Indica e de variedade Japónica. Para este último caso tem de se verificar se existem diferenças entre região para saber se os dados se podem agrupar (se não existirem diferenças) ou não (caso hajam diferenças significativas). Em síntese, são comparados e analisados sempre estes 3 casos. É importante frisar, que para fazer estas caracterizações das diferentes populações, foi usado o *Excel,* e para confirmação a função "*Descriptive statistics*" que pode ser encontrada no menu "*Basic Statistics*" do *Statistica* (podendo ser igualmente usado o *SPSS*).

O passo seguinte foi verificar os pressupostos da ANOVA (estatística paramétrica): a normalidade e a homogeneidade da variância. Para esta verificação e com base em referências bibliográficas, foram escolhidos os melhores testes para os dados em questão. Para a normalidade usaram-se os testes de Shapiro-Wilk, de Anderson-Darling e ainda, o teste de Kolmogorov-Smirnov, e para a homogeneidade da variância usaram-se os testes de Levene, o de Brown-Forsythe e o teste *F* de Fisher que vem junto com o teste *t* de *Student*. Quanto ao uso das ferramentas para tais verificações, o teste de Shapiro-Wilk e Kolmogorov-Smirnov foram executados no *SPSS* e *Statistica*. No *SPSS* é usada a opção "Gráficos de normalidade com testes" que se encontra dentro da função "Explorar" do menu e submenu "Analisar" "Estatísticas descritivas", respetivamente. Já no *Statistica*, os testes encontram-se na secção "*Normality*" que está dentro da lista "*Descriptive statistics*" que, por sua vez, se encontra no menu "*Basic Statistics*". O teste de Anderson-Darling foi realizado no *Matlab* através da função "*adtest*", pois os restantes não continham este teste. Este teste foi realizado de uma forma muito rápida através de um *script* criado antecipadamente. O *script* é apresentado parcialmente de seguida em forma de caixa de texto na Figura 4.7.

```
% Script para o teste de Normalidade de Anderson-Darling
%[h,p,adstat,cv] = adtest(x); função para este teste no matlab

clc;
clear all;
close all;
load ArrozIntegral.txt;
load ArrozBranco.txt;
load ArrozIntegralBiologico.txt;
load ArrozIntegralNaoBiologico.txt;
load ArrozBrancoIndico.txt;
load ArrozBrancoJaponico.txt;
load ArrozBrancoRibatejo.txt;
load ArrozBrancoSado.txt;

%p_values
%ad_stats
[x_integral,y_integral]=size(ArrozIntegral)
[x_branco,y_branco]=size(ArrozBranco)
[x_biologico,y_biologico]=size(ArrozIntegralBiologico)
[x_naobiologico,y_naobiologico]=size(ArrozIntegralNaoBiologico)
[x_indico,y_indico]=size(ArrozBrancoIndico)
[x_japonico,y_japonico]=size(ArrozBrancoJaponico)
[x_ribatejo,y_ribatejo]=size(ArrozBrancoRibatejo)
[x_sado,y_sado]=size(ArrozBrancoSado)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                          ArrozIntegral

for i=1:1:y_integral
    [h,p,adstat,cv] = adtest(ArrozIntegral(:,i));
    p_values_ArrozIntegral(i)=p
    ad_stats_ArrozIntegral(i)=adstat
end
ArrozIntegral(x_integral+1,:)=p_values_ArrozIntegral;
ArrozIntegral(x_integral+2,:)=ad_stats_ArrozIntegral;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Figura 4.7 - Script criado para execução do teste de normalidade de *Anderson-Darling*

Importa, no entanto, salientar que o código apresentado apenas permitir extrair os valores-p e a estatística de teste para o arroz integral, sendo análoga para os restantes tipos de arroz. Por último, os testes de Levene e Brown-Forsythe foram desempenhados no *Statistica*, encontrando-se estes, dentro da lista de tabelas exequíveis pelo *software,* no submenu "*Breakdown & one-way ANOVA"* que está dentro das "*Basic Statistics"*.

Executados os testes, e verificados (ou não) os pressupostos para todas as variáveis em todos os casos do estudo, passa-se finalmente à comparação de médias, um dos objetivos do estudo estatístico que ajuda na caracterização das populações. Para esta, usaram-se quer a estatística paramétrica (ANOVA e *t* de *Student*) que depende da validação dos pressupostos, quer estatística não-paramétrica (Kruskal-Wallis) que não depende. A *ANOVA* foi executada no *SPSS* (Analisar > Comparação de médias > Médias) e os resultados foram confirmados pelo *Statistica* (*Basic Statistics > Breakdown & one-way ANOVA*). Já o *t* de *Student* e o Kruskal-Wallis foram apenas desempenhado no *Statistica*, devido à forma mais fácil e percetível com que os resultados são apresentados.

Outro dos objetivos da dissertação (o terceiro) é a pesquisa de correlações que possam existir entre os diversos aminoácidos e o arsénio, pelos motivos que já foram referidos anteriormente. Para tal, foi usada a correlação de Spearman, por ser um tipo de estatística não paramétrica (para poder abarcar todos os dados presentes no estudo). Esta técnica foi feita nos dois *softwares*, mas o *Statistica* só assinala as correlações cuja hipótese nula se rejeita e não dá a conhecer o nível real de significância, pelo que foram eleitos os resultados dados pelo *SPSS*, provenientes da função "Correlacionar bivariável" que se encontra dentro do menu "Analisar".

Segue-se então a análise de *clusters*. Esta análise foi repartida em dois: pelas variáveis (aminoácidos) e pelos casos (amostras) e, em cada um deles, foram usados três algoritmos de *clustering*: a ligação média entre grupos, o método do centróide e o método de Ward. Para além disso, foram experimentadas hipóteses que se foram formando com o avançar do estudo. Toda a análise de *clusters* foi feita no *SPSS*, onde foram extraídos os dendrogramas resultantes.

Por fim, o modelo *k-NN*, criado e avaliado para o que os dados permitirem. Toda esta análise foi desempenhada no *software Matlab.* O modelo está divido em classificar os dados e em avaliar a potência do mesmo. Para a classificação foi usada a função *knnclassify* em que se define os dados que são de treino (os restantes são para teste),e as respetivas categorias do treino. Quando se corre o modelo, este dá as categorias que mais se adequam aos dados do teste. De notar, que esta parte não é apresentada na análise de resultados, pois não faz sentido mostrar as categorias dadas pelo modelo, quando todas elas são previamente conhecidas. O que faz sentido é avaliar a potência do modelo e a percentagem de erros que o modelo comete para os dados disponíveis. Para esta avaliação é usada a função *ClassificationKNN.fit,* com o método de validação cruzada *Leaveout* e com o parâmetro *NumNeighbors* a variar durante a pesquisa. O *script* construído para a avaliação do modelo pode ser visto na Figura 4.8.

```
% Script para a percentagem de erros do modelo

clc;
clear all;
close all;

dados = load ('Tudo.txt');
label = load('label.txt');

cvmd1 = ClassificationKNN.fit(dados,label,'CrossVal','on',
'Leaveout','on','NumNeighbors',5)

loss = kfoldLoss(cvmd1);

loss*100
```

Figura 4.8 - Script criado para avaliação do modelo *k-NN* criado

CAPÍTULO 5 – RESULTADOS E DISCUSSÃO

5.1. Estatística Descritiva

O primeiro passo deste estudo consiste em caracterizar as amostras através da estatística descritiva, nomeadamente através da média e do desvio padrão. Existe a distinção, nos subcapítulos seguintes, entre os aminoácidos e o arsénio, com as diversas separações por casos (com base nas características do arroz). De salientar, que nas tabelas daqui em diante apresentadas, os aminoácidos essenciais encontram-se diferenciados, estando a negrito e sublinhados com o intuito de se tornar mais compreensível. Os dados completos, usados para a construção das tabelas apresentadas neste ponto, podem ser consultados na Tabela I.1 e Tabela I.2, apresentado no Anexo I do presente documento.

5.1.1. Aminoácidos

A primeira caracterização efetuada ao perfil de aminoácidos do arroz considerou a totalidade das amostras analisadas como um todo, não fazendo distinção das suas diferentes características.

Tabela 5.1 - Média e desvio padrão para toda a população

Aminoácido	Abreviatura	Arroz (n = 39) (mg/100g)
Histidina	**His**	279,08 ± 107,85
Serina	Ser	370,58 ± 61,66
Arginina	Arg	639,57 ± 108,56
Glicina	Gly	341,02 ± 57,11
Ácido Aspártico	Asp	607,11 ± 86,66
Ácido Glutâmico	Glu	1250,42 ± 178,34
Treonina	**Thr**	234,12 ± 50,44
Alanina	Ala	363,57 ± 46,90
Prolina	Pro	317,81 ± 55,43
Cisteína	Cys	117,70 ± 93,2
Lisina	**Lys**	110,61 ± 38,88
Tirosina	Tyr	442,73 ± 143,48
Metionina	**Met**	224,29 ± 89,78
Valina	**Val**	332,42 ± 49,86
Isoleucina	**Ile**	243,9 ± 34,59
Leucina	**Leu**	530,39 ± 66,7
Fenilalanina	**Phe**	481,81 ± 110,42

Os valores da média e desvio padrão, expressos em mg/100g, encontram-se sumarizados na Tabela 5.1.

A análise da Tabela 5.1 permite verificar que para alguns aminoácidos, o desvio padrão é bastante alto quando comparado com a respetiva média. Os casos mais evidentes dizem respeito à cisteína, com um coeficiente de variação (CV) de cerca de 80%, e à histidina e metionina com um CV de cerca de 40% em ambas. Isto deve-se ao facto de conter todo o arroz, onde existem amostras com diferentes características que fazem aumentar os desvios.

- **Comparativo entre o Arroz Integral (n=17) e o Arroz Branco (n=22)**

De seguida, procedeu-se igualmente ao cálculo das médias e desvios padrão, com a devida separação dos dados entre arroz branco e arroz integral. Esses valores são apresentados de seguida, na Tabela 5.2.

Tabela 5.2 - Comparativo das concentrações (média e desvio padrão) dos aminoácidos no arroz integral e do arroz branco

Aminoácido	Abreviatura	Arroz Integral (n = 17) (mg/100g)	Arroz Branco (n = 22) (mg/100g)
Histidina	**His**	388,70 ± 38,19	194,37 ± 51,88
Serina	Ser	419,16 ± 28,78	333,03 ± 53,49
Arginina	Arg	736,54 ± 42,10	564,63 ± 80,52
Glicina	Gly	393,92 ± 25,47	300,14 ± 37,32
Ácido Aspártico	Asp	642,38 ± 66,76	579,86 ± 91,70
Ácido Glutâmico	Glu	1232,86 ± 121,52	1263,99 ± 214,14
Treonina	**Thr**	279,56 ± 28,22	199,01 ± 32,14
Alanina	Ala	389,67 ± 29,85	343,40 ± 48,20
Prolina	Pro	363,95 ± 30,51	282,15 ± 42,36
Cisteína	Cys	215,28 ± 51,62	42,31 ± 4,89
Lisina	**Lys**	87,24 ± 22,26	128,67 ± 39,68
Tirosina	Tyr	586,24 ± 49,29	331,84 ± 76,46
Metionina	**Met**	310,46 ± 44,66	157,70 ± 49,06
Valina	**Val**	359,73 ± 38,58	311,31 ± 47,90
Isoleucina	**Ile**	258,47 ± 31,21	232,65 ± 33,43
Leucina	**Leu**	565,07 ± 44,61	503,59 ± 69,34
Fenilalanina	**Phe**	589,22 ± 44,82	398,80 ± 63,05

Nesta divisão dos dados, os desvios padrão baixaram significativamente, não excedendo o coeficiente de variação de 30%. Relativamente à cisteína, uma observação direta sobre os valores das médias permite perceber o coeficiente de variação de 80% anteriormente evidenciado. Facilmente e, antes de executar qualquer teste estatístico, se percebe da existência de médias que irão ser significativamente diferentes e, que quase todos os aminoácidos essenciais (à exceção da lisina) estão em maior concentração no arroz integral do que no arroz branco. Estes valores de certa forma eram esperados pelas diferenças nutricionais que o arroz branco e integral apresentam, e o estudo feito por Walter, Marchezan e Avila comprova tais resultados (Walter et al., 2008).

- **Comparativo entre o Arroz Integral Biológico (n=9) e Não Biológico (n=8)**

Depois desta primeira distinção e, focando no arroz com maior concentração na maioria dos aminoácidos – o arroz integral, dividiu-se igualmente pelas características do mesmo. O arroz integral analisado era proveniente de dois tipos de agricultura diferente: a agricultura biológica e a agricultura tradicional, ou não biológica. Esta diferenciação pode ver vista na Tabela 5.3, bem como as respetivas médias e desvios padrão para todos os aminoácidos.

Tabela 5.3 - Comparativo das concentrações (média e desvio padrão) dos aminoácidos no arroz integral biológico e não biológico

Aminoácido	Abreviatura	Biológico (n = 8) (mg/100g)	Não Biológico (n = 9) (mg/100g)
Histidina	**His**	382,37 ± 45,46	394,33 ± 32,14
Serina	Ser	415,05 ± 31,43	422,82 ± 27,58
Arginina	Arg	734,24 ± 50,86	738,59 ± 35,64
Glicina	Gly	387,24 ± 26,29	399,85 ± 25,66
Ácido Aspártico	Asp	655,34 ± 77,57	630,86 ± 57,72
Ácido Glutâmico	Glu	1229,47 ± 138,15	1235,87 ± 113,20
Treonina	**Thr**	271,70 ± 30,98	286,55 ± 25,24
Alanina	Ala	381,25 ± 29,93	397,15 ± 29,40
Prolina	Pro	355,06 ± 31,11	371,85 ± 29,42
Cisteína	Cys	203,64 ± 56,73	225,62 ± 47,51
Lisina	**Lys**	85,26 ± 29,36	89,00 ± 15,15
Tirosina	Tyr	578,55 ± 61,11	593,08 ± 38,48
Metionina	**Met**	286,98 ± 37,89	331,34 ± 41,12
Valina	**Val**	361,26 ± 50,47	358,37 ± 27,27
Isoleucina	**Ile**	257,73 ± 40,33	259,13 ± 22,91
Leucina	**Leu**	557,25 ± 51,45	572,01 ± 39,36
Fenilalanina	**Phe**	583,82 ± 53,39	594,03 ± 38,31

As diferenças não são tão facilmente visualizáveis, já que os valores, quer para o arroz integral de origem biológica quer para o arroz cultivado tradicionalmente, são muito semelhantes. Num ponto posterior do presente documento serão testadas as significâncias das diferenças entre estas médias.

- **Comparativo entre o Arroz Branco da região do Ribatejo (n=12) e do Sado (n=10)**

Tomando por base o outro tipo de arroz, o arroz branco, a separação pode ser feita de duas formas, como já vem sendo referido ao longo do documento, ficando a primeira pela região de cultivo do arroz. O arroz branco, proveniente diretamente dos produtores para o laboratório, veio das áreas da região do Sado e do Ribatejo, contudo dentro de cada região existem amostras das duas variedades pelo que se têm de separar.

Na Tabela 5.4, são exibidas as médias e desvios padrão dos aminoácidos para ambas as regiões do arroz de variedade indica, onde se vê que todos os valores médios dos aminoácidos presentes no arroz branco proveniente da região Sado são superiores aos da região do Ribatejo. No entanto, sem a aplicação de um teste estatístico para comparação das médias, ainda nada se pode afirmar

relativamente à existência de diferenças significativas. De notar que as dimensões das amostras (da região do Ribatejo e do Sado) são bastante díspares.

Tabela 5.4 - Comparativo das concentrações (média e desvio padrão) dos aminoácidos no arroz branco da região do Ribatejo e do Sado – variedade indica

Aminoácido	Abreviatura	Ribatejo (n = 8) (mg/100g)	Sado (n = 4) (mg/100g)
Histidina	**His**	186,56 ± 50,44	204,97 ± 32,27
Serina	Ser	312,95 ± 46,29	388,63 ± 65,84
Arginina	Arg	540,88 ± 73,43	647,09 ± 85,03
Glicina	Gly	292,77 ± 36,33	335,69 ± 29,1
Ácido Aspártico	Asp	536,51 ± 87,79	676,87 ± 126,58
Ácido Glutâmico	Glu	1170,46 ± 202,45	1523,43 ± 274,65
Treonina	**Thr**	191,39 ± 33,35	219,27 ± 25,33
Alanina	Ala	325,7 ± 48,54	397,97 ± 59,09
Prolina	Pro	272,12 ± 41,87	317,48 ± 42,1
Cisteína	Cys	40,58 ± 4,52	45,15 ± 5,85
Lisina	**Lys**	111,94 ± 33,34	169,09 ± 53,97
Tirosina	Tyr	327,15 ± 69,2	351,87 ± 46,86
Metionina	**Met**	157,78 ± 47,19	167,54 ± 12,07
Valina	**Val**	298,95 ± 51,05	361,8 ± 54,54
Isoleucina	**Ile**	227,23 ± 35,92	265,55 ± 35,95
Leucina	**Leu**	484,99 ± 76,56	559,12 ± 88,12
Fenilalanina	**Phe**	389,66 ± 54,09	443,4 ± 55,44

Por sua vez, na Tabela 5.5 são apresentadas as médias e desvio padrão dos aminoácidos para ambas as regiões do arroz de variedade japónica. Nesta variedade já não existe uma tendência para a região como acontecia na variedade indica. Para possíveis diferenças significativas terá de se recorrer a um teste estatístico, sendo que as dimensões são bastante diferentes entre si.

Tabela 5.5 - Comparativo das concentrações (média e desvio padrão) dos aminoácidos no arroz branco da região do Ribatejo e do Sado - variedade japónica

Aminoácido	Abreviatura	Ribatejo (n = 7) (mg/100g)	Sado (n = 3) (mg/100g)
Histidina	**His**	201,97 ± 74,87	183,34 ± 9,17
Serina	Ser	322,79 ± 49,28	336,33 ± 19,56
Arginina	Arg	536,34 ± 76,34	584,09 ± 33,12
Glicina	Gly	284,36 ± 39,77	309,22 ± 12,89
Ácido Aspártico	Asp	576,89 ± 51,49	573,01 ± 43,94
Ácido Glutâmico	Glu	1203,11 ± 89,12	1309,5 ± 101,25
Treonina	**Thr**	197,95 ± 39,18	194,77 ± 14,2
Alanina	Ala	332,71 ± 28,51	342,79 ± 23,9
Prolina	Pro	274,36 ± 46,39	279,97 ± 17,74
Cisteína	Cys	42,38 ± 5,58	42,96 ± 2,11
Lisina	**Lys**	133,51 ± 28,4	108,11 ± 22,36
Tirosina	Tyr	324,16 ± 113,38	335,6 ± 33,19
Metionina	**Met**	158,35 ± 74,11	142,82 ± 13,01
Valina	**Val**	296,68 ± 29,87	311,1 ± 30,25
Isoleucina	**Ile**	216,83 ± 21,29	240,15 ± 21,28
Leucina	**Leu**	488,21 ± 50,6	515,05 ± 37,19
Fenilalanina	**Phe**	378,16 ± 80,99	411,9 ± 27,69

- **Comparativo entre o arroz branco de variedade indica (n=12) e japónica (n=10)**

Segue-se então o comparativo entre variedades de arroz, contudo torna-se necessário ir verificar previamente se as médias por região de cultivo apresentam diferenças significativas. Como as dimensões das amostras eram bastantes díspares recorreu-se directamente à estatística não-paramétrica (sem sequer verificar os pressupostos da estatística paramétrica)

Tabela 5.6 - Testes de comparação de médias das concentrações do arroz branco, por região

AA	Indico	Japónico
	Sado e Ribatejo	Sado e Ribatejo
	Kruskal Wallis (Valor-p)	Kruskal Wallis (Valor-p)
His	0,214	0,383
Ser	0,109	0,667
Arg	0,073	0,267
Gly	0,109	0,383
Asp	0,109	1,000
Glu	**0,048**	0,267
Thr	0,283	0,833
Ala	**0,048**	0,833
Pro	0,109	0,667
Cys	0,214	1,000
Lys	0,109	0,183
Tyr	0,283	0,383
Met	0,368	0,667
Val	**0,048**	0,517
Ile	0,109	0,183
Leu	0,109	0,517
Phe	0,109	0,383

Pela Tabela 5.6 retira-se que para a variedade japónica as regiões não apresentam qualquer diferença significativa, ou seja, pode-se usar o arroz branco de variedade japónica não interessando qual a região já que entre elas não existem diferenças. No arroz branco de variedade indica, o ácido glutâmico, a alanina e a valina apresentam valores-p de 0,048. Como estes valores-p são tão próximo do nível de significância (num arredondamento a duas casas decimais seria igual) admite-se também que não existem diferenças significativas entre as regiões para esses aminoácidos. Para os restantes aminoácidos essa discussão nem se coloca pois os valores-p são superiores ao nível de significância. Em suma, as regiões dentro de cada variedade não apresentam diferenças significativas o que faz com que seja possível estudar as amostras de cada variedade agrupando as duas regiões. Para além disso, de agora em diante não faz mais sentido o estudo por regiões já que com o teste de Kruskal-Wallis se obtiveram as conclusões relativas a esse caso.

Na Tabela 5.7 são então apresentados os dados do arroz branco tomando por base a variedade do mesmo (incluindo ambas as regiões): a variedade Indica e a variedade Japónica, geralmente designadas de arroz agulha e arroz carolino, respetivamente.

À semelhança dos pontos anteriores, também neste caso nada se consegue afirmar quanto à parecença entre as médias. Na comparação de médias com recurso a testes estatísticos, é possível que se consigam retirar mais informações.

Tabela 5.7 - Comparativo das concentrações (média e desvio padrão) dos aminoácidos no arroz branco de variedade indica e japónica

Aminoácido	Abreviatura	Índico (n = 12) (mg/100g)	Japónico (n = 10) (mg/100g)
Histidina	**His**	192,70 ± 44,56	196,38 ± 62,02
Serina	Ser	338,18 ± 62,73	326,85 ± 42,30
Arginina	Arg	576,28 ± 90,21	550,67 ± 69,16
Glicina	Gly	307,08 ± 38,95	291,82 ± 35,41
Ácido Aspártico	Asp	583,30 ± 118,54	575,73 ± 49,14
Ácido Glutâmico	Glu	1288,12 ± 277,23	1235,03 ± 106,56
Treonina	**Thr**	200,68 ± 32,73	197,00 ± 33,06
Alanina	Ala	349,79 ± 60,98	335,73 ± 27,49
Prolina	Pro	287,24 ± 45,80	276,04 ± 39,33
Cisteína	Cys	42,10 ± 5,23	42,55 ± 4,73
Lisina	**Lys**	130,99 ± 47,89	125,89 ± 29,24
Tirosina	Tyr	335,39 ± 61,60	327,59 ± 94,70
Metionina	**Met**	161,03 ± 38,47	153,69 ± 61,43
Valina	**Val**	319,90 ± 58,54	301,01 ± 30,79
Isoleucina	**Ile**	240,00 ± 39,11	223,83 ± 24,08
Leucina	**Leu**	509,70 ± 84,73	496,26 ± 48,33
Fenilalanina	**Phe**	407,57 ± 58,31	388,28 ± 69,96

5.1.2. Arsénio

Após apresentar todos os casos de comparação para os aminoácidos, é igualmente importante apresentar os mesmos para o elemento-traço exposto no presente trabalho – o arsénio. Por se tratar de apenas um elemento, foi possível apresentar todas as circunstâncias para posterior análise em separado num só quadro, na Tabela 5.8.

Tabela 5.8 - Comparativo das médias e desvios-padrão da concentração de arsénio para todos os casos

Circunstância				Arsénio (ppb)	Arsénio (mg/100g)
Amostras de Arroz	Conjunto			199,19 ± 102,63	1,99E-6 ± 1,03E-6
	Integral			173,78 ± 53,55	1,74E-6 ± 5,36E-7
	Branco			218,83 ± 126,28	2,19E-6 ± 1,26E-6
	Integral		Biológico	177,06 ± 65,94	1,77E-6 ± 6,59E-7
			Não biológico	170,86 ± 43,72	1,71E-6 ± 4,37E-7
	Branco	Var. Indica	Região Sado	283,73 ± 147,78	2,84E-6 ± 1,48E-6
			Região Ribatejo	168,33 ± 58,89	1,68E-6 ± 5,89E-7
		Var. Japónica	Região Sado	237,25 ± 71,24	2,37E-6 ± 7,12E-7
			Região Ribatejo	231,56 ± 180,53	2,32E-6 ± 1,81E-6
		Variedade Índica		206,80 ± 106,73	2,07E-6 ± 1,07E-6
		Variedade Japónica		233,26 ± 151,20	2,33E-6 ± 1,51E-6

Como já explicado, um elemento-traço é um elemento que se apresenta em quantidades muitos pequenas, e como tal vem expresso em ppb (partes por bilião). No entanto, e para estar nas mesmas unidades que os aminoácidos para posteriores cálculos, estes valores serão expressos igualmente noutra coluna em mg/100g.

À semelhança dos dados apresentados na Tabela 5.1, também para o arsénio foram apresentados os parâmetros para todas as amostras de arroz ("conjunto"), apenas para dar conhecimento dos mesmos, já que estatisticamente não têm importância ou significado no estudo atual. Atendendo às diminutas concentrações de arsénio presentes nas amostras, é possível identificar desvios significativos entre algumas leituras, podendo estes originar coeficientes de variação superiores a 50%, como se pode testemunhar na Tabela 5.8. Ainda assim, no arroz branco a quantidade de arsénio é ligeiramente superior ao arroz integral, tal como acontece, entre o arroz branco da região do Sado em relação à região do Ribatejo (em ambas as variedades).

Em conformidade com o que tinha sido feito nos aminoácidos, dentro de cada variedade são apresentados os valores médios e desvios padrão de cada região. Contudo as médias foram comparadas aqui neste ponto para verificar se existiam diferenças significativas entre as regiões. Isto serve para perceber se é possível estudar as duas variedades incluindo todas as regiões. Foi igualmente usado o teste de Kruskal-Wallis (estatística não-paramétrica) para a comparação de médias entre regiões dentro de cada variedade, onde obtiveram valores-p de 0,235 e 0,569 para a variedade indica e japónica, respectivamente. Como se verificou que ao nível do arsénio também não se verificam diferenças significativas nas regiões por variedade, pode-se estudar o arroz branco por variedade (últimas duas linhas da Tabela 5.8) independentemente da região a que pertençam. Numa fase mais avançada (após verificação dos pressupostos) irá ser feita uma comparação entre as médias das variedades e aí, se possível, afirmar com um maior grau de confiança sobre a diferença de médias.

5.2. Testes de normalidade

Uma vez concluída a caracterização do perfil de aminoácidos e arsénio, importa avaliar, sob o ponto de vista estatístico, se existem diferenças significativas entre as várias características (integral/branco, biológico/não biológico, índico/japónico). No entanto, antes de se proceder aos testes estatísticos é necessário verificar os pressupostos da normalidade das amostras e homogeneidade das variâncias das mesmas.

5.2.1. Aminoácidos

À semelhança do ponto anterior, o estudo dos aminoácidos e o arsénio foi feito em separado, tendo-se iniciado pelo teste de normalidade nos aminoácidos que se apresentam de seguida, e no ponto seguinte (Subcapítulo 5.2.2) da dissertação são mostrados então os testes para o arsénio.

- **Arroz branco**

Os dados para o primeiro teste foram apenas os do arroz branco, tendo sido deixados de parte os de arroz integral. Foram executados os três testes mencionados no capítulo anterior, para verificar a normalidade dos dados. Na Tabela 5.9 são apresentados os valores da estatística e respetivo valor-p

para cada um dos testes efetuado, onde se destacam a sombreado os casos em que a hipótese nula, referente à normalidade dos dados, não é respeitada.

Tabela 5.9 - Testes de normalidade às concentrações dos aminoácidos presentes no arroz branco

Aminoácido	Shapiro-Wilk		Anderson-Darling		Kolmogorov-Smirnov	
	Estatística (W)	Valor-p	Estatística (AD)	Valor-p	Estatística (KS)	Valor-p
His	**0,797**	**0,000**	**1,750**	**0,001**	**0,237**	**0,002**
Ser	0,968	0,684	0,304	0,570	0,102	0,200*
Arg	0,971	0,724	0,251	0,729	0,111	0,200*
Gly	0,981	0,935	0,167	0,946	0,095	0,200*
Asp	0,946	0,258	0,353	0,440	0,102	0,200*
Glu	0,914	0,057	0,709	0,055	0,179	0,065
Thr	0,973	0,776	0,296	0,593	0,148	0,200*
Ala	0,952	0,340	0,378	0,384	0,128	0,200*
Pro	0,965	0,600	0,383	0,374	0,128	0,200*
Cys	0,982	0,948	0,147	0,971	0,085	0,200*
Lys	0,924	0,093	0,430	0,288	0,113	0,200*
Tyr	0,930	0,122	0,590	0,111	0,182	0,056
Met	**0,876**	**0,010**	**0,967**	**0,012**	0,181	0,059
Val	0,945	0,256	0,604	0,103	0,162	0,136
Ile	0,939	0,187	0,534	0,156	0,156	0,175
Leu	0,962	0,525	0,327	0,506	0,115	0,200*
Phe	0,964	0,570	0,343	0,465	0,128	0,200*

* corresponde a um valor-p superior ou igual a 0,200

A primeira conclusão, com base no que foi descrito sobre os testes de normalidade no capítulo 4, é que os dois testes mais poderosos (Shapiro-Wilk e Anderson-Darling) estão de acordo uma vez que ambos mostram que, para um nível de significância de 5%, a histidina e a metionina não seguem uma distribuição normal. Já o teste de Kolmogorov-Smirnov está de acordo na histidina, mas talvez por ser um teste com menos poder não rejeita a hipótese de os dados seguirem uma distribuição normal no caso da metionina. No entanto, existem alguns casos em que o valor-p está muito perto de 0,05, ou seja, não se rejeita a hipótese nula mas para um nível de significância mais elevado esta seria rejeitada, o que faz com que não seja "forte" o suficiente a hipótese de que os dados seguem uma distribuição normal. Nesta situação tem-se o ácido glutâmico nos 3 testes e a tirosina para o teste de KS.

- **Arroz integral**

Após o estudo da normalidade nos aminoácidos no arroz branco, o foco vira-se para os dados referentes ao arroz integral. Foram igualmente feitos os testes para testar a normalidade dos dados do arroz que, com base nas médias apresentadas num dos pontos anteriores, tem valores de concentrações mais elevadas para quase todos os aminoácidos (à exceção da lisina e do ácido glutâmico). Os valores da estatística de teste, bem como os valores-p de cada teste para os dados do arroz integral, podem ser consultados na Tabela 5.10, que é exibida em seguida.

Tabela 5.10 - Testes de normalidade às concentrações dos aminoácidos presentes no arroz integral

Aminoácido	Shapiro-Wilk		Anderson-Darling		Kolmogorov-Smirnov	
	Estatística (W)	Valor-p	Estatística (AD)	Valor-p	Estatística (KS)	Valor-p
His	0,893	0,053	0,708	0,052	0,202	0,063
Ser	0,934	0,257	0,427	0,284	0,148	0,200*
Arg	0,963	0,682	0,271	0,661	0,152	0,200*
Gly	0,956	0,555	0,324	0,506	0,148	0,200*
Asp	0,896	0,057	**0,740**	**0,043**	0,209	0,046
Glu	0,925	0,181	0,489	0,198	0,164	0,200*
Thr	0,898	0,062	**0,737**	**0,044**	0,179	0,150
Ala	0,933	0,245	0,386	0,359	0,159	0,200*
Pro	0,898	0,063	**0,739**	**0,043**	0,199	0,074
Cys	**0,735**	**0,000**	**2,089**	**0,001**	**0,332**	**0,000**
Lys	0,918	0,136	0,467	0,225	0,168	0,200*
Tyr	0,932	0,234	0,590	0,107	0,204	0,057
Met	0,938	0,299	0,440	0,263	0,168	0,200*
Val	0,921	0,155	0,648	0,075	**0,229**	**0,018**
Ile	**0,870**	**0,022**	**1,079**	**0,006**	**0,246**	**0,008**
Leu	0,970	0,824	0,196	0,881	0,108	0,200*
Phe	0,961	0,647	0,252	0,720	0,131	0,200*

* corresponde a um valor-p superior ou igual a 0,200

No arroz integral, os testes já são mais controversos, estando apenas de acordo na cisteína e na isoleucina. A valina é dada como variável que não segue uma distribuição normal pelo teste de Kolmogorov-Smirnov, no entanto, e por se tratar do teste com menor capacidade, pode não se dar muita enfâse a este. Contudo, o teste de Anderson-Darling dá ainda mais 3 aminoácidos (ácido aspártico, treonina e prolina) onde se rejeita a hipótese nula e, onde o teste de Shapiro-Wilk dá valores-p muito próximos do nível de significância. Pode então, isto querer dizer que de facto se rejeita a hipótese nula.

- **Arroz integral biológico**

Segue então a subdivisão de cada tipo de arroz pelas suas características. Inicia-se pelo arroz que nos testes de normalidade gerou alguma contestação entre eles, o arroz integral. Uma das comparações que fará todo o sentido, se os testes o consentirem, é entre o arroz integral de cultivo biológico e de cultivo tradicional, até porque na apresentação dos parâmetros destas populações não pareceram existir diferenças significativas à "vista desarmada". Os testes de normalidade para o arroz integral biológico são expostos na Tabela 5.11, bem como as ocorrências em que se verifica a rejeição da hipótese nula.

A cisteína é o único aminoácido cujo resultado se mantém consistente com o alcançado no estudo do arroz integral biológico e não biológico. Já o teste do Kolmogorov-Smirnov, tal como já tinha acontecido anteriormente, é o único a enunciar que as concentrações de lisina não assentam numa distribuição normal. Todavia, os valores-p dos restantes testes não se encontram muitos distantes do limiar da rejeição, ficando uma ressalva para casos como este.

Tabela 5.11 - Testes de normalidade às concentrações dos aminoácidos presentes no arroz integral biológico

Aminoácido	Shapiro-Wilk		Anderson-Darling		Kolmogorov-Smirnov	
	Estatística (W)	Valor-p	Estatística (AD)	Valor-p	Estatística (KS)	Valor-p
His	0,919	0,421	0,300	0,533	0,174	0,200*
Ser	0,937	0,585	0,260	0,662	0,181	0,200*
Arg	0,914	0,387	0,303	0,525	0,204	0,200*
Gly	0,895	0,259	0,400	0,292	0,198	0,200*
Asp	0,899	0,283	0,403	0,286	0,213	0,200*
Glu	0,922	0,448	0,311	0,502	0,203	0,200*
Thr	0,925	0,469	0,297	0,544	0,178	0,200*
Ala	0,908	0,337	0,361	0,373	0,209	0,200*
Pro	0,874	0,166	0,462	0,196	0,210	0,200*
Cys	**0,787**	**0,021**	**0,763**	**0,026**	**0,296**	**0,037**
Lys	0,843	0,081	0,601	0,078	**0,302**	**0,031**
Tyr	0,954	0,752	0,259	0,664	0,220	0,200*
Met	0,969	0,889	0,207	0,835	0,185	0,200*
Val	0,919	0,423	0,335	0,437	0,211	0,200*
Ile	0,876	0,171	0,559	0,104	0,294	0,040
Leu	0,952	0,729	0,242	0,721	0,178	0,200*
Phe	0,973	0,924	0,149	0,970	0,119	0,200*

* corresponde a um valor-p superior ou igual a 0,200

- **Arroz integral não biológico**

Após análise à normalidade do arroz integral biológico, procedeu-se ao estudo para o arroz integral de cultivo tradicional ou, como no presente estudo foi intitulado, arroz integral não biológico.

Tabela 5.12 - Testes de normalidade às concentrações dos aminoácidos presentes no arroz integral não biológico

Aminoácido	Shapiro-Wilk		Anderson-Darling		Kolmogorov-Smirnov	
	Estatística (W)	Valor-p	Estatística (AD)	Valor-p	Estatística (KS)	Valor-p
His	0,899	0,245	0,467	0,198	0,234	0,170
Ser	0,926	0,444	0,415	0,274	0,210	0,200*
Arg	0,926	0,444	0,335	0,444	0,209	0,200*
Gly	0,953	0,718	0,207	0,837	0,148	0,200*
Asp	0,885	0,177	0,524	0,137	0,229	0,191
Glu	0,915	0,354	0,375	0,350	0,170	0,200*
Thr	0,863	0,102	0,658	0,057	**0,285**	**0,034**
Ala	0,944	0,627	0,242	0,726	0,145	0,200*
Pro	0,903	0,268	0,464	0,201	0,240	0,143
Cys	**0,669**	**0,001**	**1,510**	**0,001**	**0,398**	**0,000**
Lys	0,903	0,268	0,451	0,219	0,260	0,080
Tyr	0,893	0,214	0,510	0,150	0,210	0,200*
Met	0,861	0,098	0,607	0,080	**0,299**	**0,020**
Val	0,890	0,201	0,480	0,182	0,250	0,110
Ile	0,856	0,086	0,563	0,106	0,209	0,200*
Leu	0,968	0,877	0,231	0,763	0,174	0,200*
Phe	0,960	0,798	0,245	0,719	0,173	0,200*

* corresponde a um valor-p superior ou igual a 0,200

Da Tabela 5.12, retira-se que as concentrações de cisteína de facto não seguem uma distribuição normal no arroz integral, já que tanto no arroz integral como um todo, como nos casos de arroz integral só biológico ou só não biológico, os testes de normalidade estão de acordo em rejeitar essa hipótese. Mais uma vez, aparece o teste de Kolmogorov-Smirnov a servir de alerta para a metionina e treonina. Porém, desta vez, não parece ter qualquer significado já que o teste de Shapiro-Wilk tem um valor-p bastante mais elevado.

- **Arroz Branco de variedade índica**

Para terminar o estudo da normalidade, falta apenas testar as variedades do arroz branco, indica e japónica. Todos os testes para o arroz branco de variedade indica podem ser vistos na Tabela 5.13.

A aplicação dos três testes permitiu concluir de forma unânime e para os 17 aminoácidos estudados, que a hipótese nula não é rejeitada, ou seja, todos os aminoácidos seguem uma distribuição aproximadamente normal. No entanto, para se poder comparar as médias das populações do arroz branco Indico com o Japónico, recorrendo a estatística paramétrica, o mesmo teria de acontecer com o arroz branco Japónico.

Tabela 5.13 - Testes de normalidade às concentrações dos aminoácidos presentes no arroz branco de variedade indica

Aminoácido	Shapiro-Wilk		Anderson-Darling		Kolmogorov-Smirnov	
	Estatística (W)	Valor-p	Estatística (AD)	Valor-p	Estatística (KS)	Valor-p
His	0,865	0,057	0,673	0,058	0,193	0,200*
Ser	0,955	0,715	0,297	0,568	0,149	0,200*
Arg	0,966	0,870	0,233	0,766	0,146	0,200*
Gly	0,969	0,898	0,245	0,729	0,133	0,200*
Asp	0,919	0,274	0,360	0,402	0,129	0,200*
Glu	0,908	0,202	0,435	0,258	0,165	0,200*
Thr	0,979	0,980	0,185	0,903	0,129	0,200*
Ala	0,953	0,675	0,274	0,638	0,154	0,200*
Pro	0,964	0,840	0,260	0,683	0,125	0,200*
Cys	0,979	0,980	0,175	0,927	0,132	0,200*
Lys	0,888	0,112	0,510	0,162	0,181	0,200*
Tyr	0,945	0,559	0,365	0,390	0,189	0,200*
Met	0,894	0,134	0,514	0,159	0,175	0,200*
Val	0,951	0,652	0,342	0,445	0,164	0,200*
Ile	0,958	0,757	0,293	0,579	0,166	0,200*
Leu	0,955	0,708	0,298	0,566	0,161	0,200*
Phe	0,979	0,977	0,193	0,883	0,146	0,200*

* corresponde a um valor-p superior ou igual a 0,200

- **Arroz Branco de variedade japónica**

Por fim, a Tabela 5.14 apresenta os testes de normalidade para as amostras de arroz branco de variedade japónica que incluem, como já referido, amostras de ambas regiões que não apresentam diferenças significativas entre si.

Tabela 5.14 - Testes de normalidade às concentrações dos aminoácidos presentes no arroz branco de variedade japónica

Aminoácido	Shapiro-Wilk		Anderson-Darling		Kolmogorov-Smirnov	
	Estatística (W)	Valor-p	Estatística (AD)	Valor-p	Estatística (KS)	Valor-p
His	**0,746**	**0,003**	**1,203**	**0,002**	**0,323**	**0,004**
Ser	0,963	0,821	0,202	0,857	0,116	0,200*
Arg	0,985	0,985	0,161	0,953	0,141	0,200*
Gly	0,989	0,995	0,149	0,969	0,128	0,200*
Asp	0,908	0,265	0,417	0,277	0,233	0,132
Glu	0,941	0,568	0,334	0,455	0,188	0,200*
Thr	0,912	0,295	0,443	0,236	0,215	0,200*
Ala	0,964	0,826	0,192	0,885	0,120	0,200*
Pro	0,951	0,686	0,252	0,700	0,162	0,200*
Cys	0,959	0,777	0,177	0,922	0,132	0,200*
Lys	0,962	0,806	0,236	0,751	0,172	0,200*
Tyr	0,882	0,137	0,551	0,119	0,254	0,066
Met	**0,805**	**0,016**	**0,892**	**0,013**	**0,268**	**0,041**
Val	0,942	0,574	0,357	0,398	0,224	0,167
Ile	0,962	0,804	0,294	0,567	0,160	0,200*
Leu	0,944	0,596	0,236	0,750	0,141	0,200*
Phe	0,943	0,588	0,280	0,609	0,164	0,200*

* corresponde a um valor-p superior ou igual a 0,200

A leitura da Tabela 5.14 permitiu verificar que as concentrações de histidina e metionina não seguem uma distribuição normal, de acordo com os três testes realizados, o que para comparar esta variedade com a Indica terão de ser equacionadas várias hipóteses.

5.2.2. Arsénio

Por sua vez, e de igual forma, os testes foram executados para o arsénio, que são apresentados na Tabela 5.15. De notar, que por se tratar apenas de um elemento-traço, todos os casos estão presentes nas diversas linhas com os respetivos resultados aos testes nas colunas.

Tabela 5.15 - Testes de normalidade às concentrações de arsénio para as várias hipóteses em estudo

Arsénio			Shapiro-Wilk		Anderson-Darling		*Kolmogorov-Smirnov*	
			Estatística (W)	Valor-p	Estatística (AD)	Valor-p	Estatística (KS)	Valor-p
Arroz	Integral		0,966	0,753	0,184	0,911	0,116	0,200*
	Branco		**0,859**	**0,005**	**0,925**	**0,015**	**0,186**	**0,046**
	Integral	Biológico	0,936	0,573	0,241	0,725	0,160	0,200*
		Não biológico	0,982	0,975	0,138	0,982	0,115	0,200*
	Branco	Var. Indica	0,889	0,113	0,525	0,149	0,208	0,160
		Var. Japónica	**0,814**	**0,022**	**0,731**	**0,037**	**0,268**	**0,040**

* corresponde a um valor-p superior ou igual a 0,200

Percebe-se rapidamente que, dentro do arroz integral, quer o arroz integral biológico, quer o arroz integral não biológico apresentam seguir uma distribuição aproximadamente normal para os três

testes efectuados. Resta saber se o outro pressuposto é ou não verificado com vista a usar estatística paramétrica. Os testes estão sempre de acordo, no entanto, nos restantes casos existem sempre uns dados que seguem uma distribuição normal e outros que não.

5.3. Testes de homogeneidade de variância

Com vista à comparação de médias das diversas variáveis, é necessário ainda testar a homogeneidade da variância (outro pressuposto da ANOVA). Em sequência do que vem a ser feito, aqui na execução dos testes da homogeneidade, foram testadas as homogeneidades entre as 2 hipóteses que existem em cada um dos três casos.

5.3.1. Aminoácidos

Com o objetivo de facilitar a interpretação, foi concebida a Tabela 5.16, onde estão concentrados os resultados aos dois testes de homogeneidade de variância para todos os casos (colunas) com todas as respetivas variáveis (linhas).

Tabela 5.16 - Testes de homogeneidade da variância às concentrações de aminoácidos para as várias hipóteses em estudo

AA	Arroz Integral e Branco		Arroz Branco Indico e Japónico		Arroz Integral Biológico e não Biológico	
	Levene (Valor-p)	Brown Forsythe (Valor-p)	Levene (Valor-p)	Brown Forsythe (Valor-p)	Levene (Valor-p)	Brown Forsythe (Valor-p)
His	0,446	0,633	0,371	0,711	0,212	0,280
Ser	**0,029**	**0,047**	0,195	0,196	0,724	0,733
Arg	**0,018**	**0,022**	0,396	0,374	0,155	0,146
Gly	0,199	0,202	0,746	0,714	0,593	0,720
Asp	0,296	0,261	**0,030**	0,053	0,358	0,333
Glu	0,101	0,140	**0,005**	**0,023**	0,579	0,580
Thr	0,611	0,515	0,927	0,776	0,409	0,406
Ala	0,142	0,174	**0,023**	0,056	0,909	0,949
Pro	0,158	0,261	0,389	0,375	0,594	0,497
Cys	**0,000**	**0,002**	0,767	0,784	0,259	0,214
Lys	**0,031**	**0,036**	0,184	0,245	**0,025**	0,268
Tyr	0,075	0,171	0,177	0,442	0,138	0,295
Met	0,655	0,703	0,178	0,410	0,862	0,999
Val	0,288	0,337	0,063	0,086	0,134	0,153
Ile	0,634	0,604	0,118	0,141	0,174	0,324
Leu	0,183	0,196	0,185	0,237	0,578	0,595
Phe	0,064	0,117	0,531	0,684	0,368	0,358

Em análise à referida tabela, analisa-se caso a caso, começando pelos resultados da homogeneidade de variância entre as amostras de arroz branco e de arroz integral. De acordo com a Tabela 5.9, a histidina e a metionina não respeitam a normalidade, bem como na Tabela 5.10, o mesmo acontece com o ácido aspártico, a treonina, a alanina, a prolina, a cisteína e a isoleucina; então, e como na serina, arginina, cisteína e lisina se rejeita a hipótese de variâncias homogéneas em ambos os testes, conclui-se que apenas faz sentido usar estatística paramétrica em variáveis como a glicina, o ácido glutâmico, a tirosina, a valina, a leucina e a fenilalanina. De notar que pela Tabela 5.19, onde é

apresentado também o teste *F* de Fisher, este refere que para além dos assinalados, o ácido glutâmico também não possui variâncias homogéneas entre os 2 tipos de arroz. Em suma, 11 das 17 variáveis não permitem a utilização estatística paramétrica, no entanto, e como já referido, todas elas vão ser analisadas pelos dois tipos de estatística (não paramétrica e paramétrica) e retiradas as respetivas conclusões.

Por sua vez, dentro do arroz branco, na comparação entre as duas variedades, foram consultadas a Tabela 5.13 e a Tabela 5.14, onde apenas a histidina e a metionina da variedade japónica não verificam a normalidade. Pela Tabela 5.16, claramente o ácido glutâmico não tem variância homogénea e, o ácido aspártico e a alanina para um dos testes (Levene) é rejeitada a hipótese nula e no outro (Brown-Forsythe) não, ainda que por ter um valor-p próximo de 0,05 é considerado que estas variáveis não possuem variância homogénea. Na Tabela II.3 do Anexo II, o teste *F* de Fisher apresenta resultados que vêm corroborar os resultados obtidos no teste de Levene da Tabela 5.16. Resumindo, na histidina, na metionina, no ácido glutâmico, ácido aspártico e na alanina os pressupostos são violados, o que os remete para estatística não paramétrica.

Para finalizar, falta apenas relatar sobre os tipos de agricultura que é feita no arroz integral. Apenas a cisteína (Tabela 5.11 e Tabela 5.12) não segue uma distribuição normal e, apenas a lisina (Tabela 5.16) está assinalada quanto violação do pressuposto da homogeneidade de variâncias. Aqui os testes que testam a homogeneidade da variância estão bastante controversos, já que os valores-p dão bastante discrepantes. O teste *F* de Fisher (Tabela II.5 do Anexo II) apresenta resultados semelhantes aos do teste de Brown-Forsythe, contudo nada se conclui e, irão ser comparadas as médias pelos dois tipos de estatísticas na expectativa de conseguir tirar uma conclusão mais clara.

5.3.2. Arsénio

Também para o estudo do arsénio as amostras foram submetidas aos mesmos testes para semelhança das variâncias. Obviamente os casos para estudo foram os mesmos, na sequência do que tem vindo a ser feito. A Tabela 5.17 é mais simples de analisar devido a ser apenas uma variável em estudo e análise.

Tabela 5.17 - Testes de homogeneidade da variância às concentrações de arsénio para as várias hipóteses em estudo

	Arroz		Arroz Branco		Arroz Integral	
	Integral e Branco		Indico e Japónico		Biológico e não Biológico	
As	Levene (Valor-p)	Brown Forsythe (Valor-p)	Levene (Valor-p)	Brown Forsythe (Valor-p)	Levene (Valor-p)	Brown Forsythe (Valor-p)
	0,024	0,077	0,779	0,773	0,171	0,180

Na Tabela 5.15, em quase todos os casos (à exceção de no arroz integral) existia sempre uma das hipóteses que não verificava a normalidade. Então só faz sentido usar a estatística paramétrica dentro do arroz integral, já que pela Tabela 5.17 não existem factos que apontem em sentido contrário.

5.4. Comparação de médias

Até aqui, já foram apresentadas as médias (subcapítulo 4.1), já foi testada a normalidade (subcapítulo 4.2) e a homogeneidade da variância (subcapítulo 4.3) para os demais casos e variáveis. Finalmente chega-se a um dos objetivos deste estudo, a comparação das médias das amostras. A comparação das regiões por variedade já foi executada no subcapítulo onde foram apresentadas as médias, pois eram necessárias para se saber a viabilidade do estudo do arroz branco por variedade.

5.4.1. Aminoácidos

Inicia-se o estudo pelos aminoácidos, e com base no que foi referido anteriormente, foi concebida a Figura 5.1, onde são expostas as variáveis que pela violação dos pressupostos atrás verificados são remetidas para estatística não-paramétrica. No entanto, todas as variáveis para todos os casos, foram testadas quer por estatística paramétrica (ANOVA, e *t* de *Student*), quer por estatística não-paramétrica (Kruskal-Wallis). Todavia, a Figura 5.1 tem o propósito de facilitar durante a análise aos resultados de ambos os testes. As ANOVA's completas e os testes de *t* de *Student* acompanhados dos respetivos testes *F* de Fisher encontram-se no Anexo II (da Tabela II.1 à Tabela II.5).

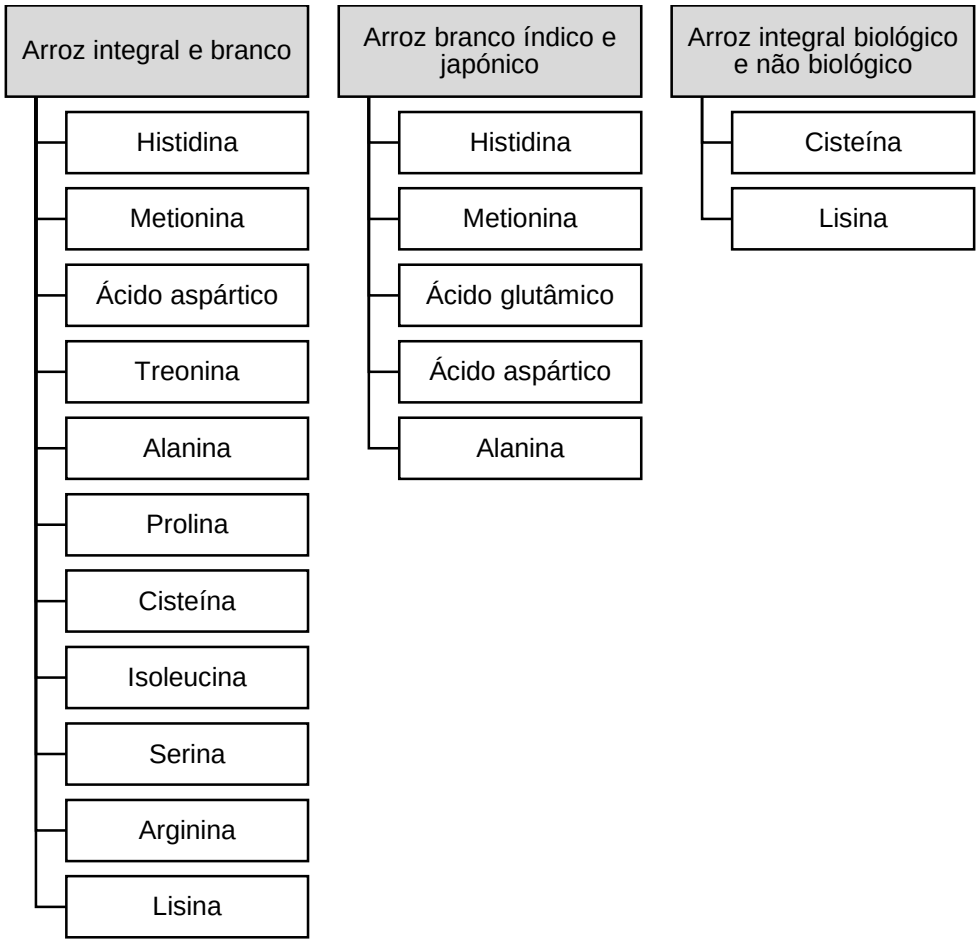

Figura 5.1 - Variáveis (aminoácidos) remetidas para estatística não-paramétrica

Na Tabela 5.18, são apresentados os resultados (valores-p), e estão assinaladas as variáveis em cada comparação em estudo que possuem médias significativamente diferentes umas das outras.

A primeira conclusão que se retira pela observação da tabela é que onde se encontram mais variáveis com diferenças significativas é na comparação de arroz por tipo (arroz branco – arroz integral), e como tal, a análise vai ter especial incidência nesta comparação.

Tabela 5.18 - Testes de comparação de médias às respetivas concentrações de aminoácidos para as várias hipóteses em estudo

AA	Arroz Integral e Branco		Arroz Branco Indico e Japónico		Arroz Integral Biológico e não Biológico	
	ANOVA (Valor-p)	Kruskal Wallis (Valor-p)	ANOVA (Valor-p)	Kruskal Wallis (Valor-p)	ANOVA (Valor-p)	Kruskal Wallis (Valor-p)
His	**0,000**	**0,003**	0,873	0,644	0,537	0,700
Ser	**0,000**	**0,000**	0,633	0,792	0,595	0,564
Arg	**0,000**	**0,000**	0,471	0,553	0,839	0,923
Gly	**0,000**	**0,000**	0,352	0,356	0,324	0,336
Asp	**0,023**	**0,014**	0,852	0,843	0,468	0,700
Glu	0,596	0,910	0,575	0,843	0,918	0,847
Thr	**0,000**	**0,000**	0,796	0,742	0,293	0,211
Ala	**0,001**	**0,001**	0,509	0,742	0,287	0,248
Pro	**0,000**	**0,000**	0,550	0,598	0,271	0,211
Cys	**0,000**	**0,000**	0,835	0,692	0,398	0,923
Lys	**0,000**	**0,001**	0,772	0,895	0,741	1,000
Tyr	**0,000**	**0,000**	0,818	0,291	0,561	0,630
Met	**0,000**	**0,000**	0,736	0,262	**0,036**	0,054
Val	**0,002**	**0,003**	0,370	0,644	0,883	0,700
Ile	**0,019**	**0,025**	0,269	0,429	0,930	0,773
Leu	**0,003**	**0,002**	0,662	0,895	0,514	0,336
Phe	**0,000**	**0,000**	0,488	0,356	0,654	0,847

Aquando da construção da Tabela 5.2, tinha-se notado uma clara diferença entre as amostras do arroz integral e do arroz branco. E agora demonstra-se isso mesmo pela Tabela 5.18, onde a única média que não é significativamente diferente, para um nível de significância de 5%, é o ácido glutâmico. Resultados semelhantes são apresentados na Tabela 5.19 pelo teste *t* de *Student*. No entanto, nos testes de homogeneidade de variância existia controvérsia entre os testes de Levene e Brown-Forsythe e o teste *F* de Fisher (apresenta um valor-p de 0,024). Com o intuito de verificar visualmente isso, foi construída a Figura 5.2, gráfico de *box and whiskers*. Pelo gráfico conclui-se de facto que as variâncias não são homogéneas, ou seja, embora as médias não sejam significativamente diferentes, não se pode afirmar que ambas as amostras pertencem à mesma população, uma vez que as variâncias evidenciam diferenças significativas. De notar, que os resultados foram semelhantes pelas duas estatísticas, quer para as 11 variáveis em que os pressupostos foram violados, quer para as restantes 6.

Na comparação das variedades do arroz branco, ambas as estatísticas deram, mais uma vez, resultados concordantes, evidenciando a não existência de diferenças significativas. Conclui-se então, que a nível proteico o arroz carolino e agulha são semelhantes. Contudo em variáveis como o ácido aspártico, ácido glutâmico e alanina pelas diferenças significativas que as variâncias apresentam (Tabela 5.16 e Tabela II.3 do Anexo II) não se pode afirmar que pertençam à mesma população.

Tabela 5.19 - Teste *t* de *Student* e teste *F* de Fisher para o tipo de arroz (branco e integral)

AA	Média Arroz Integral	Média Arroz Branco	*t* de *Student*			Desvio Padrão Integral	Desvio Padrão Branco	*F* de Fisher	
			t	g.l.	Valor-p			*F*	Valor-p
His	388,70	194,37	12,953	37	0,000	38,19	51,88	1,846	0,215
Ser	419,16	333,03	5,991	37	0,000	28,78	53,49	3,454	0,014
Arg	736,54	564,64	7,984	37	0,000	42,10	80,52	3,659	0,011
Gly	393,92	300,14	8,874	37	0,000	25,47	37,32	2,147	0,123
Asp	642,38	579,86	2,365	37	0,023	66,76	91,70	1,887	0,199
Glu	1232,86	1263,99	-0,535	37	0,596	121,52	214,14	3,105	0,025
Thr	279,56	199,01	8,177	37	0,000	28,22	32,14	1,297	0,602
Ala	389,67	343,40	3,471	37	0,001	29,85	48,20	2,607	0,055
Pro	363,95	282,15	6,720	37	0,000	30,51	42,36	1,927	0,185
Cys	215,28	42,31	15,687	37	0,000	51,62	4,89	111,210	0,000
Lys	87,24	128,67	-3,854	37	0,000	22,26	39,68	3,177	0,022
Tyr	586,24	331,84	11,919	37	0,000	49,29	76,46	2,406	0,078
Met	310,47	157,70	10,021	37	0,000	44,66	49,06	1,207	0,710
Val	359,73	311,31	3,399	37	0,002	38,58	47,90	1,542	0,381
Ile	258,47	232,65	2,461	37	0,019	31,22	33,43	1,147	0,790
Leu	565,07	503,59	3,177	37	0,003	44,61	69,35	2,416	0,077
Phe	589,22	398,80	10,549	37	0,000	44,82	63,05	1,979	0,168

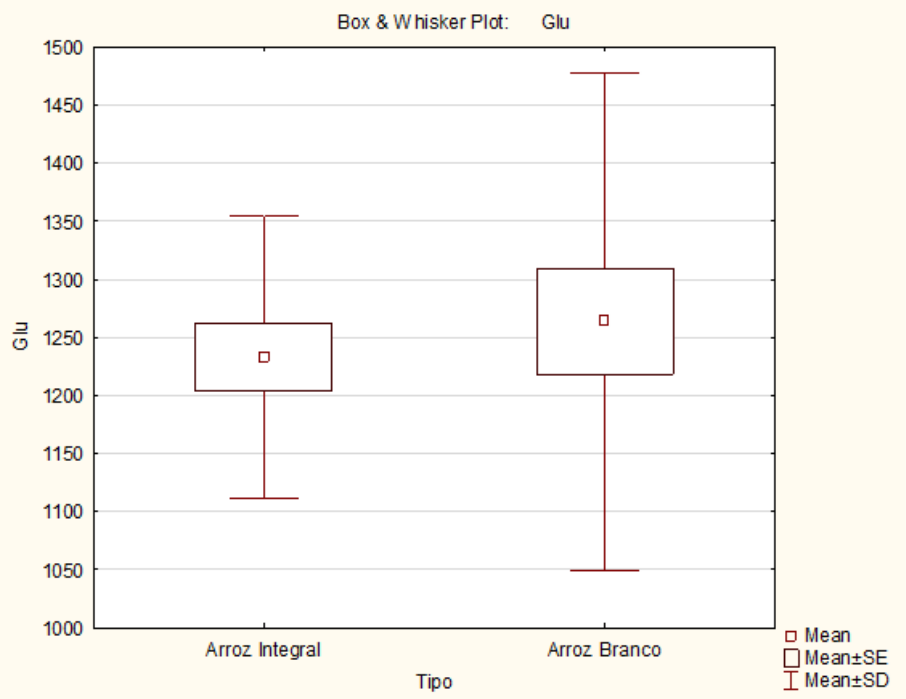

Figura 5.2 - Gráfico *Box and Whiskers* do ácido glutâmico (Glu) por tipos de arroz

Para concluir a comparação de médias, resta o caso do arroz integral biológico comparado com o não biológico. O único aminoácido assinalado é a metionina, cujos pressupostos não foram violados por esta, então o valor-p assinalado da ANOVA não é rejeitado, concluindo que a metionina possui médias significativamente diferentes entre diferentes tipos de agricultura de arroz integral. Dos restantes aminoácidos apenas a lisina possui diferenças significativas entre as variâncias (Tabela 5.16).

Sintetizando todo este estudo, o único caso em que as amostras de arroz são diferentes para quase todos os aminoácidos, é quando se comparam os dois tipos de arroz, o integral e o branco. Isto vai ser importante posteriormente para outro dos objetivos da dissertação.

5.4.2. Arsénio

Analogamente, foram comparadas as médias para o arsénio pelos dois tipos de estatística. Conclui-se pela Tabela 5.20, que não existem diferenças significativas (os testes estão sempre concordantes). Talvez isto aconteça, como foi referido na interpretação da Tabela 5.8, pelos desvios-padrão altos que as amostras para o arsénio possuem.

Tabela 5.20 - Testes de comparação de médias às respetivas concentrações de arsénio para as várias hipóteses em estudo

	Arroz		Arroz Branco		Arroz Integral	
	Integral e Branco		Indico e Japónico		Biológico e não Biológico	
As	ANOVA (Valor-p)	Kruskal Wallis (Valor-p)	ANOVA (Valor-p)	Kruskal Wallis (Valor-p)	ANOVA (Valor-p)	Kruskal Wallis (Valor-p)
	0,177	0,428	0,701	0,863	0,821	0,630

Conclui-se então que o arroz, seja qual o tipo/variedade/tipo de cultivo, tem concentrações de arsénio semelhantes a rondar os 200 ppb. No estudo feito por Simões, em 2014, as amostras de origem portuguesa tinham uma concentração entre os 114 e os 285 ppb. Ou seja, os valores obtidos na Tabela 5.8 encontram-se dentro desse intervalo de valores obtidos (Simões, 2014).

5.5. Correlação entre aminoácidos e arsénio

Outro dos objetivos era averiguar se existem correlações entre os diversos aminoácidos e o arsénio (contaminante da cadeia alimentar) no arroz, devido à importância que o mesmo tem na alimentação mundial, já que é dos principais alimentos no fornecimento de energia a nível mundial.

Foi construída a Tabela 5.21 com as devidas correlações de *Spearman* entre os aminoácidos (linhas) e o arsénio. À medida que se foram desdobrando os dados do arroz integral (onde existia correlação) foram sendo descobertas novas correlações, e como tal o mesmo foi feito para o arroz branco que quando analisado em separado não apresentava correlações nenhumas.

A primeira grande conclusão que se retira, quando se analisa a tabela, é que as amostras de arroz integral (seja qual for a hipótese ou a variável em estudo) se correlacionam positivamente com o arsénio, e as amostras de arroz branco correlacionam-se negativamente. Tal situação deve-se ao facto do arsénio não apresentar diferenças significativas em nenhuma das comparações (Tabela 5.20), ou seja, todo o arroz apresenta um nível de arsénio semelhante (a rondar os 200 ppb). Quando o arsénio é colocado ao lado das concentrações de aminoácidos correlaciona-se positivamente com o arroz que possui maior teor (arroz integral) e negativamente com o que possui menor (arroz branco). Para cimentar esta conclusão, os valores de correlação entre os vários aminoácidos para o arroz

integral e o arroz branco são aproximadamente inversos entre si, sendo positivos para o arroz integral e negativos para o arroz branco.

No arroz integral tem-se uma correlação média-forte (0,694) para um nível de significância de 0,002 entre a lisina e o arsénio. A lisina é o aminoácido limitante do arroz (Walter et al., 2008), o que com este resultado leva a conclusão de que a concentração de arsénio no arroz integral está relacionada positivamente com a concentração de lisina. A correlação com esta variável já não aparece em mais nenhuma hipótese do estudo.

Tabela 5.21 - Correlação de *Spearman* entre os diversos aminoácidos e o arsénio

Correlação AA e As		Arroz Integral	Arroz Branco				
			Tipo			Variedade	
		Agrupado (n=17)	Não biológico (n=9)	Biológico (n=8)	Agrupado (n=22)	Indica (n=12)	Japónica (n=10)
His	ρ	-0,086	-0,517	0,286	-0,132	0,021	-0,358
	Sig.	0,743	0,154	0,493	0,559	0,948	0,310
Ser	ρ	0,157	-0,450	0,571	-0,117	0,105	-0,588
	Sig.	0,548	0,224	0,139	0,604	0,746	0,074
Arg	ρ	0,140	-0,417	0,500	0,003	0,203	-0,333
	Sig.	0,593	0,265	0,207	0,990	0,527	0,347
Gly	ρ	-0,069	-0,533	0,286	0,008	0,350	-0,418
	Sig.	0,794	0,139	0,493	0,970	0,265	0,229
Asp	ρ	0,341	-0,433	0,833	-0,107	0,210	-0,661
	Sig.	0,181	0,244	0,010	0,636	0,513	0,038
Glu	ρ	0,277	-0,483	0,738	0,004	0,175	-0,661
	Sig.	0,282	0,187	0,037	0,986	0,587	0,038
Thr	ρ	0,002	-0,367	0,429	-0,127	0,175	-0,624
	Sig.	0,993	0,332	0,289	0,573	0,587	0,054
Ala	ρ	0,287	-0,283	0,810	-0,123	0,210	-0,733
	Sig.	0,264	0,460	0,015	0,587	0,513	0,016
Pro	ρ	0,059	-0,450	0,476	-0,082	0,189	-0,515
	Sig.	0,823	0,224	0,233	0,717	0,557	0,128
Cys	ρ	0,005	-0,367	0,190	-0,320	-0,014	-0,758
	Sig.	0,985	0,332	0,651	0,146	0,966	0,011
Lys	ρ	0,694	0,650	0,667	0,117	0,168	-0,042
	Sig.	0,002	0,058	0,071	0,604	0,602	0,907
Tyr	ρ	-0,044	-0,533	0,333	-0,160	-0,014	-0,333
	Sig.	0,866	0,139	0,420	0,477	0,966	0,347
Met	ρ	-0,343	-0,533	0,143	-0,303	-0,147	-0,564
	Sig.	0,178	0,139	0,736	0,170	0,649	0,090
Val	ρ	0,471	-0,167	0,714	-0,016	0,287	-0,636
	Sig.	0,057	0,668	0,047	0,942	0,366	0,048
Ile	ρ	0,373	-0,333	0,690	0,100	0,385	-0,370
	Sig.	0,141	0,381	0,058	0,658	0,217	0,293
Leu	ρ	0,196	-0,367	0,500	-0,001	0,406	-0,648
	Sig.	0,451	0,332	0,207	0,998	0,191	0,043
Phe	ρ	0,029	-0,517	0,429	-0,115	0,070	-0,358
	Sig.	0,911	0,154	0,289	0,612	0,829	0,310

Outra conclusão que se retira da Tabela 5.21, é que dos aminoácidos que se correlacionam com o arsénio aproximadamente metade são não essenciais ou condicionalmente essenciais, e a outra metade são essenciais como é o caso da lisina, que se correlaciona com o arsénio no arroz integral como um todo, a leucina com o arroz branco de variedade japónica e a valina com o arroz integral biológico.

5.6. Análise de *clusters*

Até ao presente ponto os vários estudos que têm vindo a ser feitos, tinham sempre duas componentes, os aminoácidos e o arsénio, no entanto a partir daqui o arsénio é posto de lado, pois o objetivo do mesmo já foi apresentado e explicado no ponto anterior. A análise de *clusters* foi dividida em duas partes, por variáveis (aminoácidos) e por casos (amostras), com o objetivo de perceber como é feito o agrupamento em cada caso. Em ambos os casos foram usados os três métodos escolhidos que são apresentados de seguida.

5.6.1. Variáveis (Aminoácidos)

Iniciou-se o estudo fazendo o dendrograma para as variáveis pelos três métodos: a ligação média entre grupos (Figura III.1 do Anexo III), o método do centróide (Figura III.2 do Anexo III) e o método de Ward (Figura III.3 do Anexo III) onde se destacou a separação do ácido glutâmico. Para certificar que as restantes variáveis se agrupavam de facto em 2 *clusters*, os três métodos foram novamente aplicadas e criados novos dendrogramas, desta vez sem o ácido glutâmico (Figura III.4, Figura III.5 e Figura III.6 do Anexo III). Claramente se conclui que se formam 2 *clusters* com a ausência dessa variável. Deste modo, e após validação, pode considerar-se a formação de três *clusters* cuja composição se apresenta na Tabela 5.22. Essa tabela foi criada com base nos 6 dendrogramas construídos e serve de síntese ao agrupamento das variáveis e aos intervalos das mesmas.

Tabela 5.22 - Composição dos *clusters* formados pelas variáveis (aminoácidos)

Clusters (Variáveis)	Método de Ward	Método do centróide	Ligação média entre grupos	Intervalo de valores no *cluster*
Cluster 1	Gly	Gly	Gly	Concentrações entre 110,61 e 370,58 mg/100g
	Pro	Pro	Pro	
	Val	Val	Val	
	Ser	Ser	Ser	
	Ala	Ala	Ala	
	Thr	Thr	Thr	
	Ile	Ile	Ile	
	His	Met	His	
	Met	His	Met	
	Cys	Cys	Cys	
	Lys	Lys	Lys	
Cluster 2	Tyr	Tyr	Tyr	Concentrações entre 442,73 e 639,57 mg/100g
	Phe	Phe	Phe	
	Arg	Arg	Arg	
	Asp	Asp	Asp	
	Leu	Leu	Leu	
Cluster 3	Glu	Glu	Glu	1250,42 mg/100g

Por se tratarem de variáveis o agrupamento é feito obviamente pelas distâncias, mas em termos práticos, os *clusters* são intervalos de valores onde as médias das diversas variáveis encaixam. Particularmente, pela Tabela 5.22, formam-se 3 *clusters*, em que um admite variáveis com concentrações entre, aproximadamente, os 100 e os 400 mg/100g; outro que se forma com variáveis cujas concentrações andam entre os 400 e os 650 mg/100g; e ainda outro, que só possui o ácido glutâmico cuja concentração é 1250 mg/100g.

5.6.2. Casos (Amostras)

Agora que se sabe como as variáveis em estudo se agrupam, é necessário saber o mesmo para as amostras e inferir a partir dos resultados obtidos. Durante a análise foram aplicados os três métodos previamente escolhidos, e as amostras foram identificadas pelo tipo de arroz, respetiva variedade e região quando aplicável. Esta identificação tem o intuito de facilitar a análise visual e identificar os tipos, variedades ou regiões presentes em cada *cluster.*

Remetendo para a Tabela 5.18, verifica-se que o ácido glutâmico era a única variável entre os dois tipos de arroz que não apresentava diferenças significativas. Desta forma, a análise de *clusters* foi feita com todas as variáveis e foi repetida excluindo esta variável. Esta hipótese foi formulada para perceber se o ácido glutâmico por não ser significativamente diferente, tem algum peso em determinadas amostras. As variáveis em análise (os 17 aminoácidos constituintes da proteína) são igualmente relevantes neste estudo e, como tal, recorreu-se à padronização das variáveis. Esta é uma técnica normalmente usada quando as variáveis apresentam unidades de medida diferentes, no entanto, como se obtiveram gamas de variáveis diferentes (como apresentado na Tabela 5.22) foi aplicada com o intuito de remover a influência de cada uma, e perceber se existem diferenças pela padronização dos dados.

Em suma, foram usados 3 métodos para os dados recolhidos e para os dados padronizados, tendo sido executados, em cada um, para as duas hipóteses (todas as variáveis e retirando o ácido glutâmico). No total das hipóteses estudadas foram obtidos 12 análises (dendrogramas). Os cortes não foram realizados sempre à mesma distância, tendo sido baseados numa análise visual onde os resultados aparentavam ser interpretáveis (os dendrogramas que não estão presentes na análise podem ser consultados no Anexo III do presente documento – pontos III.3 e III.4).

A primeira conclusão que se retira ao observar os 12 dendrogramas obtidos é que com ou sem padronização dos dados os resultados são semelhantes. Apenas nas distâncias dos ramos ocorrem modificações, o que pode fazer alterar o corte e, por conseguinte, o número de *clusters*. De notar, que o método do centróide é aquele onde os resultados obtidos são mais difíceis de interpretar do ponto de vista prático.

Analisando, numa primeira fase, os dendrogramas mais interpretáveis (resultantes da análise sem o ácido glutâmico: quer para dados recolhidos, quer para os dados padronizados), obtêm-se 4 clusters na maioria dos métodos (podendo a ordem dos mesmos ser diferente no dendrograma): um onde só

se encontram amostras de arroz integral, outro em que se misturam amostras dos 2 tipos de arroz, e os restantes dois de arroz branco (Figura 5.4 e Figura 5.5). Os restantes dendrogramas podem ser analisados na Figura III.7, Figura III.8, Figura III.9 do Anexo III. O único em que ocorre uma ligeira diferença é o método do centróide nos dados padronizados, que, apesar de apresentar 4 *clusters*, estes são diferentes (Figura III.10 do Anexo III).

Da análise aos 6 dendrogramas da hipótese em estudo (eliminação do Glu), retira-se a constituição dos 4 *clusters* resultantes, seguidamente apresentados na Tabela 5.23.

Tabela 5.23 - Constituição de cada *cluster*

Cluster 1	*Cluster* 2	*Cluster* 3	*Cluster* 4
Arroz branco I	Mistura	Arroz integral	Arroz branco II
Amostra 1.1	Amostra 3	Amostra 8	Amostra 1.2
Amostra 1.6	Amostra 4	Amostra 9	Amostra 1.3
Amostra 1.9	Amostra 5	Amostra 10	Amostra 1.4
Amostra 1.10	Amostra 6	Amostra 11	Amostra 1.5
Amostra 2.8	Amostra 7	Amostra 12	Amostra 1.7
	Amostra 1.12	Amostra 13	Amostra 1.8
	Amostra 2.9	Amostra 14	Amostra 1.11
	Amostra 2.10	Amostra 15	Amostra 2.1
		Amostra 16	Amostra 2.2
		Amostra 19	Amostra 2.3
		Amostra 20	Amostra 2.4
		Amostra 21	Amostra 2.5
			Amostra 2.6
			Amostra 2.7

Quando se passa à análise dos dendrogramas com as 17 variáveis presentes, os resultados são idênticos aos obtidos na análise anterior. Pelo método de Ward (Figura III.11 e Figura III.12 do Anexo III) resultam os mesmos 4 *clusters* (para ambos os dados) encontrados na Tabela 5.23.

No método de ligação média entre grupos são obtidos 5 *clusters* (para ambos os dados: padronizados e como recolhidos), em que 3 correspondem a amostras exclusivamente de arroz branco. Os restantes 2 contêm, respetivamente, apenas arroz integral e as amostras dos dois tipos de arroz misturados (Figura 5.6). Contudo, com os dados padronizados na Figura III.13 do Anexo III, denota-se que 2 *clusters* de arroz branco presentes na Figura 5.6 se agrupam, gerando o mesmo resultado da Tabela 5.23.

Por fim, segue-se a análise dos dados obtidos pelo método de centróide. Para os dados normais (Figura III.14 do Anexo III), pelo corte feito são gerados 6 *clusters*: 3 que contêm apenas arroz branco, 2 que contêm apenas arroz integral e 1 que contém as amostras que se misturam. Para os dados padronizados (Figura III.15 do Anexo III) obtiveram-se igualmente 6 *clusters* um pouco diferentes: 4 de arroz branco - um deles é idêntico ao obtido nos dados normais, e os restantes 3 são iguais ao agrupamento de 2 *clusters* obtidos no mesmo método com os dados sem modificação. Obtêm-se ainda 1 *cluster* de arroz integral e 1 de mistura.

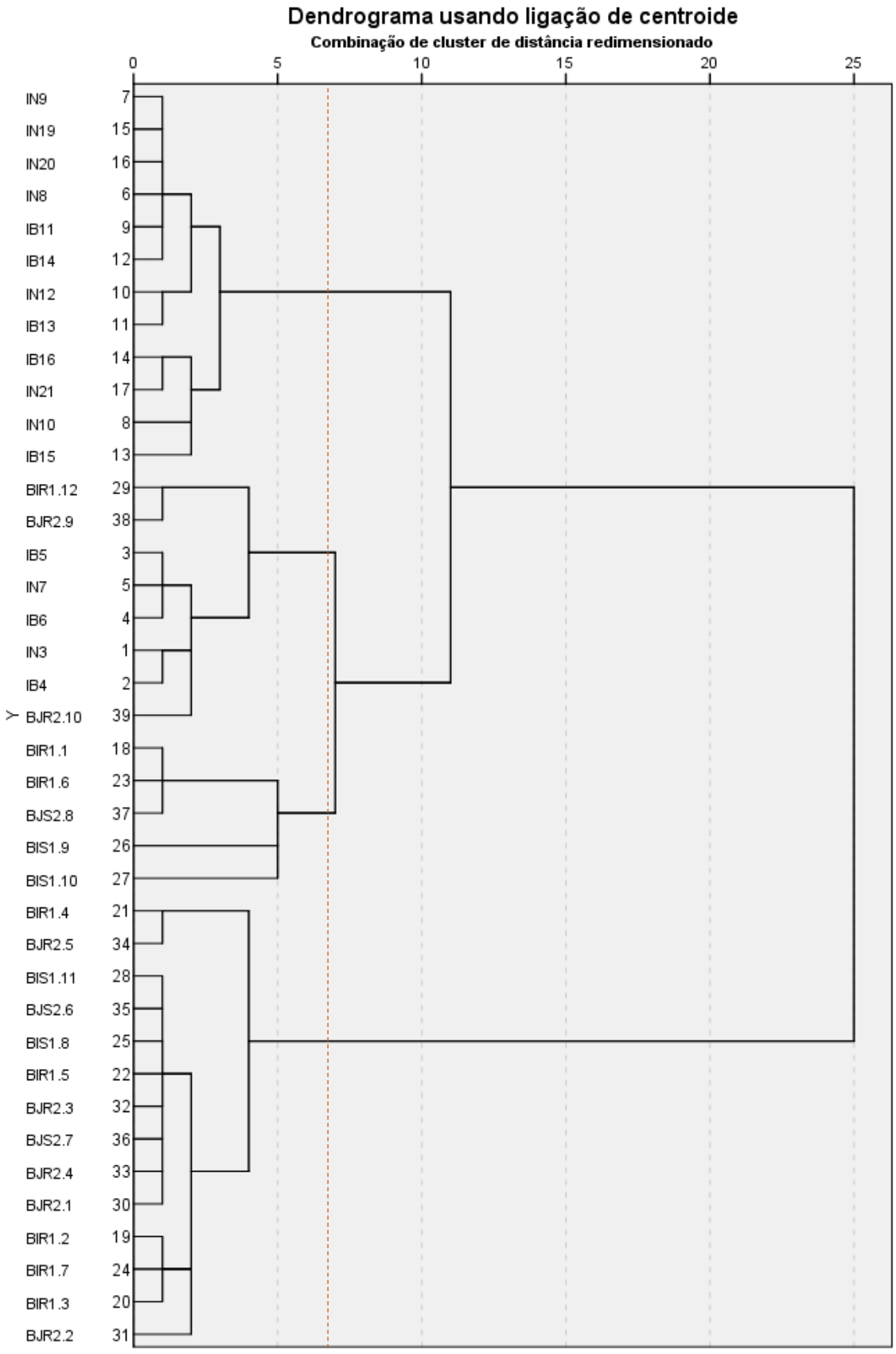

Figura 5.3 - Dendrograma das amostras retirando o ácido glutâmico (glu) para os dados recolhidos com o algoritmo do método do centróide

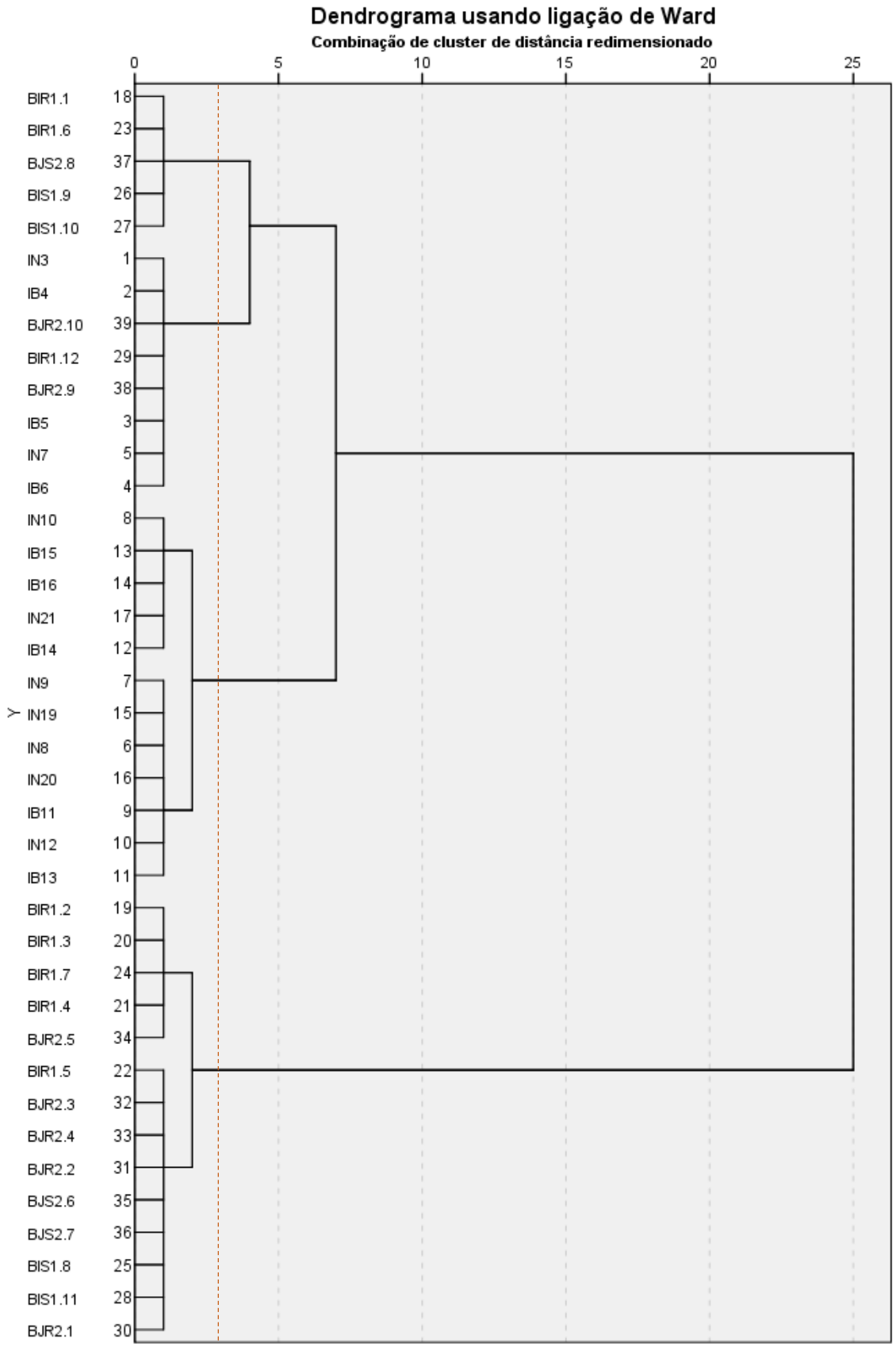

Figura 5.4 - Dendrograma das amostras retirando o ácido glutâmico (glu) para os dados padronizados com o algoritmo do método de Ward

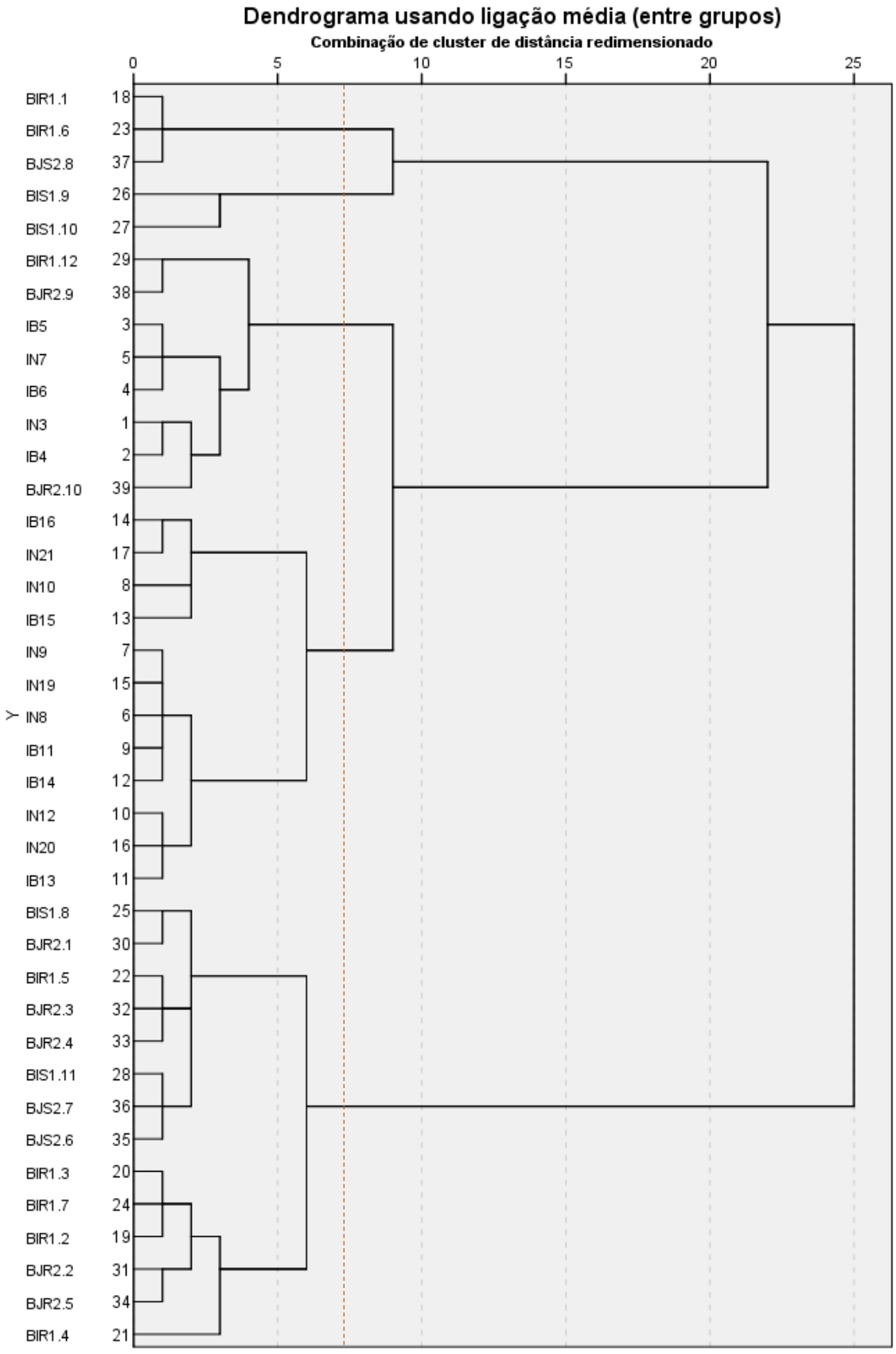

Figura 5.5 - Dendrograma das amostras com todas as variáveis para os dados recolhidos com o algoritmo da ligação média entre grupos

Para concluir, verificam-se resultados iguais para os diversos dados e hipóteses, obtidos através dos 3 métodos escolhidos, são eles:

- Sem a variável Glu, método de Ward para ambos os dados (em bruto e padronizados);
- Sem a variável Glu, método da ligação média entre grupos para ambos os dados (em bruto e padronizados);
- Sem a variável Glu, método do centróide para os dados em bruto;
- Com todas as variáveis, método de Ward para ambos os dados (em bruto e padronizados);
- Com todas as variáveis, método da ligação média entre grupos para ambos os dados (em bruto e padronizados).

Os restantes métodos não originaram resultados exatamente iguais, no entanto foram bastante semelhantes, já que por vezes 2 *clusters* obtidos nesses métodos correspondem a um outro resultante de outros métodos.

De seguida é apresentada a Tabela 5.24 que caracteriza os 4 *clusters* identificados na Tabela 5.23, para todas as variáveis (como os resultados sem o ácido glutâmico e com este foram bastante idênticos, é usado também na caracterização).

Tabela 5.24 - Caracterização dos *clusters* obtidos

AA	Cluster 1	Cluster 2	Cluster 3	Cluster 4
	Arroz branco I	Mistura	Arroz integral	Arroz branco II
His	209,03 ± 22,79	326,06 ± 24,88	410,1 ± 15,63	164,94 ± 13,41
Ser	397,62 ± 43,68	382,62 ± 17,2	432,27 ± 23,02	301,15 ± 29,07
Arg	668,02 ± 48,47	668,16 ± 55,15	751,5 ± 39,25	517,13 ± 47,58
Gly	344,28 ± 18,23	351,21 ± 28,14	404,55 ± 22,44	279,57 ± 26,85
Asp	704,27 ± 74,35	601,87 ± 34,65	660,21 ± 68,59	529,89 ± 54,34
Glu	1585,24 ± 159,16	1164,8 ± 74,48	1273,36 ± 117,01	1160,1 ± 117,04
Thr	227,06 ± 16,69	242,72 ± 11,67	295,67 ± 12,46	178,97 ± 17,56
Ala	412,35 ± 35,25	359,37 ± 14,72	400,16 ± 28,52	317,18 ± 27,6
Pro	330,02 ± 24,53	329,19 ± 11,59	378,11 ± 24,48	255,25 ± 22,55
Cys	46,78 ± 3,4	105,27 ± 49,37	246,27 ± 17,3	39,93 ± 4,17
Lys	177,71 ± 36,06	98,75 ± 13,39	82,34 ± 23,13	117,65 ± 29,34
Tyr	375,46 ± 17,83	504,53 ± 34,36	613,75 ± 23,07	284,86 ± 32,59
Met	171,27 ± 12,9	257,97 ± 14,62	333,2 ± 29,82	130,63 ± 21,57
Val	382,74 ± 27,97	322,54 ± 34,93	375,07 ± 24,24	283,52 ± 24,17
Ile	282,51 ± 20,57	226,64 ± 25,1	273,57 ± 15,38	214,55 ± 18,18
Leu	591,82 ± 64,98	520,32 ± 30,56	584,55 ± 33,43	467,78 ± 42,59
Phe	461,83 ± 28,05	513,35 ± 39,24	612,36 ± 26,97	359,01 ± 34,9

Para facilitar o entendimento dos resultados obtidos na Tabela 5.24, foi construído um gráfico apresentado na Figura 5.6, que apresenta o perfil de cada *cluster* para o conjunto dos aminoácidos.

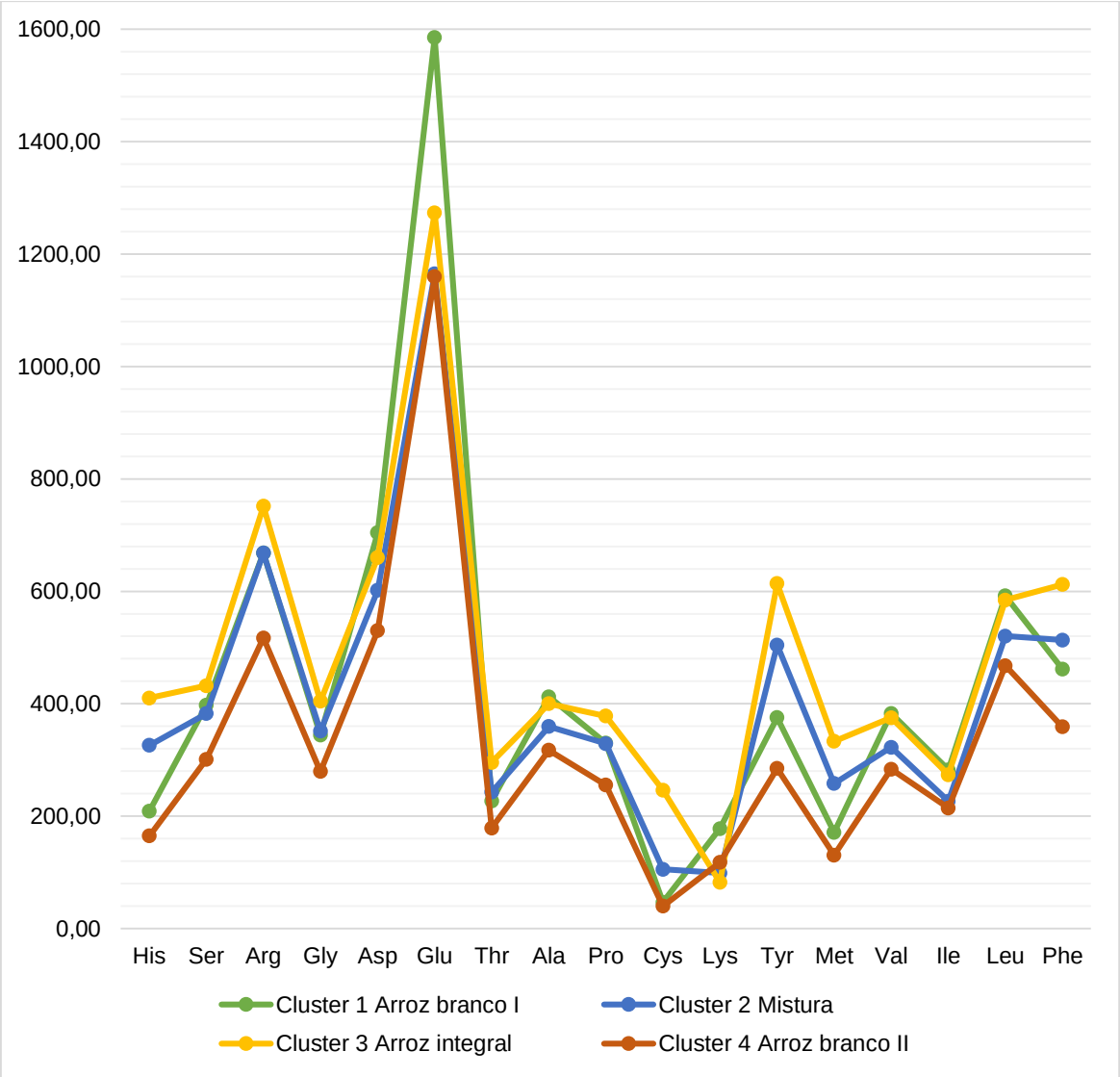

Figura 5.6 - Gráfico da caracterização feita ao arroz pela análise de *clusters*

Como seria de esperar o *cluster* de arroz integral é o que apresenta maiores concentrações para todos os aminoácidos, à exceção do ácido glutâmico. Entre os *clusters* de arroz branco há uma ligeira diferença para todos os aminoácidos, sendo que o cluster designado “Arroz branco I” possui concentrações mais elevadas, ou seja, no cluster 1 estão presentes as amostras de arroz branco mais ricas em proteína.

5.7. *k-Nearest Neighbors*

Para finalizar a análise dos dados, tem-se a criação de um modelo não-paramétrico (por tudo o que já foi visto e analisado sobre os dados) e a avaliação do mesmo. Para começar, são realidades diferentes, o modelo analisa os dados e classifica os restantes (conhecida ou não a sua origem), a avaliação quantifica a quantidade de erros feitos pelo modelo em dados conhecidos. Pelo agrupamento dos dados (amostras) que foram feitas, o modelo foi construído para classificar o arroz como integral ou branco. É importante fazer aqui um parenteses, a nível prático classificar o arroz

branco ou integral não faz sentido, porque visualmente se consegue notar as diferenças, como na Figura 2.3 (fazia muito mais sentido classificar o arroz integral como biológico ou não biológico para verificar se existiam rotulagens erradas ou adulteração) mas é o que os dados permitem (pela Tabela 5.18) e pela maneira como se agrupam (pela Tabela 5.23).

No capítulo da metodologia (Capítulo 3) está explicita a função e o modelo para classificar, no entanto aqui, pela quantidade das amostras e o pré-conhecimento do tipo de arroz a que pertencem, é apresentada apenas a avaliação feita ao modelo.

O método usado para a avaliação do modelo é o cálculo da perda (ou percentagem de errados aquando da validação cruzada (*cross validation leave-one-out*). O número de vizinhos para o modelo classificar pode ser definido pelo utilizador, então o estudo passa por fazer variar esse mesmo número de vizinhos e experimentar com e sem as amostras assinaladas como *cluster* de mistura na Tabela 5.23.

Tabela 5.25 - Resultados da avaliação feita ao modelo

kNN Leave-one-out			
n=39		n=31	
Nº de vizinhos	%Perda	Nº de vizinhos	%Perda
1	2.564	1	0
2	0	2	0
3	5.128	3	0
4	2.564	4	0
5	5.128	5	0
6	5.128	6	0
7	5.128	7	0

Os resultados da avaliação estão presentes na Tabela 5.25, e verifica-se claramente que as amostras do *cluster* nomeado “Mistura” geravam confusão e ruído para o modelo. Ou seja, a análise de *clusters* feita previamente tem todo o interesse para a criação de um modelo, já que se fica com a noção de como os dados se comportam. A grande conclusão que se retira daqui é que se viessem novas amostras do laboratório e fossem classificadas com o modelo (com as 31 amostras) possivelmente iriam ser classificadas corretamente, já que a percentagem de erros na avaliação para todos os casos foi de zero.

CAPÍTULO 6 – CONCLUSÕES E RECOMENDAÇÕES

6.1. Conclusões

A quimiometria, aplicação de métodos estatísticos e matemáticos com dados de origem química, tem um interesse e uma aplicação tão elevados ao ponto de ter uma designação própria. Como tal, a aplicação de estatística multivariada comprovou ser uma poderosa ferramenta para, quando possível, caracterizar, comparar, agrupar e avaliar dados provenientes de um produto alimentício – o arroz.

O arroz para análise é proveniente de produtores (arroz branco) e de estabelecimentos comerciais (arroz integral), e todas as fases até à conceção dos dados para a presente dissertação foram realizadas no INSA, laboratório de referência nacional.

As características químicas escolhidas do cereal em estudo, o arroz, foram as concentrações dos diversos aminoácidos que constituem as proteínas, requisito de uma boa alimentação, e as concentrações de arsénio, por se tratar de um elemento tóxico. Para facilitar e conseguir aprofundar o estudo exaustivo das amostras, foram feitas divisões tendo em conta as características do arroz. Isto fez com que fossem encontradas evidências sobre determinadas suposições que à partida não se viam.

Por vezes, e por obrigatoriedade, o estudo teve de ser remetido para estatística não-paramétrica uma vez que os dados possuem um nível de complexidade superior ao que por vezes é estudado nos exemplos designados a aprendizagem, em que estes últimos geralmente seguem uma distribuição normal.

Anteriormente à caracterização das populações dos vários tipos de arroz pelas médias e desvios padrão esta, foram verificados os pressupostos da estatística paramétrica (normalidade e homoegeneidade da variância) nas diversas variáveis e casos. Contudo existe uma excepção que teve de ser verificada previamente aos pressupostos: existência de diferenças significativas entre as regiões para cada variedade de arroz branco. Essa verificação ocorreu para que se pudesse posteriormente agrupar os dados de arroz branco por variedade, já que entre regiões não foram encontradas diferenças significativas.

Foram usadas a estatística paramétrica e não-paramétrica em tudo, sendo os resultados destas bastante semelhantes. A escolha de usar os dois tipos de estatísticas deveu-se à mistura de variáveis que em cada hipótese eram remetidas para os diferentes tipos de estatística, como se pode verificar na Tabela 6.1.

Tabela 6.1 - Quadro resumo das variáveis por hipótese remetidas para cada tipo de estatística

Arroz integral e branco		Arroz branco índico e japónico		Arroz integral biológico e não biológico	
Estatística não-paramétrica	Histidina	Estatística não-paramétrica	Histidina	Estatística não-paramétrica	Cisteína
	Metionina		Metionina		Lisina
	Ácido aspártico		Ácido glutâmico	Estatística paramétrica	Histidina
	Treonina		Alanina		Metionina
	Alanina	Estatística paramétrica	Treonina		Ácido aspártico
	Prolina		Prolina		Treonina
	Cisteína		Cisteína		Alanina
	Isoleucina		Isoleucina		Prolina
	Serina		Serina		Isoleucina
	Arginina		Arginina		Serina
	Lisina		Lisina		Arginina
Estatística paramétrica	Glicina		Glicina		Glicina
	Ácido glutâmico		Ácido aspártico		Ácido glutâmico
	Tirosina		Tirosina		Tirosina
	Valina		Valina		Valina
	Leucina		Leucina		Leucina
	Fenilalanina		Fenilalanina		Fenilalanina

Na comparação de médias descobriu-se que, a nível proteico (aminoácidos), existem diferenças significativas entre arroz branco e arroz integral em todos os aminoácidos com excepção do ácido glutâmico. Estas diferenças devem-se sobretudo ao processo de branqueamento a que o arroz branco é submetido. Nas restantes hipóteses testadas não foram encontradas mais diferenças significativas entre aminoácidos. No que toca ao elemento arsénio, nada se pode afirmar sobre diferenças significativas entre tipos , variedade de arroz branco ou tipos de agricultura do arroz integral, o que acaba por ser interessante do ponto de vista do consumidor.

Ao correlacionar os aminoácidos com o arsénio, as amostras arroz integral correlacionam-se positivamente, e as amostras de arroz branco negativamente. Tal situação deve-se ao facto do arsénio não apresentar diferenças significativas nas várias hipóteses, e quando comparado com as concentrações de aminoácidos correlaciona-se positivamente com o arroz que possui maior teor (arroz integral) e negativamente com o que possui menor (arroz branco), gerando valores aproximadamente inversos entre si. Os valores das correlações encontradas podem ser vistos na Tabela 6.2, apresentada de seguida.

Tabela 6.2 - Quadro resumo das correlações encontradas no estudo

Arroz		Arroz Branco		Arroz Integral	
Branco	Integral	Indico	Japónico	Não biológico	Biológico
Lisina	Sem correlações encontradas	Sem correlações encontradas	Ácido aspártico	Sem correlações encontradas	Ácido aspártico
			Ácido glutâmico		Ácido glutâmico
			Alanina		Alanina
			Cisteína		Valina
			Valina		
			Leucina		

Na análise de *clusters*, recorreu-se a três algoritmos distintos com o intuito de encontrar concordância e simplificar a escolha dos clusters, por sua vez dividiu-se esta análise em duas partes, analisar como se agrupavam os aminoácidos (variáveis do estudo) e as amostras (casos). Na análise às variáveis, através de um processo iterativo, verificou-se que estas agrupam-se em três: em que um admite variáveis com concentrações entre os 100 e os 400 mg/100g; o segundo que agrupa variáveis cujas concentrações andam entre os 400 e os 650 mg/100g; e por fim, um que apenas contempla o ácido glutâmico cuja concentração é 1250 mg/100g. Na segunda fase, análise aos clusters, várias hipóteses foram fundadas: estudar o agrupamento com todas as variáveis e sem o ácido glutâmico (que não dava significativamente diferente), e para cada uma estudar com os dados que foram recolhidos e os dados após padronização. A padronização foi feita para tentar remover a influência que cada variável tinha no agrupamento. A conclusão, após a análise de 12 dendrogramas, é que se formam 4 *clusters*: um de amostras de arroz integral, 2 de amostras de arroz branco (sendo que um contém um maior teor proteico que o outro) e ainda um onde estão presentes amostras dos dois tipos de arroz. Este último cluster foi designado “mistura” e será importante a sua remoção para o que se segue, a criação e avaliação de um modelo que permita distinguir o arroz com base na leitura cromatográfica. A dimensão de cada cluster pode ser consultada na Tabela 6.3, sendo que o *cluster* 2 é formado por 3 amostras de arroz branco e 5 de arroz integral.

Tabela 6.3 - Dimensão dos diferentes clusters formados

Cluster 1	*Cluster* 2	*Cluster* 3	*Cluster* 4
Arroz branco I	Mistura	Arroz integral	Arroz branco II
5 amostras	8 amostras	12 amostras	14 amostras

Por fim, o modelo k-NN, concebido com o intuito de classificar as amostras pelo tipo com base nas concentrações medidas pela análise cromatográfica, tem duas vertentes: a classificação e a veracidade dessa mesma classificação. Por se ter previamente a classificação de todas as amostras,

foi avaliada a potência do modelo com base na percentagem de erros. E conseguiu atingir-se um modelo cuja percentagem de erros fosse nula, modelo esse onde as 8 amostras que se encontravam no *cluster* “mistura”.

6.2. Recomendações

No que toca a recomendações e melhorias para o futuro estas são feitas por áreas. No laboratório, a grande dificuldade passou pela análise das folhas de cálculo, pelo que seria conveniente os dados fossem tratados quando são recolhidos para que não aconteça possuírem dados que não podem ser analisados estatisticamente por não estarem extrapolados dos valores dados pelos aparelhos para as unidades usadas.

Para trabalhos futuros, está aqui um ponto de partida ou um paralelismo, já que o mesmo pode ser feito em outros cereais ou produtos alimentares. Para quem tome este trabalho como base, será interessante conseguir através de outros dados químicos (já que estes não permitiram) criar um modelo para verificação de adulteração no arroz de cultivo biológico, ou para classificação do arroz por região de produção. Por outro lado, pode analisar-se igualmente o perfil de aminoácidos e sua correlação com o arsénio em arroz cozinhado, com o intuito de verificar se a concentração de arsénio (e suas correlações) variam, já que durante a cozedura a água pode apresentar elevados níveis de arsénio que façam com que a concentração deste se agudize no grão.

BIBLIOGRAFIA

Abelquist, E. W. (2001). *Decommissioning Health Physics: A Handbook for MARSSIM Users* (2nd Ed., pp 364–365). Taylor & Francis.

Adriano, D. C. (2001). *Trace Elements in Terrestrial Environments: Biogeochemistry, Bioavailability, and Risks of Metals* (2nd Ed., pp 2–5). Springer.

Agarwal, B. L. (2006). *Basic Statistics* (4th Ed., p 218). New Age International.

Almeida, A., Elian, S., & Nobre, J. (2008). Modificações e alternativas aos testes de Levene e de Brown e Forsythe para igualdade de variâncias e médias. *Revista Colombiana de Estadística*, *31*(2), 241–260.

Almeida, & Marques, P. (2013). A importância da cultura do arroz em Portugal e no Mundo. Em *Boas Práticas no Cultivo de Arroz por Alagamento, em Portugal*. INIAV - Instituto Nacional de Investigação Agrária e Verinária, I.P.

APARROZ. (2007). Agricultura biológica é mercado a explorar. *APARROZ Website*. Obtido 29 de Abril de 2014, de http://www.aparroz.org/index.php?option=com_content&task=view&id=30&Itemid=1

Arrozeiras Mundiarroz. (sem data). Tipos e Variedades de Arroz. Obtido 3 de Março de 2014, de http://arrozeiras-mundiarroz.pai.pt/ms/ms/arrozeiras-mundiarroz-sa-tipos-e-variedades-de-arroz-2100-051-coruche/ms-90048939-p-8/

Atkins, P. W., & Jones, L. (1999). *Princípios de química: questionando a vida moderna e o meio ambiente* (5ª Edição., pp 381–382). Bookman.

ATSDR, & EPA. (2007). *Toxicological Profile for Arsenic* (pp 1–2). Atlanta, USA.

Balch, P. A. (2006). *Prescription for Nutritional Healing* (Fifth., pp 51–61). Avery.

Bayarri, M. J., & Berger, J. O. (2000). P Values for Composite Null Models. *Journal of the American Statistical Association*, *95*(452), 1127.

Bekiro, N. (2001). Multiple t tests or ANOVA (analysis of variance)? *Turkish Respiratory Journal*, *2*(1), 21–22.

Berrueta, L. a, Alonso-Salces, R. M., & Héberger, K. (2007). Supervised pattern recognition in food analysis. *Journal of Chromatography. A*, *1158*(1-2), 196–214.

Boogers, I., Plugge, W., Stokkermans, Y. Q., & Duchateau, A. L. L. (2008). Ultra-performance liquid chromatographic analysis of amino acids in protein hydrolysates using an automated pre-column derivatisation method. *Journal of Chromatography A*, *1189*(1-2), 406–409.

Borradaile, G. J. (2003). *Statistics of Earth Science Data: Their Distribution in Time, Space and Orientation* (First., p 159). Springer.

Bradley, T. (2007). *Essential Statistics for Economics, Business and Management* (1st ed, p 439). John Wiley & Sons.

Brown, M. B., & Forsythe, A. B. (1974). Robust Tests for the Equality of Variances. *Journal of the American Statistical Association*, *69*(346), 364–367.

Campo, B. do. (sem data). Cultivo do Arroz. Obtido 2 de Março de 2013, de http://bordadocampo.com/arroz/cultivo-arroz/

Campos, L. S. (2009). *Entender a Bioquímica, 5ª Edição* (5ª Edição., p 683). Lisboa: Escolar Editora.

Cheajesadagul, P., Arnaudguilhem, C., Shiowatana, J., Siripinyanond, A., & Szpunar, J. (2013). Discrimination of geographical origin of rice based on multi-element fingerprinting by high resolution inductively coupled plasma mass spectrometry. *Food Chemistry*, *141*(4), 3504–9.

Cios, K. J., Pedrycz, W., Swiniarski, R. W., & Kurgan, L. A. (2007). *Data Mining: A Knowledge Discovery Approach* (Third., p 337). Springer.

Corrar, L. J., Paulo, E., & Filho, J. M. D. (2007). *Análise Multivariada para os cursos de Administração, Ciências Contábeis e Economia*. (Editora Atlas, Ed) (1ª Edição., p 568).

Cotarroz. (sem data-a). Cultivo e Transformação. Obtido 3 de Março de 2014, de http://www.cotarroz.pt/rubrica.aspx?id_rubrica=61&id_seccao=35

Cotarroz. (sem data-b). Tipos e Utilização de Arroz. Obtido 3 de Março de 2014, de http://www.cotarroz.pt/rubrica.aspx?id_rubrica=65&id_seccao=35

Croux, C., & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications*, *19*(4), 497–515.

De Muth, J. E. (2006). *Basic Statistics and Pharmaceutical Statistical Applications* (Second., pp 107–112). Taylor & Francis.

Decreto-Lei n. 62/2000 de 19 de Abril. Diário da República nº 93/00 - I Série A (2000). Portugal.

Devore, J. (2011). *Probability and Statistics for Engineering and the Sciences* (p 329). Cengage Learning.

Diniz, V. W. B., Filho, H. A. D., Müller, R. C. S., Fernandes, K. G., & Palheta, D. C. (2013). Classificação multivariada de ervas medicinais da Região Amazônica e suas infusões de acordo com sua composição mineral. *Química Nova*, *36*(2), 257–261.

Drumond, V. L. M. M. (2012). *Dissertação de mestrado em Tecnologia e Segurança Alimentar: Presença de aflatoxinas em arroz e cereais Importados na União Europeia - Revisão bibliográfica e análise de dados RASFF*. Faculdade de Ciências e Tecnologia - Universidade Nova de Lisboa.

Dwivedi, S., Mishra, A., Tripathi, P., Dave, R., Kumar, A., Srivastava, S., … Nautiyal, C. S. (2012). Arsenic affects essential and non-essential amino acids differentially in rice grains: inadequacy of amino acids in rice based diet. *Environment International*, *46*, 16–22.

Engmann, S., & Cousineau, D. (2011). Comparing Distributions: The Two Sample Anderson-Darling Test as an alternative to the Kolmogorov-Smirnoff test. *Journal of Applied Quantitativa Methods*, *6*(3), 1–17.

Estivill-Castro, V. (2003). Why so many clustering algorithms — A Position Paper. *SIGKDD Explorations*, *4*(1), 65–75.

FAO/WHO. (2012). *Proposed draft Maximum Levels for Arsenic in Rice - CX/CF 12/6/8* (p 1). Maastricht, Netherlands.

FAOSTAT. (2012). Food and Agricultural commodities production. Obtido 3 de Março de 2014, de http://faostat.fao.org/site/339/default.aspx

Feldman, R. M., & Valdez-Flores, C. (2009). *Applied Probability and Stochastic Processes* (Second., pp 95–97). Springer-Verlag.

Fennema, O. R. (1996). *Food Chemistry, Third Edition* (Third., pp 1–8). Taylor & Francis. Obtido de http://books.google.pt/books?id=1OhFPZ7tFz8C

Ferreira, L., & Hitchcock, D. B. (2009). A comparison of hierarchical for clustering functional data. *Communications in Statistics—Simulation and Computation*, *38*, 1925–1949.

Filho, D. B. F., & Júnior, J. A. da S. (2009). Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r)*. *Revista Política Hoje*, *18*(1), 115–146.

Gonzálvez, A., Armenta, S., & Guardia, M. d. la. (2011). Geographical traceability of «Arròs de Valencia» rice grain based on mineral element composition. *Food Chemistry*, *126*(3), 1254–1260.

Gravetter, F., & Wallnau, L. (2010). *Essentials of Statistics for the Behavioral Sciences* (8th ed, pp 472–477). Cengage Learning.

Gredilla, A., Fdez-Ortiz de Vallejuelo, S., Diego, A., Madariaga, J. M., & Amigo, J. M. (2013). Unsupervised pattern-recognition techniques to investigate metal pollution in estuaries. *Trends in Analytical Chemistry*, *46*, 59–69.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate Data Analysis* (Sixth., pp 37–40, 79–88, 408–410, 555–567). Pearson Prentice Hall.

Hauke, J., & Kossowski, T. (2011). Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. *Quaestiones Geographicae*, *30*(2), 87–93.

Hecke, T. Van. (2012). Power study of anova versus Kruskal-Wallis test. *Journal of Statistics Ans Management Systems*, *15*(2&3), 241–257.

Heftmann, E. (2004). *Chromatography: Fundamentals and applications of chromatography and related differential migration methods - Part A: Fundamentals and techniques* (First., pp 96–98). Elsevier Science. Obtido de http://books.google.pt/books?id=BOlzy9W6l6oC

Heikens, A. (2006). Arsenic contamination of irrigation water, soil and crops in Bangladesh: Risk implications for sustainable agriculture and food safety in Asia. *FAO - RAP Publication 2006/20*, *20*, 2.

Howell, D. C. (2012). *Statistical Methods for Psychology* (8th ed, pp 346–352). Cengage Learning.

Huang, G., & Paes, Â. T. (2009). Por dentro da estatística. *Einstein: Educação Continuada Em Saúde*, *7*(2), 63–64.

INSA. (sem data-a). Alimentação e Nutrição. Obtido 29 de Abril de 2014, de http://www.insa.pt/sites/INSA/Portugues/AreasCientificas/AlimentNutricao/Paginas/inicio.aspx

INSA. (sem data-b). INSA. Obtido 29 de Abril de 2014, de http://www.insa.pt/sites/INSA/Portugues/QuemSomos/Paginas/INSA.aspx

Insel, P. M., Turner, R. E., & Ross, D. (2004). *Nutrition* (2nd Ed., pp 207–208). Jones and Bartlett.

Islam, T. U. (2011). Normality Testing- A New Direction. *International Journal of Business and Social Science*, *2*(3), 115–118.

João, L., & Azambuja, M. O. De. (2005). Controle Biológico de Pragas e Invasoras do Arroz Irrigado com o Marreco-de-Pequim. *Extensão Rural E Desenvolvimento Sustentável*, *1*(4), 21–25.

Krull, I. S. (1991). *Trace Metal Analysis and Speciation* (First., pp 50–55). Elsevier Science.

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, *47*(260), 583–621.

Kumar, N., Bansal, A., Sarma, G. S., & Rawal, R. K. (2014). Chemometrics tools used in analytical chemistry: An overview. *Talanta*, *123*, 186–199.

Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics (Oxford, England)*, *24*(5), 719–20.

Lee, I., We, G. J., Kim, D. E., Cho, Y.-S., Yoon, M.-R., Shin, M., & Ko, S. (2012). Classification of rice cultivars based on cluster analysis of hydration and pasting properties of their starches. *LWT - Food Science and Technology*, *48*(2), 164–168.

Lino, M. M. R. M. de O. (2009). INE 7001 Introdução e Análise Exploratória de Dados. *Universidade Federal de Santa Catarina - Departamento de Informática e Estatística*. Obtido 27 de Março de 2014, de http://www.inf.ufsc.br/~marcelo/Caps1_e_2.pdf

Liu, B. (2011). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data* (2nd ed, pp 124–125). Springer.

Lomax, R. G., & Hahs-Vaughn, D. L. (2013). *Statistical Concepts: A Second Course* (Third., pp 192–193). Taylor & Francis.

Maimon, O., & Rokach, L. (2006). *Data Mining and Knowledge Discovery Handbook* (First., pp 330–332). Springer.

Mardia, K. V, Kent, J. T., & Bibby, J. M. (1980). *Multivariate Analysis* (1st Ed., Vol 97, pp 1–4). Paperback.

Marques, P. (Cotarroz). (2009). A cultura do arroz. *Voz Do Campo*, *Setembro*, 12–13.

Martins, I. (2012). Carolino ou Agulha, a escolha que faz a diferença. *Oleiros Magazine*. Obtido 3 de Março de 2014, de http://www.oleirosmagazine.com/oleiros-magazine/abril-2012/opiniao/carolino-ou-agulha,-a-escolha-que-faz-a-diferenca.aspx

Matsushita, K., Puri, M. L., & Hayakawa, T. (1993). *Statistical Sciences and Data Analysis: Proceedings of the Third Pacific Area Statistical Conference* (1st Ed.). VSP International Science.

Montgomery, D. C., & Runger, G. C. (2003). *Applied Statistics and Probability for Engineers* (Third Ed., pp 337, 470–473).

Mooi, E., & Sarstedt, M. (2011). *A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics* (1st Ed., pp 237–240). Springer.

Morais, C. M. (2005). Escalas de Medida , Estatística Descritiva e Inferência Estatística. *Biblioteca Digital Instituto Politécnico de Bragança*. Obtido 14 de Março de 2014, de https://bibliotecadigital.ipb.pt/bitstream/10198/7325/1/estdescr.pdf

Moret, S., Prevarin, A., & Tubaro, F. (2011). Levels of creatine, organic contaminants and heavy metals in creatine dietary supplements. *Food Chemistry*, *126*(3), 1232–1238.

Nelson, D. L., Lehninger, A. L., & Cox, M. M. (2008). *Lehninger, Principles of Biochemistry* (Fifth., pp 14, 72–82). W. H. Freeman.

Neto, P. V. (2004). Estatística Descritiva : Conceitos Básicos. *Unieducacional*. São Paulo. Obtido 4 de Março de 2014, de http://uni.educacional.com.br/up/59960001/3103751/Apos_Est_I_Fev04_C1.pdf

Nollet, L. M. L., & Toldra, F. (2012). *Food Analysis by HPLC, Third Edition* (Third Ed., p 737). Taylor & Francis.

Novarroz. (sem data-a). A produção de Arroz em Portugal. Obtido 3 de Março de 2014, de http://novarroz.pt/mundo-do-arroz/historia-do-arroz/a-producao-de-arroz-em-portugal/

Novarroz. (sem data-b). Arroz, o amigo inseparável. Obtido 3 de Março de 2014, de http://novarroz.pt/mundo-do-arroz/historia-do-arroz/arroz-o-amigo-inseparavel/

Novarroz. (sem data-c). História do Arroz em Portugal. Obtido 3 de Março de 2014, de http://novarroz.pt/mundo-do-arroz/historia-do-arroz/historia-do-arroz-em-portugal/

Novarroz. (sem data-d). Tipos e Variedades de Arroz. Obtido 3 de Março de 2014, de http://novarroz.pt/mundo-do-arroz/arroz-no-mundo/tipos-e-variedades-de-arroz/

Pais, I., & Jones, J. B. (1997). *The Handbook of Trace Elements* (1st Ed., pp 1–3). Taylor & Francis.

Park, H. M. (2009). *Comparing Group Means: T-tests and One-way ANOVA Using Stata, SAS, R, and SPSS** (pp 1–51). Working Paper. Obtido de http://www.indiana.edu/~statmath/stat/all/ttest

Pereira, Z. L., & Requeijo, J. G. (2012). *Qualidade: Planeamento e Controlo Estatístico de Processos*. (FFCT - Fundação da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, Ed) (2ª Edição., pp 156–160).

Razali, N. M., Wah, Y. B., & Sciences, M. (2011). Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov , Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 21–33.

Reaño, R., Sackville, R., & Romero, G. (2008). Directrizes para regeneração - Arroz. Em *Crop specific regeneration guidelines (CGIAR). System-wide Genetic Resource Programme (SGRP)* (pp 1–12). Dulloo M.E., Thormann I., Jorge M.A. and Hanson J., editors.

Reddy, T. A. (2011). *Applied Data Analysis and Modeling for Energy Engineers and Scientists* (1st ed). Springer.

Regulamento (CE) Nº 834/2007 do Conselho de 28 de Junho. (2007). Regulamento (CE) Nº 834/2007. *Jornal Oficial Da União Europeia*.

Rencher, A. C. (2005). *Methods of Multivariate Analysis, Second Edition*. *IIE Transactions* (Second Ed., Vol 37, pp 1083–1085). Taylor & Francis.

Saraçli, S., Dogan, N., & Dogan, I. (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications*, *1*(203), 1–8.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p Values for Testing Precise Null Hypotheses. *The American Statistician*, *55*(1), 62–71.

Sen, A., & Srivastava, S. (1990). *Regression Analysis: Theory, Methods, and Applications* (First Ed., p 105). Springer.

Shakir, L., Hussain, M., Javeed, A., Ashraf, M., & Riaz, A. (2011). Artemisinins and immune system. *European Journal of Pharmacology*, *668*(1-2), 6–14.

Shapiro, S. S., & Wilk, M. B. (1965). Biometrika Trust An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, *52*(3/4), 591–611.

Shen, F., Ying, Y., Li, B., Zheng, Y., & Zhuge, Q. (2011). Multivariate classification of rice wines according to ageing time and brand based on amino acid profiles. *Food Chemistry*.

Sheskin, D. J. (2003). *Handbook of Parametric and Nonparametric Statistical Procedures: Third Edition* (Third Edit., pp 1073–1078). Taylor & Francis.

Simões, A. C. P. (2014). *Dissertação de Mestrado em Engenharia Alimentar: Avaliação da presença de arsénio em arroz e produtos derivados de arroz*. Instituto Superior de Agronomia - Universidade de Lisboa.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining* (First Ed., pp 487–568). Addison-Wesley Longman Publishing Co., Inc.

Taylor, J. M. G. (1987). Kendall's and Spearman's Correlation Coefficients in the Presence of a Blocking Variable. *Biometrics*, *43*(2), 409–416.

Tokalıoğlu, Ş. (2012). Determination of trace elements in commonly consumed medicinal herbs by ICP-MS and multivariate analysis. *Food Chemistry*, *134*(4), 2504–8.

Trumbo, P., Schlicker, S., Yates, A. a, & Poos, M. (2002). Dietary reference intakes for energy, carbohydrate, fiber, fat, fatty acids, cholesterol, protein and amino acids. *Journal of the American Dietetic Association*, *102*(11), 1621–30.

Vorapongsathorn, T., Taejaroenkul, S., & Viwatwongkasem, C. (2004). A comparison of type I error and power of Bartlett's test, Levene's test and Cochran's test under violation of assumptions. *Songklanakarin J. Sci. Technol.*, *26*(4), 537–547.

Walter, M., Marchezan, E., & Avila, L. A. De. (2008). Arroz : composição e características nutricionais. *Ciência Rural*, *38*(4), 1184–1192.

Ward, J. J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244.

WHO. (2010). Exposure to Arsenic: A major public health concern. Obtido 22 de Abril de 2014, de http://www.who.int/ipcs/features/arsenic.pdf

WHO. (2011). Arsenic in Drinking-water. Obtido 22 de Abril de 2014, de http://www.who.int/water_sanitation_health/dwq/chemicals/arsenic.pdf

WHO, FAO, & UNU. (2007). Protein and Amino Acid Requirements in Human Nutrition. Obtido 20 de Abril de 2014, de http://whqlibdoc.who.int/trs/who_trs_935_eng.pdf

Wilson, K., & Walker, J. (2010). *Principles and Techniques of Biochemistry and Molecular Biology* (7th ed, p 328). Cambridge University Press.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques* (Second., pp 135, 149). Elsevier Science.

Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, *81*(12), 2141–2155.

ANEXOS

Anexo I – Tabelas da recolha por amostra dos aminoácidos

Tabela I.1 - Tabela dos dados recolhidos do Arroz Branco

Amostra	Código	His (mg/100g)	Ser (mg/100g)	Arg (mg/100g)	Gly (mg/100g)	Asp (mg/100g)	Glu (mg/100g)	Thr (mg/100g)	Ala (mg/100g)	Pro (mg/100g)
1.1 (n=2)	CD1	193,06 ± 0,14	368,41 ± 8,63	632,02 ± 13,17	341,82 ± 7,39	657,25 ± 1,72	1488,68 ± 26,27	225,61 ± 2,23	397,45 ± 0,6	322,9 ± 4,63
1.2 (n=2)	CD4	163,17 ± 2,18	285,6 ± 9,18	501,1 ± 3,47	276,48 ± 9,44	454,7 ± 35,57	1026,32 ± 25,33	174,25 ± 3,22	293,67 ± 10,88	243,23 ± 6,76
1.3 (n=2)	ML2	167 ± 14,57	291,93 ± 18,66	514,08 ± 39,53	276,88 ± 11,65	485,16 ± 3,37	1069,65 ± 55,66	172,73 ± 9,31	293,61 ± 15,37	244,45 ± 0,7
1.4 (n=2)	ML4	138,58 ± 3,11	242,13 ± 16,66	424,69 ± 20,56	232,72 ± 12,27	440,22 ± 50,02	946,59 ± 47,04	138,58 ± 9,65	265,08 ± 5,79	213,68 ± 6,52
1.5 (n=3)	QF1	161,46 ± 10,33	299,37 ± 5,3	509,53 ± 29,36	277,09 ± 8,25	551,57 ± 47,46	1170,12 ± 83,77	179,09 ± 14,08	324,69 ± 21,56	263,16 ± 14,43
1.6 (n=4)	QF2	199,04 ± 12,09	371,64 ± 24,88	644,59 ± 43,35	340,4 ± 21,36	659,32 ± 57,48	1470,45 ± 102,24	222,69 ± 15,13	394,77 ± 23,92	323,45 ± 17,17
1.7 (n=2)	3	167,87 ± 10,68	291,3 ± 11,76	521,23 ± 25,1	290,28 ± 13,22	471,73 ± 23,53	1054,16 ± 39,06	182,38 ± 10,04	299,04 ± 6,47	253,64 ± 0,32
1.8 (n=2)	HDL1	183,34 ± 2,15	351,72 ± 0,54	588,29 ± 1,66	322,3 ± 6,58	594,61 ± 7,44	1335,81 ± 72,07	204,22 ± 2,16	359,73 ± 21,44	297,39 ± 15,21
1.9 (n=4)	HDL2	213,28 ± 18,91	420,02 ± 26,81	705,54 ± 57,39	374,82 ± 23,86	792,39 ± 54,89	1749,51 ± 113,77	255,84 ± 26,22	466,51 ± 29,43	370,76 ± 24,2
1.10 (n=2)	HDL4	246,98 ± 10,03	464,34 ± 38,97	733,32 ± 35,12	338,61 ± 16,77	777,29 ± 14,99	1768,06 ± 10,01	216,54 ± 2,95	426,73 ± 3,99	328,48 ± 8,73
1.11 (n=2)	31	176,29 ± 4,42	318,45 ± 6,92	561,19 ± 8,65	307,03 ± 7,42	543,19 ± 34,81	1240,33 ± 65,16	200,46 ± 6,26	338,91 ± 15,88	273,3 ± 10,27
1.12 (n=2)	25	302,32 ± 14,48	353,21 ± 3,13	579,77 ± 7,42	306,49 ± 3,57	572,14 ± 78,93	1137,74 ± 112,95	235,79 ± 1,19	337,26 ± 28,41	312,41 ± 7,74
2.1 (n=2)	2	169,24 ± 10,66	336,67 ± 10,46	538,54 ± 34,92	298,99 ± 20,1	606,28 ± 8,54	1320,31 ± 20,97	198,67 ± 2,57	360,17 ± 1,79	279,79 ± 21,23
2.2 (n=4)	5	155,63 ± 1,65	272,38 ± 10,34	497,43 ± 11,74	258,86 ± 4,77	500,38 ± 11,16	1078,91 ± 45,71	164,57 ± 2,74	292,06 ± 7,29	231,34 ± 4,47
2.3 (n=2)	6	166,81 ± 2,38	306,65 ± 18,86	525,2 ± 15,37	274,48 ± 4,93	561,56 ± 21,7	1185,08 ± 61,14	181,73 ± 9,47	325,25 ± 11,84	258,11 ± 1,68
2.4 (n=2)	8	160,59 ± 9,96	306,85 ± 16,2	497,84 ± 36,48	267,13 ± 14,69	609,06 ± 29,19	1241,13 ± 59,19	182,76 ± 15,49	336,62 ± 12,24	255,22 ± 0,21
2.5 (n=2)	9	142,01 ± 2,21	267,76 ± 3,84	433,04 ± 11,39	229,88 ± 14,6	516,03 ± 7,57	1093,98 ± 28,61	156,48 ± 1,37	299,66 ± 4,93	224,85 ± 8,5
2.6 (n=3)	1	186,34 ± 25,4	326,12 ± 11,29	584,09 ± 66,19	307,6 ± 24,55	539,58 ± 18,38	1265,45 ± 63,99	182,33 ± 14,64	329,83 ± 9,25	272,22 ± 12,22
2.7 (n=2)	30	170,91 ± 18,55	319,17 ± 15,04	543,53 ± 55,19	294,31 ± 16,66	544,36 ± 42,95	1213,53 ± 16,49	187,35 ± 16,23	322,24 ± 2,27	263,18 ± 5,02
2.8 (n=2)	32	192,76 ± 14,14	363,7 ± 19,04	624,65 ± 30,58	325,75 ± 13,45	635,09 ± 2,75	1449,52 ± 4,11	214,64 ± 12,67	376,3 ± 1,71	304,5 ± 2,52
2.9 (n=2)	24	291,79 ± 18,89	366,56 ± 22,44	591,23 ± 44,39	308,36 ± 21,33	616,78 ± 14,31	1241,29 ± 0,2	238,26 ± 14,85	351,75 ± 4,29	320,66 ± 17,96
2.10 (n=2)	26	327,68 ± 16,09	402,67 ± 9,7	671,1 ± 9,7	352,82 ± 0,34	628,13 ± 71,72	1261,07 ± 103,39	263,2 ± 3,69	363,44 ± 22,55	350,55 ± 5,62

Amostra	Código	Cys (mg/100g)	Lys (mg/100g)	Tyr (mg/100g)	Met (mg/100g)	Val (mg/100g)	ILe (mg/100g)	Leu (mg/100g)	Phe (mg/100g)
1.1 (n=2)	CD1	43,97 ± 5,46	155,6 ± 3,11	350,29 ± 52,9	186,02 ± 17,3	373,25 ± 10,34	281,55 ± 5,41	590,66 ± 0,26	442,23 ± 0,62
1.2 (n=2)	CD4	39,11 ± 8,49	99,02 ± 14,84	292,17 ± 32,37	152,27 ± 7,88	270,64 ± 23,86	204,97 ± 17,11	442,48 ± 8,82	360,26 ± 9,86
1.3 (n=2)	ML2	36,35 ± 8,75	83,83 ± 29,33	286,3 ± 57,03	123,4 ± 11,87	270,71 ± 6,64	212,88 ± 3,59	441,56 ± 2,34	376,31 ± 43,91
1.4 (n=2)	ML4	33,59 ± 8,29	79,58 ± 2,01	242,09 ± 30,78	107,64 ± 10,17	228,22 ± 3,37	179,69 ± 4,59	379,32 ± 8,83	306,03 ± 8,61
1.5 (n=3)	QF1	43,18 ± 9,61	129,22 ± 34,96	299,43 ± 22,55	122,37 ± 16,94	285,35 ± 31,81	213,11 ± 25,53	475,22 ± 30,68	348,31 ± 5,48
1.6 (n=4)	QF2	44,84 ± 5,78	164,51 ± 19,37	376,08 ± 41,86	154,99 ± 9,55	374,31 ± 17,84	280,8 ± 13,03	602,77 ± 29,67	445,29 ± 39,46
1.7 (n=2)	3	37,64 ± 1,49	93,07 ± 9,66	305,1 ± 0,8	158,77 ± 0,99	284,16 ± 9,8	223,25 ± 6,33	454,07 ± 8,19	380,02 ± 23,3
1.8 (n=2)	HDL1	42,22 ± 8,35	117,29 ± 21,06	323,42 ± 40,34	152,58 ± 8,44	320,32 ± 30,72	242,26 ± 13,37	535,87 ± 34,04	410,99 ± 0,45
1.9 (n=4)	HDL2	52,59 ± 10,47	220,71 ± 43,41	385,55 ± 25,96	174,54 ± 14,68	428,93 ± 45,88	316,79 ± 31,12	689,39 ± 51,92	468,73 ± 28,94
1.10 (n=2)	HDL4	46,67 ± 1,49	210,34 ± 8,16	397,38 ± 7,05	179,65 ± 1,03	383,4 ± 11,28	264,18 ± 3,11	510,9 ± 3,17	508,19 ± 1,3
1.11 (n=2)	31	39,12 ± 1,49	128 ± 10,5	301,15 ± 13,47	163,39 ± 0,6	314,56 ± 20,56	238,97 ± 10,48	500,33 ± 19,87	385,68 ± 6,46
1.12 (n=2)	25	45,93 ± 2,31	90,68 ± 40,34	465,73 ± 18,75	256,79 ± 18,65	304,95 ± 10,98	221,57 ± 4,47	493,87 ± 16,25	458,83 ± 16,33
2.1 (n=2)	2	48,42 ± 0,74	155,01 ± 11,34	285,28 ± 1,03	143,24 ± 6,76	311,6 ± 5,55	221,47 ± 8,88	526,69 ± 11,01	362,72 ± 16,4
2.2 (n=4)	5	34 ± 3,94	174,26 ± 87,35	275,99 ± 11,93	108,3 ± 1,77	276,09 ± 8,65	207,75 ± 5,02	429,39 ± 9,29	326,45 ± 18,01
2.3 (n=2)	6	39,43 ± 3,64	142,24 ± 16,6	257,31 ± 16,1	105,96 ± 7,42	286,16 ± 15,35	213,7 ± 10,76	476,48 ± 23,8	345,62 ± 14,31
2.4 (n=2)	8	45,48 ± 2,9	145,82 ± 6,95	258,76 ± 30,08	128,85 ± 17,73	288,95 ± 21,05	213,22 ± 17,77	483,45 ± 32,42	332,58 ± 29,24
2.5 (n=2)	9	37,49 ± 3,13	112,88 ± 2,13	222,29 ± 2	94,7 ± 3,62	253,11 ± 1,23	181,22 ± 6,7	424,28 ± 6,28	300,2 ± 8,19
2.6 (n=3)	1	42,28 ± 1,3	83,16 ± 23,97	348,8 ± 61,97	132,28 ± 10,35	291,25 ± 19,11	232,43 ± 32,13	503,15 ± 34,73	414 ± 44,63
2.7 (n=2)	30	40,78 ± 2,89	103,77 ± 3,69	289,99 ± 24,3	135,03 ± 31,32	288,2 ± 33,3	218,82 ± 34,1	476,62 ± 23,91	376,99 ± 50,58
2.8 (n=2)	32	45,82 ± 3,5	137,4 ± 5,92	368 ± 11,6	161,14 ± 4	353,84 ± 21,21	269,21 ± 8,31	565,37 ± 10,79	444,71 ± 32,22
2.9 (n=2)	24	43,35 ± 2,83	109,28 ± 7,85	448,63 ± 33,59	250,37 ± 14,35	317,02 ± 16,53	229,48 ± 15,3	515,2 ± 30,6	455,66 ± 33,34
2.10 (n=2)	26	48,47 ± 2,78	95,07 ± 33,83	520,83 ± 17,35	277,04 ± 12,91	343,83 ± 8,38	250,95 ± 2,7	561,99 ± 13,74	523,85 ± 10,4

Tabela I.2 - Tabela dos dados recolhidos do Arroz Integral (n=2)

Amostra	His (mg/100g)	Ser (mg/100g)	Arg (mg/100g)	Gly (mg/100g)	Asp (mg/100g)	Glu (mg/100g)	Thr (mg/100g)	Ala (mg/100g)	Pro (mg/100g)
Amostra 3	357,25 ± 45,58	380,96 ± 9,51	717,99 ± 25,04	372,69 ± 11,05	601,59 ± 32,65	1149,73 ± 49,98	251,43 ± 6,29	378,42 ± 9,06	333,20 ± 4,74
Amostra 4	328,84 ± 23,43	404,58 ± 29,99	720,65 ± 16,70	368,65 ± 13,40	665,73 ± 6,40	1245,66 ± 19,68	251,84 ± 5,96	380,98 ± 3,65	337,81 ± 3,78
Amostra 5	361,87 ± 16,49	385,09 ± 9,16	706,91 ± 5,05	375,37 ± 5,04	572,25 ± 3,59	1076,35 ± 19,95	231,55 ± 8,51	348,69 ± 10,33	326,57 ± 10,02
Amostra 6	308,40 ± 13,90	378,87 ± 0,61	661,73 ± 7,71	355,84 ± 0,61	596,85 ± 20,56	1123,51 ± 42,50	229,49 ± 1,28	357,41 ± 4,42	323,47 ± 0,91
Amostra 7	330,30 ± 34,11	389,06 ± 16,14	695,90 ± 53,49	369,48 ± 36,81	561,50 ± 70,23	1083,02 ± 115,82	240,20 ± 10,40	357,00 ± 21,21	328,82 ± 5,67
Amostra 8	432,29 ± 36,94	436,42 ± 28,96	781,69 ± 74,41	425,42 ± 44,31	621,80 ± 9,37	1237,55 ± 18,46	300,25 ± 18,36	401,23 ± 11,65	392,56 ± 21,60
Amostra 9	407,26 ± 23,80	425,80 ± 19,13	732,44 ± 51,47	399,36 ± 30,95	638,16 ± 29,71	1266,73 ± 13,12	294,12 ± 13,04	407,15 ± 6,15	378,83 ± 13,60
Amostra 10	424,56 ± 20,46	468,97 ± 1,51	805,46 ± 8,10	443,22 ± 8,70	713,74 ± 71,20	1403,96 ± 52,80	317,71 ± 1,44	436,81 ± 23,80	413,34 ± 0,77
Amostra 11	427,44 ± 8,84	431,14 ± 10,04	782,55 ± 8,96	416,84 ± 10,32	611,71 ± 0,27	1188,36 ± 10,10	298,89 ± 3,08	373,72 ± 7,75	380,59 ± 2,82
Amostra 12	389,75 ± 18,14	401,3 ± 14,99	702,38 ± 39,71	377,8 ± 23,64	611,77 ± 29,74	1176,25 ± 63,66	279,83 ± 14,13	365,05 ± 14,87	344,38 ± 32,84
Amostra 13	400,9 ± 1,79	398,32 ± 6,83	676,77 ± 43,97	362,65 ± 11,44	597,01 ± 32,85	1108,04 ± 68,16	274,3 ± 7,2	355,59 ± 25,14	322,15 ± 13,18
Amostra14	411,69 ± 19,23	410,6 ± 2,39	746,17 ± 27,31	383,93 ± 10,09	667,07 ± 67,34	1244,51 ± 115,25	286,04 ± 0,8	384,33 ± 25,81	366,55 ± 2,07
Amostra 15	431,78 ± 12,91	474,62 ± 12	792,99 ± 17,29	423,03 ± 1,58	795,45 ± 55,3	1478,5 ± 102,53	312,76 ± 9,37	433,34 ± 21,15	399,45 ± 18,56
Amostra 16	388,00 ± 49,00	437,19 ± 40,94	786,1 ± 27,04	411,63 ± 14,54	736,65 ± 138,46	1370,83 ± 254,81	288,73 ± 43,36	415,93 ± 45,77	383,89 ± 40,57
Amostra 19	407,89 ± 4,73	437,09 ± 1,04	748,86 ± 12,26	409,68 ± 10,16	615,54 ± 62,06	1239,35 ± 97,03	300,08 ± 0,14	404,29 ± 20,07	388,33 ± 0,32
Amostra 20	399,75 ± 3,37	432,33 ± 16,7	734,83 ± 12,22	407,00 ± 20,69	579,28 ± 48,37	1151,79 ± 58,66	294,41 ± 6,13	382,8 ± 7,49	377,3 ± 0,09
Amostra 21	399,91 ± 12,04	433,46 ± 8,95	727,75 ± 14,56	394,01 ± 6,4	734,4 ± 40,75	1414,41 ± 65,55	300,93 ± 7,69	441,61 ± 19,72	389,91 ± 8,74

Amostra	Cys (mg/100g)	Lys (mg/100g)	Tyr (mg/100g)	Met (mg/100g)	Val (mg/100g)	ILe (mg/100g)	Leu (mg/100g)	Phe (mg/100g)
Amostra 3	137,40 ± 1,90	107,54 ± 19,02	525,08 ± 12,54	252,14 ± 7,07	356,81 ± 2,52	247,18 ± 5,35	537,82 ± 6,32	546,91 ± 17,02
Amostra 4	139,13 ± 14,90	114,13 ± 20,08	526,37 ± 34,83	232,65 ± 12,33	382,81 ± 20,63	262,21 ± 10,74	564,42 ± 15,68	554,44 ± 30,17
Amostra 5	145,89 ± 0,35	71,65 ± 22,90	531,67 ± 7,07	263,72 ± 1,93	292,21 ± 10,85	200,35 ± 8,32	495,88 ± 15,68	541,00 ± 7,11
Amostra 6	133,97 ± 1,70	98,72 ± 3,48	479,87 ± 18,43	254,31 ± 4,63	285,17 ± 4,35	195,09 ± 1,38	489,40 ± 2,17	493,20 ± 11,42
Amostra 7	148,03 ± 7,94	102,95 ± 9,84	538,05 ± 47,50	276,74 ± 19,37	297,54 ± 1,61	206,33 ± 0,18	503,94 ± 3,89	532,93 ± 51,85
Amostra 8	263,97 ± 17,83	70,43 ± 18,89	642,66± 72,93	375,33 ± 30,03	361,58 ± 15,26	265,52 ± 14,01	580,85 ± 27,36	635,85 ± 74,37
Amostra 9	250,82 ± 11,21	82,37 ± 24,19	612,95 ± 43,81	354,17 ± 23,42	356,49 ± 9,47	261,58 ± 5,70	572,71 ± 13,33	604,08 ± 53,36
Amostra 10	246,67 ± 9,50	84,43 ± 12,64	620,01 ± 11,34	370,76 ± 5,52	393,24 ± 11,53	284,78 ± 3,78	632,26 ± 3,84	653,54 ± 19,55
Amostra 11	257,51 ± 0,37	46,19 ± 0,28	636,05 ± 0,92	349,68 ± 9,91	363,03 ± 5,87	266,55 ± 1,64	566,56 ± 2,52	641,80 ± 3,60
Amostra 12	240,15 ± 4,42	79,63 ± 2,85	586,05 ± 11,6	333,53 ± 11,71	341,26 ± 11,95	251,5 ± 9,13	537,91 ± 20,13	580,76 ± 16,8
Amostra 13	250,67 ± 3,22	38,75 ± 7,35	593,24 ± 10,24	288,59 ± 0,15	356,26 ± 12,45	265,99 ± 6,33	535,07 ± 22,86	583,07 ± 4,34
Amostra14	252,36 ± 17,82	110,86 ± 18,52	595,33 ± 42,12	293,23 ± 5,06	374,4 ± 1,25	275,1 ± 3,83	560,64 ± 3,62	592,42 ± 42,83
Amostra 15	254,78 ± 0,63	100,98 ± 5,92	662,59 ± 38,3	325,94 ± 4,98	426,68 ± 5,26	312,67 ± 2,65	643,84 ± 5,03	653,61 ± 39,44
Amostra 16	194,8 ± 85,45	100,77 ± 13,29	603,25 ± 45,62	287,75 ± 49,03	409,5 ± 29,57	283,88 ± 38,06	602,2 ± 63,91	610,99 ± 20,83
Amostra 19	246,49 ± 2,06	74,94 ± 18,83	611,32 ± 10,01	345,3 ± 3,81	369,02 ± 2,27	267,9 ± 2,45	594,7 ± 1,89	601,75 ± 13,15
Amostra 20	250,44 ± 0,75	85,41 ± 20,63	608,36 ± 34,22	337,42 ± 7,93	368,58 ± 18,65	272,51 ± 18,46	583 ± 30,09	604,18 ± 31,82
Amostra 21	246,6 ± 8,19	113,31 ± 3,89	593,19 ± 11,92	336,66 ± 7,36	380,82 ± 11,95	274,86 ± 8,82	604,91 ± 15,04	586,28 ± 14,85

Anexo II – ANOVA's e *t-Student* para comparação de médias

Tabela II.1 - ANOVA a um fator para o tipo de arroz (branco e integral)

Fonte de Variação (Integral – Branco)		SS	g.l.	MS	F	p
His	Entre os níveis	362140,275	1	362140,275	167,783	0,000
	Dentro dos níveis (Erro)	79860,302	37	2158,387		
	Total	442000,577	38			
Ser	Entre os níveis	71146,057	1	71146,057	35,892	0,000
	Dentro dos níveis (Erro)	73341,728	37	1982,209		
	Total	144487,785	38			
Arg	Entre os níveis	283382,425	1	283382,425	63,743	0,000
	Dentro dos níveis (Erro)	164492,187	37	4445,735		
	Total	447874,612	38			
Gly	Entre os níveis	84330,774	1	84330,774	78,750	0,000
	Dentro dos níveis (Erro)	39622,204	37	1070,870		
	Total	123952,977	38			
Asp	Entre os níveis	37491,960	1	37491,960	5,596	0,023
	Dentro dos níveis (Erro)	247910,933	37	6700,295		
	Total	285402,893	38			
Glu	Entre os níveis	9292,862	1	9292,862	0,287	0,596
	Dentro dos níveis (Erro)	1199267,922	37	32412,647		
	Total	1208560,784	38			
Thr	Entre os níveis	62226,824	1	62226,824	66,856	0,000
	Dentro dos níveis (Erro)	34438,086	37	930,759		
	Total	96664,910	38			
Ala	Entre os níveis	20531,047	1	20531,047	12,051	0,001
	Dentro dos níveis (Erro)	63037,467	37	1703,715		
	Total	83568,514	38			
Pro	Entre os níveis	64169,101	1	64169,101	45,156	0,000
	Dentro dos níveis (Erro)	52578,857	37	1421,050		
	Total	116747,958	38			
Cys	Entre os níveis	286911,200	1	286911,200	246,097	0,000
	Dentro dos níveis (Erro)	43136,372	37	1165,848		
	Total	330047,572	38			
Lys	Entre os níveis	16460,726	1	16460,726	14,857	0,000
	Dentro dos níveis (Erro)	40994,967	37	1107,972		
	Total	57455,694	38			
Tyr	Entre os níveis	620612,126	1	620612,126	142,053	0,000
	Dentro dos níveis (Erro)	161648,058	37	4368,866		
	Total	782260,184	38			
Met	Entre os níveis	223809,485	1	223809,485	100,428	0,000
	Dentro dos níveis (Erro)	82456,738	37	2228,560		
	Total	306266,223	38			
Val	Entre os níveis	22482,207	1	22482,207	11,555	0,002
	Dentro dos níveis (Erro)	71992,176	37	1945,734		
	Total	94474,383	38			
Ile	Entre os níveis	6393,685	1	6393,685	6,056	0,019
	Dentro dos níveis (Erro)	39061,457	37	1055,715		
	Total	45455,141	38			
Leu	Entre os níveis	36237,363	1	36237,363	10,094	0,003
	Dentro dos níveis (Erro)	132828,410	37	3589,957		
	Total	169065,773	38			
Phe	Entre os níveis	347725,414	1	347725,414	111,274	0,000
	Dentro dos níveis (Erro)	115623,151	37	3124,950		
	Total	463348,566	38			

Tabela II.2 - ANOVA a um fator para a variedade de arroz branco (japónico e índico)

Fonte de Variação (Japónico - Índico)		SS	g.l.	MS	F	p
His	Entre os níveis	73,772	1	73,772	0,026	0,873
	Dentro dos níveis (Erro)	56453,920	20	2822,696		
	Total	56527,692	21			
Ser	Entre os níveis	699,581	1	699,581	0,236	0,633
	Dentro dos níveis (Erro)	59386,694	20	2969,335		
	Total	60086,275	21			
Arg	Entre os níveis	3578,295	1	3578,295	0,540	0,471
	Dentro dos níveis (Erro)	132561,894	20	6628,095		
	Total	136140,190	21			
Gly	Entre os níveis	1269,955	1	1269,955	0,908	0,352
	Dentro dos níveis (Erro)	27975,727	20	1398,786		
	Total	29245,681	21			
Asp	Entre os níveis	312,739	1	312,739	0,035	0,852
	Dentro dos níveis (Erro)	176289,698	20	8814,485		
	Total	176602,436	21			
Glu	Entre os níveis	15374,081	1	15374,081	0,324	0,575
	Dentro dos níveis (Erro)	947618,462	20	47380,923		
	Total	962992,542	21			
Thr	Entre os níveis	74,066	1	74,066	0,069	0,796
	Dentro dos níveis (Erro)	21619,274	20	1080,964		
	Total	21693,341	21			
Ala	Entre os níveis	1077,384	1	1077,384	0,452	0,509
	Dentro dos níveis (Erro)	47702,631	20	2385,132		
	Total	48780,015	21			
Pro	Entre os níveis	683,767	1	683,767	0,370	0,550
	Dentro dos níveis (Erro)	36998,313	20	1849,916		
	Total	37682,079	21			
Cys	Entre os níveis	1,118	1	1,118	0,045	0,835
	Dentro dos níveis (Erro)	502,040	20	25,102		
	Total	503,158	21			
Lys	Entre os níveis	141,813	1	141,813	0,086	0,772
	Dentro dos níveis (Erro)	32922,930	20	1646,147		
	Total	33064,743	21			
Tyr	Entre os níveis	332,043	1	332,043	0,054	0,818
	Dentro dos níveis (Erro)	122438,185	20	6121,909		
	Total	122770,228	21			
Met	Entre os níveis	293,929	1	293,929	0,117	0,736
	Dentro dos níveis (Erro)	50248,166	20	2512,408		
	Total	50542,095	21			
Val	Entre os níveis	1947,001	1	1947,001	0,842	0,370
	Dentro dos níveis (Erro)	46233,338	20	2311,667		
	Total	48180,339	21			
Ile	Entre os níveis	1427,304	1	1427,304	1,295	0,269
	Dentro dos níveis (Erro)	22042,163	20	1102,108		
	Total	23469,467	21			
Leu	Entre os níveis	985,199	1	985,199	0,197	0,662
	Dentro dos níveis (Erro)	99998,806	20	4999,940		
	Total	100984,005	21			
Phe	Entre os níveis	2030,763	1	2030,763	0,499	0,488
	Dentro dos níveis (Erro)	81451,832	20	4072,592		
	Total	83482,596	21			

Tabela II.3 - Teste *t* de *Student* e teste *F* de Fisher para a variedade de arroz branco (japónico e índico)

AA	Média Arroz Branco Indico	Média Arroz Branco Japónico	*t* de *Student*			Desvio Padrão Indico	Desvio Padrão Japónico	*F* de Fisher	
			t	g.l.	Valor-p			*F*	Valor-p
His	192,70	196,38	-0,162	20	0,873	44,56	62,02	1,937	0,299
Ser	338,18	326,85	0,485	20	0,633	62,73	42,30	2,199	0,247
Arg	576,28	550,67	0,735	20	0,471	90,21	69,16	1,701	0,434
Gly	307,08	291,82	0,953	20	0,352	38,95	35,41	1,210	0,787
Asp	583,30	575,72	0,188	20	0,852	118,54	49,14	5,819	0,013
Glu	1288,12	1235,03	0,570	20	0,575	277,23	106,56	6,769	0,008
Thr	200,68	197,00	0,262	20	0,796	32,73	33,06	1,020	0,958
Ala	349,79	335,73	0,672	20	0,509	60,98	27,49	4,918	0,024
Pro	287,24	276,04	0,608	20	0,550	45,80	39,33	1,357	0,658
Cys	42,10	42,55	-0,211	20	0,835	5,23	4,73	1,225	0,772
Lys	130,99	125,89	0,294	20	0,772	47,89	29,24	2,682	0,149
Tyr	335,39	327,59	0,233	20	0,818	61,60	94,70	2,364	0,180
Met	161,03	153,69	0,342	20	0,736	38,47	61,43	2,550	0,146
Val	319,90	301,01	0,918	20	0,370	58,54	30,79	3,615	0,064
Ile	240,00	223,83	1,138	20	0,269	39,11	24,08	2,637	0,156
Leu	509,70	496,26	0,444	20	0,662	84,73	48,33	3,074	0,103
Phe	407,57	388,28	0,706	20	0,488	58,31	69,96	1,439	0,561

Tabela II.4 – ANOVA a um fator para o tipo de arroz integral (Biológico-Não biológico)

Fonte de Variação (Biológico-Não biológico)		SS	g.l.	MS	F	p
His	Entre os níveis	606,144	1	606,144	0,400	0,537
	Dentro dos níveis (Erro)	22726,466	15	1515,098		
	Total	23332,610	16			
Ser	Entre os níveis	255,734	1	255,734	0,295	0,595
	Dentro dos níveis (Erro)	12999,719	15	866,648		
	Total	13255,453	16			
Arg	Entre os níveis	80,296	1	80,296	0,043	0,839
	Dentro dos níveis (Erro)	28271,702	15	1884,780		
	Total	28351,998	16			
Gly	Entre os níveis	673,658	1	673,658	1,041	0,324
	Dentro dos níveis (Erro)	9702,864	15	646,858		
	Total	10376,523	16			
Asp	Entre os níveis	2537,257	1	2537,257	0,553	0,468
	Dentro dos níveis (Erro)	68771,240	15	4584,749		
	Total	71308,496	16			
Glu	Entre os níveis	173,239	1	173,239	0,011	0,918
	Dentro dos níveis (Erro)	236102,141	15	15740,143		
	Total	236275,380	16			
Thr	Entre os níveis	934,017	1	934,017	1,186	0,293
	Dentro dos níveis (Erro)	11810,728	15	787,382		
	Total	12744,745	16			
Ala	Entre os níveis	1071,355	1	1071,355	1,219	0,287
	Dentro dos níveis (Erro)	13186,096	15	879,073		
	Total	14257,452	16			
Pro	Entre os níveis	1194,710	1	1194,710	1,308	0,271
	Dentro dos níveis (Erro)	13702,068	15	913,471		
	Total	14896,778	16			
Cys	Entre os níveis	2046,142	1	2046,142	0,756	0,398
	Dentro dos níveis (Erro)	40587,072	15	2705,805		
	Total	42633,214	16			
Lys	Entre os níveis	59,413	1	59,413	0,113	0,741
	Dentro dos níveis (Erro)	7870,811	15	524,721		
	Total	7930,224	16			
Tyr	Entre os níveis	894,056	1	894,056	0,353	0,561
	Dentro dos níveis (Erro)	37983,774	15	2532,252		
	Total	38877,829	16			
Met	Entre os níveis	8332,978	1	8332,978	5,301	0,036
	Dentro dos níveis (Erro)	23581,664	15	1572,111		
	Total	31914,642	16			
Val	Entre os níveis	35,366	1	35,366	0,022	0,883
	Dentro dos níveis (Erro)	23776,471	15	1585,098		
	Total	23811,837	16			
Ile	Entre os níveis	8,299	1	8,299	0,008	0,930
	Dentro dos níveis (Erro)	15583,690	15	1038,913		
	Total	15591,989	16			
Leu	Entre os níveis	922,803	1	922,803	0,448	0,514
	Dentro dos níveis (Erro)	30921,602	15	2061,440		
	Total	31844,405	16			
Phe	Entre os níveis	441,943	1	441,943	0,209	0,654
	Dentro dos níveis (Erro)	31698,613	15	2113,241		
	Total	32140,556	16			

Tabela II.5 - Teste *t* de *Student* e teste *F* de Fisher para o tipo de arroz integral (Biológico-Não biológico)

AA	Média Arroz Integral Não Biológico	Média Arroz Integral Biológico	*t* de *Student*			Desvio Padrão Não Biológico	Desvio Padrão Biológico	*F* de Fisher	
			t	g.l.	Valor-p			*F*	Valor-p
His	394,33	382,36	0,633	15	0,537	32,14	45,46	2,001	0,352
Ser	422,82	415,05	0,543	15	0,595	27,58	31,43	1,298	0,717
Arg	738,59	734,23	0,206	15	0,839	35,64	50,86	2,036	0,340
Gly	399,85	387,24	1,021	15	0,324	24,66	26,29	1,137	0,852
Asp	630,86	655,34	-0,744	15	0,468	57,72	77,57	1,806	0,425
Glu	1235,87	1229,47	0,105	15	0,918	113,20	138,15	1,489	0,587
Thr	286,55	271,70	1,089	15	0,293	25,24	30,98	1,507	0,576
Ala	397,15	381,25	1,104	15	0,287	29,40	29,93	1,036	0,950
Pro	371,85	355,06	1,144	15	0,271	29,42	31,11	1,118	0,870
Cys	225,62	203,64	0,870	15	0,398	47,51	56,73	1,426	0,627
Lys	89,00	85,26	0,336	15	0,741	15,15	29,36	3,753	0,083
Tyr	593,07	578,55	0,594	15	0,561	38,48	61,11	2,521	0,219
Met	331,34	286,98	2,302	15	0,036	41,12	37,89	1,178	0,842
Val	358,37	361,26	-0,149	15	0,883	27,27	50,47	3,426	0,106
Ile	259,13	257,73	0,089	15	0,930	22,91	40,33	3,100	0,135
Leu	572,01	557,25	0,669	15	0,514	39,36	51,45	1,709	0,469
Phe	594,03	583,82	0,457	15	0,654	38,31	53,39	1,942	0,372

Anexo III –Dendrogramas (Análise de *Clusters*)

III.1. Variáveis (aminoácidos)

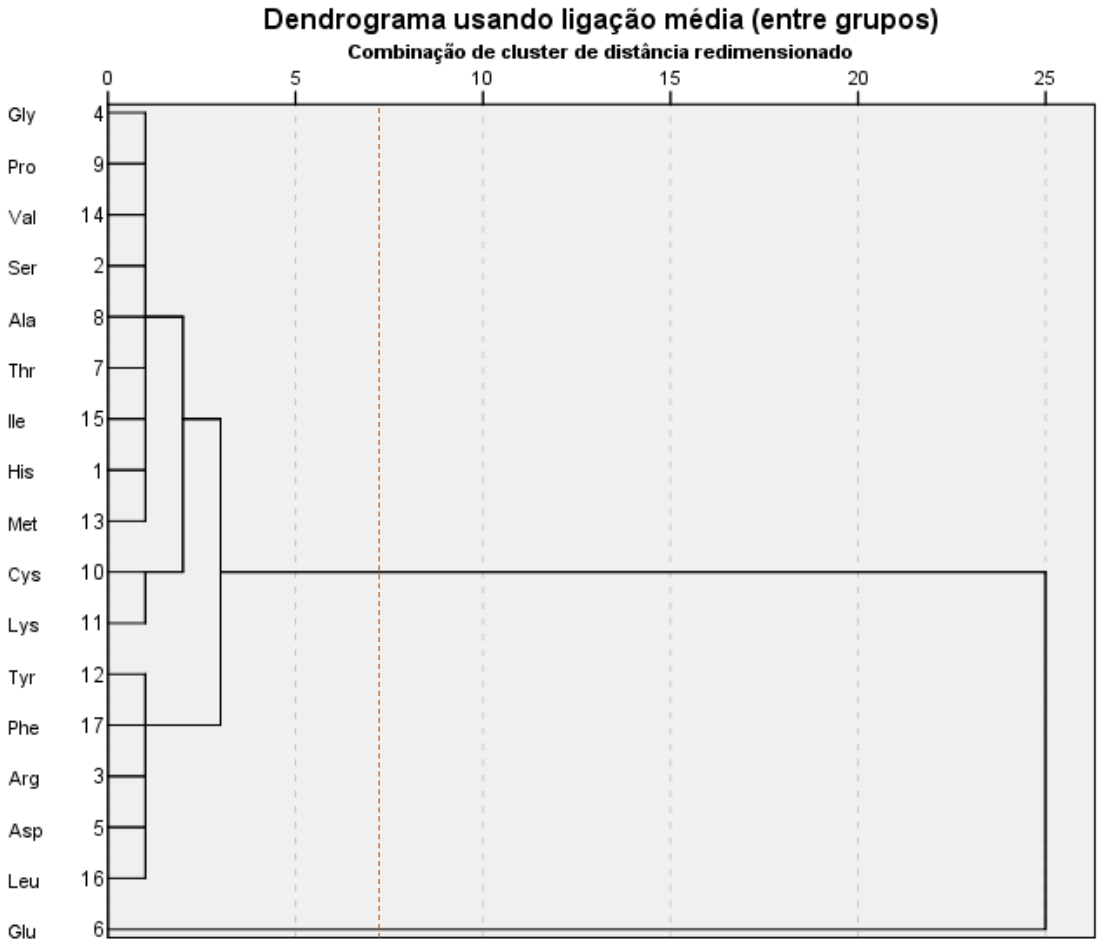

Figura III.1 - Dendrograma das variáveis utilizando o algoritmo da ligação média entre grupos

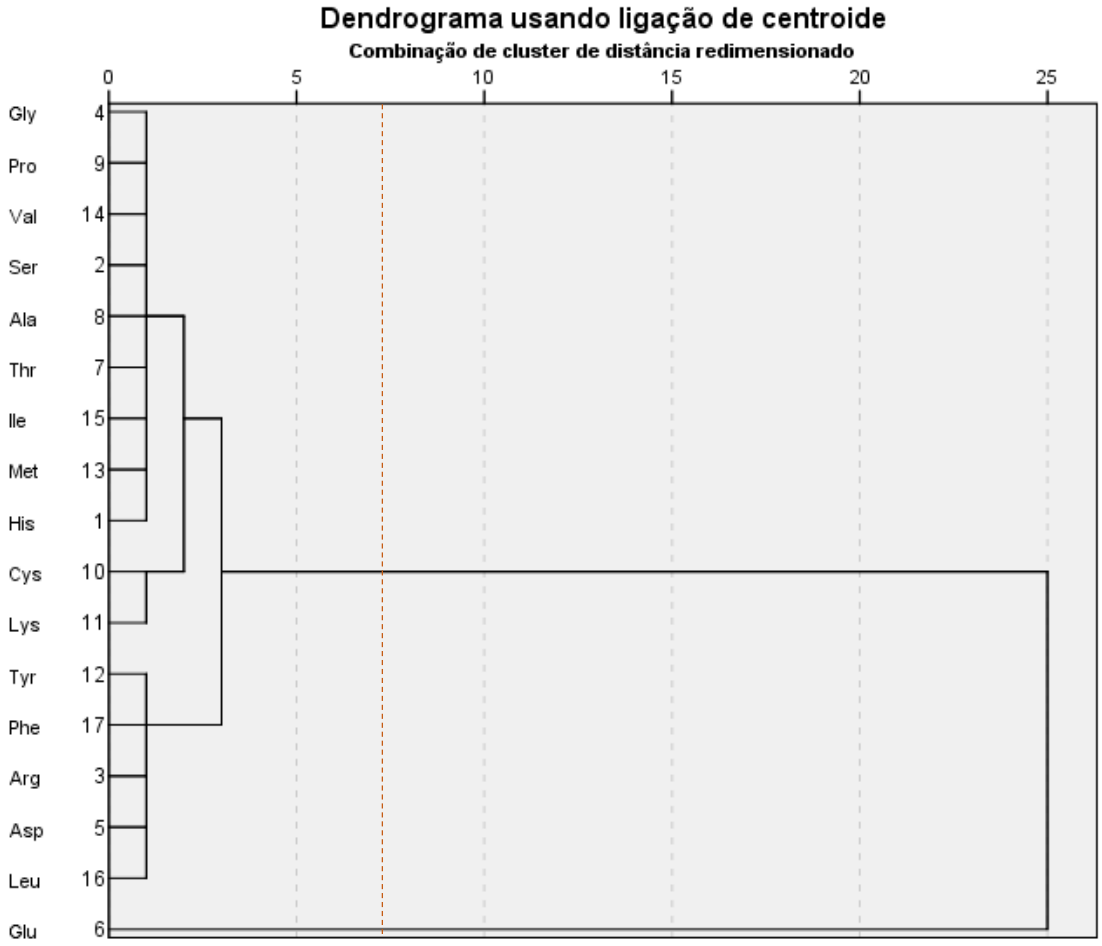

Figura III.2 - Dendrograma das variáveis utilizando o algoritmo do método do centróide

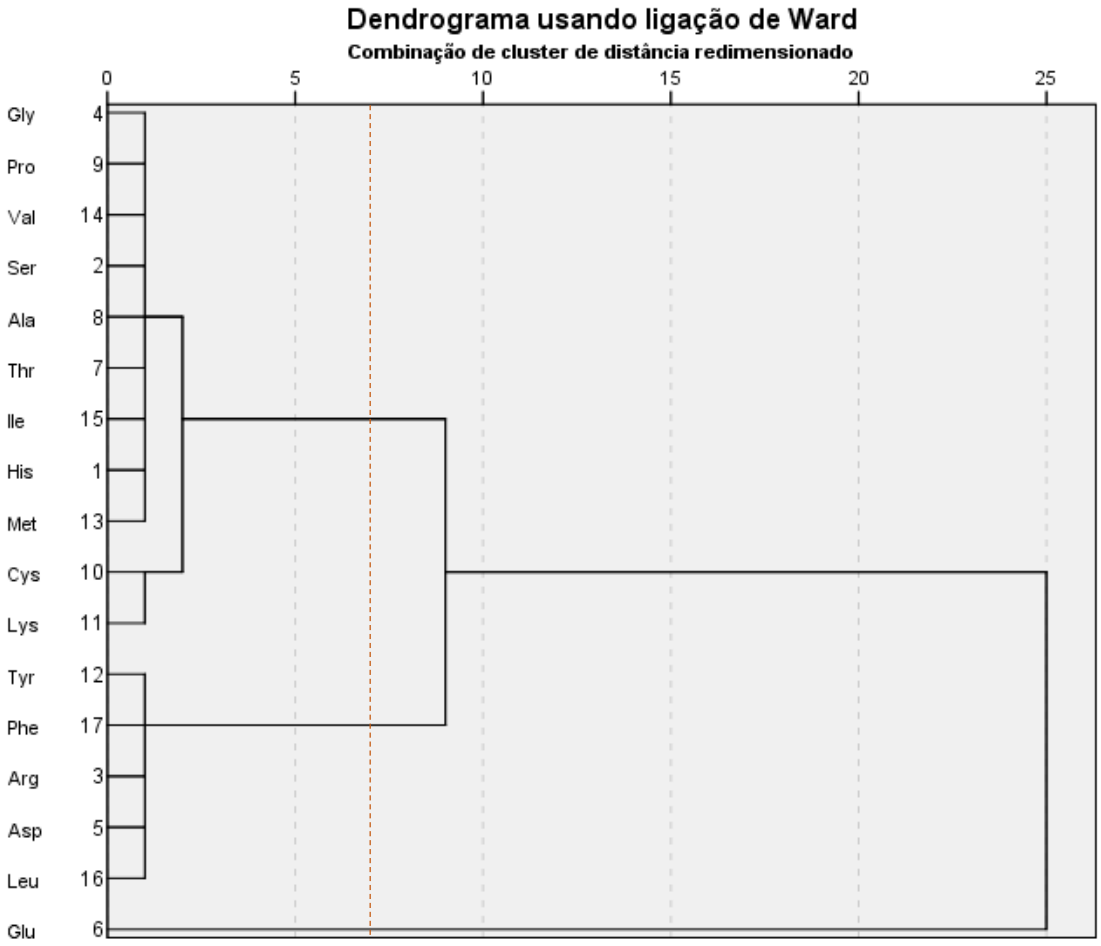

Figura III.3 - Dendrograma das variáveis utilizando o algoritmo do método de Ward

III.2. Variáveis (aminoácidos, retirando da análise o Glu – ácido glutâmico)

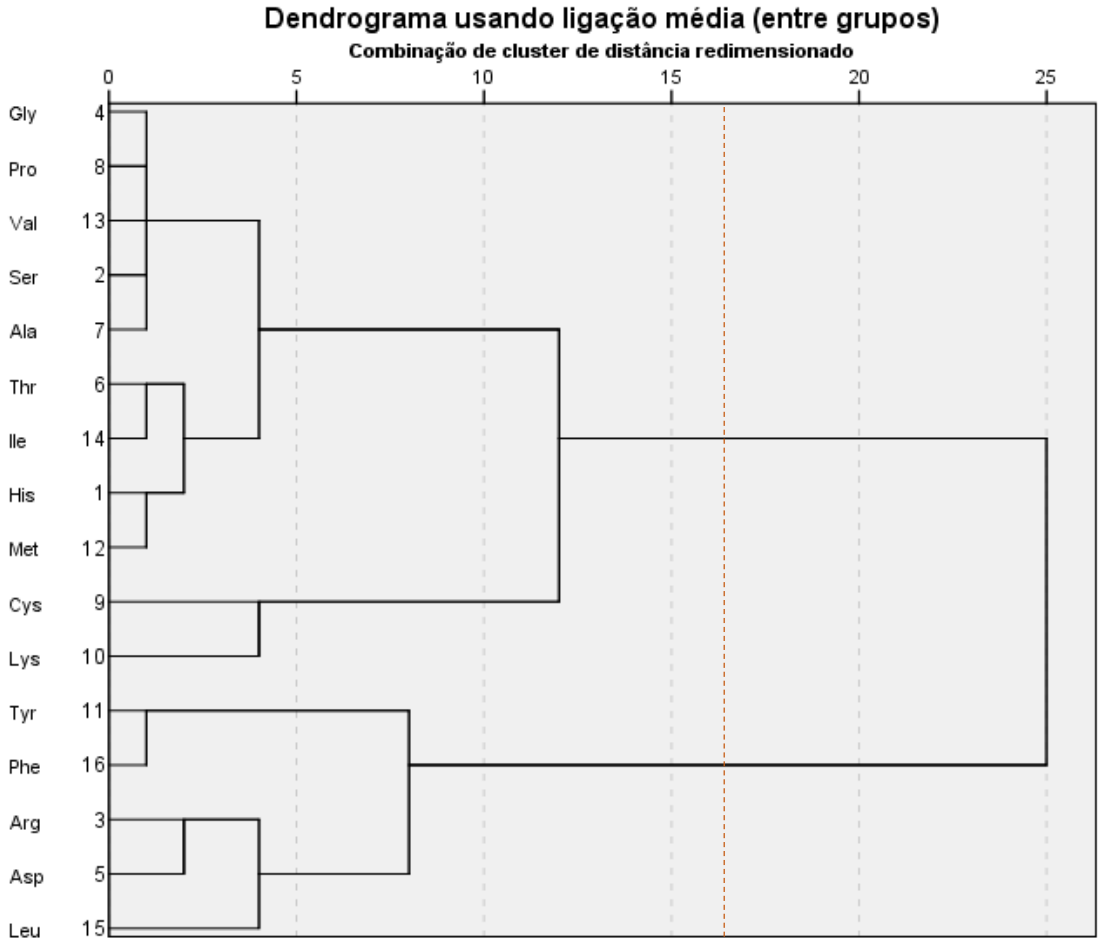

Figura III.4 - Dendrograma das variáveis (sem Glu) utilizando o algoritmo da ligação média entre grupos

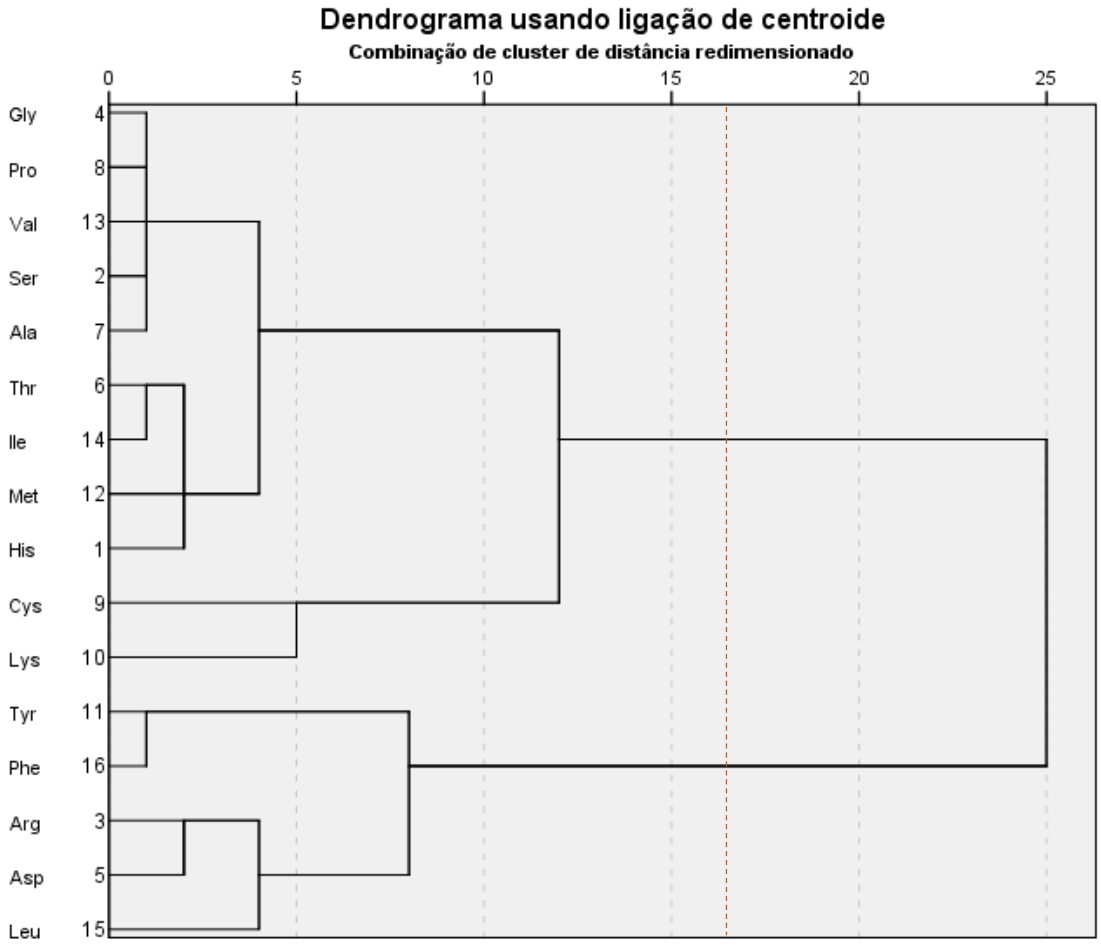

Figura III.5 - Dendrograma das variáveis (sem Glu) utilizando o algoritmo do método do centróide

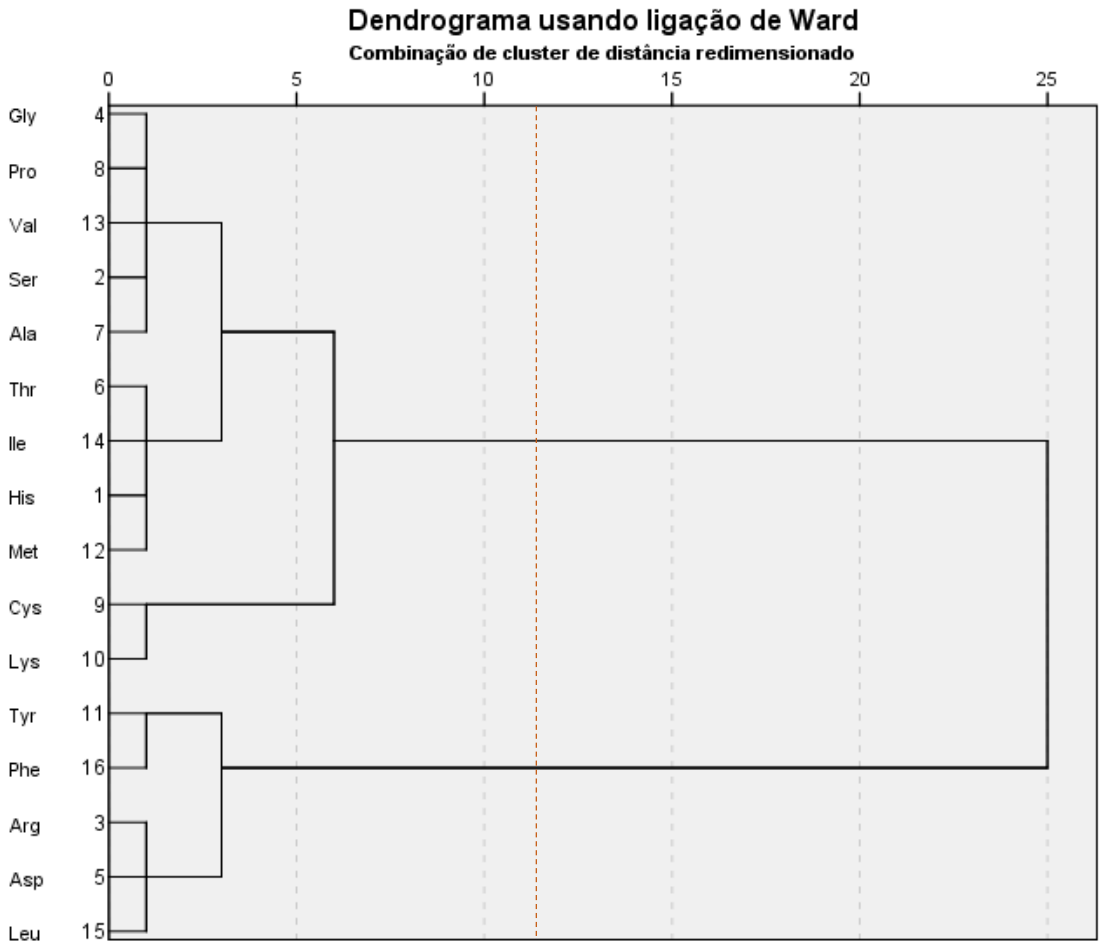

Figura III.6 - Dendrograma das variáveis (sem Glu) utilizando o algoritmo do método de Ward

III.3. Casos (amostras retirando da análise a variável Glu – ácido glutâmico)

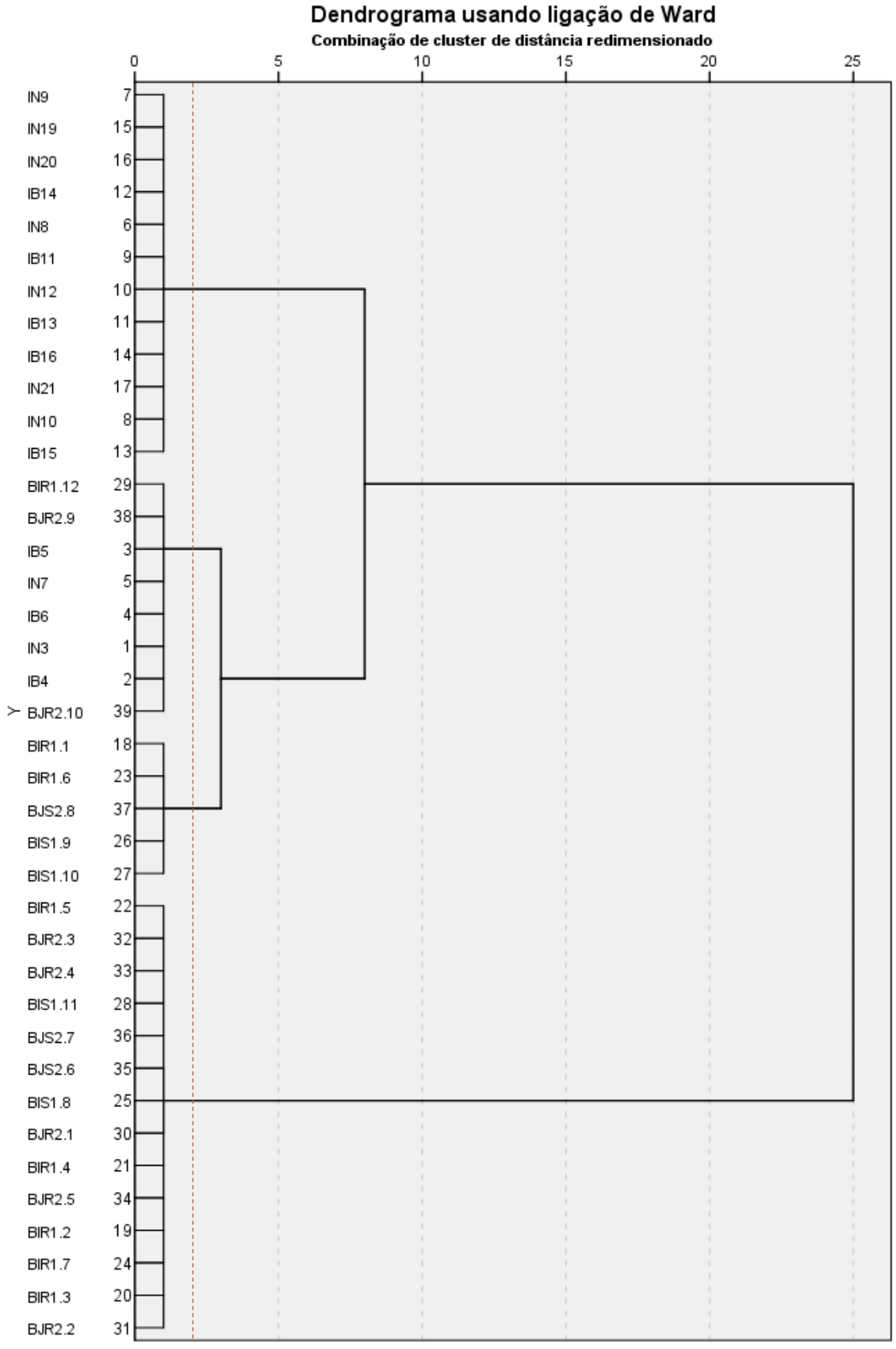

Figura III.7 - Dendrograma das amostras retirando o ácido glutâmico (glu) para os dados recolhidos com o algoritmo do método de Ward

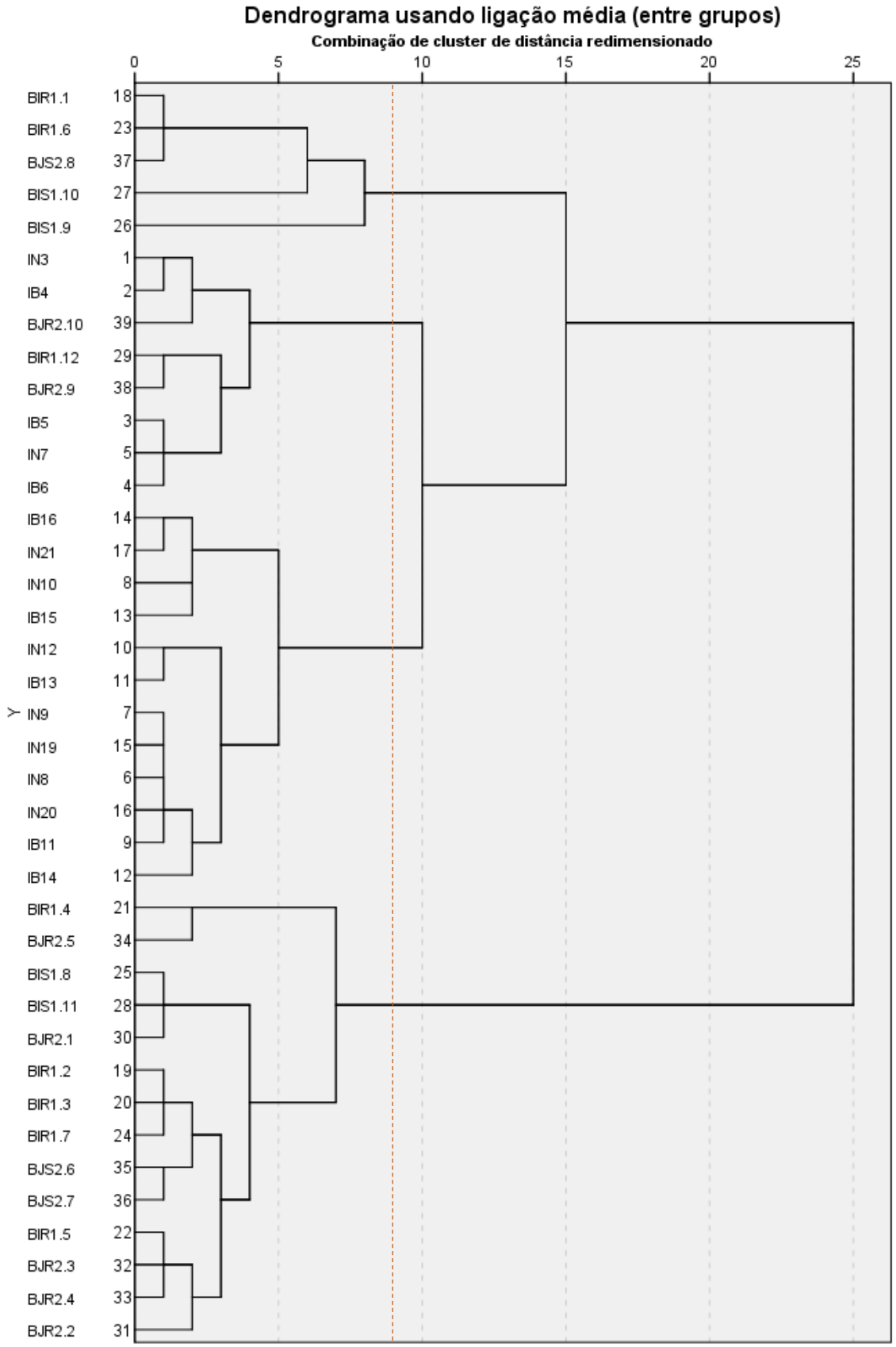

Figura III.8 - Dendrograma das amostras retirando o ácido glutâmico (glu) para os dados padronizados com o algoritmo da ligação média entre grupos

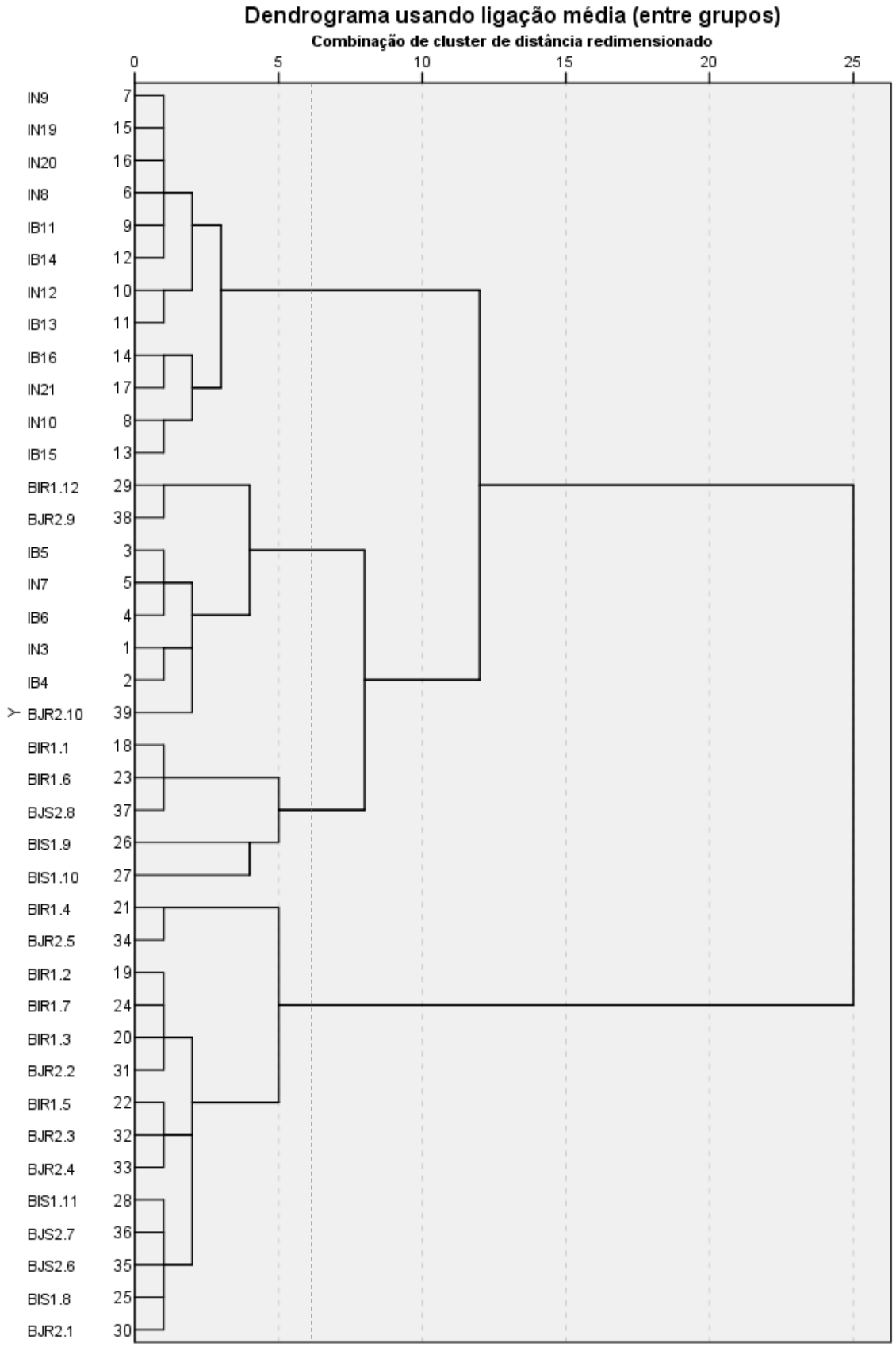

Figura III.9 - Dendrograma das amostras retirando o ácido glutâmico (glu) para os dados recolhidos com o algoritmo da ligação média entre grupos

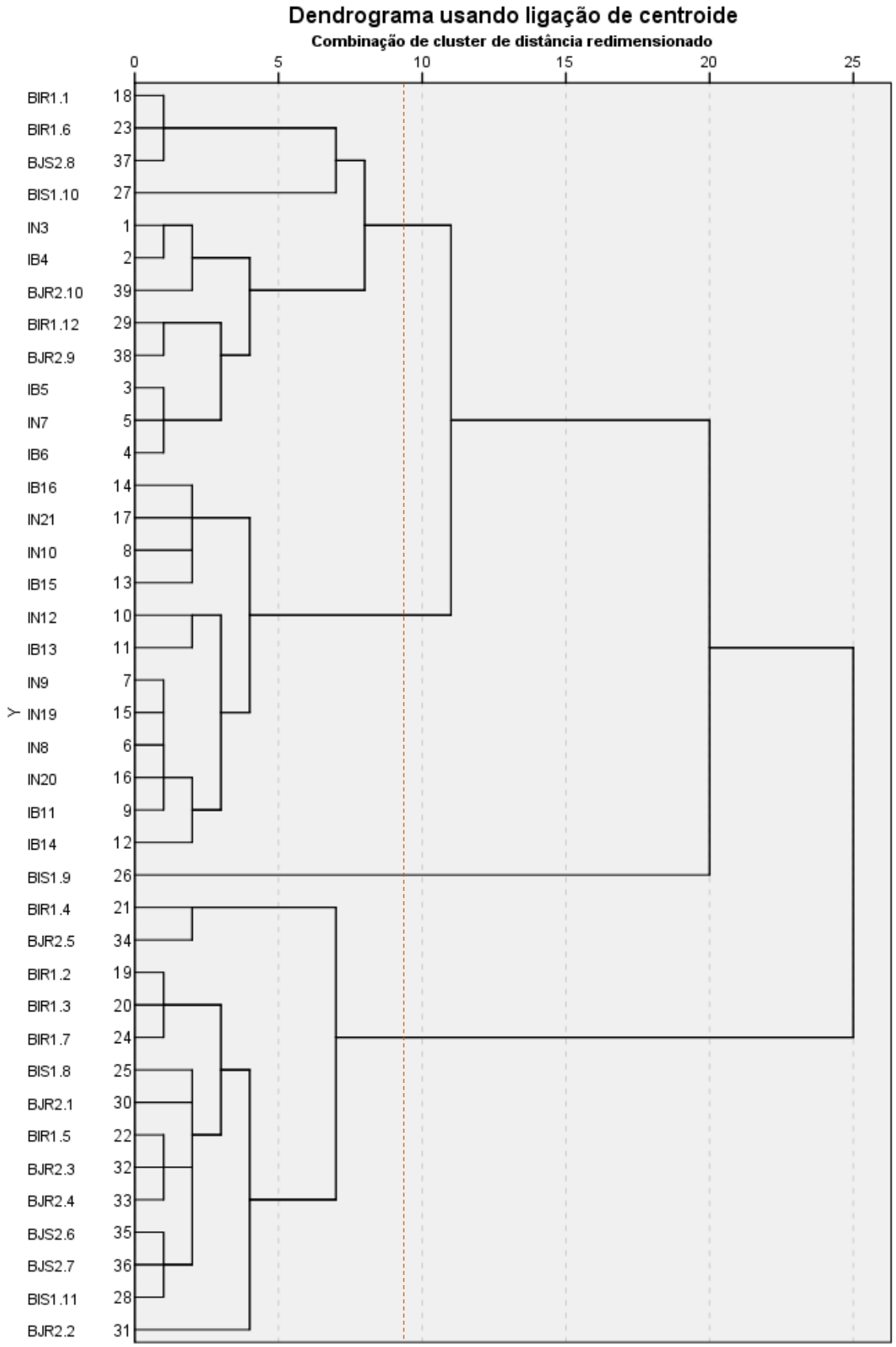

Figura III.10 - Dendrograma das amostras retirando o ácido glutâmico (glu) para os dados padronizados com o algoritmo do método do centróide

III.4. Casos (amostras com todas as variáveis)

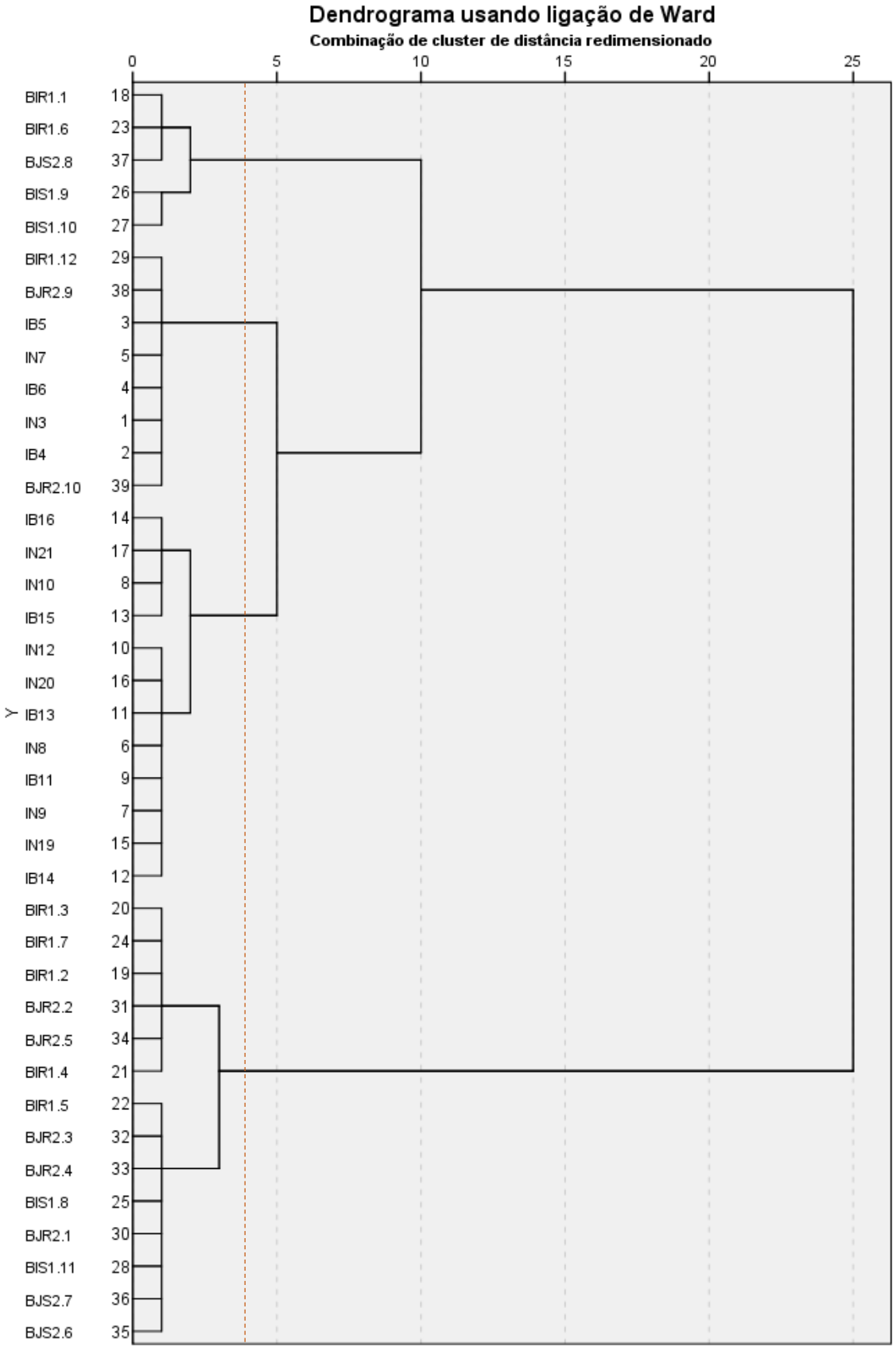

Figura III.11 - Dendrograma das amostras com todas as variáveis para os dados recolhidos com o algoritmo do método de Ward

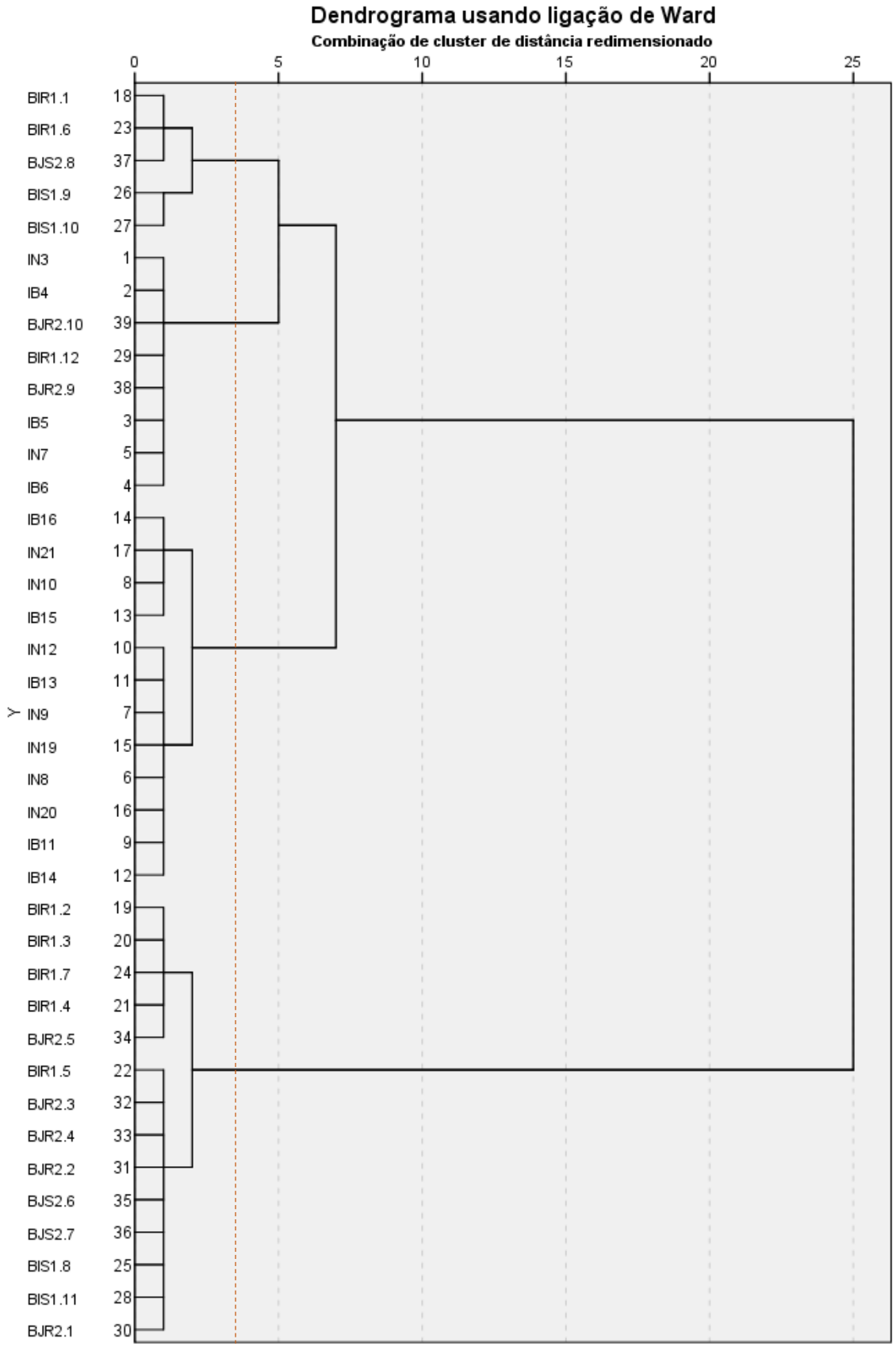

Figura III.12 - Dendrograma das amostras com todas as variáveis para os dados padronizados com o algoritmo do método de Ward

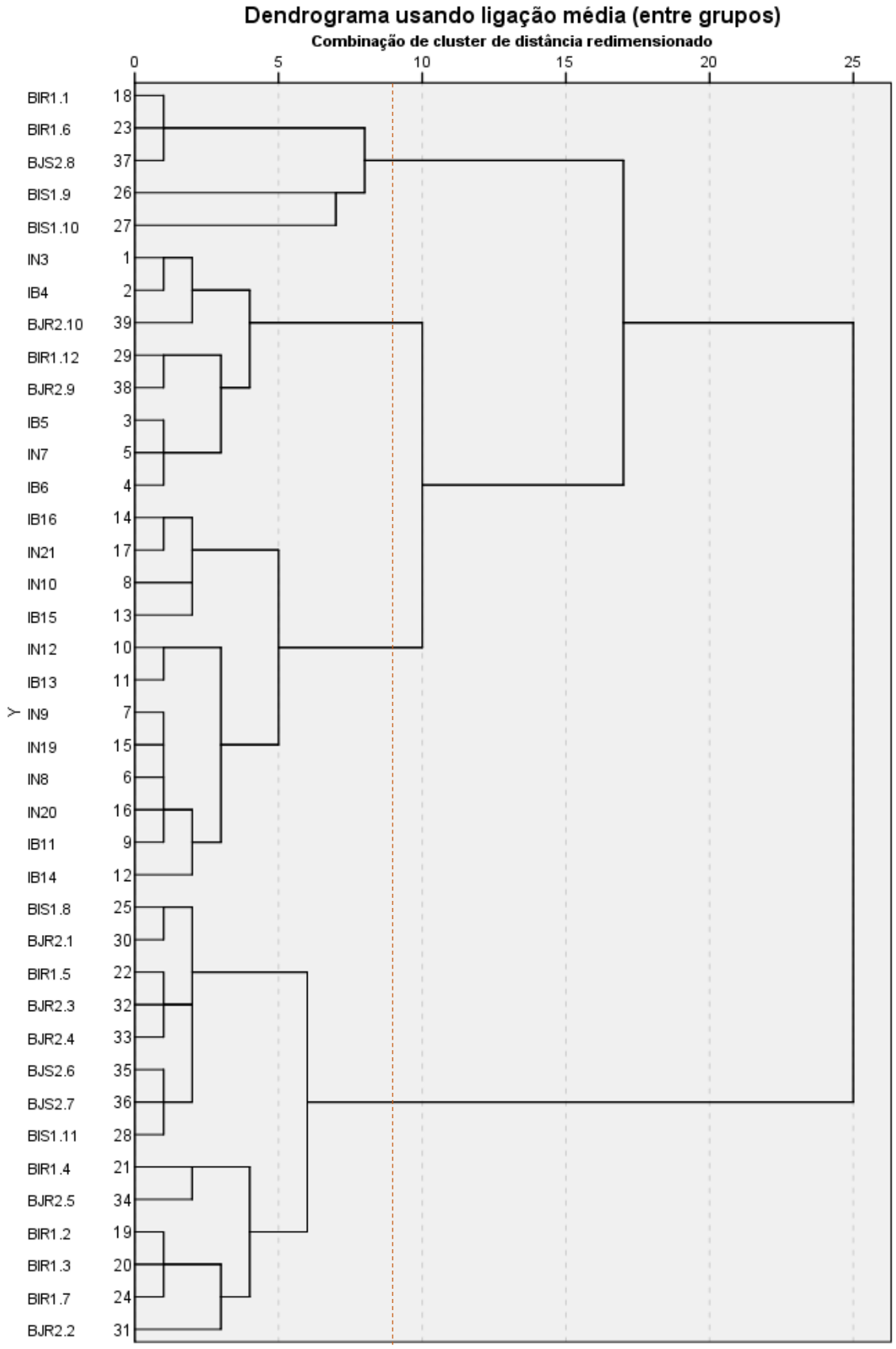

Figura III.13 - Dendrograma das amostras com todas as variáveis para os dados padronizados com o algoritmo da ligação média entre grupos

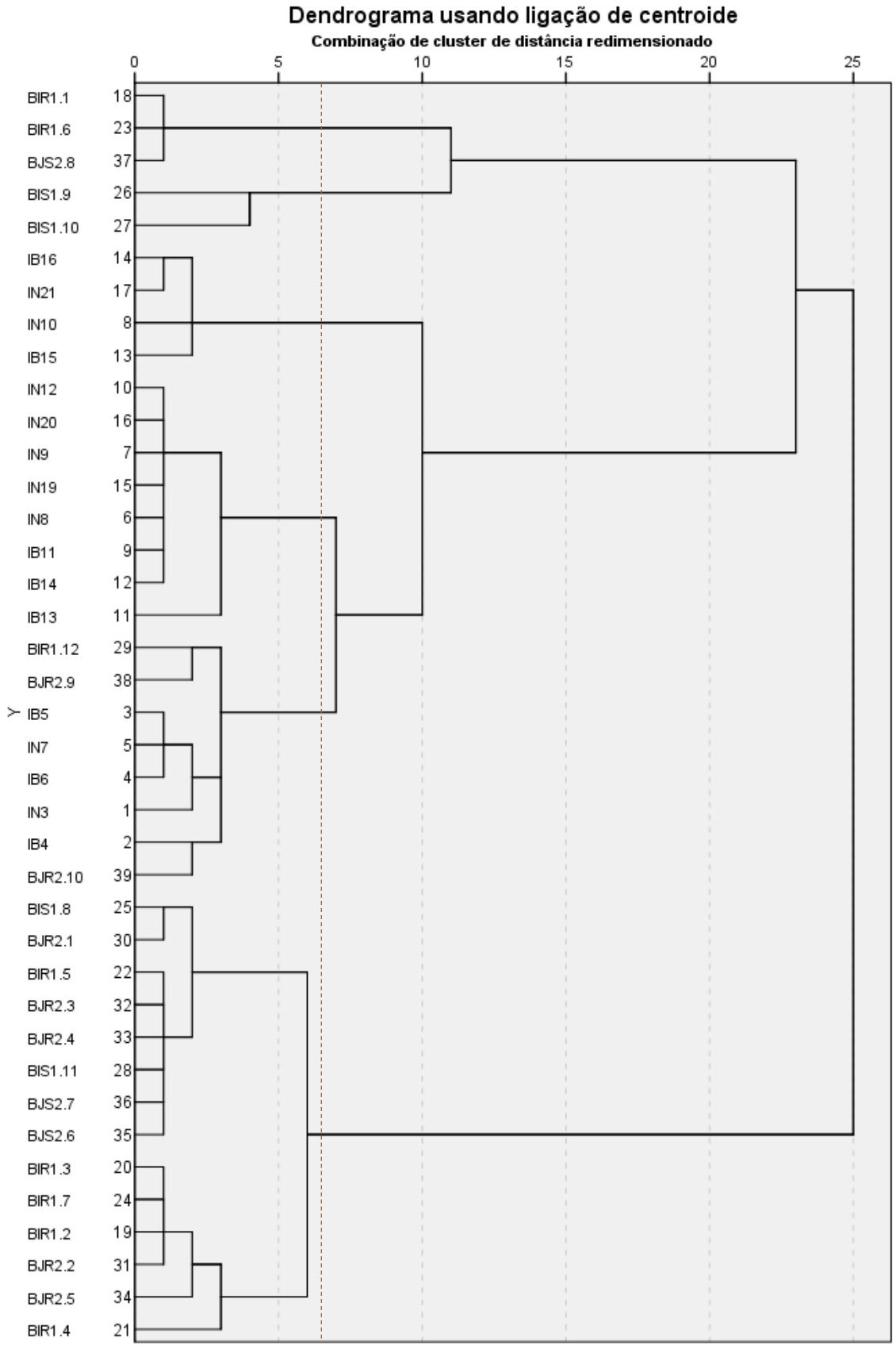

Figura III.14 - Dendrograma das amostras com todas as variáveis para os dados recolhidos com o algoritmo do método do centróide

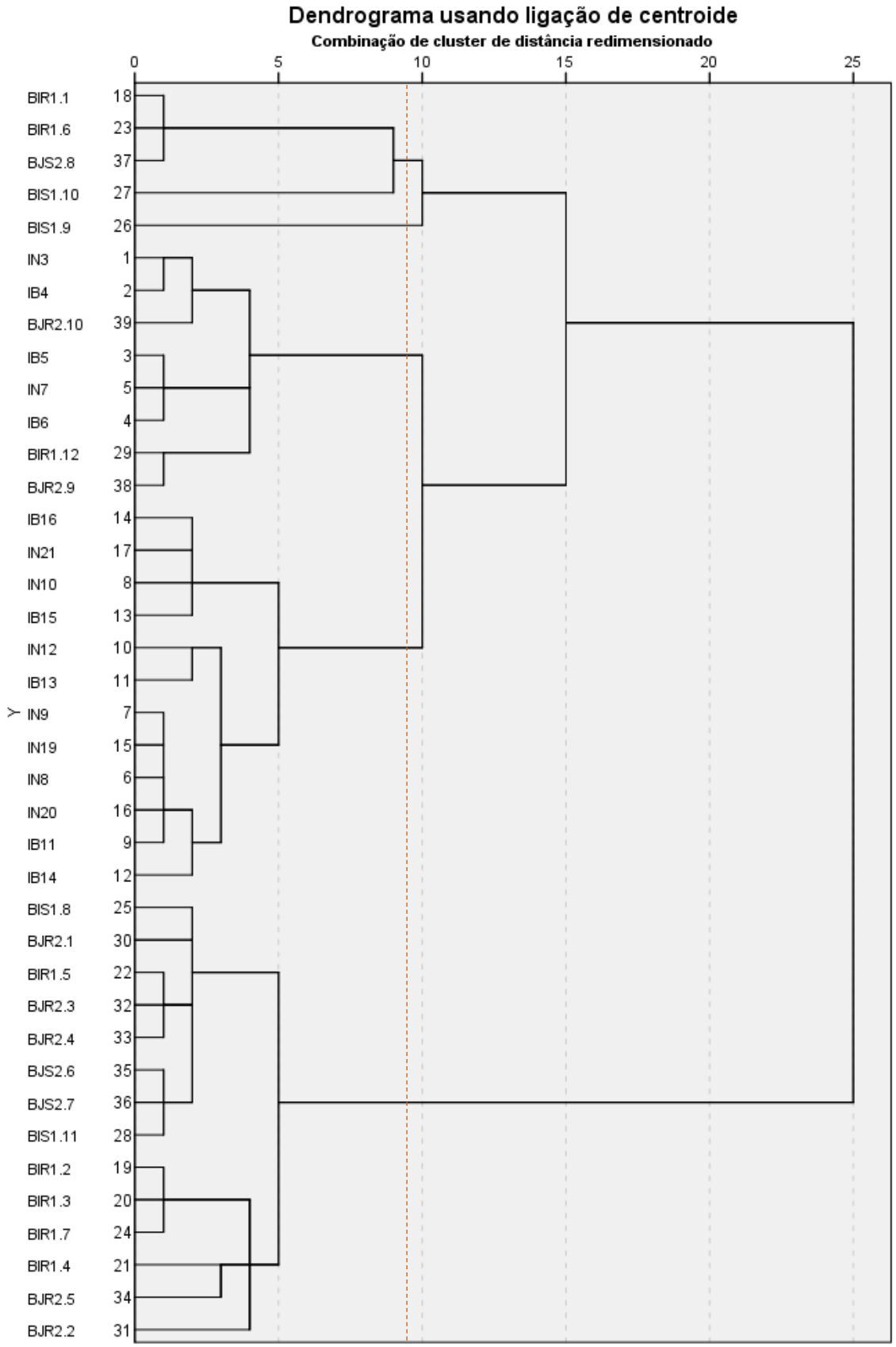

Figura III.15 - Dendrograma das amostras com todas as variáveis para os dados padronizados com o algoritmo do método do centróide