

Análise de Agrupamentos

Talvez Hamlet esteja certo. Podemos estar vivendo reclusos numa casca de noz, mas nos considerando reis do espaço infinito.

Stephen Hawking

Ao final deste capítulo, você terá condições de:

- Estabelecer as circunstâncias a partir das quais a análise de agrupamentos pode ser utilizada.
- Saber calcular, entre duas observações, as diferentes medidas de distância (dissimilaridade) para variáveis métricas e de semelhança (similaridade) para variáveis binárias.
- Compreender os diferentes esquemas de aglomeração hierárquicos em análise de agrupamentos, bem como saber fazer a interpretação de dendrogramas com foco na alocação das observações em cada grupo.
- Entender o esquema de aglomeração não hierárquico *k-means* e saber diferenciá-lo dos esquemas hierárquicos.
- Elaborar a análise de agrupamentos de maneira algébrica e por meio do IBM SPSS Statistics Software® e do Stata Statistical Software® e interpretar seus resultados.

9.1. INTRODUÇÃO

A **análise de agrupamentos** representa um conjunto de técnicas exploratórias muito úteis e que podem ser aplicadas quando há a intenção de se verificar a existência de **comportamentos semelhantes entre observações** (indivíduos, empresas, municípios, países, entre outros exemplos) em relação a determinadas variáveis e o objetivo de se criarem grupos, ou **clusters**, em que prevaleça a **homogeneidade interna**. Nesse sentido, esse conjunto de técnicas, também conhecido por **análise de conglomerados** ou **análise de clusters**, tem por objetivo principal a alocação de observações em uma quantidade relativamente pequena de agrupamentos **homogêneos internamente e heterogêneos entre si** e que representem o comportamento conjunto das observações a partir de determinadas variáveis. Ou seja, as observações de determinado grupo devem ser relativamente semelhantes entre si, em relação às variáveis inseridas na análise, e consideravelmente diferentes das observações de outros grupos.

As técnicas de análise de agrupamentos são consideradas **exploratórias**, ou de **interdependência**, uma vez que suas aplicações não apresentam caráter preditivo para outras observações não presentes inicialmente na amostra, e a inclusão de novas observações no banco de dados torna necessária a reaplicação da modelagem, para que, eventualmente, sejam gerados novos agrupamentos. Além disso, a inclusão de nova variável também pode fazer com que haja um rearranjo completo das observações nos grupos.

O pesquisador pode optar por elaborar uma análise de agrupamentos quando tiver o objetivo de **ordenar e alocar as observações em grupos** e, a partir de então, estudar qual a quantidade interessante de **clusters** formados, ou pode, *a priori*, definir a quantidade de grupos que deseja formar, embasado por determinado critério, e verificar como se comportam o ordenamento e a alocação das observações naquela quantidade especificada de grupos. Independentemente da natureza do objetivo, a análise de agrupamentos continuará exploratória. Caso um pesquisador tenha a intenção de utilizar uma técnica para, de fato, confirmar o estabelecimento dos grupos e tornar a análise preditiva, poderá fazer uso, por exemplo, de técnicas como **análise discriminante** ou **regressão logística multinomial**.

A elaboração da análise de agrupamentos não exige conhecimento de álgebra matricial ou de estatística, ao contrário de técnicas como análise fatorial e análise de correspondência. O pesquisador interessado em aplicar uma análise de agrupamentos necessita, a partir da **definição dos objetivos de pesquisa**, escolher determinada

medida de distância ou de semelhança, que servirá de base para que as observações sejam consideradas menos ou mais próximas, e determinado **esquema de aglomeração**, que deverá ser definido entre os **métodos hierárquicos e não hierárquicos**. Dessa forma, terá condições de analisar, interpretar e comparar os resultados.

É importante ressaltar que resultados obtidos por meio de esquemas de aglomeração hierárquicos e não hierárquicos podem ser comparados, e, nesse sentido, o pesquisador tem a liberdade de elaborar a técnica, fazendo uso de um ou outro método, e reaplicá-la, se julgar necessário. **Enquanto os esquemas hierárquicos permitem a identificação do ordenamento e da alocação das observações, oferecendo possibilidades para que o pesquisador estude, avalie e decida sobre a quantidade de agrupamentos formados, nos esquemas não hierárquicos, parte-se de uma quantidade conhecida de clusters e, a partir de então, é elaborada a alocação das observações nesses clusters, com posterior avaliação da representatividade de cada variável para a formação deles.** Portanto, o resultado de um método pode servir de *input* para a realização do outro, tornando a **análise cíclica**. A Figura 9.1 apresenta a lógica a partir da qual a análise de agrupamentos pode ser elaborada.

Quando da escolha da medida de distância ou de semelhança e do esquema de aglomeração, devem ser levados em consideração aspectos como a quantidade previamente desejada de agrupamentos, definida com base em algum critério de alocação de recursos, bem como determinadas restrições que podem levar o pesquisador a optar por uma solução específica. Conforme discutem Bussab *et al.* (1990), critérios diferentes a respeito de medidas de distância e de esquemas de aglomeração podem levar a formações distintas de agrupamentos, e a homogeneidade desejada pelo pesquisador depende fundamentalmente dos objetivos estipulados na pesquisa.

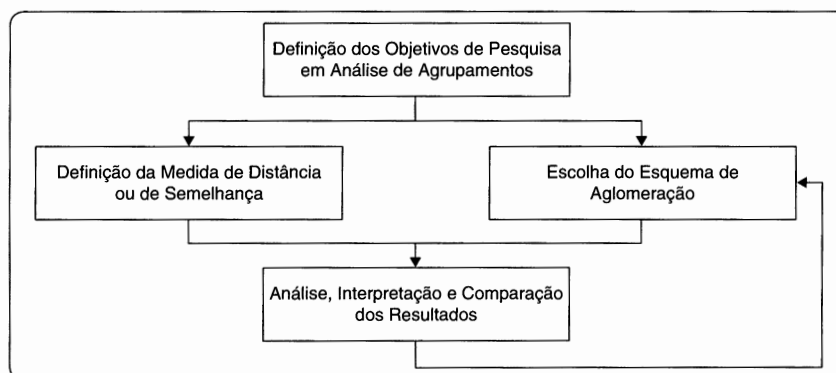


Figura 9.1 Lógica para elaboração da análise de agrupamentos.

Imagine que um pesquisador tenha interesse em estudar a relação de interdependência entre indivíduos de uma população de determinado município com base apenas em duas variáveis métricas (idade, em anos, e renda média familiar, em R\$). Seu intuito é avaliar a eficiência de programas sociais voltados à área da saúde e, com base nessas variáveis, propor uma quantidade ainda desconhecida de novos programas voltados a grupos homogêneos de pessoas. Após a coleta dos dados, o pesquisador elaborou um gráfico de dispersão, como o apresentado na Figura 9.2.

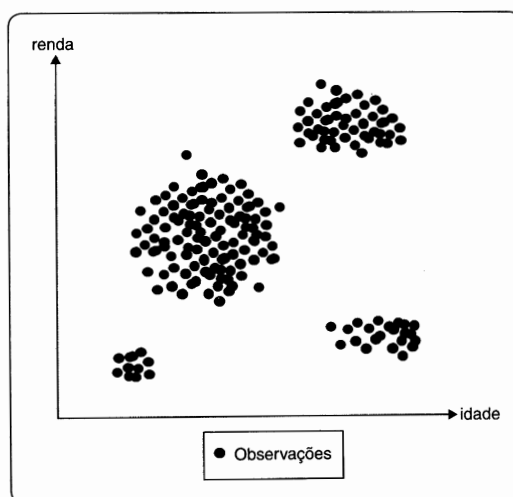


Figura 9.2 Gráfico de dispersão de indivíduos para renda e idade.

Com base no gráfico da Figura 9.2, o pesquisador identificou quatro *clusters*, destacando-os em novo gráfico (Figura 9.3).

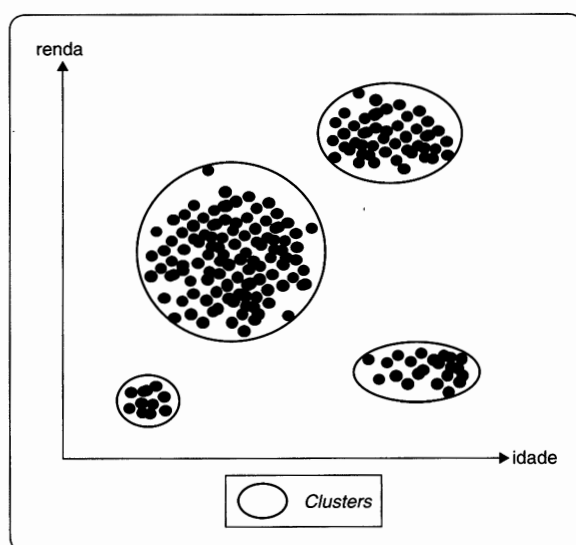


Figura 9.3 Destaque para a formação de quatro *clusters*.

A partir da formação desses *clusters*, o pesquisador resolveu elaborar uma análise acerca do comportamento das observações em cada grupo ou, mais precisamente, sobre a variabilidade existente dentro dos agrupamentos e entre eles, a fim de poder embasar, de maneira clara e consciente, sua decisão a respeito da alocação dos indivíduos nesses quatro novos programas sociais. A fim de ilustrar essa questão, o pesquisador elaborou o gráfico da Figura 9.4.

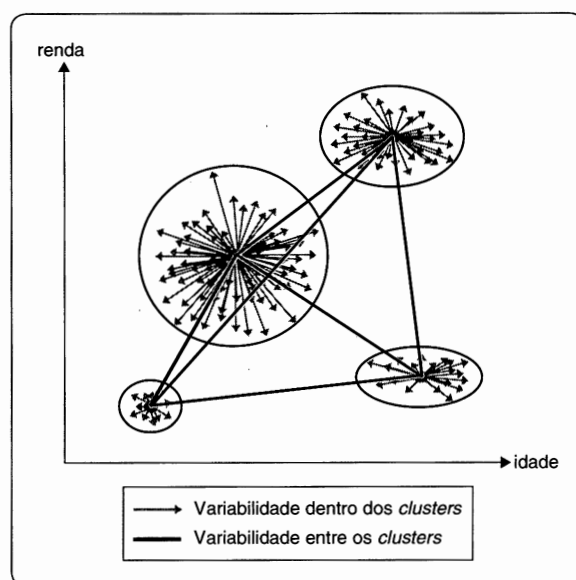


Figura 9.4 Ilustração sobre a variabilidade dentro dos *clusters* e entre eles.

Com base nesse gráfico, o pesquisador pôde perceber que os grupos formados apresentavam bastante homogeneidade interna, com determinado indivíduo apresentando maior proximidade com outros indivíduos do mesmo grupo do que com indivíduos de outros grupos. Essa é a essência fundamental da análise de agrupamentos.

Caso a quantidade de programas sociais a serem oferecidos à população (quantidade de *clusters*) já tivesse sido imposta ao pesquisador, por razões relativas a restrições orçamentárias, jurídicas ou políticas, ainda assim poderia ser utilizada a análise de agrupamentos para, apenas e tão somente, ser determinada a alocação dos indivíduos do município naquela quantidade de programas (grupos).

Tendo concluído a pesquisa e alocado os indivíduos nos diferentes programas sociais voltados à área da saúde, o pesquisador resolveu elaborar, no ano seguinte, a mesma pesquisa com os indivíduos do mesmo município. Porém, nesse ínterim, um grupo de bilionários em idade avançada resolveu se mudar para a cidade, e, ao elaborar o novo gráfico de dispersão, o pesquisador percebeu que aqueles quatro *clusters* nitidamente formados no ano anterior já não existiam mais, visto que sofreram um processo de fusão quando da inclusão dos bilionários. O novo gráfico de dispersão encontra-se na Figura 9.5.

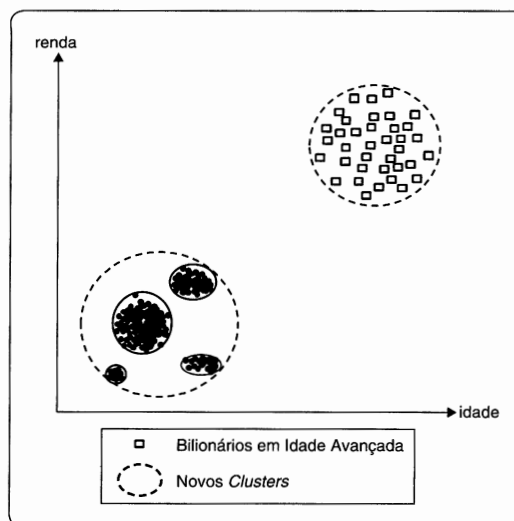


Figura 9.5 Rearranjo dos *clusters* na presença de bilionários em idade avançada.

Essa nova situação exemplifica a importância de que a **análise de agrupamentos seja sempre reaplicada quando da inclusão de novas observações** (e também novas variáveis), o que descaracteriza e inviabiliza totalmente seu poder preditivo, conforme discutimos.

Mais que isso, esse exemplo demonstra ser recomendável, antes da elaboração de qualquer análise de agrupamentos, que o pesquisador estude o comportamento dos dados e verifique a existência de observações discrepantes em relação a determinadas variáveis, visto que **a formação de clusters é bastante sensível à presença de outliers**. A **exclusão** ou a **retenção** de outliers na base, entretanto, vai depender dos objetivos de pesquisa e da natureza dos dados, já que, se determinadas observações representarem aberrações em termos de valores das variáveis, em comparação às demais observações, e acabarem por formar *clusters* pequenos, insignificantes ou até mesmo individuais, podem, de fato, ser excluídas. Por outro lado, caso essas observações representem um ou mais grupos relevantes, ainda que diferentes dos demais, devem ser consideradas na análise e, quando da reaplicação da técnica, podem ser separadas para que outras segmentações sejam mais bem estruturadas em novos grupos, formados com maior homogeneidade interna.

Ressaltamos que os métodos de análise de agrupamentos são considerados **procedimentos estáticos**, já que a inclusão de novas observações ou variáveis pode alterar os *clusters*, tornando obrigatória a elaboração de uma nova análise.

Nesse exemplo, percebemos que as variáveis originais a partir das quais são estabelecidos os grupos são métricas, visto que a análise de agrupamentos partiu do estudo do **comportamento de distâncias (medidas de dissimilaridade)** entre as observações. Em alguns casos, conforme estudaremos ao longo do capítulo, podem ser elaboradas análises de *clusters* a partir do **comportamento de semelhanças (medidas de similaridade)** entre observações que apresentam variáveis binárias. É comum, entretanto, que pesquisadores façam uso do **incorreto procedimento de ponderação arbitrária** em variáveis qualitativas como, por exemplo, variáveis em **escala Likert**, para, a partir de então, ser aplicada uma análise de agrupamentos. **Isso é um erro grave**, já que existem técnicas exploratórias destinadas exclusivamente ao estudo do comportamento de variáveis qualitativas, por exemplo, a análise de correspondência.

Historicamente, embora muitas medidas de distância e de semelhança remontem ao final do século XIX e início do século XX, a análise de agrupamentos, como conjunto de técnicas mais estruturado, teve origem na Antropologia, com Driver e Kroeber (1932), e na Psicologia, com Zubin (1938a e 1938b) e Tryon (1939),

conforme discutem Reis (2001) e Fávero *et al.* (2009). Com o reconhecimento dos procedimentos de aglomeração e classificação de observações como método científico, aliado ao profundo desenvolvimento computacional, verificado principalmente após a década de 1960, a utilização da análise de agrupamentos passa a ser mais frequente após a publicação da relevante obra de Sokal e Sneath (1963), em que são realizados procedimentos para comparar as similaridades biológicas de organismos com características semelhantes e as respectivas espécies.

Atualmente, a análise de agrupamentos apresenta vasta possibilidade de aplicação em áreas como comportamento do consumidor, segmentação de mercado, estratégia, ciência política, economia, finanças, contabilidade, atuária, engenharia, logística, ciência da computação, educação, medicina, biologia, genética, bioestatística, psicologia, antropologia, demografia, geografia, ecologia, climatologia, geologia, arqueologia, criminologia e perícia, entre outras.

Neste capítulo, trataremos das técnicas de análise de agrupamentos, com os seguintes objetivos: (1) introduzir os conceitos; (2) apresentar, de maneira algébrica e prática, o passo a passo da modelagem; (3) interpretar os resultados obtidos; e (4) propiciar a aplicação das técnicas em SPSS e Stata. Seguindo a lógica proposta no livro, será inicialmente elaborada a solução algébrica de um exemplo vinculada à apresentação dos conceitos. Somente após a introdução dos conceitos serão apresentados os procedimentos para a elaboração das técnicas em SPSS e Stata.

9.2. ANÁLISE DE AGRUPAMENTOS

Muitos são os procedimentos para que seja elaborada uma análise de agrupamentos, visto que existem diferentes medidas de distância ou de semelhança para, respectivamente, variáveis métricas ou binárias. Além disso, definida a medida de distância ou de semelhança, o pesquisador ainda precisa determinar, entre diversas possibilidades, o método de aglomeração das observações, a partir de determinados critérios hierárquicos ou não hierárquicos. Nesse sentido, o que inicialmente parece trivial, ao se desejar agrupar observações em *clusters* internamente homogêneos, pode se tornar um tanto complexo, na medida em que **há uma multiplicidade de combinações entre diferentes medidas de distância ou de semelhança e métodos de aglomeração**. É de fundamental importância, portanto, que o pesquisador defina, com base na teoria subjacente e em seus objetivos de pesquisa, bem como em sua experiência e intuição, os critérios a partir dos quais as observações serão alocadas em cada um dos grupos.

Nas seções seguintes, apresentaremos o desenvolvimento teórico da técnica, bem como a elaboração de um exemplo prático. Nas seções 9.2.1 e 9.2.2, são apresentados e discutidos os conceitos pertinentes às medidas de distância e de semelhança e aos métodos de aglomeração, respectivamente, sempre acompanhados de resoluções algébricas elaboradas a partir de um banco de dados.

9.2.1. Definição das medidas de distância ou de semelhança em análise de agrupamentos

Conforme discutimos, a primeira etapa para a elaboração de uma análise de agrupamentos consiste em definir a medida de distância (dissimilaridade) ou de semelhança (similaridade) que servirá de base para que cada observação seja alocada em determinado grupo.

As medidas de distância são frequentemente utilizadas quando as variáveis do banco de dados forem essencialmente métricas, visto que, quanto maiores as diferenças entre os valores das variáveis de duas determinadas observações, menor a similaridade entre elas ou, em outras palavras, maior a dissimilaridade.

Já as medidas de semelhança são frequentemente utilizadas quando as variáveis forem binárias, e o que interessa é a frequência dos pares de respostas convergentes 1-1 ou 0-0 de duas determinadas observações. Nesse caso, quanto maior a frequência de pares convergentes, maior a semelhança (similaridade) entre as observações.

Exceção a essa lógica está na medida de correlação de Pearson entre duas observações, calculada a partir de variáveis métricas, porém com características de similaridade, conforme veremos na próxima seção.

Enquanto estudaremos as medidas de dissimilaridade para variáveis métricas na seção 9.2.1.1, a seção 9.2.1.2 é destinada ao estudo das medidas de similaridade para variáveis binárias.

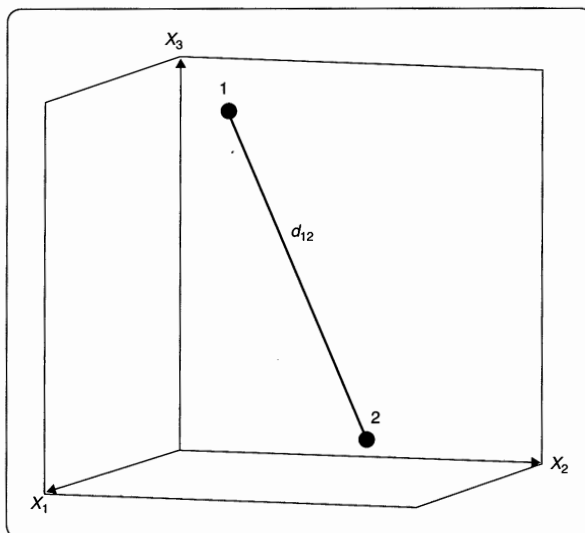
9.2.1.1. Medidas de distância (dissimilaridade) entre observações para variáveis métricas

Imagine que tenhamos a intenção de calcular, para uma situação hipotética, a distância entre duas determinadas observações i ($i = 1, 2$) provenientes de um banco de dados que apresenta três variáveis métricas (X_{1i} , X_{2i} , X_{3i}), com valores na mesma unidade de medida. Esses dados encontram-se na Tabela 9.1.

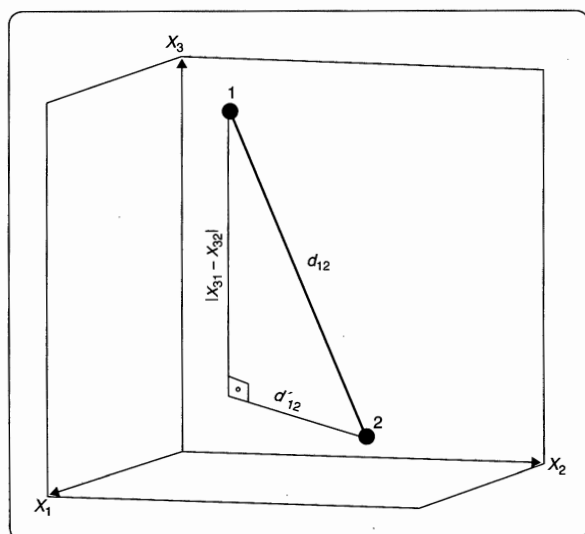
Tabela 9.1 Parte de banco de dados com duas observações e três variáveis métricas.

Observação i	X_{1i}	X_{2i}	X_{3i}
1	3,7	2,7	9,1
2	7,8	8,0	1,5

A partir desses dados, é possível ilustrarmos a configuração das duas observações em um espaço tridimensional, visto que temos exatamente três variáveis. A Figura 9.6 apresenta a posição relativa de cada observação, com destaque para a distância entre elas (d_{12}).

**Figura 9.6** Gráfico de dispersão tridimensional para situação hipotética com duas observações e três variáveis.

A distância d_{12} , que é uma medida de dissimilaridade, pode ser facilmente calculada fazendo uso, por exemplo, de sua projeção sobre o plano horizontal formado pelos eixos X_1 e X_2 , chamada de distância d'_{12} , conforme mostra a Figura 9.7.

**Figura 9.7** Gráfico tridimensional com destaque para a projeção de d_{12} sobre o plano horizontal.

Dessa forma, com base na conhecida expressão da **distância de Pitágoras** para triângulos retângulos, podemos determinar d_{12} por meio da seguinte expressão:

$$d_{12} = \sqrt{(d'_{12})^2 + (X_{31} - X_{32})^2} \quad (9.1)$$

sabendo-se que $|X_{31} - X_{32}|$ é a distância das projeções verticais (eixo X_3) dos pontos 1 e 2.

Entretanto, também não conhecemos a distância d'_{12} e, dessa forma, precisamos novamente recorrer à expressão de Pitágoras, agora fazendo uso das distâncias das projeções dos Pontos 1 e 2 sobre os outros dois eixos (X_1 e X_2), conforme mostra a Figura 9.8.

Logo, podemos escrever que:

$$d'_{12} = \sqrt{(X_{11} - X_{12})^2 + (X_{21} - X_{22})^2} \quad (9.2)$$

e, substituindo (2) em (1), temos que:

$$d_{12} = \sqrt{(X_{11} - X_{12})^2 + (X_{21} - X_{22})^2 + (X_{31} - X_{32})^2} \quad (9.3)$$

que é a expressão da distância (medida de dissimilaridade) entre os Pontos 1 e 2, também conhecida por expressão da **distância euclidiana**.

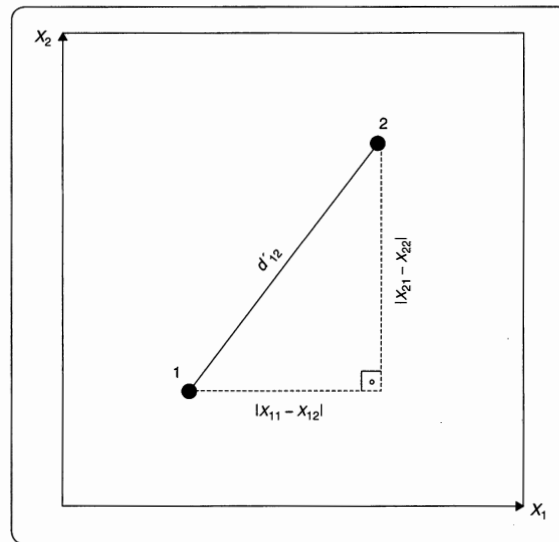


Figura 9.8 Projeção dos pontos no plano formado por X_1 e X_2 e destaque para d'_{12} .

Portanto, para os dados do nosso exemplo, temos que:

$$d_{12} = \sqrt{(3,7 - 7,8)^2 + (2,7 - 8,0)^2 + (9,1 - 1,5)^2} = 10,132$$

cuja unidade de medida é a mesma das variáveis originais do banco de dados. É importante ressaltar que, caso as variáveis não se apresentem na mesma unidade de medida, um **procedimento de padronização dos dados** precisará ser elaborado preliminarmente, conforme discutiremos mais adiante.

Podemos generalizar esse problema para uma situação em que o banco de dados apresente n observações e, para cada observação i ($i = 1, \dots, n$), valores correspondentes a cada uma das j ($j = 1, \dots, k$) variáveis métricas X , conforme mostra a Tabela 9.2.

Tabela 9.2 Modelo geral de um banco de dados para elaboração da análise de agrupamentos.

Observação i	Variável j			
	X_{1i}	X_{2i}	...	X_{ki}
1	X_{11}	X_{21}	...	X_{k1}
2	X_{12}	X_{22}		X_{k2}
\vdots	\vdots	\vdots		
p	X_{1p}	X_{2p}		X_{kp}
\vdots	\vdots	\vdots		
q	X_{1q}	X_{2q}		X_{kq}
\vdots	\vdots	\vdots		...
n	X_{1n}	X_{2n}		X_{kn}

Logo, a expressão (9.4), com base na expressão (9.3), apresenta a definição geral da distância euclidiana entre duas observações quaisquer p e q .

$$d_{pq} = \sqrt{(X_{1p} - X_{1q})^2 + (X_{2p} - X_{2q})^2 + \dots + (X_{kp} - X_{kq})^2} = \sqrt{\sum_{j=1}^k (X_{jp} - X_{jq})^2} \quad (9.4)$$

Embora a distância euclidiana seja a mais comumente utilizada em análises de agrupamentos, existem outras medidas de dissimilaridade que podem ser utilizadas, e a adoção de cada uma delas depende dos pressupostos e dos objetivos do pesquisador. Na sequência, apresentamos outras medidas de dissimilaridade que podem ser utilizadas:

- **Distância quadrática euclidiana:** alternativamente à distância euclidiana, pode ser utilizada quando as variáveis apresentarem pequena dispersão de seus valores, fazendo com que o uso da distância euclidiana ao quadrado facilite a interpretação dos *outputs* da análise e a alocação das observações nos grupos. Sua expressão é dada por:

$$d_{pq} = (X_{1p} - X_{1q})^2 + (X_{2p} - X_{2q})^2 + \dots + (X_{kp} - X_{kq})^2 = \sum_{j=1}^k (X_{jp} - X_{jq})^2 \quad (9.5)$$

- **Distância de Minkowski:** é a expressão de medida de dissimilaridade mais geral a partir da qual outras derivam. É dada por:

$$d_{pq} = \left[\sum_{j=1}^k (|X_{jp} - X_{jq}|)^m \right]^{\frac{1}{m}} \quad (9.6)$$

em que m assume valores inteiros e positivos ($m = 1, 2, \dots$). Podemos verificar que a distância euclidiana é um caso particular da distância de Minkowski, quando $m = 2$.

- **Distância de Manhattan:** também conhecida por **distância absoluta** ou **bloco**, não leva em consideração a geometria triangular inerente à expressão inicial de Pitágoras e considera apenas as diferenças entre os valores de cada variável. Sua expressão, também um caso particular da distância de Minkowski quando $m = 1$, é dada por:

$$d_{pq} = \sum_{j=1}^k |X_{jp} - X_{jq}| \quad (9.7)$$

- **Distância de Chebychev:** também conhecida por **distância infinita** ou **máxima**, é um caso particular da distância de Manhattan por considerar, para duas determinadas observações, apenas a máxima diferença entre todas as j variáveis em estudo. Sua expressão é dada por:

$$d_{pq} = \max |X_{jp} - X_{jq}| \quad (9.8)$$

também um caso particular da distância de Minkowski quando $m = \infty$.

- **Distância de Canberra:** utilizada para os casos em que as variáveis apresentam apenas valores positivos, assume valores entre 0 e j (número de variáveis). Sua expressão é dada por:

$$d_{pq} = \sum_{j=1}^k \frac{|X_{jp} - X_{jq}|}{(X_{jp} + X_{jq})} \quad (9.9)$$

Na presença de variáveis métricas, o pesquisador ainda pode fazer uso da **correlação de Pearson**, que, embora não seja uma medida de dissimilaridade (na realidade, é uma medida de similaridade), pode propiciar informações importantes quando o intuito for agrupar linhas do banco de dados. A expressão da correlação de Pearson entre os valores de duas observações quaisquer p e q pode ser escrita como:

$$\rho_{pq} = \frac{\sum_{j=1}^k (X_{jp} - \bar{X}_p) \cdot (X_{jq} - \bar{X}_q)}{\sqrt{\sum_{j=1}^k (X_{jp} - \bar{X}_p)^2} \cdot \sqrt{\sum_{j=1}^k (X_{jq} - \bar{X}_q)^2}} \quad (9.10)$$

em que \bar{X}_p e \bar{X}_q representam, respectivamente, a média de todos os valores das variáveis para as observações p e q , ou seja, a média de cada uma das linhas do banco de dados.

Podemos notar, portanto, que estamos lidando com um coeficiente de correlação entre linhas, e não entre colunas (variáveis), o mais comum em análise de dados, e seus valores variam entre -1 e 1 . **O coeficiente de correlação de Pearson pode ser utilizado como medida de similaridade entre as linhas do banco de dados em análises que envolvem, por exemplo, séries de tempo, ou seja, para os casos em que as observações representam períodos.** Nesse caso, o pesquisador pode ter a intenção de estudar correlações entre períodos distintos, para investigar, por exemplo, uma eventual **recorrência de comportamento em linha para o conjunto de variáveis**, o que pode fazer determinados períodos, não necessariamente subsequentes, serem agrupados por similaridade de comportamento.

Voltando aos dados apresentados na Tabela 9.1, podemos calcular as diferentes medidas de distância entre as observações 1 e 2, dadas pelas expressões (9.4) a (9.9), assim como a medida de similaridade correlacional, dada pela expressão (9.10). A Tabela 9.3 apresenta esses cálculos e os respectivos resultados.

Com base nesses resultados, podemos verificar que medidas diferentes geram resultados distintos, o que pode fazer as observações serem alocadas em diferentes agrupamentos homogêneos, dependendo da escolha da medida para análise, conforme discutem Vicini e Souza (2005) e Malhotra (2012). Nesse sentido, é de fundamental importância que o pesquisador sempre embase sua escolha e tenha em mente as razões que o levaram a utilizar determinada medida, em detrimento das demais. A própria utilização de mais de uma medida, quando da análise do mesmo banco de dados, pode sustentar essa decisão, visto que os resultados podem, nesse caso, ser comparados.

Tabela 9.3 Medidas de distância e de similaridade correlacional entre as observações 1 e 2.

Observação i	X_{1i}	X_{2i}	X_{3i}	Média
1	3,7	2,7	9,1	5,167
2	7,8	8,0	1,5	5,767

Distância euclidiana	
$d_{12} = \sqrt{(3,7 - 7,8)^2 + (2,7 - 8,0)^2 + (9,1 - 1,5)^2} = 10,132$	
Distância quadrática euclidiana	
$d_{12} = (3,7 - 7,8)^2 + (2,7 - 8,0)^2 + (9,1 - 1,5)^2 = 102,660$	
Distância de Manhattan	
$d_{12} = 3,7 - 7,8 + 2,7 - 8,0 + 9,1 - 1,5 = 17,000$	
Distância de Chebychev	
$d_{12} = 9,1 - 1,5 = 7,600$	
Distância de Canberra	
$d_{12} = \frac{ 3,7 - 7,8 }{(3,7 + 7,8)} + \frac{ 2,7 - 8,0 }{(2,7 + 8,0)} + \frac{ 9,1 - 1,5 }{(9,1 + 1,5)} = 1,569$	
Correlação de Pearson (similaridade)	
$\rho_{12} = \frac{(3,7 - 5,167) \cdot (7,8 - 5,767) + (2,7 - 5,167) \cdot (8,0 - 5,767) + (9,1 - 5,167) \cdot (1,5 - 5,767)}{\sqrt{(3,7 - 5,167)^2 + (2,7 - 5,167)^2 + (9,1 - 5,167)^2} \cdot \sqrt{(7,8 - 5,767)^2 + (8,0 - 5,767)^2 + (1,5 - 5,767)^2}} = -0,993$	

Esse caso fica bastante visível quando incluímos uma terceira observação na análise, conforme mostra a Tabela 9.4.

Tabela 9.4 Parte de banco de dados com três observações e três variáveis métricas.

Observação i	X_{1i}	X_{2i}	X_{3i}
1	3,7	2,7	9,1
2	7,8	8,0	1,5
3	8,9	1,0	2,7

Enquanto a distância euclidiana sugere que as observações mais similares (menor distância) são a 2 e a 3, por meio da distância de Chebychev as observações 1 e 3 são as mais similares. A Tabela 9.5 apresenta essas distâncias para cada par de observações, com destaque, em negrito, para o menor valor de cada distância.

Tabela 9.5 Distância euclidiana e de Chebychev entre os pares de observações da Tabela 9.4.

Distância	Par de Observações 1 e 2	Par de Observações 1 e 3	Par de Observações 2 e 3
Euclidiana	$d_{12} = 10,132$	$d_{13} = 8,420$	$d_{23} = \mathbf{7,187}$
Chebychev	$d_{12} = 7,600$	$d_{13} = \mathbf{6,400}$	$d_{23} = 7,000$

Portanto, em determinado esquema de aglomeração, teríamos, apenas em função da escolha da medida de dissimilaridade, agrupamentos iniciais distintos.

Além da decisão sobre a escolha da medida de distância, o pesquisador também deve verificar se os dados precisam ser preliminarmente tratados. Nos exemplos abordados até o presente momento, tomamos o cuidado de apresentar variáveis métricas sempre com valores na mesma unidade de medida (por exemplo, notas de Matemática, Física e Química, que variam de 0 a 10). Entretanto, caso as variáveis sejam medidas em unidades distintas (por exemplo, renda em R\$, escolaridade em anos de estudo e quantidade de filhos), a intensidade das distâncias entre as observações poderá ser influenciada arbitrariamente pelas variáveis que eventualmente apresentarem maior magnitude de seus valores, em detrimento das demais. Nessas situações, o pesquisador deve padronizar os dados, a fim de que a arbitrariedade das unidades de medida seja eliminada, fazendo cada variável ter a mesma contribuição sobre a medida de distância considerada.

O método mais comumente utilizado para padronização de variáveis é conhecido por **procedimento Zscores**, em que, para cada observação i , o valor de uma nova variável padronizada ZX_j é obtido pela subtração do correspondente valor da variável original X_j pela sua média e, na sequência, o valor resultante é dividido pelo seu desvio-padrão, conforme apresentado na expressão (9.11).

$$ZX_{ji} = \frac{X_{ji} - \bar{X}_j}{s_j} \quad (9.11)$$

em que \bar{X} e s representam a média e o desvio-padrão da variável X_j . Dessa forma, independentemente da magnitude dos valores e da natureza das unidades de medida das variáveis originais de um banco de dados, todas as respectivas variáveis padronizadas pelo procedimento **Zscores** terão média igual a 0 e desvio-padrão igual a 1, o que garante a eliminação de eventuais arbitrariedades das unidades de medida sobre a distância entre cada par de observações. Além disso, o procedimento **Zscores** tem a vantagem de não alterar a distribuição da variável original.

Portanto, caso as variáveis originais apresentem unidades de medida distintas, as expressões das medidas de distância (9.4) a (9.9) devem ter os termos X_{jp} e X_{jq} substituídos, respectivamente, por ZX_{jp} e ZX_{jq} . O Quadro 9.1 apresenta essas expressões, com base nas variáveis padronizadas.

Embora a correlação de Pearson não seja uma medida de dissimilaridade (na realidade, é uma medida de similaridade), é relevante comentar que seu uso também requer que as variáveis sejam padronizadas por meio do procedimento Zscores caso não apresentem as mesmas unidades de medida. Caso o intuito fosse agrupar variáveis, que é o objetivo do próximo capítulo (análise fatorial), a padronização de variáveis por meio do procedimento **Zscores** seria, de fato, irrelevante, dado que a análise consistiria em avaliar a correlação entre colunas do banco de dados. Como o objetivo do presente capítulo, por outro lado, é agrupar linhas do banco de dados que representam as observações, a padronização das variáveis faz-se necessária para a elaboração de uma correta análise de agrupamentos.

Quadro 9.1 Expressões das medidas de distância com variáveis padronizadas.

Medida de Distância (Dissimilaridade)	Expressão
Euclidiana	$d_{pq} = \sqrt{\sum_{j=1}^k (ZX_{jp} - ZX_{jq})^2}$
Quadrática euclidiana	$d_{pq} = \sum_{j=1}^k (ZX_{jp} - ZX_{jq})^2$
Minkowski	$d_{pq} = \left[\sum_{j=1}^k (ZX_{jp} - ZX_{jq})^m \right]^{\frac{1}{m}}$
Manhattan	$d_{pq} = \sum_{j=1}^k ZX_{jp} - ZX_{jq} $
Chebychev	$d_{pq} = \max ZX_{jp} - ZX_{jq} $
Canberra	$d_{pq} = \sum_{j=1}^k \frac{ ZX_{jp} - ZX_{jq} }{(ZX_{jp} + ZX_{jq})}$

9.2.1.2. Medidas de semelhança (similaridade) entre observações para variáveis binárias

Imagine agora que tenhamos a intenção de calcular a distância entre duas determinadas observações i ($i = 1, 2$) provenientes de um banco de dados que apresenta sete variáveis (X_{1i}, \dots, X_{7i}), porém, todas referentes à presença ou ausência de características. Nessa situação, é comum que a presença ou ausência de determinada característica seja representada por uma **variável binária**, ou **dummy**, que assume valor 1, caso a característica ocorra, e 0, caso contrário. Esses dados encontram-se na Tabela 9.6.

É importante ressaltar que o artifício das variáveis binárias não gera problemas de **ponderação arbitrária**, oriunda das categorias das variáveis, ao contrário do que ocorreria caso fossem atribuídos valores discretos (1, 2, 3, ...) para cada categoria de cada variável qualitativa. Nesse sentido, caso determinada variável qualitativa apresente k categorias, serão necessárias $(k-1)$ variáveis binárias que representarão a presença ou a ausência de cada uma das categorias, ficando todas as variáveis binárias iguais a 0 para o caso de ocorrer a categoria de referência.

Tabela 9.6 Parte de banco de dados com duas observações e sete variáveis binárias.

Observação i	X_{1i}	X_{2i}	X_{3i}	X_{4i}	X_{5i}	X_{6i}	X_{7i}
1	0	0	1	1	0	1	1
2	0	1	1	1	1	0	1

Portanto, fazendo uso da expressão (9.4), podemos calcular a distância quadrática euclidiana entre as observações 1 e 2, conforme segue:

$$d_{12} = \sum_{j=1}^7 (X_{j1} - X_{j2})^2 = (0-0)^2 + (0-1)^2 + (1-1)^2 + (1-1)^2 + (0-1)^2 + (1-0)^2 + (1-1)^2 = 3$$

que representa o número total de variáveis com diferenças de resposta entre as observações 1 e 2.

Logo, para duas quaisquer observações p e q , quanto maior a quantidade de respostas iguais (0-0 ou 1-1), menor a distância quadrática euclidiana entre elas, visto que:

$$(X_{jp} - X_{jq})^2 = \begin{cases} 0 & \text{se } X_{jp} = X_{jq} \\ 1 & \text{se } X_{jp} \neq X_{jq} \end{cases} \quad (9.12)$$

Conforme discutem Johnson e Wichern (2007), cada parcela da distância representada pela expressão (9.12) é considerada uma medida de dissimilaridade, uma vez que quantidades maiores de discrepâncias de resposta resultam em maiores distâncias quadráticas euclidianas. Por outro lado, os cálculos ponderam igualmente os pares de respostas 0-0 e 1-1, sem importância relativa superior ao par de respostas 1-1 que, em muitos casos, é um indicador mais forte de similaridade que o par de respostas 0-0. Por exemplo, ao se agruparem pessoas, o fato de duas delas comerem lagosta todos os dias é uma evidência mais forte de similaridade que a ausência dessa característica para ambas.

Nesse sentido, muitos autores, com o intuito de que fossem criadas medidas de semelhança entre observações, propuseram a utilização de coeficientes que levassem em consideração a similaridade de respostas 1-1 e 0-0, sem que necessariamente esses pares tivessem a mesma importância relativa. Para que possamos apresentar essas medidas, é necessário construir uma tabela de frequências absolutas de respostas 0 e 1 para cada par de observações quaisquer p e q , conforme mostra a Tabela 9.7.

Tabela 9.7 Frequências absolutas de respostas 0 e 1 para duas observações p e q .

Observação q \ Observação p	1	0	Total
1	a	b	$a + b$
0	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

Com base nessa tabela, apresentamos, a seguir, as principais medidas de semelhança existentes, lembrando que a adoção de cada uma depende dos pressupostos e dos objetivos do pesquisador.

- **Medida de emparelhamento simples:** é a medida de similaridade mais utilizada para variáveis binárias, sendo discutida e utilizada por Zubin (1938a) e Sokal e Michener (1958). Essa medida, que iguala os pesos das respostas convergentes 1-1 e 0-0, tem sua expressão dada por:

$$s_{pq} = \frac{a + d}{a + b + c + d} \quad (9.13)$$

- **Medida de Jaccard:** embora tenha sido primeiramente proposta por Gilbert (1894), levou esse nome por ter sido discutida e utilizada em dois seminais trabalhos desenvolvidos por Jaccard (1901, 1908). Essa medida não leva em conta a frequência do par de respostas 0-0, considerada irrelevante. Entretanto, é possível que ocorra uma situação em que todas as variáveis sejam iguais a 0 para duas determinadas observações, ou seja, somente exista frequência na célula d da Tabela 9.7. Nesse caso, softwares como o Stata apresentam medida de Jaccard igual a 1, o que faz sentido do ponto de vista de similaridade. Sua expressão geral é dada por:

$$s_{pq} = \frac{a}{a + b + c} \quad (9.14)$$

- **Medida de Dice:** embora conhecida apenas por esse nome, foi sugerida e discutida por Czekanowski (1932), Dice (1945) e Sørensen (1948). É similar ao coeficiente de Jaccard, porém dobra o peso sobre a frequência de pares de respostas em convergência do tipo 1-1. Assim como naquele caso, softwares como o Stata apresentam medida de Dice igual a 1 para os casos em que todas as variáveis sejam iguais a 0 para duas determinadas observações, evitando, assim, a indefinição do cálculo. Sua expressão é dada por:

$$s_{pq} = \frac{2a}{2 \cdot a + b + c} \quad (9.15)$$

- **Medida antiDice:** proposta inicialmente por Sokal e Sneath (1963) e Anderberg (1973), a nomenclatura antiDice decorre do fato de que esse coeficiente dobra o peso sobre as frequências de pares de respostas diferentes do tipo 1-1, ou seja, dobra o peso sobre as divergências de respostas. Assim como as medidas de Jaccard e de Dice, a medida antiDice também ignora a frequência de pares de respostas 0-0. Sua expressão é dada por:

$$s_{pq} = \frac{a}{a + 2 \cdot (b + c)} \quad (9.16)$$

- **Medida de Russell e Rao:** também bastante utilizada, privilegia, no cálculo de seu coeficiente, apenas as similaridades das respostas 1-1. Foi proposta por Russell e Rao (1940), tendo sua expressão dada por:

$$s_{pq} = \frac{a}{a+b+c+d} \quad (9.17)$$

- **Medida de Ochiai:** embora conhecida por esse nome, foi proposta inicialmente por Driver e Kroeber (1932), sendo utilizada posteriormente por Ochiai (1957). Esse coeficiente é indefinido quando uma ou ambas as observações estudadas apresentarem os valores de todas as variáveis iguais a 0. Entretanto, se ambos os vetores apresentarem todos os valores iguais a 0, softwares como o Stata oferecem medida de Ochiai igual a 1. Se esse fato ocorrer para apenas um dos dois vetores, a medida de Ochiai é considerada igual a 0. Sua expressão é dada por:

$$s_{pq} = \frac{a}{\sqrt{(a+b) \cdot (a+c)}} \quad (9.18)$$

- **Medida de Yule:** proposta por Yule (1900) e utilizada por Yule e Kendall (1950), essa medida de semelhança para variáveis binárias oferece como resposta um coeficiente que varia de -1 a 1. Conforme podemos verificar, por meio de sua expressão apresentada a seguir, o coeficiente gerado é indefinido se um ou ambos os vetores comparados apresentarem todos os valores iguais a 0 ou 1. Softwares como o Stata geram medida de Yule igual a 1, se $b = c = 0$ (convergência total de respostas), e igual a -1, se $a = d = 0$ (divergência total de respostas).

$$s_{pq} = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c} \quad (9.19)$$

- **Medida de Rogers e Tanimoto:** essa medida, que dobra o peso das respostas discrepantes 0-1 e 1-0 em relação ao peso das combinações de respostas convergentes do tipo 1-1 e 0-0, foi inicialmente proposta por Rogers e Tanimoto (1960). Sua expressão, que passa a ser igual à da medida antiDice quando a frequência de respostas 0-0 for igual a 0 ($d = 0$), é dada por:

$$s_{pq} = \frac{a+d}{a+d+2 \cdot (b+c)} \quad (9.20)$$

- **Medida de Sneath e Sokal:** ao contrário da medida de Rogers e Tanimoto, essa medida, proposta por Sneath e Sokal (1962), dobra o peso das respostas convergentes do tipo 1-1 e 0-0 em relação ao das demais combinações de respostas (1-0 e 0-1). Sua expressão, que passa a ser igual à da medida Dice quando a frequência de respostas do tipo 0-0 for igual a 0 ($d = 0$), é dada por:

$$s_{pq} = \frac{2 \cdot (a+d)}{2 \cdot (a+d) + b+c} \quad (9.21)$$

- **Medida de Hamann:** Hamann (1961) propõe essa medida de semelhança para variáveis binárias com o intuito de que fossem subtraídas as frequências de respostas discrepantes (1-0 e 0-1) do total de respostas convergentes (1-1 e 0-0). Esse coeficiente, que varia de -1 (divergência total de repostas) a 1 (convergência total de respostas), é igual a duas vezes a medida de emparelhamento simples menos 1. Sua expressão é dada por:

$$s_{pq} = \frac{(a+d) - (b+c)}{a+b+c+d} \quad (9.22)$$

Assim como o elaborado na seção 9.2.1.1 em relação às medidas de dissimilaridade aplicadas a variáveis métricas, vamos voltar aos dados apresentados na Tabela 9.6, com o intuito de calcular as diferentes medidas de similaridade entre as observações 1 e 2, que apresentam apenas variáveis binárias. Para tanto, devemos, a partir daquela tabela, construir a tabela de frequências absolutas de respostas 0 e 1 para as referidas observações (Tabela 9.8).

Tabela 9.8 Frequências absolutas de respostas 0 e 1 para as observações 1 e 2.

Observação 2 \ Observação 1	1	0	Total
1	3	2	5
0	1	1	2
Total	4	3	7

Logo, fazendo uso das expressões (9.13) a (9.22), temos condições de calcular as medidas de similaridade propriamente ditas. A Tabela 9.9 apresenta os cálculos e os resultados de cada medida.

Tabela 9.9 Medidas de semelhança (similaridade) entre as observações 1 e 2.

Emparelhamento Simples $s_{12} = \frac{3+1}{7} = 0,571$	Jaccard $s_{12} = \frac{3}{6} = 0,500$
Dice $s_{12} = \frac{2 \cdot (3)}{2 \cdot (3) + 2 + 1} = 0,667$	AntiDice $s_{12} = \frac{3}{3 + 2 \cdot (2 + 1)} = 0,333$
Russell e Rao $s_{12} = \frac{3}{7} = 0,429$	Ochiai $s_{12} = \frac{3}{\sqrt{(3+2) \cdot (3+1)}} = 0,671$
Yule $s_{12} = \frac{3 \cdot 1 - 2 \cdot 1}{3 \cdot 1 + 2 \cdot 1} = 0,200$	Rogers e Tanimoto $s_{12} = \frac{3+1}{3+1+2 \cdot (2+1)} = 0,400$
Sneath e Sokal $s_{12} = \frac{2 \cdot (3+1)}{2 \cdot (3+1) + 2 + 1} = 0,727$	Hamann $s_{12} = \frac{(3+1) - (2+1)}{7} = 0,143$

Analogamente ao discutido quando do cálculo das medidas de dissimilaridade, é visível que medidas de similaridade diferentes geram resultados distintos, o que pode fazer, quando da elaboração do método de aglomeração, que as observações sejam alocadas em diferentes agrupamentos homogêneos, dependendo da escolha da medida para análise.

Lembramos que não faz sentido algum aplicar o procedimento de padronização Zscores para o cálculo das medidas de semelhança discutidas nesta seção, visto que as variáveis utilizadas para a análise de agrupamentos são binárias.

Neste momento, é importante ressaltar que, em vez de serem utilizadas medidas de semelhança para a definição de *clusters* quando da presença de variáveis binárias, é bastante comum que se definam agrupamentos a partir de coordenadas de cada observação, que podem ser geradas quando da elaboração de uma **análise de correspondência** (simples ou múltipla), técnica exploratória aplicada apenas e tão somente a bancos de dados que oferecem variáveis qualitativas, com o intuito de elaborar **mapas perceptuais** construídos com base nas frequências das categorias de cada uma das variáveis em análise. Essa técnica será estudada no Capítulo 11.

Definida a medida a ser utilizada, com base nos objetivos de pesquisa, na teoria subjacente e em sua experiência e intuição, o pesquisador deve partir para a definição do esquema de aglomeração. Os principais esquemas em análise de agrupamentos serão estudados na próxima seção.

9.2.2. Esquemas de aglomeração em análise de agrupamentos

Conforme discutem Vicini e Souza (2005) e Johnson e Wichern (2007), na análise de agrupamentos, a escolha do método de aglomeração, também conhecido como **esquema de aglomeração**, é tão importante quanto a definição da medida de distância (ou de semelhança), e essa decisão também precisa ser tomada com base naquilo que o pesquisador pretende em termos de objetivos de pesquisa.

Os esquemas de aglomeração podem ser classificados, basicamente, em dois tipos, conhecidos por **hierárquicos** e **não hierárquicos**. Enquanto os primeiros caracterizam-se por privilegiar uma estrutura hierárquica (passo a passo) para a formação dos agrupamentos, os esquemas não hierárquicos utilizam algoritmos para maximizar a homogeneidade dentro de cada agrupamento, sem que haja um processo hierárquico para tal.

Os esquemas de aglomeração hierárquicos podem ser **aglomerativos** ou **divisivos**, dependendo do modo como é iniciado o processo. Caso todas as observações sejam consideradas separadas e, a partir de suas distâncias (ou semelhanças), sejam formados grupos até que se chegue a um estágio final com apenas um agrupamento, então esse processo é conhecido como aglomerativo. Dentre os esquemas hierárquicos aglomerativos, são mais comumente utilizados aqueles que apresentam **método de encadeamento** do tipo **único** (*nearest neighbor* ou *single linkage*), **completo** (*furthest neighbor* ou *complete linkage*) ou **médio** (*between groups* ou *average linkage*). Por outro lado, caso todas as observações sejam consideradas agrupadas e, estágio após estágio, sejam formados grupos menores pela separação de cada observação, até que essas subdivisões gerem grupos individuais (ou seja, observações totalmente separadas), então, estaremos diante de um processo divisivo.

Já os esquemas de aglomeração não hierárquicos, entre os quais o mais popular é o procedimento **k-means**, ou **k-médias**, referem-se a processos em que são definidos centros de aglomeração a partir dos quais são alocadas as observações pela proximidade a eles. Ao contrário dos esquemas hierárquicos, em que o pesquisador pode estudar as diversas possibilidades de alocação das observações e até definir uma quantidade interessante de *clusters* com base em cada um dos estágios de agrupamento, um esquema de aglomeração não hierárquico requer a estipulação, *a priori*, da quantidade de *clusters* a partir da qual serão definidos os centros de aglomeração e alocadas as observações. É por essa razão que se recomenda a elaboração de um esquema de aglomeração hierárquico preliminarmente à de um esquema não hierárquico, quando não há uma estimativa razoável da quantidade de *clusters* que podem ser formados a partir das observações do banco de dados e com base nas variáveis em estudo.

A Figura 9.9 apresenta a lógica dos esquemas de aglomeração em análise de agrupamentos.

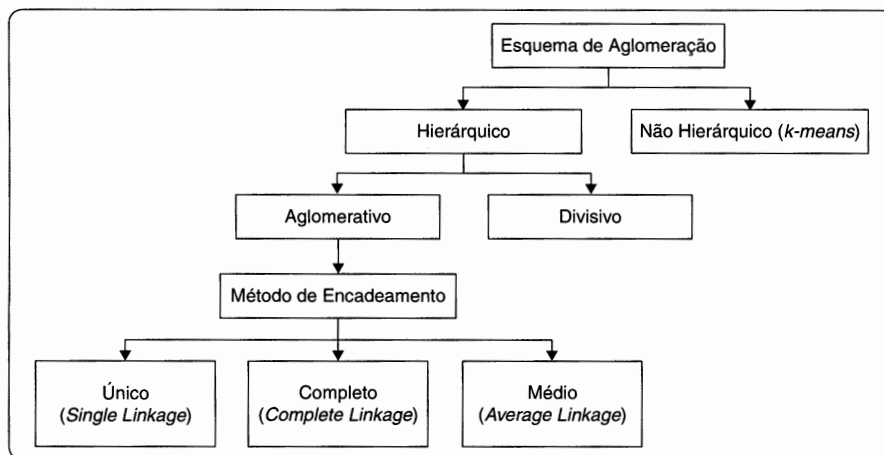


Figura 9.9 Esquemas de aglomeração em análise de agrupamentos.

Enquanto estudaremos os esquemas de aglomeração hierárquicos na seção 9.2.2.1, a seção 9.2.2.2 é destinada ao estudo do esquema de aglomeração não hierárquico *k-means*.

9.2.2.1. Esquemas de aglomeração hierárquicos

Nesta seção, apresentaremos os principais esquemas hierárquicos aglomerativos, em que são formados agrupamentos cada vez maiores a cada estágio de aglomeração pela junção de novas observações ou grupos, em função de determinado critério (método de encadeamento) e com base na medida de distância escolhida. Na seção 9.2.2.1.1 serão apresentados os principais conceitos pertinentes a esses esquemas, e na seção 9.2.2.1.2 será elaborado um exemplo prático resolvido algebricamente.

9.2.2.1.1. Notação

Três são os principais métodos de encadeamento em esquemas hierárquicos aglomerativos, conforme mostra a Figura 9.9: método de encadeamento único (*nearest neighbor* ou *single linkage*), completo (*furthest neighbor* ou *complete linkage*) e médio (*between groups* ou *average linkage*).

A Tabela 9.10 apresenta, de forma ilustrativa, a distância a ser considerada em cada estágio de aglomeração, em função do método de encadeamento escolhido.

Tabela 9.10 Distância a ser considerada em função do método de encadeamento.

Método de Encadeamento	Ilustração	Distância (Dissimilaridade)
Único (<i>Nearest Neighbor</i> ou <i>Single Linkage</i>)		d_{23}
Completo (<i>Furthest Neighbor</i> ou <i>Complete Linkage</i>)		d_{15}
Médio (<i>Between Groups</i> ou <i>Average Linkage</i>)		$\frac{d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25}}{6}$

O método de encadeamento único privilegia as menores distâncias (daí vem a nomenclatura *nearest neighbor*) para que sejam formados novos agrupamentos a cada estágio de aglomeração pela incorporação de observações ou grupos. Nesse sentido, **sua aplicação é recomendável para os casos em que as observações sejam relativamente afastadas**, isto é, diferentes, e deseja-se formar agrupamentos levando-se em consideração um mínimo de homogeneidade. Por outro lado, sua análise fica prejudicada quando da existência de observações ou agrupamentos pouco afastados entre si, conforme mostra a Figura 9.10.

Já o método de encadeamento completo vai em direção contrária, ou seja, privilegia as maiores distâncias entre as observações ou grupos para que sejam formados novos agrupamentos (daí, a nomenclatura *furthest neighbor*) e, dessa maneira, sua adoção é **recomendável para os casos em que não exista considerável afastamento entre as observações** e o pesquisador tenha a necessidade de identificar heterogeneidades entre elas.

Por fim, no método de encadeamento médio dois grupos sofrem fusão com base na **distância média entre todos os pares de observações pertencentes a esses grupos** (daí, a nomenclatura *average linkage*). Dessa forma, embora ocorram alterações no cálculo das medidas de distância entre os agrupamentos, o método de encadeamento médio acaba por preservar a solução de ordenamento das observações em cada grupo, oferecida pelo método de encadeamento único, caso haja um considerável afastamento entre as observações. O mesmo vale em relação à solução de ordenamento oferecida pelo método de encadeamento completo, caso as observações sejam bastante próximas entre si.

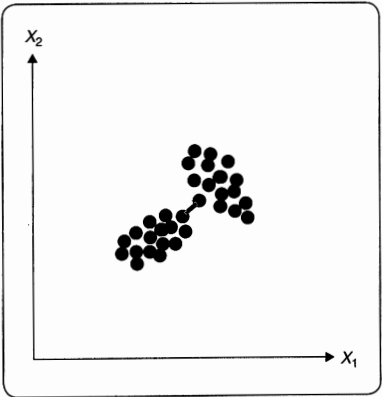


Figura 9.10 Método de encadeamento único – Análise prejudicada na existência de observações ou agrupamentos pouco afastados.

Johnson e Wichern (2007) propõem uma sequência lógica de passos para que se facilite o entendimento da análise de agrupamentos, elaborada por meio de determinado método hierárquico aglomerativo:

1. Sendo n a quantidade de observações de um banco de dados, devemos dar início ao esquema de aglomeração com exatamente n grupos individuais (estágio 0), de modo que teremos inicialmente uma matriz de distâncias (ou de semelhanças) \mathbf{D}_0 composta pelas distâncias entre cada par de observações.
2. No primeiro estágio, devemos escolher a menor distância entre todas as que compõem a matriz \mathbf{D}_0 , ou seja, aquela que une as duas observações mais similares. Nesse exato momento, deixamos de ter n grupos individuais para termos $(n - 1)$ grupos, sendo um deles formado por duas observações.
3. No estágio de aglomeração seguinte, devemos repetir o estágio anterior, porém agora levando em consideração a distância entre cada par de observações e entre o primeiro grupo já formado e cada uma das demais observações, com base em um dos métodos de encadeamento adotado. Em outras palavras, teremos, após o primeiro estágio de aglomeração, uma matriz \mathbf{D}_1 , com dimensões $(n - 1) \times (n - 1)$, em que uma das linhas será representada pelo primeiro par agrupado de observações. No segundo estágio, consequentemente, um novo grupo será formado pelo agrupamento de duas novas observações ou pela junção de determinada observação ao primeiro grupo já formado anteriormente, no primeiro estágio.
4. O processo anterior deve ser repetido $(n - 1)$ vezes, até que reste apenas um único grupo formado por todas as observações. Em outras palavras, no estágio $(n - 2)$ teremos uma matriz \mathbf{D}_{n-2} que conterá apenas a distância entre os dois últimos grupos remanescentes, antes da fusão final.
5. Por fim, a partir dos estágios de aglomeração e das distâncias entre os agrupamentos formados, é possível construir um gráfico em formato de árvore, que resume o processo de aglomeração e explicita a alocação de cada observação em cada agrupamento. Esse gráfico é conhecido como **dendrograma** ou **fenograma**.

Portanto, os valores que compõem as matrizes \mathbf{D} de cada um dos estágios serão função da medida de distância escolhida e do método de encadeamento adotado. Imagine, em determinado estágio de aglomeração s , que um pesquisador agrupe dois *clusters* M e N já formados anteriormente, contendo, respectivamente, m e n observações, a fim de que seja formado o *cluster* MN . Na sequência, tem a intenção de agrupar MN com outro *cluster* W , com w observações. Como sabemos que a decisão de escolha do próximo agrupamento será sempre a menor distância entre cada par de observações ou grupos nos métodos hierárquicos aglomerativos, o esquema de aglomeração será de fundamental importância para que sejam analisadas as distâncias que compõem cada matriz \mathbf{D}_s . A partir dessa lógica, e com base na Tabela 9.10, apresentamos, a seguir, o critério de cálculo da distância, inserida na matriz \mathbf{D}_s , entre os *clusters* MN e W , em função do método de encadeamento:

- **Método de Encadeamento Único (Nearest Neighbor ou Single Linkage)**

$$d_{(MN)W} = \min\{d_{MW}; d_{NW}\} \quad (9.23)$$

em que d_{MW} e d_{NW} são as distâncias entre as observações mais próximas dos *clusters* M e W e dos *clusters* N e W , respectivamente.

- **Método de Encadeamento Completo (Furthest Neighbor ou Complete Linkage)**

$$d_{(MN)W} = \max\{d_{MW}; d_{NW}\} \quad (9.24)$$

em que d_{MW} e d_{NW} são as distâncias entre as observações mais distantes dos *clusters* M e W e dos *clusters* N e W , respectivamente.

- **Método de Encadeamento Médio (Between Groups ou Average Linkage)**

$$d_{(MN)W} = \frac{\sum_{p=1}^{m+n} \sum_{q=1}^w d_{pq}}{(m+n) \cdot (w)} \quad (9.25)$$

em que d_{pq} representa a distância entre qualquer observação p do *cluster* MN e qualquer observação q do *cluster* W , e $m+n$ e w representam, respectivamente, a quantidade de observações nos *clusters* MN e W .

Na próxima seção, apresentaremos um exemplo prático que será resolvido algebricamente, a partir do qual os conceitos referentes aos métodos hierárquicos aglomerativos poderão ser fixados.

9.2.2.1.2. Exemplo prático de análise de agrupamentos com esquemas de aglomeração hierárquicos

Imagine que o professor de uma faculdade, bastante preocupado com a capacidade de aprendizado dos alunos em sua disciplina de métodos quantitativos, tenha o interesse em alocá-los em grupos com a maior homogeneidade possível, com base nas notas obtidas no vestibular em disciplinas consideradas quantitativas (Matemática, Física e Química).

Nesse sentido, o professor fez um levantamento sobre essas notas, que variam de 0 a 10, e, dado que realizará uma análise de agrupamentos inicialmente de maneira algébrica, resolveu trabalhar, para efeitos didáticos, apenas com cinco alunos. O banco de dados encontra-se na Tabela 9.11.

Tabela 9.11 Exemplo: Notas de Matemática, Física e Química no vestibular.

Estudante (Observação)	Nota de Matemática (X_{1i})	Nota de Física (X_{2i})	Nota de Química (X_{3i})
Gabriela	3,7	2,7	9,1
Luiz Felipe	7,8	8,0	1,5
Patrícia	8,9	1,0	2,7
Ovídio	7,0	1,0	9,0
Leonor	3,4	2,0	5,0

Com base nos dados obtidos, é construído o gráfico da Figura 9.11, e, como as variáveis são métricas, será adotada a medida de dissimilaridade conhecida por distância euclidiana para a análise de agrupamentos. Além disso, **como todas as variáveis apresentam valores na mesma unidade de medida (notas de 0 a 10), não será necessária, nesse caso, a elaboração da padronização pelo procedimento Zscores.**

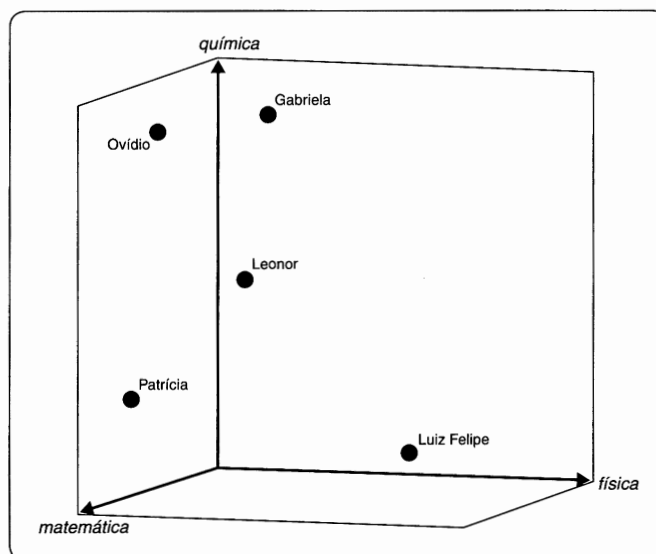


Figura 9.11 Gráfico tridimensional com posição relativa dos cinco estudantes.

Nas próximas seções, serão elaborados os esquemas hierárquicos aglomerativos com base na distância euclidiana, por meio dos três métodos de encadeamento estudados.

9.2.2.1.2.1. Método de encadeamento único (*nearest neighbor* ou *single linkage*)

A partir dos dados apresentados na Tabela 9.11, iremos, neste momento, elaborar uma análise de agrupamentos por meio de um esquema de aglomeração hierárquico com método de encadeamento único. Inicialmente, definimos a matriz D_0 , composta pelas distâncias euclidianas (dissimilaridades) entre cada par de observações, conforme segue:

	Gabriela	Luiz Felipe	Patrícia	Ovídio	Leonor
$D_0 =$ Gabriela	0,000				
Luiz Felipe	10,132	0,000			
Patrícia	8,420	7,187	0,000		
Ovídio	3,713	10,290	6,580	0,000	
Leonor	4,170	8,223	6,045	5,474	0,000

É importante mencionar que, neste momento inicial, cada observação é considerada um *cluster* individual, ou seja, no estágio 0, temos 5 *clusters* (tamanho da amostra). Em destaque, na matriz D_0 , está a menor distância entre todas as observações e, portanto, serão inicialmente agrupadas, no primeiro estágio, as observações **Gabriela** e **Ovídio**, que passam a formar um novo *cluster*.

Para que seja elaborado o próximo estágio de aglomeração, devemos construir a matriz D_1 , em que são calculadas as distâncias entre o *cluster* **Gabriela-Ovídio** e as demais observações, ainda isoladas. Dessa forma, por meio do método de encadeamento único e com base na expressão (9.23), temos que:

$$d_{(\text{Gabriela-Ovídio})\text{Luiz Felipe}} = \text{mín} \{10,132; 10,290\} = 10,132$$

$$d_{(\text{Gabriela-Ovídio})\text{Patrícia}} = \text{mín} \{8,420; 6,580\} = 6,580$$

$$d_{(\text{Gabriela-Ovídio})\text{Leonor}} = \text{mín} \{4,170; 5,474\} = 4,170$$

A matriz D_1 encontra-se a seguir:

	Gabriela Ovídio	Luiz Felipe	Patrícia	Leonor
$D_1 =$ Gabriela	0,000			
Ovídio				
Luiz Felipe	10,132	0,000		
Patrícia	6,580	7,187	0,000	
Leonor	4,170	8,223	6,045	0,000

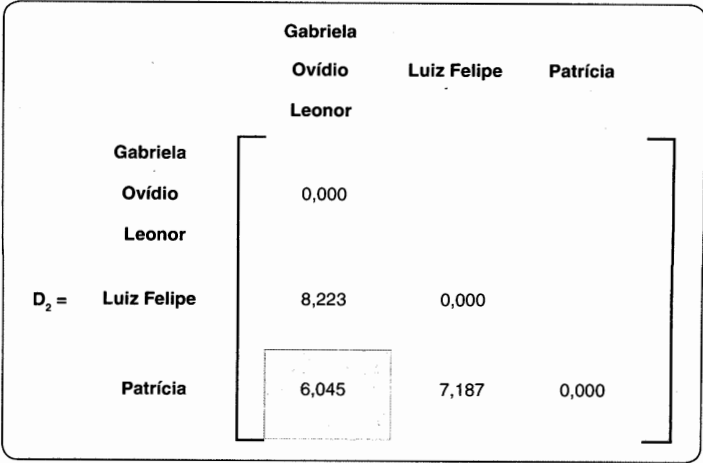
Da mesma forma, na matriz D_1 está em destaque a menor distância entre todas. Portanto, no segundo estágio, é inserida a observação **Leonor** no *cluster* já formado **Gabriela-Ovídio**. As observações **Luiz Felipe** e **Patrícia** permanecem ainda isoladas.

Para que possamos dar o próximo passo, devemos construir a matriz D_2 , em que são calculadas as distâncias entre o *cluster* **Gabriela-Ovídio-Leonor** e as duas observações remanescentes. Analogamente, temos que:

$$d_{(\text{Gabriela-Ovídio-Leonor})\text{Luiz Felipe}} = \text{mín} \{10,132; 8,223\} = 8,223$$

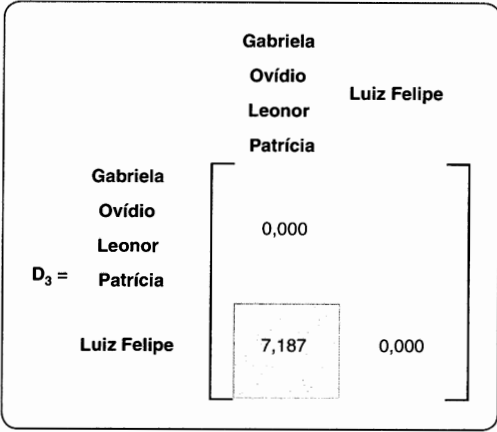
$$d_{(\text{Gabriela-Ovídio-Leonor})\text{Patrícia}} = \text{mín} \{6,580; 6,045\} = 6,045$$

A matriz D_2 pode ser escrita como:



No terceiro estágio de aglomeração, é incorporada a observação **Patrícia** no *cluster* **Gabriela-Ovídio-Leonor**, visto que a correspondente distância é a menor entre todas as apresentadas na matriz **D₂**. Portanto, podemos escrever a matriz **D₃**, que se encontra na sequência, levando em consideração o seguinte critério:

$d_{(Gabriela-Ovídio-Leonor-Patrícia) \text{ Luiz Felipe}} = \text{mín} \{8,223; 7,187\} = 7,187$



Por fim, no quarto e último estágio, todas as observações estão alocadas no mesmo agrupamento, encerrando-se, assim, o processo hierárquico. A Tabela 9.12 apresenta um resumo desse esquema de aglomeração elaborado por meio do método de encadeamento único.

Tabela 9.12 Esquema de aglomeração pelo método de encadeamento único.

Estágio	Agrupamento	Observação Agrupada	Menor Distância Euclidiana
1	Gabriela	Ovídio	3,713
2	Gabriela – Ovídio	Leonor	4,170
3	Gabriela – Ovídio – Leonor	Patrícia	6,045
4	Gabriela – Ovídio – Leonor – Patrícia	Luiz Felipe	7,187

Com base nesse esquema de aglomeração, podemos construir um gráfico em formato de árvore, conhecido como **dendrograma** ou **fenograma**, cujo intuito é ilustrar o passo a passo dos agrupamentos e facilitar a visualização da alocação de cada observação em cada estágio. O dendrograma encontra-se na Figura 9.12.

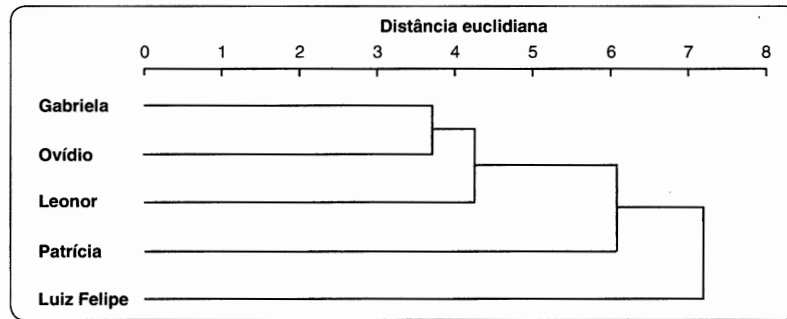


Figura 9.12 Dendrograma – Método de encadeamento único.

Por meio das Figuras 9.13 e 9.14, temos condições de interpretar o dendrograma construído.

Inicialmente, traçamos três linhas (I, II e III) ortogonais às linhas do dendrograma, conforme mostra a Figura 9.13, que permitem identificar as quantidades de agrupamentos em cada estágio de aglomeração, bem como as observações em cada *cluster*.

Assim, a linha I “corta” o dendrograma imediatamente após o primeiro estágio de aglomeração e, neste momento, podemos verificar que existem quatro *clusters* (quatro encontros com as linhas horizontais do dendrograma), um deles formado pelas observações **Gabriela e Ovídio**, e os demais, pelas observações individuais.

Já a linha II encontra três linhas horizontais do dendrograma, o que significa que, após o segundo estágio, em que foi incorporada a observação **Leonor** ao agrupamento já formado **Gabriela-Ovídio**, existem três *clusters*.

Por fim, a linha III é desenhada imediatamente após o terceiro estágio, em que ocorre o agrupamento da observação **Patrícia** com o *cluster* **Gabriela-Ovídio-Leonor**. Como são identificados dois encontros entre essa linha e as linhas horizontais do dendrograma, verificamos que a observação **Luiz Felipe** permanece isolada, enquanto as demais formam um único agrupamento.

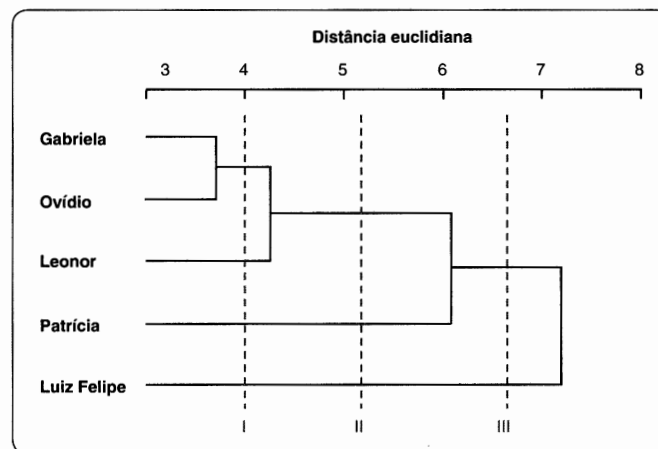


Figura 9.13 Interpretação do dendrograma – Quantidade de *clusters* e alocação das observações.

Além de propiciar o estudo sobre a quantidade de *clusters* em cada estágio de aglomeração, bem como sobre a alocação das observações, o dendrograma também permite que o pesquisador analise a magnitude dos saltos de distância para que se estabeleçam os agrupamentos. Um salto com magnitude elevada, em comparação aos demais, pode indicar que determinada observação ou *cluster* consideravelmente distintos estejam incorporados a agrupamentos já formados, o que fornece subsídios ao estabelecimento de uma solução da quantidade de agrupamentos sem a necessidade de um próximo estágio de aglomeração.

Embora se saiba que a determinação taxativa de uma solução da quantidade de *clusters* pode prejudicar a análise, o estabelecimento de um indício dessa quantidade, dados a medida de distância utilizada e o método de encadeamento adotado, pode fazer o pesquisador compreender mais razoavelmente as características das observações que levaram a esse fato. Além disso, como a quantidade de agrupamentos é importante para a elaboração de esquemas de aglomeração não hierárquicos, essa informação (considerada *output* do esquema hierárquico) pode servir de *input* para o procedimento *k-means*.

A Figura 9.14 apresenta três saltos de distância (A, B e C), referentes a cada um dos estágios de aglomeração, e, a partir de sua análise, podemos verificar que o salto B, que representa a incorporação da observação **Patrícia** ao *cluster* já formado **Gabriela-Ovídio-Leonor**, é o maior dos três. Assim, caso haja a intenção de definir uma quantidade interessante de agrupamentos nesse exemplo, o pesquisador pode optar pela solução com três *clusters* (linha II da Figura 9.13), sem o estágio em que é incorporada a observação **Patrícia**, visto que possivelmente apresenta características não tão homogêneas que inviabilizam sua inclusão no *cluster* já formado, dado o grande salto de distância. Nesse caso, portanto, teríamos um agrupamento formado por **Gabriela, Ovídio e Leonor**, outro formado apenas por **Patrícia** e um terceiro formado apenas por **Luiz Felipe**.

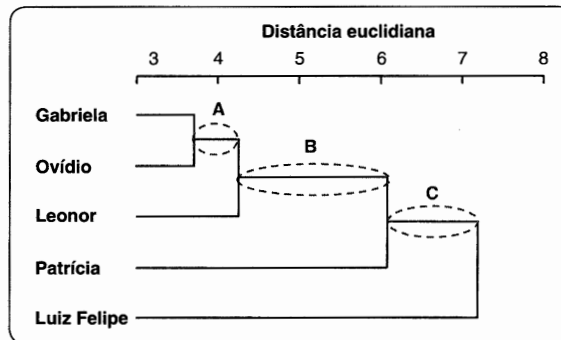


Figura 9.14 Interpretação do dendrograma – Saltos de distância.

Um **critério muito útil para a identificação da quantidade de *clusters***, quando do uso de medidas de dissimilaridade em métodos aglomerativos, consiste em **identificar um considerável salto de distância** (quando possível) e definir a quantidade de agrupamentos formados no estágio de aglomeração imediatamente anterior ao grande salto, visto que **saltos muito elevados podem incorporar observações com características não tão homogêneas**.

Além disso, é relevante também comentar que, caso os saltos de distância de um estágio para outro sejam pequenos, pela existência de variáveis com valores muito próximos para as observações, o que pode dificultar a leitura do dendrograma, **o pesquisador poderá fazer uso da distância quadrática euclidiana, a fim de que os saltos fiquem mais nítidos e explicitados**, facilitando a identificação dos agrupamentos no dendrograma e propiciando melhores argumentos para a tomada de decisão.

Softwares como o SPSS apresentam dendrogramas com medidas de distância rescalonadas, a fim de facilitar a interpretação da alocação de cada observação e a visualização dos grandes saltos de distância.

A Figura 9.15 apresenta, de forma ilustrativa, como podem ser estabelecidos os agrupamentos após a elaboração do método de encadeamento único.

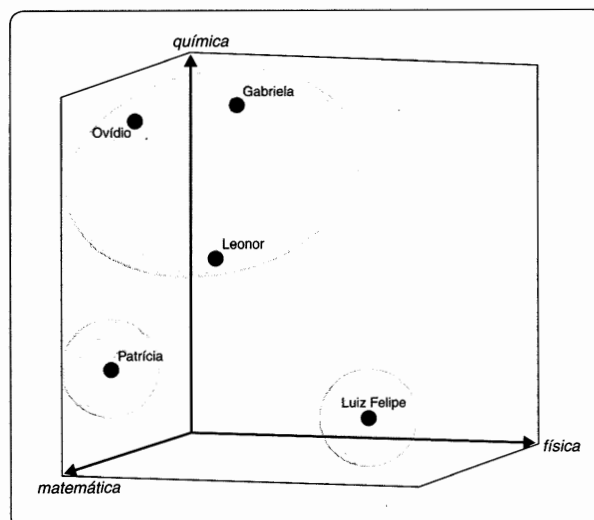


Figura 9.15 Sugestão de agrupamentos formados após o método de encadeamento único.

Na sequência, elaboraremos o mesmo exemplo, porém fazendo uso dos métodos de encadeamento completo e médio, a fim de que possam ser comparados os ordenamentos das observações e os saltos de distância.

9.2.2.1.2.2. Método de encadeamento completo (*furthest neighbor ou complete linkage*)

A matriz D_0 , reproduzida a seguir, é obviamente a mesma, e a menor distância euclidiana, em destaque, ocorre entre as observações **Gabriela** e **Ovídio**, que passam a formar o primeiro agrupamento. Ressalta-se que o primeiro agrupamento será sempre o mesmo, independentemente do método de encadeamento adotado, visto que o primeiro estágio sempre levará em consideração a menor distância entre dois pares de observações ainda isoladas.

	Gabriela	Luiz Felipe	Patrícia	Ovídio	Leonor
Gabriela	0,000				
Luiz Felipe	10,132	0,000			
$D_0 =$ Patrícia	8,420	7,187	0,000		
Ovídio	3,713	10,290	6,580	0,000	
Leonor	4,170	8,223	6,045	5,474	0,000

No método de encadeamento completo, devemos fazer uso da expressão (9.24), a fim de que possa ser construída a matriz D_1 , conforme segue:

$$d_{(\text{Gabriela-Ovídio})\text{Luiz Felipe}} = \max \{10,132; 10,290\} = 10,290$$

$$d_{(\text{Gabriela-Ovídio})\text{Patrícia}} = \max \{8,420; 6,580\} = 8,420$$

$$d_{(\text{Gabriela-Ovídio})\text{Leonor}} = \max \{4,170; 5,474\} = 5,474$$

A matriz D_1 encontra-se a seguir, e, por meio dela, podemos verificar que a observação **Leonor** será incorporada ao *cluster* formado por **Gabriela** e **Ovídio**. Novamente, o menor valor, entre todos apresentados na matriz D_1 , encontra-se em destaque.

	Gabriela Ovídio	Luiz Felipe	Patrícia	Leonor
Gabriela Ovídio	0,000			
Luiz Felipe	10,290	0,000		
$D_1 =$ Patrícia	8,420	7,187	0,000	
Leonor	5,474	8,223	6,045	0,000

Assim como o verificado quando da elaboração do método de encadeamento único, aqui, as observações **Luiz Felipe** e **Patrícia** também permanecem isoladas neste estágio. As diferenças entre os métodos começam a surgir na sequência. Vamos, portanto, construir a matriz D_2 , fazendo uso dos seguintes critérios:

$$d_{(\text{Gabriela-Ovídio-Leonor})\text{Luiz Felipe}} = \text{máx} \{10,290; 8,223\} = 10,290$$

$$d_{(\text{Gabriela-Ovídio-Leonor})\text{Patrícia}} = \text{máx} \{8,420; 6,045\} = 8,420$$

A matriz \mathbf{D}_2 pode ser escrita como:

		Gabriela		
		Ovídio	Luiz Felipe	Patrícia
		Leonor		
$\mathbf{D}_2 =$	Gabriela	[
	Ovídio		0,000	
	Leonor			
	Luiz Felipe		10,290	0,000
	Patrícia		8,420	7,187
				0,000

No terceiro estágio de aglomeração, um novo agrupamento é formado pela fusão das observações **Patrícia** e **Luiz Felipe**, visto que o critério *furthest neighbor* adotado pelo método de encadeamento completo faz a distância entre essas duas observações ser a menor entre todas calculadas para a construção da matriz \mathbf{D}_2 . Note, portanto, que, nesse estágio, ocorrem diferenças em relação ao método de encadeamento único no que diz respeito ao ordenamento e à alocação das observações em grupos.

Para a construção da matriz \mathbf{D}_3 , portanto, devemos levar em consideração o seguinte critério:

$$d_{(\text{Gabriela-Ovídio-Leonor}) (\text{Luiz Felipe-Patrícia})} = \text{máx} \{10,290; 8,420\} = 10,290$$

		Gabriela		Luiz Felipe
		Ovídio		Patrícia
		Leonor		
$\mathbf{D}_3 =$	Gabriela	[
	Ovídio		0,000	
	Leonor			
	Luiz Felipe		10,290	0,000
	Patrícia			

Da mesma forma, no quarto e último estágio, todas as observações estão alocadas no mesmo *cluster*, visto que há o agrupamento de **Gabriela-Ovídio-Leonor** com **Luiz Felipe-Patrícia**. A Tabela 9.13 apresenta um resumo desse esquema de aglomeração, elaborado por meio do método de encadeamento completo.

Tabela 9.13 Esquema de aglomeração pelo método de encadeamento completo.

Estágio	Agrupamento	Observação Agrupada	Menor Distância Euclidiana
1	Gabriela	Ovídio	3,713
2	Gabriela – Ovídio	Leonor	5,474
3	Luiz Felipe	Patrícia	7,187
4	Gabriela – Ovídio – Leonor	Luiz Felipe – Patrícia	10,290

O dendrograma desse esquema de aglomeração encontra-se na Figura 9.16. Podemos inicialmente verificar que o ordenamento das observações é diferente do observado no dendrograma da Figura 9.12.

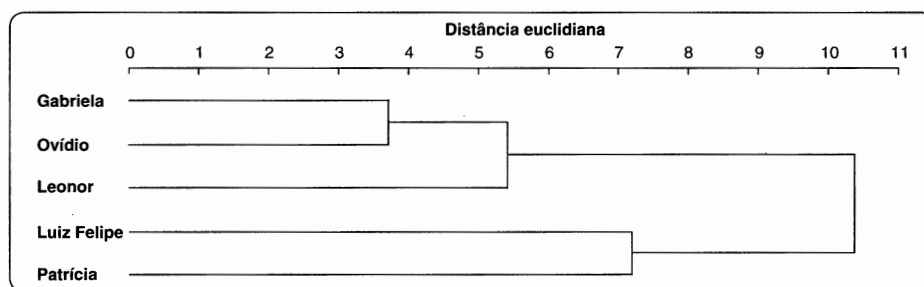


Figura 9.16 Dendrograma – Método de encadeamento completo.

Analogamente ao realizado no método anterior, optamos por desenhar duas linhas verticais (I e II) sobre o maior salto de distância, conforme podemos observar na Figura 9.17.

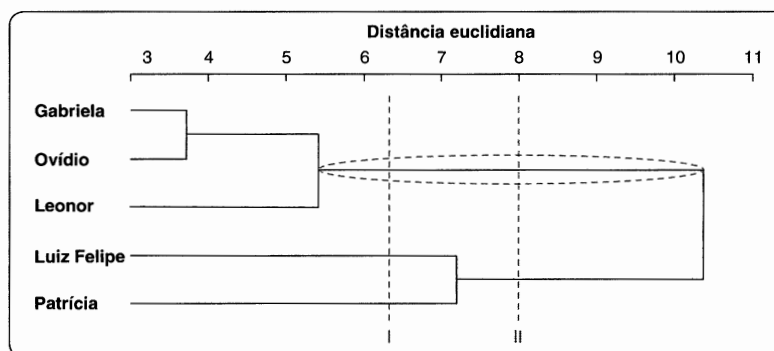


Figura 9.17 Interpretação do dendrograma – *Clusters* e salto de distância.

Logo, caso o pesquisador opte por considerar três *clusters*, a solução ficará igual àquela encontrada anteriormente pelo método de encadeamento único, sendo um composto por **Gabriela**, **Ovídio** e **Leonor**, outro, por **Luiz Felipe**, e um terceiro, por **Patrícia** (linha I da Figura 9.17). Entretanto, caso opte por definir dois agrupamentos (linha II), a solução será diferente, visto que, nesse caso, o segundo *cluster* será formado por **Luiz Felipe** e **Patrícia**, enquanto no caso anterior, era formado apenas por **Luiz Felipe**, já que a observação **Patrícia** fora alocada no primeiro *cluster*.

Analogamente ao realizado no método anterior, a Figura 9.18 apresenta, de forma ilustrativa, como podem ser estabelecidos os agrupamentos após a elaboração do método de encadeamento completo.

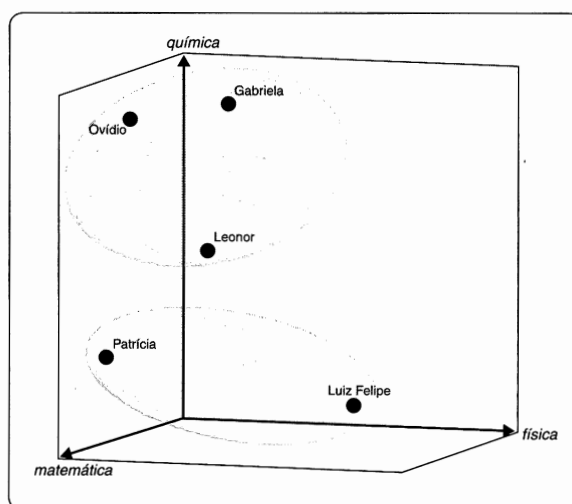


Figura 9.18 Sugestão de agrupamentos formados após o método de encadeamento completo.

A definição do método de aglomeração pode ser embasada pela aplicação do método de encadeamento médio, em que dois grupos sofrem fusão com base na distância média entre todos os pares de observações

pertencentes a esses grupos. Portanto, conforme discutimos, **caso o método mais adequado seja o de encadeamento único pela existência de observações com considerável afastamento, o ordenamento e a alocação das observações serão mantidos pelo método de encadeamento médio**. Por outro lado, **os outputs desse método apresentarão consistência com a solução obtida pelo método de encadeamento completo no que diz respeito ao ordenamento e à alocação das observações, caso estas sejam bastante similares nas variáveis em estudo**.

Neste sentido, é recomendável que o pesquisador aplique os três métodos de encadeamento quando da elaboração de análise de agrupamento por meio de esquemas de aglomeração hierárquicos. Vamos, portanto, ao método de encadeamento médio.

9.2.2.1.2.3. Método de encadeamento médio (*between groups* ou *average linkage*)

Inicialmente, reproduzimos a seguir a matriz de distâncias euclidianas entre cada par de observações (matriz D_0), com destaque novamente para a menor distância entre elas.

	Gabriela	Luiz Felipe	Patrícia	Ovídio	Leonor
$D_0 =$ Gabriela	0,000				
Luiz Felipe	10,132	0,000			
Patrícia	8,420	7,187	0,000		
Ovídio	3,713	10,290	6,580	0,000	
Leonor	4,170	8,223	6,045	5,474	0,000

Com base na expressão (9.25), temos condições de calcular os termos da matriz D_1 , dado que já é formado o primeiro *cluster* **Gabriela-Ovídio**. Assim, temos que:

$$d_{(\text{Gabriela-Ovídio})\text{Luiz Felipe}} = \frac{10,132 + 10,290}{2} = 10,211$$

$$d_{(\text{Gabriela-Ovídio})\text{Patrícia}} = \frac{8,420 + 6,580}{2} = 7,500$$

$$d_{(\text{Gabriela-Ovídio})\text{Leonor}} = \frac{4,170 + 5,474}{2} = 4,822$$

A matriz D_1 encontra-se a seguir, e, por meio dela, podemos verificar que a observação **Leonor** é novamente incorporada ao *cluster* formado por **Gabriela** e **Ovídio**. O menor valor, entre todos apresentados na matriz D_1 , também se encontra em destaque.

	Gabriela Ovídio	Luiz Felipe	Patrícia	Leonor
$D_1 =$ Gabriela	0,000			
Ovídio				
Luiz Felipe	10,211	0,000		
Patrícia	7,500	7,187	0,000	
Leonor	4,822	8,223	6,045	0,000

Para a construção da matriz D_2 , em que são calculadas as distâncias entre o *cluster* **Gabriela-Ovídio-Leonor** e as duas observações remanescentes, devemos elaborar os seguintes cálculos:

$$d_{(\text{Gabriela-Ovídio-Leonor}) \text{ Luiz Felipe}} = \frac{10,132 + 10,290 + 8,223}{3} = 9,548$$

$$d_{(\text{Gabriela-Ovídio-Leonor}) \text{ Patrícia}} = \frac{8,420 + 6,580 + 6,045}{3} = 7,015$$

Note que as distâncias utilizadas para o cálculo das dissimilaridades a serem inseridas na matriz D_2 são as medidas euclidianas originais entre cada par de observações, ou seja, são provenientes da matriz D_0 . A matriz D_2 encontra-se a seguir:

		Gabriela Ovídio Leonor Luiz Felipe Patrícia		
$D_2 =$	Gabriela			
	Ovídio	0,000		
	Leonor			
	Luiz Felipe	9,548	0,000	
	Patrícia	7,015	7,187	0,000

Assim como verificado quando da elaboração do método de encadeamento único, aqui, a observação **Patrícia** também é incorporada ao *cluster* já formado por **Gabriela, Ovídio e Leonor**, permanecendo isolada a observação **Luiz Felipe**. Por fim, a matriz D_3 pode ser construída a partir do seguinte cálculo:

$$d_{(\text{Gabriela-Ovídio-Leonor-Patrícia}) \text{ Luiz Felipe}} = \frac{10,132 + 10,290 + 8,223 + 7,187}{4} = 8,958$$

		Gabriela Ovídio Leonor Patrícia Luiz Felipe	
$D_3 =$	Gabriela		
	Ovídio	0,000	
	Leonor		
	Patrícia		
	Luiz Felipe	8,958	0,000

Novamente, no quarto e último estágio, todas as observações estão no mesmo agrupamento. A Tabela 9.14 e a Figura 9.19 apresentam, respectivamente, o resumo desse esquema de aglomeração e o correspondente dendrograma resultante desse método de encadeamento médio.

Tabela 9.14 Esquema de aglomeração pelo método de encadeamento médio.

Estágio	Agrupamento	Observação Agrupada	Menor Distância Euclidiana
1	Gabriela	Ovídio	3,713
2	Gabriela – Ovídio	Leonor	4,822
3	Gabriela – Ovídio – Leonor	Patrícia	7,015
4	Gabriela – Ovídio – Leonor – Patrícia	Luiz Felipe	8,958

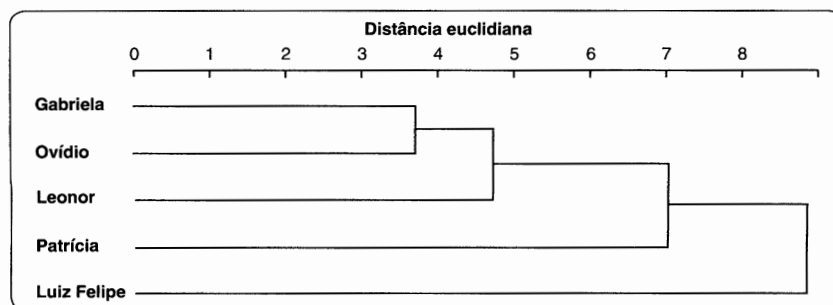


Figura 9.19 Dendrograma – Método de encadeamento médio.

Podemos verificar que a Tabela 9.14 e a Figura 9.19, embora com outros valores de distância, apresentam o mesmo ordenamento e a mesma alocação de observações nos agrupamentos que os apresentados, respectivamente, na Tabela 9.12 e na Figura 9.12, obtidos quando da elaboração do método de encadeamento único.

Nesse sentido, podemos afirmar que as observações são consideravelmente distintas em relação às variáveis estudadas, fato comprovado pela consistência de respostas obtidas pelos métodos de encadeamento único e médio. Caso as observações fossem mais similares, fato não observado no gráfico da Figura 9.11, a consistência de respostas ocorreria entre os métodos de encadeamento completo e médio, conforme já discutido. Portanto, **a elaboração inicial de gráficos de dispersão, quando possível, pode auxiliar o pesquisador, ainda que de forma preliminar, na escolha do método a ser adotado.**

Os esquemas de aglomeração hierárquicos são bastante úteis para oferecer uma possibilidade de que seja analisada, de forma exploratória, a similaridade entre observações com base no comportamento de determinadas variáveis. É de fundamental importância, todavia, que o pesquisador compreenda que **esses métodos não são conclusivos** em si mesmos e mais de uma resposta pode ser obtida, dependendo do que se deseja e do comportamento dos dados.

Além disso, é preciso que o pesquisador tenha consciência sobre a sensibilidade desses métodos em relação à presença de *outliers*. **A existência de uma observação muito discrepante pode fazer outras observações, não tão similares entre si, serem alocadas em um mesmo agrupamento pelo fato de se diferenciarem mais substancialmente da considerada outlier.** Portanto, é recomendável que sucessivas aplicações de esquemas hierárquicos aglomerativos com o método de encadeamento escolhido sejam elaboradas, e, em cada aplicação, seja identificadas uma ou mais observações consideradas *outliers*. Esse procedimento tornará a análise de agrupamentos mais confiável, visto que poderão ser formados *clusters* cada vez mais homogêneos. O pesquisador tem a liberdade de caracterizar a observação mais discrepante como aquela que acabou por ficar isolada após o penúltimo estágio de aglomeração, caso aconteça, ou seja, antes da fusão total. Porém, muitos são os métodos para que se defina um *outlier*. Barnett e Lewis (1994), por exemplo, citam quase 1.000 artigos provenientes da literatura sobre *outliers*, e, para efeitos didáticos, discutiremos, no apêndice deste capítulo, um efetivo procedimento em Stata para a detecção de *outliers* quando de uma análise multivariada de dados.

É relevante também enfatizar, conforme discutimos na presente seção, que diferentes métodos de encadeamento, quando da elaboração de esquemas hierárquicos aglomerativos, devem ser aplicados ao mesmo banco de dados, e os **dendrogramas resultantes, comparados**. Esse procedimento auxiliará o pesquisador em sua tomada de decisão, tanto em relação à escolha de uma interessante quantidade de agrupamentos quanto em relação ao ordenamento das observações e à alocação de cada uma nos diferentes *clusters* formados. Isso propiciará inclusive que se tome uma decisão coerente em relação à quantidade de agrupamentos que poderá ser considerada *input* de uma eventual análise não hierárquica.

Por fim, mas não menos importante, vale a pena comentar que os esquemas de aglomeração apresentados nesta seção (Tabelas 9.12, 9.13 e 9.14) oferecem **valores crescentes das medidas de agrupamento pelo fato de ter sido adotada uma medida de dissimilaridade** (distância euclidiana) como critério de comparação entre as observações. Caso tivéssemos escolhido a correlação de Pearson entre as observações, medida de similaridade também utilizada para variáveis métricas, conforme discutimos na seção 9.2.1.1, **os valores das medidas de agrupamento nos esquemas de aglomeração seriam decrescentes**. Este último fato também ocorre para análises de agrupamento em que são utilizadas medidas de semelhança (similaridade), como as estudadas na seção 9.2.1.2, para avaliar o comportamento de observações com base em variáveis binárias.

Na próxima seção elaboraremos, de forma algébrica, o mesmo exemplo por meio da aplicação do esquema de aglomeração não hierárquico *k-means*.

9.2.2.2. Esquema de aglomeração não hierárquico *k-means*

Dentre os esquemas de aglomeração não hierárquicos, o procedimento *k-means* é o mais utilizado por pesquisadores em diversos campos do conhecimento. Dado que a quantidade de *clusters* é definida preliminarmente pelo pesquisador, esse procedimento pode ser elaborado após a aplicação de um esquema hierárquico aglomerativo quando não se tem ideia da quantidade de *clusters* que podem ser formados e, nessa situação, o *output* obtido por esse procedimento pode servir de *input* para o não hierárquico.

9.2.2.2.1. Notação

Assim como a elaborada na seção 9.2.2.1.1, apresentamos, a seguir, uma sequência lógica de passos, com base em Johnson e Wichern (2007), para que seja facilitado o entendimento da análise de agrupamentos, elaborada por meio do procedimento *k-means*:

1. Definimos a quantidade inicial de *clusters* e os respectivos centroides. O objetivo é dividir as observações do banco de dados em K *clusters*, de modo que aquelas dentro de cada *cluster* estejam mais próximas entre si se comparadas a qualquer outra pertencente a um diferente. Para tal, as observações precisam arbitrariamente ser alocadas nos K *clusters*, a fim de que possam ser calculados os respectivos centroides.
2. Devemos selecionar determinada observação que se encontra mais próxima de um centroide e realocá-la nesse *cluster*. Neste momento, outro *cluster* acaba de perder aquela observação, e, portanto, devem ser recalculados os centroides do *cluster* que a recebe e os do *cluster* que a perde.
3. Devemos proceder com o passo anterior até que não seja mais possível realocar observação alguma por maior proximidade a um centroide de outro *cluster*.

A coordenada \bar{x} de um centroide deve ser recalculada quando da inclusão ou exclusão de determinada observação p no respectivo *cluster*, com base nas seguintes expressões:

$$\bar{x}_{\text{nov}} = \frac{N \cdot \bar{x} + x_p}{N + 1}, \text{ caso a observação } p \text{ seja inserida no } \textit{cluster} \text{ em análise} \quad (9.26)$$

$$\bar{x}_{\text{nov}} = \frac{N \cdot \bar{x} + x_p}{N - 1}, \text{ caso a observação } p \text{ seja excluída do } \textit{cluster} \text{ em análise} \quad (9.27)$$

em que N e \bar{x} referem-se, respectivamente, à quantidade de observações no *cluster* e à coordenada de seu centroide antes da realocação daquela observação. Além disso, x_p refere-se à coordenada da observação p que sofreu mudança de *cluster*.

A Figura 9.20 apresenta, para duas variáveis (X_1 e X_2), uma situação hipotética que representa o término do procedimento *k-means*, em que não é mais possível realocar observação alguma pelo fato de não mais haver maiores proximidades a centroides de outros agrupamentos.

A matriz de distâncias entre as observações não precisa ser definida a cada passo, ao contrário dos esquemas de aglomeração hierárquicos, o que reduz a exigência em relação à capacidade computacional, permitindo que os esquemas de aglomeração não hierárquicos possam ser aplicados a bancos de dados consideravelmente maiores que aqueles tradicionalmente estudados por meio de esquemas hierárquicos.

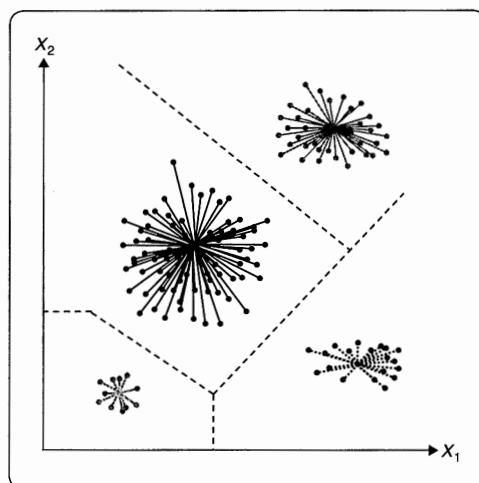


Figura 9.20 Situação hipotética que representa o término do procedimento *k-means*.

Além disso, lembramos que as variáveis devem ser padronizadas antes da elaboração do procedimento *k-means*, assim como nos esquemas de aglomeração hierárquicos, caso os respectivos valores não estejam na mesma unidade de medida.

Finalmente, após a conclusão desse procedimento, é importante que o pesquisador estude se os valores de determinada variável métrica diferem-se entre os grupos definidos, ou seja, se a variabilidade entre os *clusters* é significativamente superior à variabilidade interna a cada *cluster*. O teste *F* da análise de variância de um fator (em inglês, *one-way analysis of variance* ou *one-way ANOVA*) permite que seja elaborada essa análise, sendo que suas hipóteses nula e alternativa podem ser definidas da seguinte maneira:

H_0 : a variável em análise apresenta a mesma média em todos os grupos formados.

H_1 : a variável em análise apresenta média diferente em pelo menos um dos grupos em relação aos demais.

Dessa forma, um único teste *F* pode ser aplicado para cada variável, com o intuito de se avaliar a existência de pelo menos uma diferença entre todas as possibilidades de comparações, e, nesse sentido, a principal vantagem de sua aplicação reside no fato de que não precisam ser elaborados ajustes em relação a dimensões discrepantes dos grupos para se analisarem diversas comparações. Por outro lado, a rejeição da hipótese nula, a determinado nível de significância, não permite que o pesquisador saiba qual(is) grupo(s) é(são) estatisticamente diferente(s) dos demais em relação à variável em análise.

A expressão da estatística *F*, correspondente a esse teste, é dada pela seguinte expressão:

$$F = \frac{\text{variabilidade entre os grupos}}{\text{variabilidade dentro dos grupos}} = \frac{\sum_{k=1}^K N_k \cdot (\bar{X}_k - \bar{X})^2}{\frac{\sum_{ki} (X_{ki} - \bar{X}_k)^2}{n - K}} \quad (9.28)$$

em que N representa a quantidade de observações no k -ésimo *cluster*, \bar{X}_k é a média da variável X no mesmo k -ésimo *cluster*, \bar{X} é a média geral da variável X e X_{ki} é o valor que a variável X assume para determinada observação i presente no k -ésimo *cluster*. Além disso, K representa a quantidade de grupos (*clusters*) a serem comparados, e n , o tamanho da amostra.

Fazendo uso da estatística *F*, o pesquisador terá condições de identificar as variáveis cujas médias mais se diferem entre os grupos, ou seja, aquelas que mais contribuem para a formação de pelo menos um dos K *clusters* (maior estatística *F*), bem como aquelas que não contribuem para a formação da quantidade sugerida de agrupamentos, a determinado nível de significância.

Na próxima seção, apresentaremos um exemplo prático que será resolvido por meio de solução algébrica, a partir do qual os conceitos referentes ao procedimento *k-means* poderão ser fixados.

9.2.2.2.2. Exemplo prático de análise de agrupamentos com esquema de aglomeração não hierárquico *k-means*

Para resolução algébrica do esquema de aglomeração não hierárquico *k-means*, faremos uso dos dados de nosso próprio exemplo, que se encontram na Tabela 9.11 e são reproduzidos na Tabela 9.15.

Tabela 9.15 Exemplo: Notas de Matemática, Física e Química no vestibular.

Estudante (Observação)	Nota de Matemática (X_{1i})	Nota de Física (X_{2i})	Nota de Química (X_{3i})
Gabriela	3,7	2,7	9,1
Luiz Felipe	7,8	8,0	1,5
Patrícia	8,9	1,0	2,7
Ovídio	7,0	1,0	9,0
Leonor	3,4	2,0	5,0

Softwares como o SPSS utilizam a distância euclidiana como padrão de medida de dissimilaridade, razão pela qual elaboraremos os procedimentos algébricos com base nessa medida. Esse critério inclusive permitirá que os resultados obtidos sejam comparados com os encontrados quando da elaboração dos esquemas de aglomeração hierárquicos na seção 9.2.2.1.2, visto que, naquelas situações, também foi utilizada a distância euclidiana. Da mesma forma, não será também necessária a padronização das variáveis pelo procedimento *Zscores*, já que apresentam valores na mesma unidade de medida (notas de 0 a 10). **Caso contrário, é de fundamental importância que o pesquisador padronize as variáveis antes da elaboração do procedimento *k-means*.**

Fazendo uso da sequência lógica apresentada na seção 9.2.2.2.1, vamos elaborar o procedimento *k-means* com $K = 3$ clusters. Essa quantidade de agrupamentos pode ser oriunda de uma decisão do pesquisador pautada por determinado critério preliminar ou escolhida com base nos *outputs* dos esquemas de aglomeração hierárquicos. No nosso caso, a decisão foi tomada com base na comparação dos dendrogramas já elaborados e pela semelhança dos *outputs* obtidos pelos métodos de encadeamento único e médio.

Assim, precisamos alocar arbitrariamente as observações em três clusters, a fim de que possam ser calculados os respectivos centroides. Portanto, podemos definir que as observações **Gabriela** e **Luiz Felipe** formam o primeiro cluster, **Patrícia** e **Ovídio**, o segundo, e **Leonor**, o terceiro. A Tabela 9.16 apresenta a formação arbitrária desses clusters preliminares, bem como o cálculo das coordenadas dos respectivos centroides, o que possibilita o passo inicial do algoritmo do procedimento *k-means*.

Tabela 9.16 Alocação arbitrária das observações em $K = 3$ clusters e cálculo das coordenadas dos centroides – Passo inicial do procedimento *k-means*.

Agrupamento	Coordenadas dos Centroides		
	Variável		
	Nota de Matemática	Nota de Física	Nota de Química
Gabriela Luiz Felipe	$\frac{3,7 + 7,8}{2} = 5,75$	$\frac{2,7 + 8,0}{2} = 5,35$	$\frac{9,1 + 1,5}{2} = 5,30$
Patrícia Ovídio	$\frac{8,9 + 7,0}{2} = 7,95$	$\frac{1,0 + 1,0}{2} = 1,00$	$\frac{2,7 + 9,0}{2} = 5,85$
Leonor	3,40	2,00	5,00

Com base nessas coordenadas, construímos o gráfico da Figura 9.21, que apresenta a alocação arbitrária de cada observação em seu cluster, bem como os respectivos centroides.

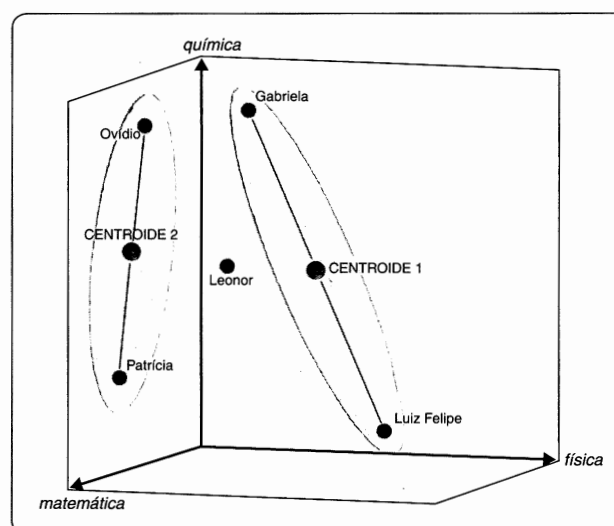


Figura 9.21 Alocação arbitrária das observações em $K = 3$ clusters e respectivos centroides – Passo inicial do procedimento *k-means*.

Com base no segundo passo da sequência lógica apresentada na seção 9.2.2.2.1, devemos escolher determinada observação e calcular a distância entre ela e os centroides de todos os agrupamentos, supondo que seja ou não realocada em cada cluster. Selecionando, por exemplo, a primeira observação (**Gabriela**), vamos calcular as

distâncias entre ela e os centroides dos agrupamentos já formados (**Gabriela-Luiz Felipe**, **Patrícia-Ovídio** e **Leonor**) e, na sequência, supor que ela deixe seu *cluster* (**Gabriela-Luiz Felipe**) e seja inserida em um dos outros dois agrupamentos, formando o *cluster* **Gabriela-Patrícia-Ovídio** ou o **Gabriela-Leonor**. Assim, a partir das expressões (9.26) e (9.27), devemos recalculas as coordenadas dos novos centroides, simulando que, de fato, ocorra a realocação de **Gabriela** para um dos dois *clusters*, conforme mostra a Tabela 9.17.

Tabela 9.17 Simulação de realocação de Gabriela e cálculo das coordenadas dos novos centroides.

Agrupamento	Simulação	Coordenadas dos Centroides		
		Variável		
		Nota de Matemática	Nota de Física	Nota de Química
Luiz Felipe	Exclusão de Gabriela	$\frac{2 \cdot (5,75) - 3,70}{2-1} = 7,80$	$\frac{2 \cdot (5,35) - 2,70}{2-1} = 8,00$	$\frac{2 \cdot (5,30) - 9,10}{2-1} = 1,50$
Gabriela Patrícia Ovídio	Inclusão de Gabriela	$\frac{2 \cdot (7,95) + 3,70}{2+1} = 6,53$	$\frac{2 \cdot (1,00) + 2,70}{2+1} = 1,57$	$\frac{2 \cdot (5,85) + 9,10}{2+1} = 6,93$
Gabriela Leonor	Inclusão de Gabriela	$\frac{1 \cdot (3,40) + 3,70}{1+1} = 3,55$	$\frac{1 \cdot (2,00) + 2,70}{1+1} = 2,35$	$\frac{1 \cdot (5,00) + 9,10}{1+1} = 7,05$

Obs.: Note que os valores calculados das coordenadas do centróide de **Luiz Felipe** são exatamente iguais às coordenadas originais dessa observação, conforme mostra a Tabela 9.15.

Nesse sentido, a partir das Tabelas 9.15, 9.16 e 9.17, podemos calcular as seguintes distâncias euclidianas:

- **Suposição de que Gabriela não seja realocada:**

$$d_{\text{Gabriela}-(\text{Gabriela-Luiz Felipe})} = \sqrt{(3,70-5,75)^2 + (2,70-5,35)^2 + (9,10-5,30)^2} = 5,066$$

$$d_{\text{Gabriela}-(\text{Patrícia-Ovídio})} = \sqrt{(3,70-7,95)^2 + (2,70-1,00)^2 + (9,10-5,85)^2} = 5,614$$

$$d_{\text{Gabriela-Leonor}} = \sqrt{(3,70-3,40)^2 + (2,70-2,00)^2 + (9,10-5,00)^2} = 4,170$$

- **Suposição de que Gabriela seja realocada:**

$$d_{\text{Gabriela-Luiz Felipe}} = \sqrt{(3,70-7,80)^2 + (2,70-8,00)^2 + (9,10-1,50)^2} = 10,132$$

$$d_{\text{Gabriela}-(\text{Gabriela-Patrícia-Ovídio})} = \sqrt{(3,70-6,53)^2 + (2,70-1,57)^2 + (9,10-6,93)^2} = 3,743$$

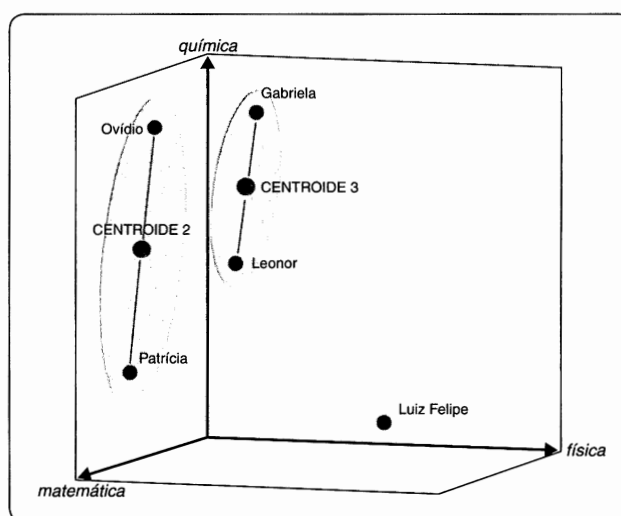
$$d_{\text{Gabriela}-(\text{Gabriela-Leonor})} = \sqrt{(3,70-3,55)^2 + (2,70-2,35)^2 + (9,10-7,05)^2} = 2,085$$

Como **Gabriela** encontra-se mais próxima do centróide de **Gabriela-Leonor** (menor distância euclidiana), devemos realocar essa observação no *cluster* formado inicialmente apenas pela observação **Leonor**. Logo, o *cluster* em que a observação **Gabriela** estava inicialmente (**Gabriela-Luiz Felipe**) acaba de perdê-la, passando a observação **Luiz Felipe** a compor um *cluster* individual; portanto, devem ser recalculados os centroides do *cluster* que a recebe e do que a perde. A Tabela 9.18 apresenta a formação dos novos *clusters*, assim como o cálculo das coordenadas dos respectivos centroides.

Tabela 9.18 Novos centroides com realocação de Gabriela.

Agrupamento	Coordenadas dos Centroides		
	Variável		
	Nota de Matemática	Nota de Física	Nota de Química
Luiz Felipe	7,80	8,00	1,50
Patrícia Ovídio	7,95	1,00	5,85
Gabriela Leonor	$\frac{3,7+3,4}{2}=3,55$	$\frac{2,7+2,0}{2}=2,35$	$\frac{9,1+5,0}{2}=7,05$

Com base nessas novas coordenadas, podemos construir o gráfico que se encontra na Figura 9.22.

**Figura 9.22** Novos clusters e respectivos centroides – Realocação de Gabriela.

Vamos proceder novamente com o passo anterior. Como a observação **Luiz Felipe** está, neste momento, isolada, vamos simular a realocação da terceira observação (**Patrícia**). Devemos calcular as distâncias entre ela e os centroides dos agrupamentos já formados (**Luiz Felipe, Patrícia-Ovídio** e **Gabriela-Leonor**) e, na sequência, supor que ela deixe seu cluster (**Patrícia-Ovídio**) e seja inserida em um dos outros dois agrupamentos, formando o cluster **Luiz Felipe-Patrícia** ou o **Gabriela-Patrícia-Leonor**. Também com base nas expressões (9.26) e (9.27), devemos recalculas as coordenadas dos novos centroides, simulando que de fato ocorra a realocação de **Patrícia** para um desses dois clusters, conforme mostra a Tabela 9.19.

Tabela 9.19 Simulação de realocação de Patrícia – Passo seguinte do algoritmo do procedimento *k-means*.

Agrupamento	Simulação	Coordenadas dos Centroides		
		Variável		
		Nota de Matemática	Nota de Física	Nota de Química
Luiz Felipe Patrícia	Inclusão de Patrícia	$\frac{1 \cdot (7,80) + 8,90}{1+1} = 8,35$	$\frac{1 \cdot (8,00) + 1,00}{1+1} = 4,50$	$\frac{1 \cdot (1,50) + 2,70}{1+1} = 2,10$
Ovídio	Exclusão de Patrícia	$\frac{2 \cdot (7,95) - 8,90}{2-1} = 7,00$	$\frac{2 \cdot (1,00) - 1,00}{2-1} = 1,00$	$\frac{2 \cdot (5,85) - 2,70}{2-1} = 9,00$
Gabriela Patrícia Leonor	Inclusão de Patrícia	$\frac{2 \cdot (3,55) + 8,90}{2+1} = 5,33$	$\frac{2 \cdot (2,35) + 1,00}{2+1} = 1,90$	$\frac{2 \cdot (7,05) + 2,70}{2+1} = 5,60$

Obs.: Note que os valores calculados das coordenadas do centroide de **Ovídio** são exatamente iguais às originais dessa observação, conforme mostra a Tabela 9.15.

Analogamente ao realizado quando da simulação de realocação de **Gabriela**, vamos calcular, com base nas Tabelas 9.15, 9.18 e 9.19, as distâncias euclidianas entre **Patrícia** e cada um dos centroides:

• **Suposição de que Patrícia não seja realocada:**

$$d_{\text{Patrícia-Luiz Felipe}} = \sqrt{(8,90-7,80)^2 + (1,00-8,00)^2 + (2,70-1,50)^2} = 7,187$$

$$d_{\text{Patrícia-(Patrícia-Ovídio)}} = \sqrt{(8,90-7,95)^2 + (1,00-1,00)^2 + (2,70-5,85)^2} = 3,290$$

$$d_{\text{Patrícia-(Gabriela-Leonor)}} = \sqrt{(8,90-3,55)^2 + (1,00-2,35)^2 + (2,70-7,05)^2} = 7,026$$

• **Suposição de que Patrícia seja realocada:**

$$d_{\text{Patrícia-(Luiz Felipe-Patrícia)}} = \sqrt{(8,90-8,35)^2 + (1,00-4,50)^2 + (2,70-2,10)^2} = 3,593$$

$$d_{\text{Patrícia-Ovídio}} = \sqrt{(8,90-7,00)^2 + (1,00-1,00)^2 + (2,70-9,00)^2} = 6,580$$

$$d_{\text{Patrícia-(Gabriela-Patrícia-Leonor)}} = \sqrt{(8,90-5,33)^2 + (1,00-1,90)^2 + (2,70-5,60)^2} = 4,684$$

Tendo em vista que a distância euclidiana entre **Patrícia** e o *cluster* **Patrícia-Ovídio** é a menor, não iremos realocá-la para outro agrupamento e manteremos, nesse momento, a solução apresentada na Tabela 9.18 e na Figura 9.22.

Na sequência, vamos elaborar o mesmo procedimento, porém simulando a realocação da quarta observação (**Ovídio**). Analogamente, devemos, portanto, calcular as distâncias entre essa observação e os centroides dos agrupamentos já formados (**Luiz Felipe**, **Patrícia-Ovídio** e **Gabriela-Leonor**) e, em seguida, fazer a suposição de que ela deixe seu *cluster* (**Patrícia-Ovídio**) e seja inserida em um dos outros dois agrupamentos, formando o *cluster* **Luiz Felipe-Ovídio** ou o **Gabriela-Ovídio-Leonor**. Novamente por meio das expressões (9.26) e (9.27), podemos recalcular as coordenadas dos novos centroides, simulando que de fato ocorra a realocação de **Ovídio** para um desses dois *clusters*, conforme mostra a Tabela 9.20.

Tabela 9.20 Simulação de realocação de Ovídio – Novo passo do algoritmo do procedimento *k-means*.

Agrupamento	Simulação	Coordenadas dos Centroides		
		Variável		
		Nota de Matemática	Nota de Física	Nota de Química
Luiz Felipe Ovídio	Inclusão de Ovídio	$\frac{1 \cdot (7,80) + 7,00}{1+1} = 7,40$	$\frac{1 \cdot (8,00) + 1,00}{1+1} = 4,50$	$\frac{1 \cdot (1,50) + 9,00}{1+1} = 5,25$
Patrícia	Exclusão de Ovídio	$\frac{2 \cdot (7,95) - 7,00}{2-1} = 8,90$	$\frac{2 \cdot (1,00) - 1,00}{2-1} = 1,00$	$\frac{2 \cdot (5,85) - 9,00}{2-1} = 2,70$
Gabriela Ovídio Leonor	Inclusão de Ovídio	$\frac{2 \cdot (3,55) + 7,00}{2+1} = 4,70$	$\frac{2 \cdot (2,35) + 1,00}{2+1} = 1,90$	$\frac{2 \cdot (7,05) + 9,00}{2+1} = 7,70$

Obs.: Note que os valores calculados das coordenadas do centroide de **Patrícia** são exatamente iguais às originais dessa observação, conforme mostra a Tabela 9.15.

A seguir, encontram-se os cálculos das distâncias euclidianas entre **Ovídio** e cada um dos centroides, elaborados a partir das Tabelas 9.15, 9.18 e 9.20:

• **Suposição de que Ovídio não seja realocado:**

$$d_{\text{Ovídio-Luiz Felipe}} = \sqrt{(7,00-7,80)^2 + (1,00-8,00)^2 + (9,00-1,50)^2} = 10,290$$

$$d_{\text{Ovídio-(Patrícia-Ovídio)}} = \sqrt{(7,00-7,95)^2 + (1,00-1,00)^2 + (9,00-5,85)^2} = 3,290$$

$$d_{\text{Ovídio-(Gabriela-Leonor)}} = \sqrt{(7,00-3,55)^2 + (1,00-2,35)^2 + (9,00-7,05)^2} = 4,187$$

• **Suposição de que Ovídio seja realocado:**

$$d_{\text{Ovídio-(Luiz Felipe-Ovídio)}} = \sqrt{(7,00-7,40)^2 + (1,00-4,50)^2 + (9,00-5,25)^2} = 5,145$$

$$d_{\text{Ovídio-Patrícia}} = \sqrt{(7,00-8,90)^2 + (1,00-1,00)^2 + (9,00-2,70)^2} = 6,580$$

$$d_{\text{Ovídio-(Gabriela-Ovídio-Leonor)}} = \sqrt{(7,00-4,70)^2 + (1,00-1,90)^2 + (9,00-7,70)^2} = 2,791$$

Nesse caso, como a observação **Ovídio** encontra-se mais próxima do centroide de **Gabriela-Ovídio-Leonor** (menor distância euclidiana), devemos realocar essa observação no *cluster* formado inicialmente por **Gabriela** e **Leonor**. Portanto, a observação **Patrícia** passa a formar um *cluster* individual. A Tabela 9.21 apresenta as coordenadas dos centróides dos *clusters* **Luiz Felipe**, **Patrícia** e **Gabriela-Ovídio-Leonor**.

Tabela 9.21 Novos centróides com realocação de Ovídio.

Agrupamento	Coordenadas dos Centroides		
	Variável		
	Nota de Matemática	Nota de Física	Nota de Química
Luiz Felipe	7,80	8,00	1,50
Patrícia	8,90	1,00	2,70
Gabriela Ovídio Leonor	4,70	1,90	7,70

Não iremos elaborar o procedimento proposto para a quinta observação (**Leonor**), visto que ela já sofreu fusão com a observação **Gabriela** logo no primeiro passo do algoritmo. Podemos considerar que o procedimento *k-means* esteja encerrado, uma vez que não é mais possível realocar qualquer observação por maior proximidade a um centróide de outro *cluster*. A Figura 9.23 apresenta a alocação de cada observação em seu *cluster*, bem como os respectivos centróides. Note que a solução obtida é igual à encontrada por meio dos métodos de encadeamento único (Figura 9.15) e médio, quando da elaboração dos esquemas de aglomeração hierárquicos.

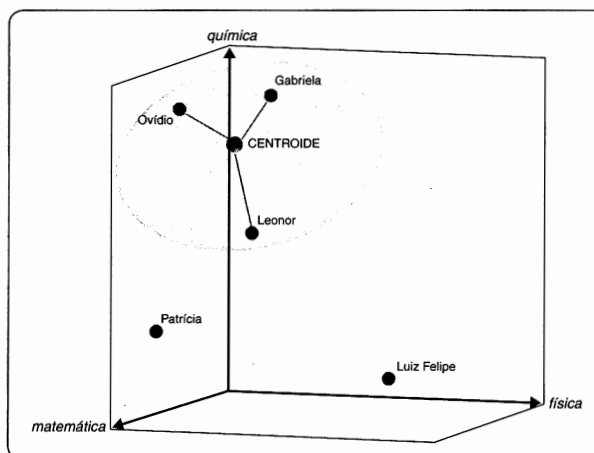


Figura 9.23 Solução do procedimento *k-means*.

Conforme já discutimos, podemos verificar que a matriz de distâncias entre as observações não precisa ser definida a cada passo do algoritmo referente ao procedimento *k-means*, ao contrário dos esquemas de aglomeração hierárquicos, o que reduz a exigência em relação à capacidade computacional, permitindo que os esquemas de aglomeração não hierárquicos possam ser aplicados a bancos de dados consideravelmente maiores que os tradicionalmente estudados por meio de esquemas hierárquicos.

A Tabela 9.22 apresenta as distâncias euclidianas entre cada observação do banco de dados original e os centroides de cada um dos *clusters* formados.

Tabela 9.22 Distâncias euclidianas entre observações e centroides dos *clusters*.

Estudante (Observação)	Agrupamento		
	Luiz Felipe	Patrícia	Gabriela Ovídio Leonor
Gabriela	10,132	8,420	1,897
Luiz Felipe	0,000	7,187	9,234
Patrícia	7,187	0,000	6,592
Ovídio	10,290	6,580	2,791
Leonor	8,223	6,045	2,998

Ressaltamos que esse algoritmo pode ser elaborado com outra alocação preliminar das observações nos *clusters* além da escolhida nesse exemplo. **A reaplicação do procedimento *k-means* com diversas escolhas arbitrárias, dada a quantidade *K* de *clusters*, permite que o pesquisador avalie a estabilidade do procedimento de agrupamento e embase, de maneira consistente, a alocação das observações nos grupos.**

Após a conclusão desse procedimento, é de fundamental importância que verifiquemos, por meio do teste *F* da análise de variância de um fator (*one-way analysis of variance* ou *one-way ANOVA*), se os valores de cada uma das três variáveis consideradas na análise são estatisticamente diferentes entre os três *clusters*. Para facilitar o cálculo das estatísticas *F* correspondentes a esse teste, elaboramos as Tabelas 9.23, 9.24 e 9.25, que apresentam as médias por *cluster* e geral das variáveis *matemática*, *física* e *química*, respectivamente.

Tabela 9.23 Médias por *cluster* e geral da variável *matemática*.

Cluster 1	Cluster 2	Cluster 3
$X_{\text{Luiz Felipe}} = 7,80$	$X_{\text{Patrícia}} = 8,90$	$X_{\text{Gabriela}} = 3,70$
		$X_{\text{Ovídio}} = 7,00$
		$X_{\text{Leonor}} = 3,40$
$\bar{X}_1 = 7,80$	$\bar{X}_2 = 8,90$	$\bar{X}_3 = 4,70$
$\bar{X} = 6,16$		

Tabela 9.24 Médias por *cluster* e geral da variável *física*.

Cluster 1	Cluster 2	Cluster 3
$X_{\text{Luiz Felipe}} = 8,00$	$X_{\text{Patrícia}} = 1,00$	$X_{\text{Gabriela}} = 2,70$
		$X_{\text{Ovídio}} = 1,00$
		$X_{\text{Leonor}} = 2,00$
$\bar{X}_1 = 8,00$	$\bar{X}_2 = 1,00$	$\bar{X}_3 = 1,90$
$\bar{X} = 2,94$		

Tabela 9.25 Médias por *cluster* e geral da variável *química*.

Cluster 1	Cluster 2	Cluster 3
$X_{\text{Luiz Felipe}} = 1,50$	$X_{\text{Patrícia}} = 2,70$	$X_{\text{Gabriela}} = 9,10$
		$X_{\text{Ovídio}} = 9,00$
		$X_{\text{Leonor}} = 5,00$
$\bar{X}_1 = 1,50$	$\bar{X}_2 = 2,70$	$\bar{X}_3 = 7,70$
$\bar{X} = 5,46$		

Logo, com base nos valores apresentados nessas tabelas e fazendo uso da expressão (9.28), temos condições de calcular as variabilidades entre os grupos e dentro deles para cada uma das variáveis, bem como as respectivas estatísticas F . As Tabelas 9.26, 9.27 e 9.28 apresentam esses cálculos.

Tabela 9.26 Variabilidades e estatística F para a variável *matemática*.

Variabilidade entre os grupos	$\frac{(7,80 - 6,16)^2 + (8,90 - 6,16)^2 + 3 \cdot (4,70 - 6,16)^2}{3 - 1} = 8,296$
Variabilidade dentro dos grupos	$\frac{(3,70 - 4,70)^2 + (7,00 - 4,70)^2 + (3,40 - 4,70)^2}{5 - 3} = 3,990$
F	$\frac{8,296}{3,990} = 2,079$

Obs.: O cálculo da variabilidade dentro dos grupos levou em consideração apenas o *cluster* 3, visto que os demais apresentam variabilidade igual a 0 por serem formados por uma única observação.

Tabela 9.27 Variabilidades e estatística F para a variável *física*.

Variabilidade entre os grupos	$\frac{(8,00 - 2,94)^2 + (1,00 - 2,94)^2 + 3 \cdot (1,90 - 2,94)^2}{3 - 1} = 16,306$
Variabilidade dentro dos grupos	$\frac{(2,70 - 1,90)^2 + (1,00 - 1,90)^2 + (2,00 - 1,90)^2}{5 - 3} = 0,730$
F	$\frac{16,306}{0,730} = 22,337$

Obs.: Igual à da tabela anterior.

Tabela 9.28 Variabilidades e estatística F para a variável *química*.

Variabilidade entre os grupos	$\frac{(1,50 - 5,46)^2 + (2,70 - 5,46)^2 + 3 \cdot (7,70 - 5,46)^2}{3 - 1} = 19,176$
Variabilidade dentro dos grupos	$\frac{(9,10 - 7,70)^2 + (9,00 - 7,70)^2 + (5,00 - 7,70)^2}{5 - 3} = 5,470$
F	$\frac{19,176}{5,470} = 3,506$

Obs.: Igual à da Tabela 9.26.

Vamos agora analisar a rejeição ou não da hipótese nula dos testes F para cada uma das variáveis. Como existem dois graus de liberdade para a variabilidade entre os grupos ($K - 1 = 2$) e dois graus de liberdade para a variabilidade dentro dos grupos ($n - K = 2$), temos, por meio da Tabela A do apêndice do livro, que $F_c = 19,00$ (F crítico ao nível de significância de 5%). Dessa forma, apenas para a variável *física* podemos rejeitar a hipótese nula de que todos os grupos formados possuem a mesma média, uma vez que F calculado $F_{cal} = 22,337 > F_c = F_{2,2,5\%}$.

= 19,00, Logo, para essa variável, existe pelo menos um grupo que apresenta média estatisticamente diferente dos demais. Para as variáveis *matemática* e *química*, no entanto, não podemos rejeitar a hipótese nula do teste ao nível de significância de 5%.

Softwares como o SPSS e o Stata não oferecem o F_c para os graus de liberdade definidos e determinado nível de significância. Todavia, oferecem o nível de significância do F_{cal} para esses graus de liberdade. Assim, em vez de analisarmos se $F_{cal} > F_c$, devemos verificar se o nível de significância do F_{cal} é menor que 0,05 (5%). Portanto:

Se $Sig. F$ (ou $Prob. F$) < 0,05, existe pelo menos uma diferença entre os grupos para a variável em análise.

O nível de significância do F_{cal} pode ser obtido no Excel por meio do comando **Fórmulas** → **Inserir Função** → **DISTF**, que abrirá uma caixa de diálogo como a apresentada na Figura 9.24.

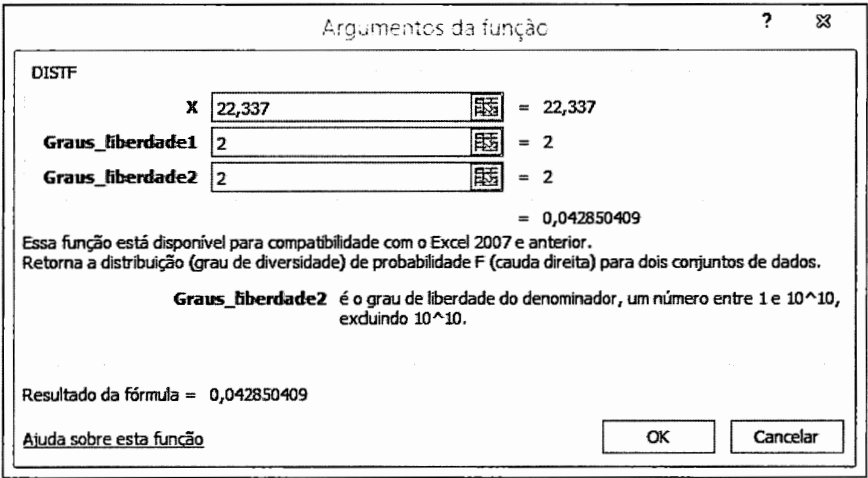


Figura 9.24 Obtenção do nível de significância de F (comando **Inserir Função**).

Conforme podemos observar por meio dessa figura, o $sig. F$ para a variável *física* é menor que 0,05 ($sig. F = 0,043$), ou seja, existe pelo menos uma diferença entre os grupos para essa variável ao nível de significância de 5%. Um pesquisador interessado poderá realizar o mesmo procedimento para as variáveis *matemática* e *química*. A Tabela 9.29 apresenta, de forma resumida, os resultados da análise de variância de um fator, com as variabilidades de cada variável, as estatísticas F e os respectivos níveis de significância.

Tabela 9.29 Análise de variância de um fator (ANOVA).

Variável	Variabilidade entre os grupos	Variabilidade dentro dos grupos	F	$Sig. F$
<i>matemática</i>	8,296	3,990	2,079	0,325
<i>física</i>	16,306	0,730	22,337	0,043
<i>química</i>	19,176	5,470	3,506	0,222

A tabela de **análise de variância de um fator** ainda permite que o pesquisador identifique as **variáveis que mais contribuem para a formação de pelo menos um dos clusters**, por possuírem média estatisticamente diferente em pelo menos um dos grupos em relação aos demais, visto que elas apresentarão maiores valores da estatística F . É relevante comentar que **os valores da estatística F são bastante sensíveis ao tamanho da amostra**, e, nesse caso, as variáveis *matemática* e *química* acabaram por não apresentar médias estatisticamente diferentes entre os três grupos, muito em função de a amostra ser reduzida (apenas cinco observações).

Ressaltamos que essa **análise de variância de um fator também pode ser realizada logo após a aplicação de determinado esquema de aglomeração hierárquico**, visto que depende apenas da classificação das observações em grupos. O único cuidado que o pesquisador deve ter, ao comparar os resultados obtidos por um esquema hierárquico com os obtidos por um esquema não hierárquico, é em relação à adoção da mesma medida de distância em ambas as situações. **Alocações diferentes das observações em uma mesma**

quantidade de clusters podem ocorrer caso sejam utilizadas medidas distintas de distância em um esquema hierárquico e em um esquema não hierárquico; portanto, podem ser calculados valores diferentes das estatísticas F nas duas situações.

De maneira geral, caso haja uma ou mais variáveis que não contribuam para a formação da quantidade sugerida de agrupamentos, recomendamos que o **procedimento seja reaplicado sem sua presença**. Nessas situações, poderá ocorrer a alteração da quantidade de agrupamentos e, caso o pesquisador veja a necessidade de embasar o *input* inicial a respeito da quantidade K de clusters, **poderá inclusive fazer uso de um esquema hierárquico aglomerativo sem a presença daquelas variáveis antes da reaplicação do procedimento k -means, o que tornará a análise cíclica.**

Além disso, a existência de *outliers* pode gerar clusters com considerável dispersão, e o **tratamento da base de dados com foco na identificação de observações muito discrepantes passa a ser um procedimento recomendável** antes da elaboração de esquemas de aglomeração não hierárquicos. No apêndice deste capítulo, será apresentado um importante procedimento em Stata para a detecção de *outliers* multivariados.

Assim como os esquemas de aglomeração hierárquicos, **o procedimento não hierárquico k -means não pode ser utilizado como técnica isolada** com a finalidade de que seja tomada uma decisão conclusiva a respeito do agrupamento de observações. **O comportamento dos dados, o tamanho da amostra e os critérios adotados pelo pesquisador podem ser bastante sensíveis para a alocação das observações e a formação de clusters.** A combinação dos *outputs* encontrados com os provenientes de outras técnicas pode mais fortemente embasar as escolhas do pesquisador e propiciar maior transparência no processo decisório.

Ao término da análise de agrupamentos, como os **clusters formados podem ser representados no banco de dados por uma nova variável qualitativa** com termos vinculados a cada observação (*cluster 1, cluster 2, ..., cluster K*), a partir dela, podem ser elaboradas outras técnicas multivariadas exploratórias, como análise de correspondência, a fim de que se estude, dependendo dos objetivos do pesquisador, uma eventual associação entre os agrupamentos e as categorias de outras variáveis qualitativas.

Essa nova variável qualitativa, que representa a alocação de cada observação, pode também ser utilizada como **explicativa** de determinado fenômeno em modelos multivariados confirmatórios, por exemplo, modelos de regressão múltipla, desde que transformada em variáveis *dummy* que representem as categorias (*clusters*) dessa nova variável gerada na análise de agrupamentos. Por outro lado, tal procedimento somente faz sentido quando há o intuito de elaborar um **diagnóstico** acerca do comportamento da variável dependente, sem que haja a intenção de previsões. Como uma nova observação não possui seu posicionamento em determinado *cluster*, a obtenção de sua alocação somente é possível ao se incluir tal observação em nova análise de agrupamentos, a fim de que seja obtida uma nova variável qualitativa e, conseqüentemente, novas *dummies*.

Ademais, essa nova variável qualitativa também pode ser considerada dependente de um modelo de regressão logística multinomial, permitindo que o pesquisador avalie as probabilidades que cada observação tem de pertencer a cada um dos clusters formados, em função do comportamento de outras variáveis explicativas não inicialmente consideradas na análise de agrupamentos. Ressaltamos, da mesma forma, que esse procedimento depende dos objetivos e do constructo estabelecido de pesquisa e apresenta caráter de diagnóstico do comportamento das variáveis na amostra para as observações existentes, sem finalidade preditiva.

Por fim, se os agrupamentos formados apresentarem **substancialidade** em relação à quantidade de observações alocadas, podem inclusive ser aplicadas, com o uso de outras variáveis, **técnicas confirmatórias específicas para cada cluster identificado**, a fim de que possam eventualmente ser gerados modelos mais bem ajustados.

Na sequência, o mesmo banco de dados será utilizado para que se elaborem análises de agrupamentos nos softwares SPSS e Stata. Enquanto na seção 9.3 serão apresentados os procedimentos para elaboração das técnicas estudadas no SPSS, assim como seus resultados, na seção 9.4 serão apresentados os comandos para realização dos procedimentos no Stata, com respectivos *outputs*.

9.3. ANÁLISE DE AGRUPAMENTOS COM ESQUEMAS DE AGLOMERAÇÃO HIERÁRQUICOS E NÃO HIERÁRQUICOS NO SOFTWARE SPSS

Nesta seção, apresentaremos o passo a passo para a elaboração do nosso exemplo no IBM SPSS Statistics Software®. O principal objetivo é propiciar ao pesquisador uma oportunidade de elaborar análises de agrupamentos com esquemas hierárquicos e não hierárquicos nesse software, dada sua facilidade de manuseio e a

didática das operações. A cada apresentação de um *output*, faremos menção ao respectivo resultado obtido quando da solução algébrica nas seções anteriores, a fim de que o pesquisador possa compará-los e formar seu conhecimento e erudição sobre o tema. A reprodução das imagens nesta seção tem autorização da International Business Machines Corporation®.

9.3.1. Elaboração de esquema de aglomeração hierárquico no software SPSS

Voltando ao exemplo apresentado na seção 9.2.2.1.2, lembremos que nosso professor tem o interesse de agrupar estudantes em *clusters* homogêneos em relação a notas (de 0 a 10) obtidas no vestibular nas disciplinas de Matemática, Física e Química. Os dados encontram-se no arquivo **Vestibular.sav** e são exatamente iguais aos apresentados na Tabela 9.11. Nesta seção, realizaremos a análise de agrupamentos fazendo uso da distância euclidiana entre as observações e levando em consideração apenas o método de encadeamento único.

Para que seja elaborada uma análise de agrupamentos por meio de um método hierárquico no SPSS, devemos clicar em **Analyze** → **Classify** → **Hierarchical Cluster...** Uma caixa de diálogo como a apresentada na Figura 9.25 será aberta.

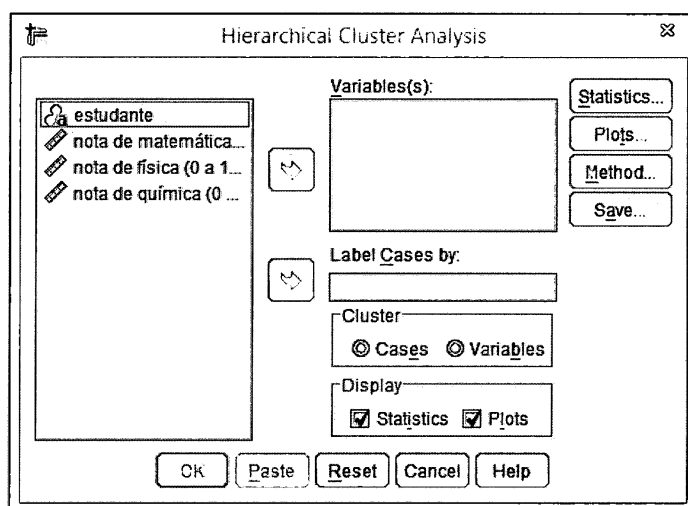


Figura 9.25 Caixa de diálogo para elaboração da análise de agrupamentos com método hierárquico no SPSS.

Na sequência, devemos inserir as variáveis originais de nosso exemplo (*matemática*, *física* e *química*) em **Variables** e a variável que identifica as observações (*estudante*) em **Label Cases by**, conforme mostra a Figura 9.26. Caso o pesquisador não possua uma variável que represente o nome das observações (neste caso, uma *string*), poderá deixar este último campo sem preenchimento.

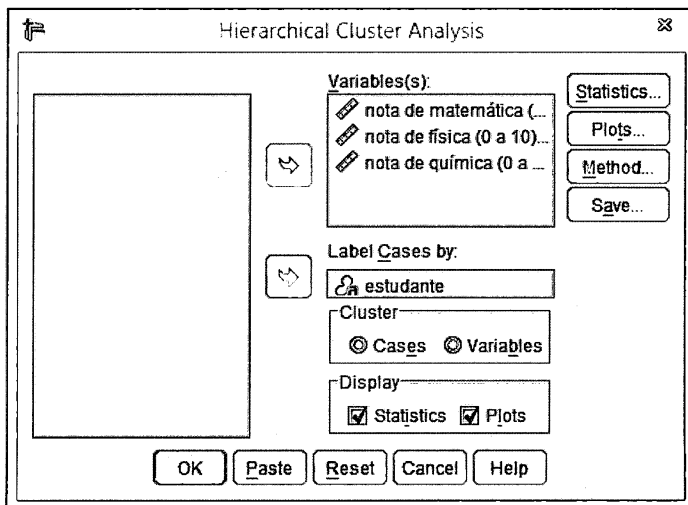


Figura 9.26 Seleção das variáveis originais.

No botão **Statistics...**, marcaremos primeiramente as opções **Agglomeration schedule** e **Proximity matrix**, que fazem com que sejam apresentados, nos *outputs*, a tabela com o esquema de aglomeração, elaborada com base na medida de distância a ser escolhida e no método de encadeamento a ser definido, e a matriz de distâncias entre cada par de observações, respectivamente. Ainda manteremos a opção **None** em **Cluster Membership**. A Figura 9.27 mostra como ficará essa caixa de diálogo.

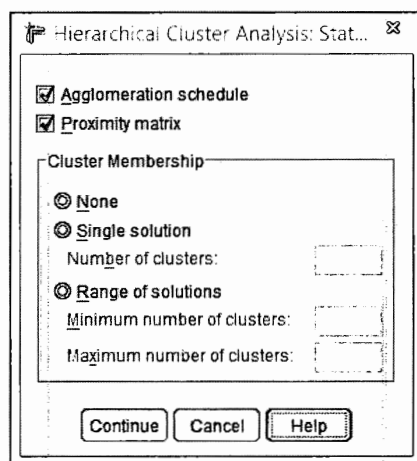


Figura 9.27 Seleção das opções que geram o esquema de aglomeração e a matriz de distâncias entre pares de observações.

Ao clicarmos em **Continue**, voltaremos para a caixa de diálogo principal da análise de agrupamentos hierárquicos. Na sequência, devemos clicar no botão **Plots...** Conforme mostra a Figura 9.28, iremos selecionar a opção **Dendrogram** e a opção **None** em **Icicle**.

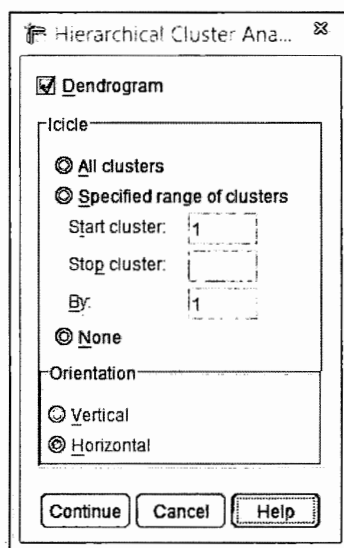


Figura 9.28 Seleção da opção que gera o dendrograma.

Da mesma forma, vamos clicar em **Continue** para que retornemos à caixa de diálogo principal.

Em **Method...**, que é a caixa de diálogo mais importante da análise de agrupamentos hierárquicos, devemos escolher o método de encadeamento único, também conhecido por *nearest neighbor* ou *single linkage*. Portanto, em **Cluster Method**, vamos selecionar a opção **Nearest neighbor**. Um curioso pesquisador poderá verificar que os métodos de encadeamento completo (**Furthest neighbor**) e médio (**Between-groups linkage**), estudados na seção 9.2.2.1, também estão disponíveis para seleção nesta opção.

Além disso, como as variáveis do banco de dados são métricas, vamos escolher uma das medidas de dissimilaridade dispostas em **Measure** → **Interval**. A fim de que seja mantida a mesma lógica utilizada quando da resolução algébrica de nosso exemplo, escolheremos a distância euclidiana como medida de dissimilaridade e, portanto, devemos selecionar a opção **Euclidean distance**. Pode-se verificar também que, nessa opção, estão dispostas as

outras medidas de dissimilaridade estudadas na seção 9.2.1.1, como a distância quadrática euclidiana, Minkowski, Manhattan (**Block**, no SPSS), Chebychev e a própria correlação de Pearson que, embora seja uma medida de similaridade, também é utilizada para variáveis métricas.

É importante mencionar que, embora não façamos uso de medidas de semelhança neste exemplo, pelo fato de não estarmos trabalhando com variáveis binárias, algumas medidas de similaridade podem ser selecionadas caso seja a situação com que se depare o pesquisador. Portanto, conforme estudamos na seção 9.2.1.2, podem ser selecionadas, em **Measure** → **Binary**, as medidas de emparelhamento simples (**Simple matching**, no SPSS), Jaccard, Dice, AntiDice (**Sokal and Sneath 2**, no SPSS), Russell e Rao, Ochiai, Yule (**Yule's Q**, no SPSS), Rogers e Tanimoto, Sneath e Sokal (**Sokal and Sneath 1**, no SPSS) e Hamann, entre outras.

Ainda na mesma caixa de diálogo, o pesquisador pode solicitar que a análise de agrupamentos seja elaborada a partir das variáveis padronizadas. Caso seja o intuito, para situações em que as variáveis originais apresentem unidades de medida distintas, pode ser selecionada a opção **Z scores** em **Transform Values** → **Standardize**, que fará todos os cálculos serem elaborados a partir da padronização das variáveis, que passarão a apresentar médias iguais a 0 e desvios-padrão iguais a 1.

Feitas essas considerações, a caixa de diálogo no nosso exemplo ficará conforme mostra a Figura 9.29.

Na sequência, podemos clicar em **Continue** e em **OK**.

O primeiro *output* (Figura 9.30) apresenta a matriz de dissimilaridades D_0 composta pelas distâncias euclidianas entre cada par de observações. Podemos notar, inclusive, que, na legenda, consta o dizer “*This is a dissimilarity matrix*”. Caso essa matriz fosse composta por medidas de semelhança, oriundas de cálculos elaborados a partir de variáveis binárias, o dizer seria “*This is a similarity matrix*”.

Figura 9.29 Caixa de diálogo para seleção do método de encadeamento e da medida de distância.

Por meio dessa matriz, que é igual àquela cujos valores foram calculados e apresentados na seção 9.2.2.1.2, podemos verificar que as observações **Gabriela** e **Ovídio** são as mais similares (menor distância euclidiana) em relação às variáveis *matemática*, *física* e *química* ($d_{\text{Gabriela-Ovídio}} = 3,713$).

Proximity Matrix					
Case	Euclidean Distance				
	1:Gabriela	2:Luiz Felipe	3:Patrícia	4:Ovídio	5:Leonor
1:Gabriela	,000	10,132	8,420	3,713	4,170
2:Luiz Felipe	10,132	,000	7,187	10,290	8,223
3:Patrícia	8,420	7,187	,000	6,580	6,045
4:Ovídio	3,713	10,290	6,580	,000	5,474
5:Leonor	4,170	8,223	6,045	5,474	,000

This is a dissimilarity matrix

Figura 9.30 Matriz de distâncias euclidianas (medidas de dissimilaridade) entre pares de observações.

Portanto, no esquema hierárquico apresentado na Figura 9.31, o primeiro estágio de aglomeração justamente ocorre pela fusão desses dois estudantes, com **Coefficient** (distância euclidiana) igual a 3,713. Note que as colunas **Cluster Combined Cluster 1** e **Cluster 2** referem-se a observações isoladas, quando ainda não incorporadas a determinado agrupamento ou a *clusters* já formados. Obviamente, no primeiro estágio de aglomeração, o primeiro *cluster* é formado pela fusão de duas observações isoladas.

Agglomeration Schedule						
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	4	3,713	0	0	2
2	1	5	4,170	1	0	3
3	1	3	6,045	2	0	4
4	1	2	7,187	3	0	0

Figura 9.31 Esquema hierárquico de aglomeração – Método de encadeamento único e distância euclidiana.

Na sequência, no segundo estágio, a observação **Leonor** (5) é incorporada ao *cluster* já formado anteriormente por **Gabriela** (1) e **Ovídio** (4). Podemos verificar que, em se tratando do método de encadeamento único, a distância considerada para a aglomeração de **Leonor** foi a menor entre essa observação e **Gabriela** ou **Ovídio**, ou seja, o critério adotado foi:

$$d_{(\text{Gabriela-Ovídio}) \text{ Leonor}} = \min \{4,170; 5,474\} = 4,170$$

Podemos notar também que, enquanto as colunas **Stage Cluster First Appears Cluster 1** e **Cluster 2** indicam em qual estágio anterior cada correspondente observação foi incorporada a determinado agrupamento, a coluna **Next Stage** mostra em qual futuro estágio o respectivo *cluster* receberá uma nova observação ou agrupamento, dado que estamos lidando com um método aglomerativo.

No terceiro estágio, ao *cluster* já formado, **Gabriela-Ovídio-Leonor**, é incorporada a observação **Patrícia** (3), respeitando-se o seguinte critério de distância:

$$d_{(\text{Gabriela-Ovídio-Leonor}) \text{ Patrícia}} = \min \{8,420; 6,580; 6,045\} = 6,045$$

E, por fim, no quarto e último estágio, dado que temos cinco observações, a observação **Luiz Felipe**, ainda isolada (note que a última observação a ser incorporada a um *cluster* corresponde ao último valor igual a 0 na coluna **Stage Cluster First Appears Cluster 2**), passa a ser incorporada ao *cluster* já formado pelas demais observações, encerrando-se o esquema aglomerativo. A distância considerada nesse estágio é dada por:

$$d_{(\text{Gabriela-Ovídio-Leonor-Patrícia}) \text{ Luiz Felipe}} = \min \{10,132; 10,290; 8,223; 7,187\} = 7,187$$

Com base na ordenação das observações no esquema de aglomeração e nas distâncias utilizadas como critério de agrupamento, pode ser construído o dendrograma, que se encontra na Figura 9.32. Note que as medidas de distância são rescalonadas para a construção dos dendrogramas no SPSS, a fim de que possa ser facilitada a interpretação da alocação de cada observação nos *clusters* e, principalmente, a visualização dos maiores saltos de distância, conforme discutimos na seção 9.2.2.1.2.1.

O ordenamento das observações no dendrograma corresponde ao que foi apresentado no esquema de aglomeração (Figura 9.31) e, a partir da análise da Figura 9.32, é possível identificar que o maior salto de distância ocorre quando da fusão de **Patrícia** com o *cluster* já formado **Gabriela-Ovídio-Leonor**. Esse salto já podia ter sido identificado no esquema de aglomeração da Figura 9.31, visto que um grande aumento de distância ocorre quando se passa do segundo para o terceiro estágio, ou seja, quando se incrementa a distância euclidiana de 4,170 para 6,045 (44,96%) para que novo *cluster* possa ser formado pela incorporação de outra observação. Portanto, podemos optar pela configuração existente ao final do segundo estágio de aglomeração, em que são formados três *clusters*. Conforme discutimos na seção 9.2.2.1.2.1, **o critério para a identificação da quantidade de clusters que leva em consideração o estágio de aglomeração imediatamente anterior a um grande salto é bastante útil e muito adotado.**

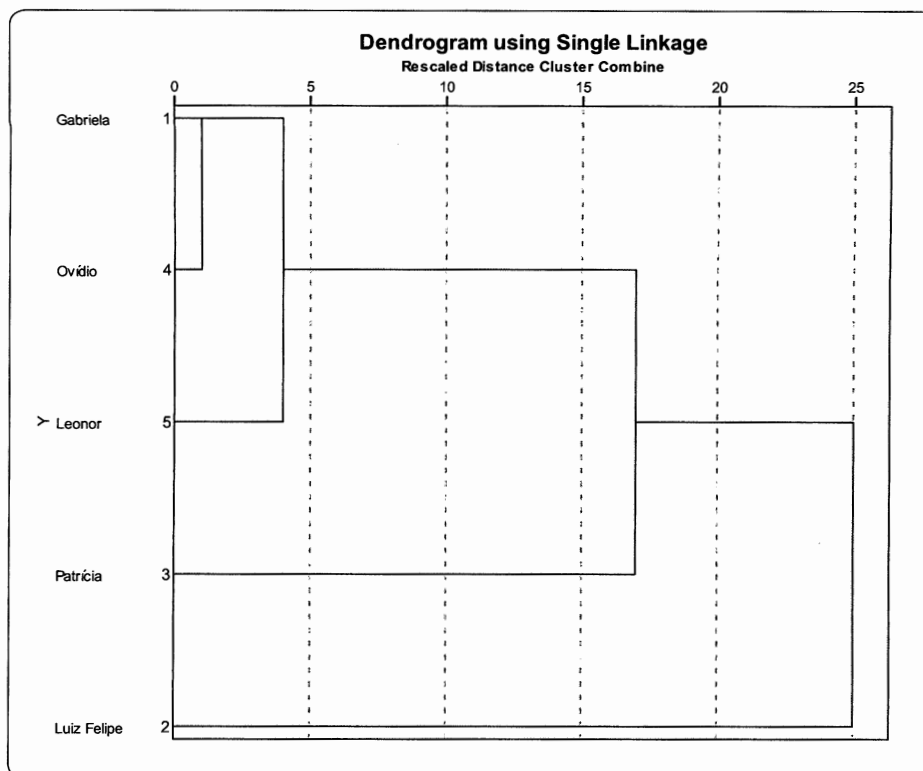


Figura 9.32 Dendrograma – Método de encadeamento único e distâncias euclidianas reescaladas no SPSS.

A Figura 9.33 apresenta uma linha vertical (tracejada) que “corta” o dendrograma na região em que ocorrem os maiores saltos. Neste momento, como acontecem três encontros com linhas do dendrograma, podemos identificar três correspondentes *clusters*, formados, respectivamente, por **Gabriela-Ovídio-Leonor**, **Patrícia** e **Luiz Felipe**.

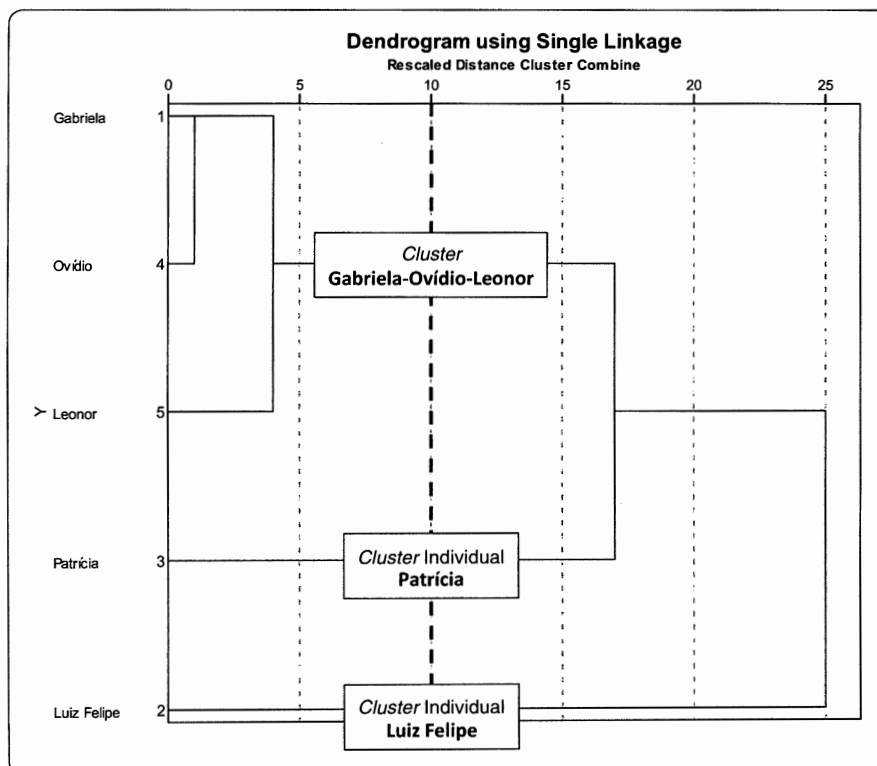


Figura 9.33 Dendrograma com identificação dos *clusters*.

Conforme discutimos, é comum encontramos dendrogramas que ofereçam certa dificuldade para que se identifiquem saltos de distância, muito em função da existência de observações consideravelmente similares no banco de dados em relação a todas as variáveis em análise. Nessas situações, é recomendável que se utilize a **medida de distância quadrática euclidiana e método de encadeamento completo** (*furthest neighbor*). Essa combinação de critérios é bastante popular em bases de dados com observações muito homogêneas.

Adotada a solução com três *clusters*, podemos novamente clicar em **Analyze → Classify → Hierarchical Cluster...** e, no botão **Statistics...**, selecionar a opção **Single solution** em **Cluster Membership**. Nessa opção, devemos inserir o número 3 em **Number of clusters**, conforme mostra a Figura 9.34.

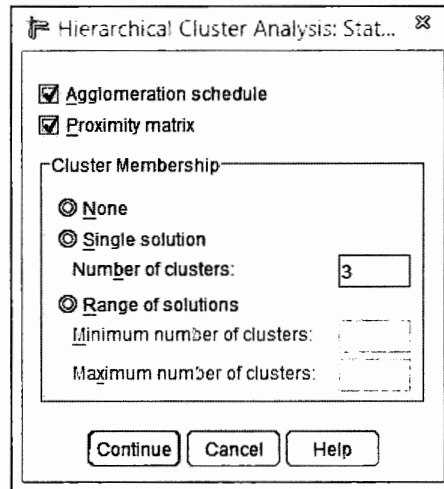


Figura 9.34 Definição da quantidade de *clusters*.

Ao clicarmos em **Continue**, retornaremos à caixa de diálogo principal da análise de agrupamentos. No botão **Save...**, vamos agora selecionar a opção **Single solution** e, da mesma forma, inserir o número 3 em **Number of clusters**, conforme mostra a Figura 9.35, a fim de que nova variável correspondente à alocação das observações nos agrupamentos seja disponibilizada no banco de dados.

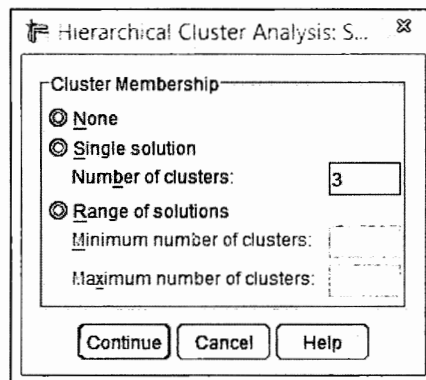


Figura 9.35 Seleção da opção para salvar a alocação das observações nos *clusters* como nova variável no banco de dados – Procedimento hierárquico.

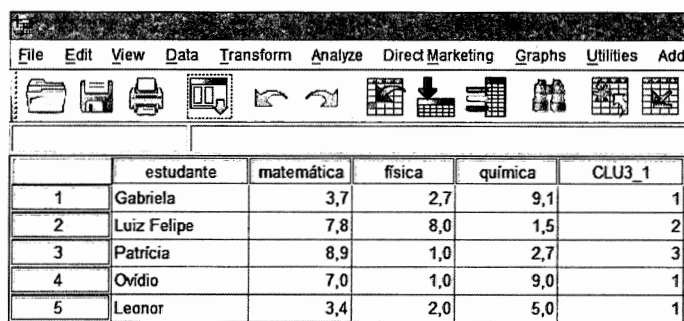
Na sequência, podemos clicar em **Continue** e em **OK**.

Embora os *outputs* gerados sejam os mesmos, é importante notar que uma nova tabela de resultados é apresentada, correspondente à alocação propriamente dita das observações nos *clusters*. A Figura 9.36 mostra, para três agrupamentos, que, enquanto as observações **Gabriela**, **Ovídio** e **Leonor** formam um único *cluster*, nomeado por 1, as observações **Luiz Felipe** e **Patrícia** formam dois *clusters* individuais, nomeados, respectivamente, por 2 e 3. Embora as nomeações sejam numéricas, é importante ressaltar que representam apenas **rótulos (categorias)** de uma variável qualitativa.

Cluster Membership	
Case	3 Clusters
1:Gabriela	1
2:Luiz Felipe	2
3:Patrícia	3
4:Ovídio	1
5:Leonor	1

Figura 9.36 Alocação das observações nos clusters.

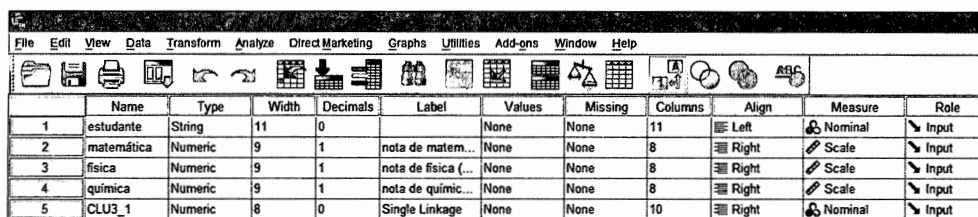
Ao elaborarmos o procedimento descrito, podemos verificar que é gerada uma nova variável no banco de dados, chamada pelo SPSS de *CLU3_1*, conforme mostra a Figura 9.37.



	estudante	matemática	física	química	CLU3_1
1	Gabriela	3,7	2,7	9,1	1
2	Luiz Felipe	7,8	8,0	1,5	2
3	Patrícia	8,9	1,0	2,7	3
4	Ovídio	7,0	1,0	9,0	1
5	Leonor	3,4	2,0	5,0	1

Figura 9.37 Banco de dados com nova variável *CLU3_1* – Alocação de cada observação.

A natureza dessa nova variável é automaticamente classificada pelo software como **Nominal**, ou seja, qualitativa, conforme podemos comprovar na Figura 9.38, que pode ser obtida ao clicarmos em **Variable View**, no canto inferior esquerdo da tela do SPSS.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	estudante	String	11	0		None	None	11	Left	Nominal	Input
2	matemática	Numeric	9	1	nota de matem...	None	None	8	Right	Scale	Input
3	física	Numeric	9	1	nota de física (...)	None	None	8	Right	Scale	Input
4	química	Numeric	9	1	nota de químic...	None	None	8	Right	Scale	Input
5	CLU3_1	Numeric	8	0	Single Linkage	None	None	10	Right	Nominal	Input

Figura 9.38 Classificação nominal (qualitativa) da variável *CLU3_1*.

Conforme discutimos, a variável *CLU3_1* pode ser utilizada em outras técnicas exploratórias, como análise de correspondência, ou em técnicas confirmatórias. Neste último caso, pode ser inserida, por exemplo, no vetor de variáveis explicativas (desde que transformada para *dummies*) de um modelo de regressão múltipla, ou como variável dependente de determinado modelo de regressão logística multinomial em que haja a intenção de estudar o comportamento de outras variáveis não inseridas na análise de agrupamentos sobre a probabilidade de inserção de cada observação em cada um dos clusters formados. Essa decisão, no entanto, depende dos objetivos e do constructo de pesquisa.

Neste momento, o pesquisador pode considerar a análise de agrupamentos com esquemas de aglomeração hierárquicos finalizada. Entretanto, com base na criação da nova variável *CLU3_1*, poderá ainda estudar, por meio da análise de variância de um fator, se os valores de determinada variável diferem-se entre os clusters formados, ou seja, se a variabilidade entre os grupos é significativamente superior à variabilidade interna a cada um deles. Mesmo que a análise não tenha sido elaborada quando da resolução algébrica dos esquemas hierárquicos, visto que optamos por realizá-la apenas após o procedimento *k-means*, na seção 9.2.2.2.2, mostraremos a seguir como pode ser aplicada neste momento, visto que já temos a alocação das observações nos grupos.

Para tanto, vamos clicar em **Analyze** → **Compare Means** → **One-Way ANOVA**.... Na caixa de diálogo que será aberta, devemos inserir as variáveis *matemática*, *física* e *química* em **Dependent List** e a variável *CLU3_1* (*Single Linkage*) em **Factor**. A caixa de diálogo ficará conforme mostra a Figura 9.39.

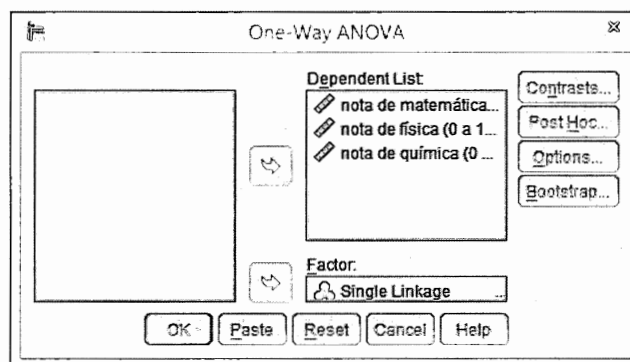


Figura 9.39 Caixa de diálogo com seleção das variáveis para elaboração da análise de variância de um fator no SPSS.

No botão **Options...**, marcaremos as opções **Descriptive** (em **Statistics**) e **Means plot**, como mostra a Figura 9.40.

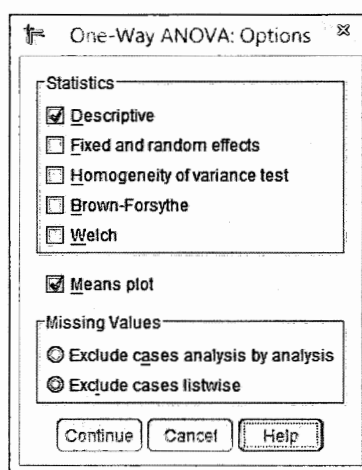


Figura 9.40 Seleção de opções para realização da análise de variância de um fator.

Na sequência, podemos clicar em **Continue** e em **OK**.

Enquanto a Figura 9.41 apresenta as estatísticas descritivas dos *clusters* por variável, de forma correspondente às Tabelas 9.23, 9.24 e 9.25, a Figura 9.42 faz uso desses valores e apresenta o cálculo das variabilidades entre os grupos (**Between Groups**) e dentro dos grupos (**Within Groups**), bem como as estatísticas *F* para cada variável e os respectivos níveis de significância. Podemos verificar que esses valores correspondem aos calculados algebricamente na seção 9.2.2.2.2 e apresentados na Tabela 9.29.

Descriptives									
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum	
					Lower Bound	Upper Bound			
nota de matemática (0 a 10)	1	3	4,700	1,9975	1,1533	-262	9,662	3,4	7,0
	2	1	7,800	7,8	7,8
	3	1	8,900	8,9	8,9
	Total	5	6,160	2,4785	1,1084	3,083	9,237	3,4	8,9
nota de física (0 a 10)	1	3	1,900	,8544	,4933	-,222	4,022	1,0	2,7
	2	1	8,000	8,0	8,0
	3	1	1,000	1,0	1,0
	Total	5	2,940	2,9186	1,3052	-,684	6,564	1,0	8,0
nota de química (0 a 10)	1	3	7,700	2,3388	1,3503	1,890	13,510	5,0	9,1
	2	1	1,500	1,5	1,5
	3	1	2,700	2,7	2,7
	Total	5	5,460	3,5104	1,5699	1,101	9,819	1,5	9,1

Figura 9.41 Estatísticas descritivas dos *clusters* por variável.

		ANOVA				
		Sum of Squares	df	Mean Square	F	Sig.
nota de matemática (0 a 10)	Between Groups	16,592	2	8,296	2,079	,325
	Within Groups	7,980	2	3,990		
	Total	24,572	4			
nota de física (0 a 10)	Between Groups	32,612	2	16,306	22,337	,043
	Within Groups	1,460	2	,730		
	Total	34,072	4			
nota de química (0 a 10)	Between Groups	38,352	2	19,176	3,506	,222
	Within Groups	10,940	2	5,470		
	Total	49,292	4			

Figura 9.42 Análise de variância de um fator – Variabilidades entre grupos e dentro dos grupos, estatísticas *F* e níveis de significância por variável.

A partir da Figura 9.42, podemos verificar que o *sig. F* para a variável *física* é menor que 0,05 (*sig. F* = 0,043), ou seja, existe pelo menos um grupo que apresenta média estatisticamente diferente dos demais ao nível de significância de 5%. Porém, o mesmo não pode ser dito em relação às variáveis *matemática* e *química*.

Embora tenhamos uma ideia acerca de qual grupo apresenta média estatisticamente diferente dos demais para a variável *física*, com base nos *outputs* da Figura 9.41, a elaboração de gráficos pode facilitar ainda mais a análise das diferenças de médias das variáveis por *cluster*. Os gráficos gerados pelo SPSS (Figuras 9.43, 9.44 e 9.45) permitem que visualizemos essas diferenças entre os grupos para cada variável analisada.

Logo, a partir do gráfico da Figura 9.44, é possível visualizar que o grupo 2, formado apenas pela observação **Luiz Felipe**, apresenta, de fato, média diferente dos demais em relação à variável *física*.

Além disso, embora notemos, a partir dos gráficos das Figuras 9.43 e 9.45, que existem diferenças de médias das variáveis *matemática* e *química* entre os grupos, essas diferenças não podem ser consideradas estatisticamente significantes, ao nível de significância de 5%, visto que estamos lidando com uma quantidade muito pequena de observações, e os valores da estatística *F* são bastante sensíveis ao tamanho da amostra. Essa análise gráfica torna-se bastante útil quando do estudo de bancos de dados com uma quantidade maior de observações e variáveis.

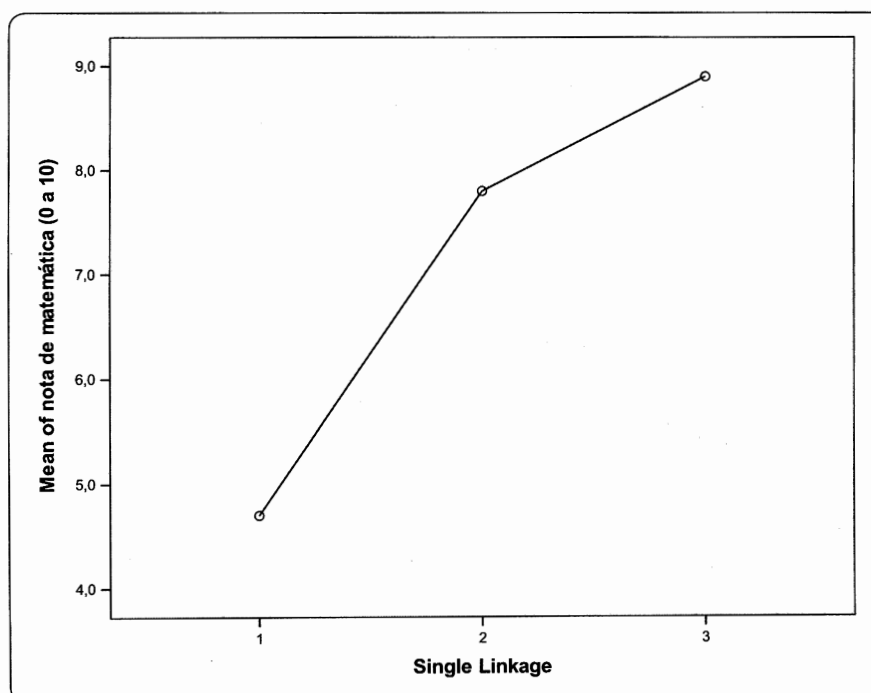


Figura 9.43 Médias da variável *matemática* nos três *clusters*.

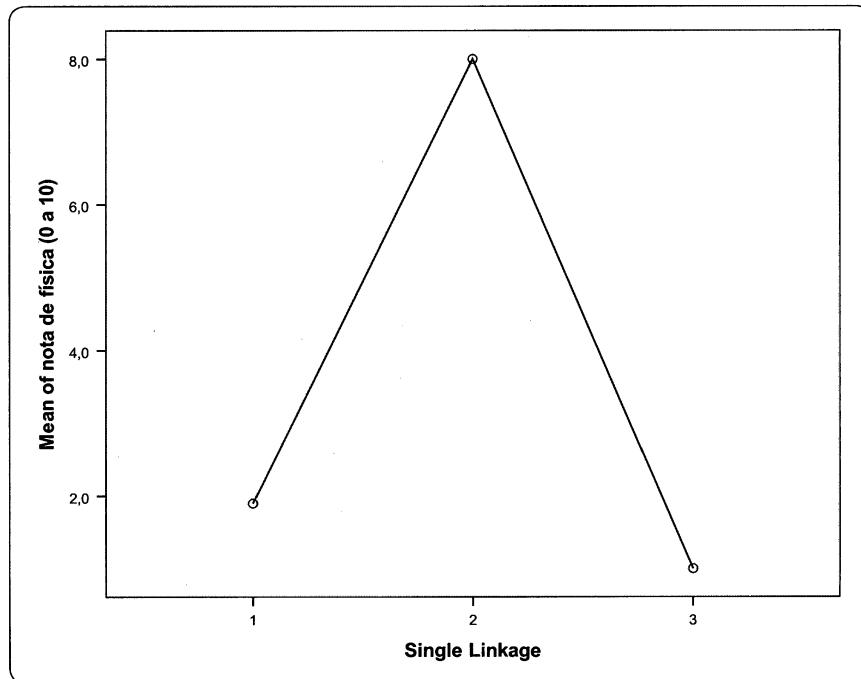


Figura 9.44 Médias da variável *física* nos três *clusters*.

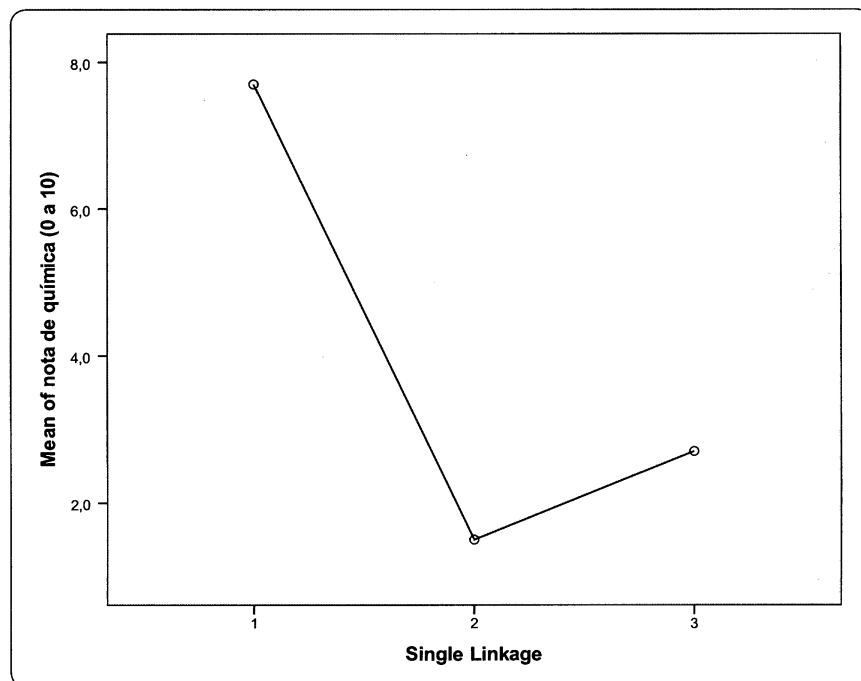
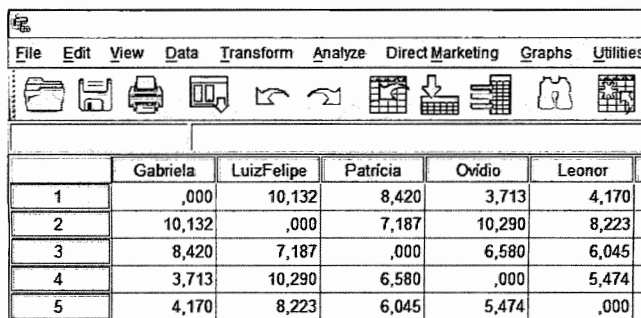


Figura 9.45 Médias da variável *química* nos três *clusters*.

Por fim, o pesquisador pode ainda complementar sua análise elaborando um procedimento conhecido por **escalonamento multidimensional**, já que o uso da matriz de distâncias pode propiciar a elaboração de um gráfico que permite a visualização das posições relativas de cada observação de forma bidimensional, independentemente da quantidade total de variáveis.

Para tanto, devemos estruturar um novo banco de dados, formado justamente pela matriz de distâncias. Para os dados de nosso exemplo, podemos abrir o arquivo **VestibularMatriz.sav**, que contém a matriz de distâncias euclidianas apresentada na Figura 9.46. Note que as colunas desse novo banco de dados se referem às observações do banco de dados original, assim como as linhas (matriz quadrada de distâncias).



	Gabriela	LuizFelipe	Patricia	Ovidio	Leonor
1	,000	10,132	8,420	3,713	4,170
2	10,132	,000	7,187	10,290	8,223
3	8,420	7,187	,000	6,580	6,045
4	3,713	10,290	6,580	,000	5,474
5	4,170	8,223	6,045	5,474	,000

Figura 9.46 Banco de dados com a matriz de distâncias euclidianas.

Vamos clicar em **Analyze** → **Scale** → **Multidimensional Scaling (ASCAI)**.... Na caixa de diálogo que será aberta, devemos inserir as variáveis que representam as observações em **Variables**, conforme mostra a Figura 9.39. Como os dados já correspondem a distâncias, nada precisará ser feito em relação ao campo **Distances**.

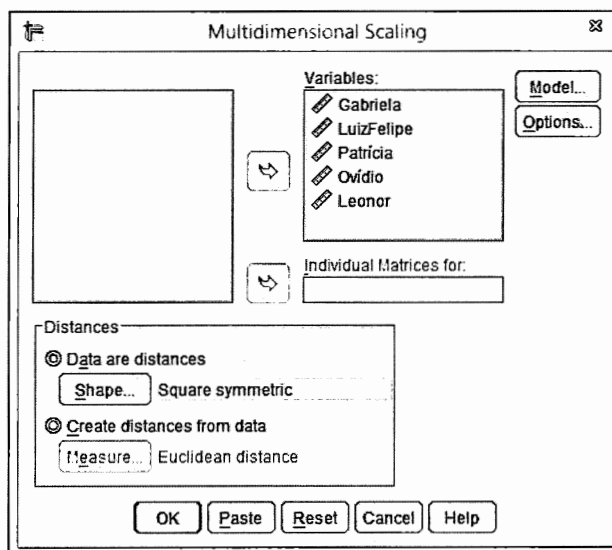


Figura 9.47 Caixa de diálogo com seleção das variáveis para elaboração de escalonamento multidimensional no SPSS.

No botão **Model...**, marcaremos a opção **Ratio** em **Level of Measurement** (note que já está selecionada a opção **Euclidean distance** em **Scaling Model**) e, no botão **Options...**, a opção **Group plots** em **Display**, conforme mostram, respectivamente, as Figuras 9.48 e 9.49.

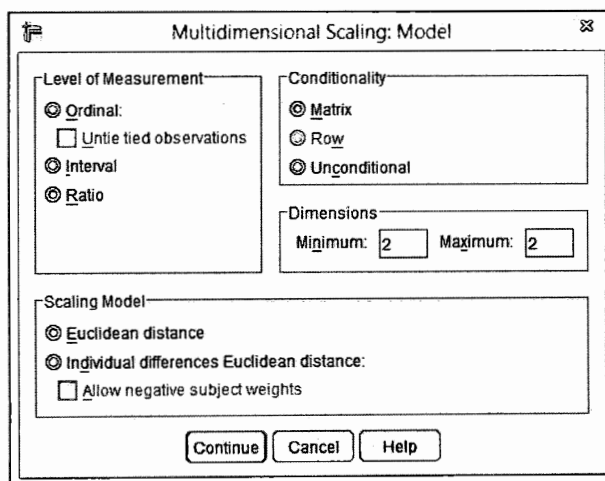


Figura 9.48 Definição da natureza da variável correspondente à medida de distância.

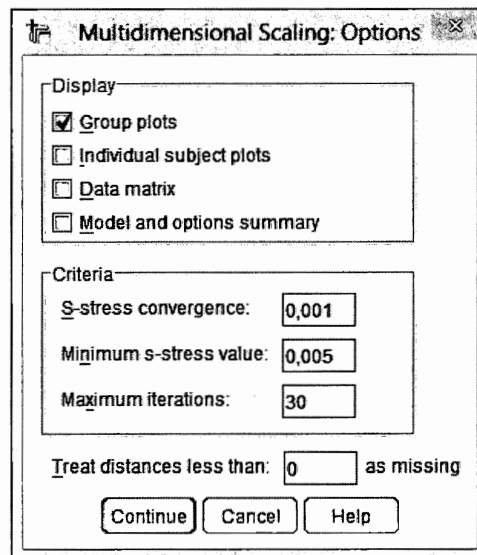


Figura 9.49 Seleção de opção para elaboração de gráfico bidimensional.

Na sequência, podemos clicar em **Continue** e em **OK**.

A Figura 9.50 apresenta o gráfico com as posições relativas das observações projetadas em um plano.

Esse tipo de gráfico é bastante útil quando se deseja elaborar apresentações didáticas sobre o agrupamento de observações (indivíduos, empresas, municípios, países, entre outros exemplos) e facilitar a interpretação dos *clusters*, principalmente quando há uma quantidade relativamente grande de variáveis no banco de dados.

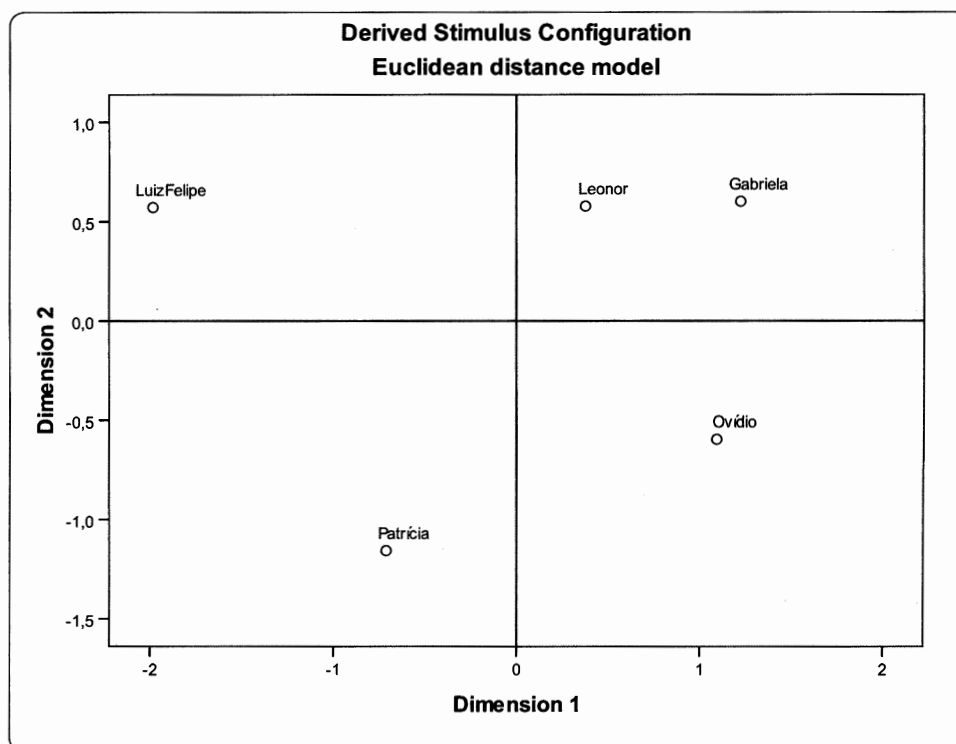


Figura 9.50 Gráfico bidimensional com as posições relativas projetadas das observações.

9.3.2. Elaboração do esquema de aglomeração não hierárquico *k-means* no software SPSS

Mantendo a lógica proposta no capítulo, elaboraremos, a partir do mesmo banco de dados, uma análise de agrupamentos com base no esquema de aglomeração não hierárquico *k-means*. Portanto, devemos novamente fazer uso do arquivo **Vestibular.sav**.

Para tanto, devemos clicar em **Analyze** → **Classify** → **K-Means Cluster...** Na caixa de diálogo que será aberta, devemos inserir as variáveis *matemática*, *física* e *química* em **Variables**, e a variável *estudante* em **Label Cases by**. A principal diferença entre essa caixa de diálogo inicial e aquela correspondente ao procedimento hierárquico refere-se à determinação da quantidade de *clusters* a partir da qual o algoritmo *k-means* será elaborado. Em nosso exemplo, vamos inserir o número 3 em **Number of Clusters**. A Figura 9.51 mostra como ficará a caixa de diálogo.

Podemos notar que inserimos as variáveis originais no campo **Variables**. Esse procedimento é aceitável, visto que, para nosso exemplo, possuem valores na mesma unidade de medida. Entretanto, caso esse fato não se verifique, o pesquisador deverá, antes de elaborar o procedimento *k-means*, padronizá-las pelo procedimento *Zscores*, em **Analyze** → **Descriptive Statistics** → **Descriptives...**, inserir as variáveis originais em **Variables** e selecionar a opção **Save standardized values as variables**. Ao clicar em **OK**, o pesquisador irá verificar que novas variáveis padronizadas passarão a compor o banco de dados.

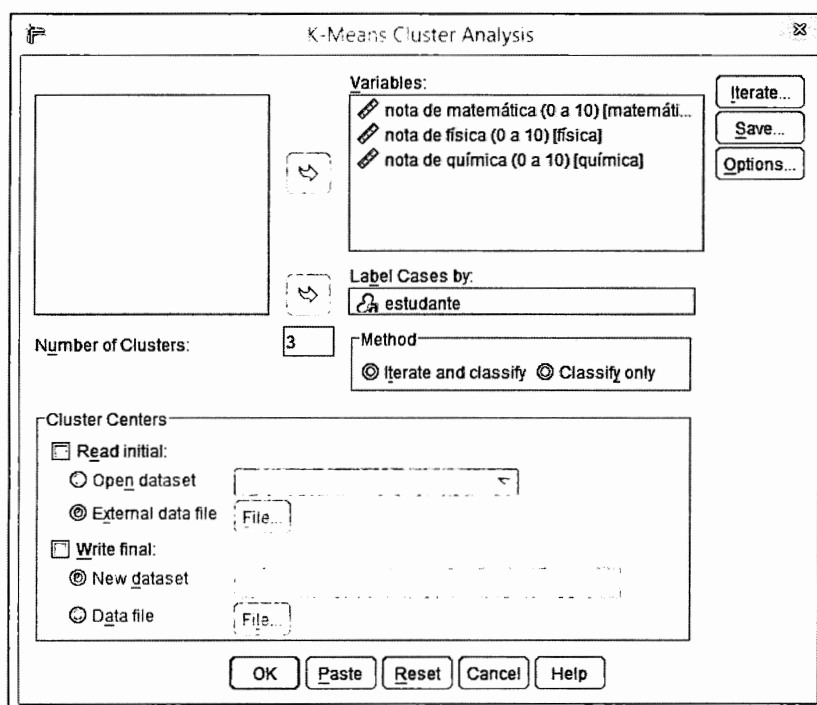


Figura 9.51 Caixa de diálogo para elaboração da análise de agrupamentos com método não hierárquico *k-means* no SPSS.

Voltando à tela inicial do procedimento *k-means*, vamos clicar no botão **Save...** Na caixa de diálogo que será aberta, devemos selecionar a opção **Cluster membership**, conforme mostra a Figura 9.52.

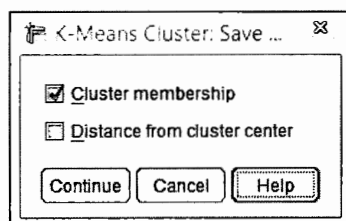


Figura 9.52 Seleção da opção para salvar a alocação das observações nos *clusters* como nova variável no banco de dados – Procedimento não hierárquico.

Ao clicarmos em **Continue**, voltaremos à caixa de diálogo anterior. No botão **Options...**, vamos selecionar as opções **Initial cluster centers**, **ANOVA table** e **Cluster information for each case**, em **Statistics**, conforme mostra a Figura 9.53.

Na sequência, podemos clicar em **Continue** e em **OK**. É importante mencionar que o SPSS já utiliza como padrão a distância euclidiana como medida de dissimilaridade quando da elaboração do procedimento *k-means*.

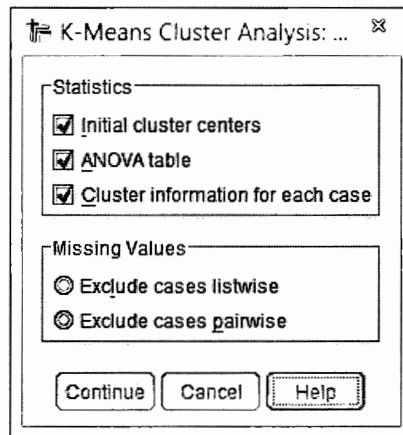


Figura 9.53 Seleção de opções para realização do procedimento *k-means*.

Os dois primeiros *outputs* gerados referem-se ao passo inicial e ao procedimento iterativo do algoritmo *k-means*. São apresentadas as coordenadas dos centroides no passo inicial e, por meio dos quais, podemos perceber que o SPSS considera que os três *clusters* sejam formados, respectivamente, pelas três primeiras observações do banco de dados. Embora essa decisão seja diferente da adotada por nós na seção 9.2.2.2.2, essa escolha é puramente arbitrária, e, conforme poderemos verificar adiante, não afetará em nada a formação dos *clusters* no passo final do algoritmo *k-means*.

Enquanto a Figura 9.54 apresenta os valores propriamente ditos das variáveis originais para as observações **Gabriela**, **Luiz Felipe** e **Patrícia** (conforme mostra a Tabela 9.15) como coordenadas dos centroides dos três grupos, na Figura 9.55 podemos verificar, após a primeira iteração do algoritmo, que a mudança de coordenada do centroide do primeiro *cluster* é de 1,897, que corresponde exatamente à distância euclidiana entre a observação **Gabriela** e o *cluster* **Gabriela-Ovídio-Leonor** (conforme mostra a Tabela 9.22). Nessa última figura, ainda é possível verificar a menção, em seu rodapé, à medida de 7,187, que corresponde à distância euclidiana entre as observações **Luiz Felipe** e **Patrícia**, que permanecem isoladas após o procedimento iterativo.

Initial Cluster Centers			
	Cluster		
	1	2	3
nota de matemática (0 a 10)	3,7	7,8	8,9
nota de física (0 a 10)	2,7	8,0	1,0
nota de química (0 a 10)	9,1	1,5	2,7

Figura 9.54 Passo inicial do algoritmo *k-means* – Centroides dos três grupos como coordenadas das observações.

Iteration History ^a			
Iteration	Change in Cluster Centers		
	1	2	3
1	1,897	,000	,000
2	,000	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 2. The minimum distance between initial centers is 7,187.

Figura 9.55 Primeira iteração do algoritmo *k-means* e mudança nas coordenadas dos centroides.

As três figuras seguintes referem-se ao estágio final do algoritmo *k-means*. Enquanto o *output* **Cluster Membership** (Figura 9.56) mostra a alocação de cada observação em cada um dos três *clusters*, bem como as distâncias

euclidianas entre cada observação e o centroide do respectivo grupo, o *output* **Distances between Final Cluster Centers** (Figura 9.58) apresenta as distâncias euclidianas entre os centroides dos grupos. Esses dois *outputs* trazem valores já calculados algebricamente na seção 9.2.2.2.2 e apresentados na Tabela 9.22. Além disso, o *output* **Final Cluster Centers** (Figura 9.57) apresenta as coordenadas dos centroides dos grupos após o estágio final desse procedimento não hierárquico, que correspondem aos valores já calculados e apresentados na Tabela 9.21.

Cluster Membership			
Case Number	estudante	Cluster	Distance
1	Gabriela	1	1,897
2	Luiz Felipe	2	,000
3	Patrícia	3	,000
4	Ovídio	1	2,791
5	Leonor	1	2,998

Figura 9.56 Estágio final do algoritmo *k-means* – Alocação das observações e distâncias a centroides de respectivos *clusters*.

	Cluster		
	1	2	3
nota de matemática (0 a 10)	4,7	7,8	8,9
nota de física (0 a 10)	1,9	8,0	1,0
nota de química (0 a 10)	7,7	1,5	2,7

Figura 9.57 Estágio final do algoritmo *k-means* – Coordenadas dos centroides dos *clusters*.

Distances between Final Cluster Centers			
Cluster	1	2	3
1		9,234	6,592
2	9,234		7,187
3	6,592	7,187	

Figura 9.58 Estágio final do algoritmo *k-means* – Distâncias entre os centroides dos *clusters*.

O *output* **ANOVA** (Figura 9.59) é análogo àquele apresentado na Tabela 9.29 da seção 9.2.2.2.2 e na Figura 9.42 da seção 9.3.1 e, por meio do qual, podemos verificar que apenas a variável *física* apresenta média estatisticamente diferente em pelo menos um dos grupos formados em relação aos demais, ao nível de 5% de significância.

Conforme discutimos anteriormente, caso uma ou mais variáveis não estejam contribuindo para a formação da quantidade sugerida de agrupamentos, sugere-se que o algoritmo seja reaplicado sem a presença dessas variáveis. O pesquisador pode inclusive fazer uso de um procedimento hierárquico sem a presença das referidas variáveis antes da reaplicação do procedimento *k-means*. Para os dados de nosso exemplo, entretanto, a análise se tornaria univariada pela exclusão das variáveis *matemática* e *química*, o que comprova o **risco que o pesquisador assume ao trabalhar com bancos de dados muito pequenos em análise de agrupamentos**.

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
nota de matemática (0 a 10)	8,296	2	3,990	2	2,079	,325
nota de física (0 a 10)	16,306	2	,730	2	22,337	,043
nota de química (0 a 10)	19,176	2	5,470	2	3,506	,222

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Figura 9.59 Análise de variância de um fator no procedimento *k-means* – Variabilidades entre grupos e dentro dos grupos, estatísticas *F* e níveis de significância por variável.

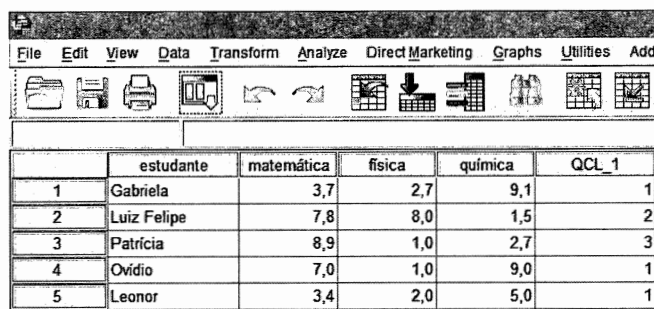
É importante mencionar que o *output* **ANOVA** deve ser utilizado apenas para o estudo das variáveis que mais contribuem para a formação da quantidade especificada de *clusters*, visto que esta é escolhida para que sejam maximizadas as diferenças entre as observações alocadas em grupos distintos. Portanto, como explicita o rodapé desse *output*, não se pode utilizar a estatística *F* com o intuito de verificar a igualdade ou não dos grupos formados. Por essa razão, não é raro que encontremos na literatura o termo *pseudo F* para essa estatística.

Por fim, a Figura 9.60 mostra a quantidade de observações em cada um dos *clusters*.

Number of Cases in each Cluster		
Cluster	1	3,000
	2	1,000
	3	1,000
Valid		5,000
Missing		,000

Figura 9.60 Quantidade de observações em cada *cluster*.

Analogamente ao procedimento hierárquico, podemos verificar que é gerada uma nova variável (obviamente qualitativa) no banco de dados após a elaboração do procedimento *k-means*, chamada pelo SPSS de *QCL_1*, conforme mostra a Figura 9.61.



	estudante	matemática	física	química	QCL_1
1	Gabriela	3,7	2,7	9,1	1
2	Luiz Felipe	7,8	8,0	1,5	2
3	Patrícia	8,9	1,0	2,7	3
4	Ovídio	7,0	1,0	9,0	1
5	Leonor	3,4	2,0	5,0	1

Figura 9.61 Banco de dados com nova variável *QCL_1* – Alocação de cada observação.

Essa variável acabou sendo idêntica à variável *CLU3_1* (Figura 9.37) neste exemplo. Porém, esse fato nem sempre acontece para uma quantidade maior de observações e nos casos em que são utilizadas medidas de dissimilaridade distintas nos procedimentos hierárquico e não hierárquico.

Apresentados os procedimentos para aplicação da análise de agrupamentos no SPSS, partiremos para a elaboração da técnica no Stata.

9.4. ANÁLISE DE AGRUPAMENTOS COM ESQUEMAS DE AGLOMERAÇÃO HIERÁRQUICOS E NÃO HIERÁRQUICOS NO SOFTWARE STATA

Apresentaremos agora o passo a passo para a elaboração de nosso exemplo no Stata Statistical Software®. Nosso objetivo, nesta seção, não é discutir novamente os conceitos pertinentes à análise de agrupamentos, mas propiciar ao pesquisador uma oportunidade de elaborar a técnica por meio dos comandos desse software. A cada apresentação de um *output*, faremos menção ao respectivo resultado obtido quando da aplicação da técnica de forma algébrica e também por meio do SPSS. A reprodução das imagens apresentadas nesta seção tem autorização da StataCorp LP®.

9.4.1. Elaboração de esquemas de aglomeração hierárquicos no software Stata

Já partiremos, portanto, para o banco de dados elaborado pelo professor a partir dos levantamentos das notas de Matemática, Física e Química obtidas no vestibular por cinco alunos. O banco de dados encontra-se no arquivo **Vestibular.dta** e é exatamente igual ao apresentado na Tabela 9.11 da seção 9.2.2.1.2.

Inicialmente, podemos digitar o comando **desc**, que possibilita a análise das características do banco de dados, como a quantidade de observações, a quantidade de variáveis e a descrição de cada uma. A Figura 9.62 apresenta o primeiro *output* do Stata.

```
. desc
```

obs:	5
vars:	4
size:	135 (99.9% of memory free)

```
-----
```

variable name	storage type	display format	value label	variable label

estudante	strll	%11s		
matemática	float	%9.1f		nota de matemática (0 a 10)
física	float	%9.1f		nota de física (0 a 10)
química	float	%9.1f		nota de química (0 a 10)

```
Sorted by:
```

Figura 9.62 Descrição do banco de dados **Vestibular.dta**.

Conforme já discutimos, como as variáveis originais apresentam valores na mesma unidade de medida, não é necessário padronizá-las pelo procedimento *Zscores* nesse exemplo. Entretanto, caso o pesquisador deseje, poderá obter as variáveis padronizadas por meio dos seguintes comandos:

```
egen zmatemática = std(matemática)
```

```
egen zfísica = std(física)
```

```
egen zquímica = std(química)
```

Inicialmente, vamos obter a matriz de distâncias entre os pares de observações. De maneira geral, a sequência de comandos para a obtenção de matrizes de distância ou de semelhança no Stata é:

```
matrix dissimilarity D = variáveis*, opção*
```

```
matrix list D
```

em que o termo **variáveis*** deverá ser substituído pela lista de variáveis a serem consideradas na análise, e o termo **opção*** deverá ser substituído pelo termo correspondente à medida de distância ou de semelhança que se deseja utilizar. Enquanto o Quadro 9.2 apresenta os termos do Stata correspondentes a cada uma das medidas para variáveis métricas estudadas na seção 9.2.1.1, o Quadro 9.3 apresenta os termos referentes às medidas utilizadas para variáveis binárias estudadas na seção 9.2.1.2.

Quadro 9.2 Termos do Stata correspondentes às medidas para variáveis métricas.

Medida para Variáveis Métricas	Termo do Stata
Euclidiana	L2
Quadrática euclidiana	L2squared
Manhattan	L1
Chebychev	Linf
Canberra	Canberra
Correlação de Pearson	corr

Quadro 9.3 Termos do Stata correspondentes às medidas para variáveis binárias.

Medida para Variáveis Binárias	Termo do Stata
Emparelhamento simples	matching
Jaccard	Jaccard
Dice	Dice
AntiDice	antiDice
Russell e Rao	Russell
Ochiai	Ochiai
Yule	Yule
Rogers e Tanimoto	Rogers
Sneath e Sokal	Sneath
Hamann	Hamann

Portanto, como desejamos obter a matriz de distâncias euclidianas entre os pares de observações, a fim de que seja mantido o critério adotado no capítulo, devemos digitar a seguinte sequência de comandos:

```
matrix dissimilarity D = matemática física química, L2
```

```
matrix list D
```

O *output* gerado, que se encontra na Figura 9.63, está em conformidade com o apresentado na matriz D_0 da seção 9.2.2.1.2.1, e também na Figura 9.30 quando da elaboração da técnica no SPSS (seção 9.3.1).

```
. matrix dissimilarity D = matemática física química, L2
. matrix list D

symmetric D[5,5]
      obs1      obs2      obs3      obs4      obs5
obs1      0
obs2  10.132127      0
obs3  8.4196199  7.1867934      0
obs4  3.7134889  10.290287  6.5802734      0
obs5  4.1701323  8.2225301  6.0448321  5.4735728      0
```

Figura 9.63 Matriz de distâncias euclidianas entre pares de observações.

Na sequência, vamos partir para a realização da análise de agrupamentos propriamente dita. O comando geral para a elaboração de uma análise de agrupamentos por meio de um esquema hierárquico no Stata é dado por:

```
cluster método* variáveis*, measure(opção*)
```

em que, além da substituição dos termos *variáveis** e *opção**, conforme discutimos anteriormente, devemos substituir o termo *método** pelo correspondente ao método de encadeamento escolhido pelo pesquisador. O Quadro 9.4 apresenta os termos do Stata referentes aos métodos estudados na seção 9.2.2.1.

Quadro 9.4 Termos do Stata correspondentes aos métodos de encadeamento em esquemas hierárquicos de aglomeração

Método de Encadeamento	Termo do Stata
Único	singlelinkage
Completo	completelinkage
Médio	averagelinkage

Portanto, para os dados de nosso exemplo e seguindo o critério adotado ao longo do capítulo (método de encadeamento único com distância euclidiana – termo **L2**), devemos digitar o seguinte comando:

```
cluster singlelinkage matemática física química, measure(L2)
```

Em seguida, podemos digitar o comando **cluster list**, que faz com que sejam apresentados, de forma resumida, os critérios utilizados pelo pesquisador para a elaboração da análise de agrupamentos hierárquicos. A Figura 9.64 apresenta os *outputs* gerados.

```
. cluster singlelinkage matemática física química, measure(L2)
cluster name: _clus_1

. cluster list
_clus_1 (type: hierarchical, method: single, dissimilarity: L2)
  vars: _clus_1_id (id variable)
        _clus_1_ord (order variable)
        _clus_1_hgt (height variable)
  other: cmd: cluster singlelinkage matemática física química, measure(L2)
         varlist: matemática física química
         range: 0 .
```

Figura 9.64 Elaboração da análise de agrupamentos hierárquicos e resumo dos critérios adotados.

A partir da Figura 9.64 e da análise do banco de dados, podemos verificar que três novas variáveis são criadas, referentes à identificação de cada observação (*_clus_1_id*), ao ordenamento das observações quando dos agrupamentos (*_clus_1_ord*) e às distâncias euclidianas utilizadas para que se agrupe nova observação em cada um dos estágios de aglomeração (*_clus_1_hgt*). A Figura 9.65 mostra como fica o banco de dados após a elaboração dessa análise de agrupamentos.

	estudante	matemática	física	química	_clus_1_id	_clus_1_ord	_clus_1_hgt
1	Gabriela	3.7	2.7	9.1	1	2	7.1867934
2	Luiz Felipe	7.8	8.0	1.5	2	3	6.0448321
3	Patrícia	8.9	1.0	2.7	3	1	3.7134889
4	Ovídio	7.0	1.0	9.0	4	4	4.1701323
5	Leonor	3.4	2.0	5.0	5	5	.

Figura 9.65 Banco de dados com as novas variáveis.

É importante mencionar que o Stata apresenta a variável *_clu_1_hgt* com valores defasados em uma linha, o que pode tornar a análise um pouco confusa. Nesse sentido, enquanto a distância de 3,713 refere-se à fusão entre as observações **Ovídio** e **Gabriela** (primeiro estágio do esquema de aglomeração), a distância de 7,187 corresponde à fusão entre **Luiz Felipe** e o *cluster* já formado por todas as demais observações (último estágio do esquema de aglomeração), conforme já mostravam a Tabela 9.12 e a Figura 9.31.

Logo, para que o pesquisador corrija este problema de defasagem e obtenha o real comportamento das distâncias em cada novo estágio de aglomeração, poderá digitar a sequência de comandos a seguir, cujo *output* se encontra na Figura 9.66. Note que uma nova variável é criada (*dist*) e corresponde à correção da defasagem da variável *_clu_1_hgt* (termo [*_n-1*]), apresentando o valor de cada distância euclidiana para que se estabeleça um novo agrupamento em cada estágio do esquema de aglomeração.

```
gen dist = _clus_1_hgt[_n-1]
replace dist=0 if dist==.
sort dist
list estudante dist
```

```
. gen dist = _clus_1_hgt[_n-1]
(1 missing value generated)

. replace dist=0 if dist==.
(1 real change made)

. sort dist

. list estudante dist
```

	estudante	dist
1.	Gabriela	0
2.	Ovídio	3.713489
3.	Leonor	4.170132
4.	Patrícia	6.044832
5.	Luiz Felipe	7.186793

Figura 9.66 Estágios do esquema de aglomeração e respectivas distâncias euclidianas.

Elaborada essa etapa, podemos solicitar que o Stata construa o dendrograma, digitando um dos dois equivalentes comandos:

```
cluster dendrogram, labels(estudante) horizontal
```

ou

```
cluster tree, labels(estudante) horizontal
```

O gráfico gerado encontra-se na Figura 9.67.

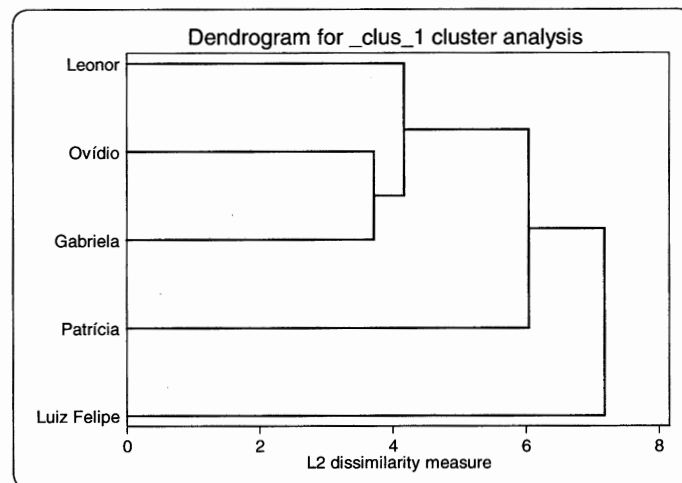


Figura 9.67 Dendrograma – Método de encadeamento único e distâncias euclidianas no Stata.

Podemos notar que o dendrograma construído pela Stata, em termos de distâncias euclidianas, é igual ao apresentado na Figura 9.12, elaborada quando da resolução algébrica da modelagem, porém difere-se daquele construído pelo SPSS (Figura 9.32) por não considerar medidas rescalonadas. Independentemente desse fato, vamos adotar como possível solução uma quantidade de três *clusters*, sendo um formado por **Leonor**, **Ovídio** e **Gabriela**, outro, por **Patrícia**, e um terceiro, por **Luiz Felipe**, já que os critérios discutidos sobre grandes saltos de distância nos levam coerentemente a essa decisão.

Para que seja gerada uma nova variável, correspondente à alocação das observações nos três *clusters*, devemos digitar a sequência de comandos a seguir. Note que nomeamos essa nova variável de *cluster*. O *output* da Figura 9.68 mostra a alocação das observações nos grupos e é equivalente ao apresentado na Figura 9.36 (SPSS).

```
cluster generate cluster = groups(3), name(_clus_1)
sort _clus_1_id
list estudante cluster
```

```
. cluster generate cluster = groups(3), name(_clus_1)
. sort _clus_1_id
. list estudante cluster
```

	estudante	cluster
1.	Gabriela	3
2.	Luiz Felipe	1
3.	Patrícia	2
4.	Ovídio	3
5.	Leonor	3

Figura 9.68 Alocação das observações nos *clusters*.

Finalmente, vamos estudar, por meio da análise de variância de um fator (ANOVA), se os valores de determinada variável diferem-se entre os grupos representados pelas categorias da nova variável qualitativa *cluster* gerada no banco de dados, ou seja, se a variabilidade entre os grupos é significativamente superior à variabilidade interna a cada um deles, seguindo a lógica proposta na seção 9.3.1. Para tanto, vamos digitar os seguintes comandos, em que são relacionadas individualmente as três variáveis métricas (*matemática*, *física* e *química*) com a variável *cluster*:

oneway matemática cluster, tabulate

oneway física cluster, tabulate

oneway química cluster, tabulate

Os resultados da ANOVA para as três variáveis estão na Figura 9.69.

. oneway matemática cluster, tabulate

Summary of nota de matemática (0 a 10)			
cluster	Mean	Std. Dev.	Freq.
1	7.8	0.0	1
2	8.9	0.0	1
3	4.7	2.0	3
Total	6.2	2.5	5

Analysis of Variance					
Source	SS	df	MS	F	Prob > F
Between groups	16.5919981	2	8.29599906	2.08	0.3248
Within groups	7.97999966	2	3.98999983		
Total	24.5719978	4	6.14299944		

. oneway física cluster, tabulate

Summary of nota de física (0 a 10)			
cluster	Mean	Std. Dev.	Freq.
1	8.0	0.0	1
2	1.0	0.0	1
3	1.9	0.9	3
Total	2.9	2.9	5

Analysis of Variance					
Source	SS	df	MS	F	Prob > F
Between groups	32.6119999	2	16.306	22.34	0.0429
Within groups	1.46000008	2	.730000038		
Total	34.072	4	8.51799999		

. oneway química cluster, tabulate

Summary of nota de química (0 a 10)			
cluster	Mean	Std. Dev.	Freq.
1	1.5	0.0	1
2	2.7	0.0	1
3	7.7	2.3	3
Total	5.5	3.5	5

Analysis of Variance					
Source	SS	df	MS	F	Prob > F
Between groups	38.3520014	2	19.1760007	3.51	0.2219
Within groups	10.9400011	2	5.47000053		
Total	49.2920025	4	12.3230006		

Figura 9.69 ANOVA para as variáveis matemática, física e química.

Os *outputs* dessa figura, que apresentam os resultados das variabilidades entre os grupos (**Between groups**) e dentro dos grupos (**Within groups**), as estatísticas *F* e os respectivos níveis de significância (*Prob. F*, ou **Prob > F** no Stata) para cada variável, são iguais aos calculados algebricamente e apresentados na Tabela 9.29 (seção 9.2.2.2.2) e também na Figura 9.42 quando da elaboração deste procedimento no SPSS (seção 9.3.1).

Portanto, conforme já discutimos, podemos verificar que, enquanto para a variável *física* existe pelo menos um *cluster* que apresenta média estatisticamente diferente dos demais, ao nível de significância de 5% (*Prob. F* = 0,0429

$< 0,05$), as variáveis *matemática* e *química* não possuem médias estatisticamente diferentes entre os três grupos formados para essa amostra e ao nível de significância estipulado.

É importante lembrar que, caso exista uma quantidade maior de variáveis que apresentem *Prob. F* menor que 0,05, aquela considerada mais discriminante dos grupos é a com maior estatística *F* (ou seja, menor nível de significância *Prob. F*).

Mesmo podendo finalizar a análise hierárquica neste momento, o pesquisador tem a opção de elaborar um escalonamento multidimensional, a fim de visualizar as projeções das posições relativas das observações em um gráfico bidimensional, assim como realizado na seção 9.3.1. Para tanto, poderá digitar o seguinte comando:

```
mds matemática física química, id(estudante) method(modern)
measure(L2) loss(sstress) config nolog
```

Os *outputs* gerados encontram-se nas Figuras 9.70 e 9.71, sendo que o gráfico desta última figura corresponde ao apresentado na Figura 9.50.

```
. mds matemática física química, id(estud) method(modern) measure(L2) loss(sstress)
config nolog
(transform(identity) assumed)

Modern multidimensional scaling
  dissimilarity: L2, computed on 3 variables

Loss criterion: sstress = raw_sstress/norm(distances^2)
Transformation: identity (no transformation)

                                     Number of obs   =           5
                                     Dimensions         =           2
                                     Loss criterion      =       0.1095

Normalization: principal

Configuration in 2-dimensional Euclidean space (principal normalization)
```

estudante	dim1	dim2
Gabriela	3.9262	1.9516
Ovídio	3.5524	-1.9206
Leonor	1.2243	1.8871
Patrícia	-2.2858	-3.7417
Luiz_Felipe	-6.4170	1.8237

Figura 9.70 Elaboração do escalonamento multidimensional no Stata.

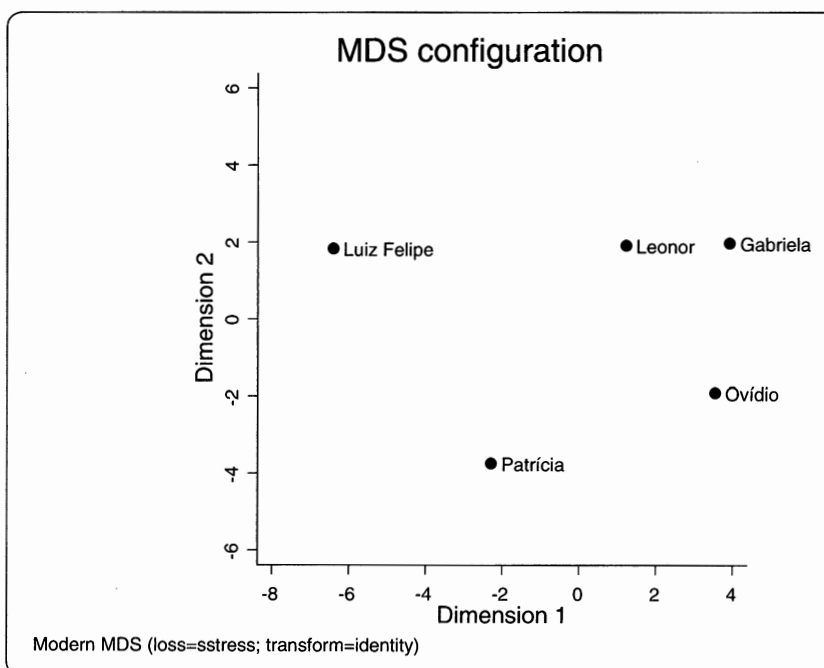


Figura 9.71 Gráfico com projeções das posições relativas das observações.

Apresentados os comandos para a realização da análise de agrupamentos com esquema de aglomeração hierárquico no Stata, partiremos para a elaboração do esquema de aglomeração não hierárquico *k-means* no mesmo software.

9.4.2. Elaboração do esquema de aglomeração não hierárquico *k-means* no software Stata

Para que realizemos o procedimento *k-means* aos dados do arquivo **Vestibular.dta**, devemos digitar o seguinte comando:

```
cluster kmeans matemática física química, k(3) name(kmeans)
measure(L2) start(firstk)
```

em que o termo **k(3)** é *input* para que o algoritmo seja elaborado com três agrupamentos. Além disso, definimos que uma nova variável com a alocação das observações nos três grupos será gerada no banco de dados com o nome *kmeans* (termo **name(kmeans)**), e a medida de distância utilizada será a distância euclidiana (termo **L2**). Além disso, o termo **firstk** especifica que as coordenadas das primeiras *k* observações da amostra serão utilizadas como centroides dos *k* clusters (no nosso caso, $k = 3$), o que corresponde exatamente ao critério adotado pelo SPSS, conforme discutimos na seção 9.3.2.

Na sequência, podemos digitar o comando **cluster list kmeans** para que sejam apresentados, de forma resumida, os critérios adotados para a elaboração do procedimento *k-means*.

Os *outputs* da Figura 9.72 mostram o que é gerado pelo Stata após a digitação dos dois últimos comandos.

```
. cluster kmeans matemática física química, k(3) name(kmeans) measure(L2) start(firstk)

. cluster list kmeans
kmeans (type: partition, method: kmeans, dissimilarity: L2)
  vars: kmeans (group variable)
  other: cmd: cluster kmeans matemática física química, k(3) name (kmeans)
measure(L2) start(firstk)
  varlist: matemática física química
  k: 3
  start: firstk
  range: 0 .
```

Figura 9.72 Elaboração do procedimento não hierárquico *k-means* e resumo dos critérios adotados.

Os dois comandos seguintes geram, nos *outputs* do software, duas tabelas referentes, respectivamente, à quantidade de observações em cada um dos três *clusters* formados, bem como a alocação de cada observação nesses grupos:

```
table kmeans
```

```
list estudante kmeans
```

A Figura 9.73 mostra esses *outputs*.

```
. table kmeans

-----+-----
kmeans |      Freq.
-----+-----
      1 |          3
      2 |          1
      3 |          1
-----+-----

. list estudante kmeans

+-----+-----+
| estudante | kmeans |
+-----+-----+
1.   Gabriela |      1 |
2.   Luiz Felipe |      2 |
3.   Patrícia |      3 |
4.   Ovídio |      1 |
5.   Leonor |      1 |
+-----+-----+
```

Figura 9.73 Quantidade de observações em cada *cluster* e alocação das observações.

Esses resultados correspondem ao encontrado quando da resolução algébrica do procedimento *k-means* na seção 9.2.2.2 (Figura 9.23) e ao obtido quando da elaboração desse procedimento por meio do SPSS na seção 9.3.2 (Figuras 9.60 e 9.61).

Embora tenhamos condições de elaborar uma análise de variância de um fator para as variáveis originais do banco de dados, a partir da nova variável qualitativa gerada (*kmeans*), optamos por não realizar esse procedimento aqui, visto que já o fizemos para a variável *cluster* gerada na seção 9.4.1 após o procedimento hierárquico, que é exatamente igual à variável *kmeans* neste caso.

Por outro lado, apresentamos, para efeitos didáticos, o seguinte comando, que permite que as médias de cada variável nos três *clusters* sejam geradas, para efeitos de comparação:

tabstat matemática física química, by(kmeans)

O *output* gerado encontra-se na Figura 9.74, e equivale ao apresentado nas Tabelas 9.23, 9.24 e 9.25.

```
. tabstat matemática física química, by(kmeans)
```

Summary statistics: mean by categories of: kmeans			
kmeans	matemática	física	química
1	4.7	1.9	7.7
2	7.8	8	1.5
3	8.9	1	2.7
Total	6.16	2.94	5.46

Figura 9.74 Médias por *cluster* e geral das variáveis *matemática*, *física* e *química*.

Por fim, o pesquisador pode ainda elaborar um gráfico que mostra as inter-relações das variáveis, duas a duas. Esse gráfico, conhecido por **matrix**, pode propiciar ao pesquisador melhor entendimento sobre como as variáveis se relacionam, oferecendo inclusive sugestões acerca do posicionamento relativo das observações de cada *cluster* nessas inter-relações. Para a construção do gráfico, que se encontra na Figura 9.75, devemos digitar o seguinte comando:

graph matrix matemática física química, mlabel(kmeans)

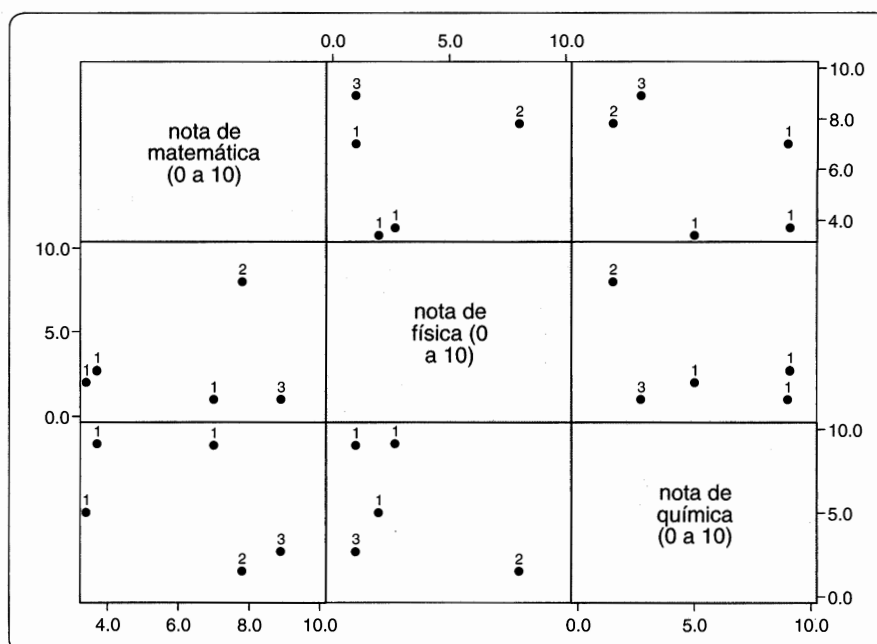


Figura 9.75 Inter-relação das variáveis e posição relativa das observações de cada *cluster* – Gráfico **matrix**.

Obviamente, este gráfico poderia também ter sido construído na seção anterior, porém optamos por apresentá-lo apenas ao término da elaboração do procedimento *k-means* no Stata. Por meio de sua análise, é possível verificarmos, entre outros fatos, que a consideração apenas das variáveis *matemática* e *química* não é suficiente para que sejam afastadas as observações **Luiz Felipe** e **Patrícia** (*clusters* 2 e 3, respectivamente), sendo necessária a consideração da variável *física* para que esses dois estudantes sejam, de fato, alocados em *clusters* distintos quando da formação de três agrupamentos. Embora seja um tanto quanto óbvio quando analisamos os dados na própria base, o gráfico torna-se bastante útil para amostras maiores e com uma quantidade considerável de variáveis, fato que multiplicaria essas inter-relações.

9.5. CONSIDERAÇÕES FINAIS

Muitas são as situações em que o pesquisador pode desejar agrupar observações (indivíduos, empresas, municípios, países, partidos políticos, espécies vegetais, entre outros exemplos) a partir de determinadas variáveis métricas ou até mesmo binárias. A criação de agrupamentos homogêneos, a redução estrutural dos dados e a verificação da validade de constructos previamente estabelecidos são algumas das principais razões que levam o pesquisador a optar por trabalhar com a análise de agrupamentos.

Esse conjunto de técnicas permite que os mecanismos de tomada de decisão sejam mais bem estruturados e justificados a partir do comportamento e da relação de interdependência entre as observações de determinado banco de dados. Como a variável que representa os *clusters* formados é qualitativa, os *outputs* da análise de agrupamentos podem servir de *inputs* em outras técnicas multivariadas, tanto exploratórias, quanto confirmatórias.

É fortemente recomendável que o pesquisador justifique, com clareza e transparência, a escolha da medida que servirá de base para que as observações sejam consideradas mais ou menos similares, bem como as razões que o levam à definição de esquemas de aglomeração não hierárquicos ou hierárquicos e, neste último caso, à determinação dos métodos de encadeamento.

A evolução da capacidade computacional e o desenvolvimento de novos softwares com recursos bastante aprimorados fizeram surgir, nos últimos anos, novas e esmeradas técnicas de análise de agrupamentos que utilizam algoritmos cada vez mais requintados e voltados à tomada de decisão nos mais diversos campos do conhecimento, sempre com o objetivo principal de agrupar observações frente a determinados critérios. Neste capítulo, entretanto, procuramos oferecer uma visão geral sobre os principais métodos de análise de agrupamentos, considerados também os mais populares.

Finalmente, ressaltamos que a aplicação desse importante conjunto de técnicas deve ser sempre feita por meio do correto e consciente uso do software escolhido para a modelagem, com base na teoria subjacente e na experiência e intuição do pesquisador.

9.6. EXERCÍCIOS

1. O departamento de concessão de bolsas de estudo de uma faculdade deseja investigar a relação de interdependência entre os estudantes ingressantes em determinado ano letivo, com base apenas em duas variáveis métricas (idade, em anos, e renda média familiar, em R\$). O objetivo é propor uma quantidade ainda desconhecida de novos programas de concessão de bolsas voltados a grupos homogêneos de alunos. Para tanto, foram coletados os dados dos 100 novos estudantes e elaborada uma base, que se encontra nos arquivos **Bolsa de Estudo.sav** e **Bolsa de Estudo.dta**, com as seguintes variáveis:

Variável	Descrição
<i>estudante</i>	Variável <i>string</i> que identifica o estudante ingressante na faculdade.
<i>idade</i>	Idade do estudante (anos).
<i>renda</i>	Renda média familiar (R\$).

Pede-se:

- a. Elabore uma análise de agrupamentos por meio de um esquema de aglomeração hierárquico, com método de encadeamento completo (*furthest neighbor*) e distância quadrática euclidiana. Apresente apenas a parte final da tabela do esquema de aglomeração e discuta os resultados. **Lembrete:** Como as variáveis possuem unidades distintas de medida, é necessária a aplicação do procedimento de padronização *Zscores* para a correta elaboração da análise de agrupamentos.

- b. Com base na tabela do item anterior e no dendrograma, pergunta-se: Há indícios de serem formados quantos agrupamentos de estudantes?
 - c. É possível identificar um ou mais estudantes muito discrepantes dos demais em relação às duas variáveis em análise?
 - d. Se a resposta do item anterior for positiva, elabore novamente a análise de agrupamentos hierárquicos com os mesmos critérios, porém, agora, sem o(s) estudante(s) considerado(s) discrepante(s). A partir da análise dos novos resultados, podem ser identificados novos agrupamentos?
 - e. Discuta como a presença de *outliers* pode prejudicar a interpretação dos resultados em análise de agrupamentos.
2. A diretoria de marketing de um grupo varejista deseja estudar eventuais discrepâncias existentes em suas 18 lojas espalhadas em três regionais distribuídas pelo território nacional. A direção da companhia, a fim de manter e preservar a imagem e a identidade da marca, deseja saber se as lojas são homogêneas em relação à percepção dos consumidores sobre atributos como atendimento, sortimento e organização. Dessa forma, foi inicialmente elaborada uma pesquisa com amostras de clientes em cada loja, a fim de que fossem coletados dados referentes a esses atributos, definidos com base na nota média obtida (0 a 100) em cada estabelecimento comercial.

Na sequência, foi elaborado o banco de dados de interesse, que contém as seguintes variáveis:

Variável	Descrição
<i>loja</i>	Variável <i>string</i> que varia de 01 a 18 e que identifica o estabelecimento comercial (loja).
<i>regional</i>	Variável <i>string</i> que identifica cada regional (Regional 1 a Regional 3).
<i>atendimento</i>	Avaliação média dos consumidores sobre o atendimento (nota de 0 a 100).
<i>sortimento</i>	Avaliação média dos consumidores sobre o sortimento (nota de 0 a 100).
<i>organização</i>	Avaliação média dos consumidores sobre a organização da loja (nota de 0 a 100).

Os dados encontram-se nos arquivos **Regional Varejista.sav** e **Regional Varejista.dta**. Pede-se:

- a. Elabore uma análise de agrupamentos por meio de um esquema de aglomeração hierárquico, com método de encadeamento único e distância euclidiana. Apresente a matriz de distâncias entre cada par de observações. **Lembrete:** Como as variáveis possuem a mesma unidade de medida, não é necessária a aplicação do procedimento de padronização *Zscores*.
 - b. Apresente e discuta a tabela do esquema de aglomeração.
 - c. Com base na tabela do item anterior e no dendrograma, pergunta-se: Há indícios de serem formados quantos agrupamentos de lojas?
 - d. Elabore um escalonamento multidimensional e, na sequência, apresente e discuta o gráfico bidimensional gerado com as posições relativas das lojas.
 - e. Elabore uma análise de agrupamentos por meio do procedimento *k-means*, com a quantidade de agrupamentos sugerida no item (c), e interprete, considerando o nível de significância de 5%, a análise de variância de um fator para cada variável considerada no estudo. Qual variável mais contribui para a formação de pelo menos um dos *clusters* formados, ou seja, qual delas é a mais discriminante dos grupos?
 - f. Existe correspondência entre as alocações das observações nos grupos obtidas pelos métodos hierárquico e não hierárquico?
 - g. É possível identificar associação entre alguma regional e determinado grupo discrepante de lojas, o que poderia justificar a preocupação da diretoria em relação à imagem e à identidade da marca? Caso a resposta seja afirmativa, elabore novamente a análise de agrupamentos hierárquicos com os mesmos critérios, porém, agora, sem esse grupo discrepante de lojas. A partir da análise dos novos resultados, pode-se visualizar, de forma mais nítida, as diferenças entre as demais lojas?
3. Um analista do mercado financeiro decide elaborar uma pesquisa com presidentes e diretores de grandes empresas atuantes nos setores de saúde, educação e transporte, a fim de investigar o modo como são realizadas as operações das companhias e os mecanismos que regem os processos decisórios. Para tanto, elaborou um questionário com 50 perguntas, cujas respostas são apenas dicotômicas, ou binárias. Após a aplicação do questionário, obteve um retorno de 35 empresas e, a partir de então, estruturou o banco de dados, presente nos arquivos **Pesquisa Binária.sav** e **Pesquisa Binária.dta**. De maneira genérica, as variáveis são:

Variável	Descrição
<i>q1 a q50</i>	50 variáveis <i>dummy</i> que se referem ao modo como são realizados as operações e os processos de tomada de decisão nas empresas.
<i>setor</i>	Setor de atuação da empresa (critério Bovespa).

O principal objetivo do analista é verificar se empresas atuantes no mesmo setor apresentam similaridades em relação ao modo como são realizados as operações e os processos de tomada de decisão, ao menos na perspectiva dos próprios gestores. Para tanto, após a coleta dos dados, pode ser elaborada uma análise de agrupamentos. Pede-se:

- a. Com base na análise de agrupamentos hierárquicos elaborada com método de encadeamento médio (*between groups*) e medida de semelhança (similaridade) de emparelhamento simples para variáveis binárias, analise o esquema de aglomeração gerado.
 - b. Interprete o dendrograma.
 - c. Verifique se existe correspondência entre as alocações das empresas nos *clusters* e os respectivos setores de atuação, ou, em outras palavras, se as empresas atuantes no mesmo setor apresentam similaridades em relação ao modo como são realizados as operações e os processos de tomada de decisão.
4. O proprietário de uma empresa hortifrúti decide monitorar as vendas de seus produtos ao longo de 16 semanas (4 meses). O objetivo principal é verificar se existe recorrência do comportamento de vendas de três principais produtos (banana, laranja e maçã) após certo período, em função das oscilações semanais de preços dos produtores, repassados aos consumidores e que podem afetar as vendas. Os dados encontram-se nos arquivos **Hortifrúti.sav** e **Hortifrúti.dta**, que apresentam as seguintes variáveis:

Variável	Descrição
<i>semana</i>	Variável <i>string</i> que varia de 1 a 16 e identifica a semana em que as vendas foram monitoradas.
<i>semana_mês</i>	Variável <i>string</i> que varia de 1 a 4 e identifica a semana de cada um dos meses.
<i>banana</i>	Quantidade de bananas vendidas na semana (un.).
<i>laranja</i>	Quantidade de laranjas vendidas na semana (un.).
<i>maçã</i>	Quantidade de maçãs vendidas na semana (un.).

Pede-se:

- a. Elabore uma análise de agrupamentos por meio de um esquema de aglomeração hierárquico, com método de encadeamento único (*nearest neighbor*) e medida de correlação de Pearson. Apresente a matriz de medidas de similaridade (correlação de Pearson) entre cada linha do banco de dados (períodos semanais). **Lembrete:** Como as variáveis possuem a mesma unidade de medida, não é necessária a aplicação do procedimento de padronização *Zscores*.
- b. Apresente e discuta a tabela do esquema de aglomeração.
- c. Com base na tabela do item anterior e no dendrograma, pergunta-se: Há indícios de recorrência do comportamento conjunto de vendas de banana, laranja e maçã em determinadas semanas?

APÊNDICE

Detecção de *outliers* multivariados

Embora a detecção de *outliers* seja extremamente importante quando da aplicação de praticamente todas as técnicas em análise multivariada de dados, optamos por inserir este apêndice no presente capítulo em razão de a análise de agrupamentos representar o primeiro conjunto estudado de técnicas exploratórias, cujos *outputs* podem ser utilizados como *inputs* de diversas outras técnicas, bem como pelo fato de observações muito discrepantes poderem interferir consideravelmente na formação dos *clusters*.

Barnett e Lewis (1994) citam quase 1.000 artigos provenientes da literatura sobre *outliers*; porém, optamos por apresentar um algoritmo bastante efetivo e computacionalmente simples e rápido para a detecção de *outliers* multivariados.

A. Breve Apresentação do Algoritmo *Blocked Adaptive Computationally Efficient Outlier Nominators*

Billor, Hadi e Velleman (2000), em seminal trabalho, apresentam um interessante algoritmo que possui a finalidade de detectar *outliers* multivariados, denominado *Blocked Adaptive Computationally Efficient Outlier Nominators*, ou simplesmente *BACON*. Esse algoritmo, explicado de forma clara e didática por Weber (2012), é definido com base na elaboração de alguns passos, descritos brevemente a seguir:

1. A partir de um banco de dados com n observações e j ($j = 1, \dots, k$) variáveis X , sendo cada observação identificada por i ($i = 1, \dots, n$), a distância entre uma observação i , que possui um vetor com dimensão k $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$, e a média geral dos valores de toda a amostra (grupo G), que também possui um vetor com dimensão k $\bar{\mathbf{x}} (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k)$, é dada pela seguinte expressão, conhecida por **distância de Mahalanobis**:

$$d_{iG} = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})' \cdot \mathbf{S}^{-1} \cdot (\mathbf{x}_i - \bar{\mathbf{x}})} \quad (9.29)$$

em que \mathbf{S} representa a matriz de covariâncias das n observações. Portanto, o passo inicial do algoritmo consiste em identificar m ($m > k$) observações homogêneas (grupo inicial M) que apresentam as menores distâncias de Mahalanobis com relação à amostra toda.

É importante mencionar que a medida de dissimilaridade conhecida por distância de Mahalanobis, não abordada ao longo do capítulo, é adotada pelos autores supramencionados por possuir a propriedade de não ser suscetível à existência de diferentes unidades de medida das variáveis.

2. Na sequência, são calculadas as distâncias de Mahalanobis entre cada observação i e a média dos valores das m observações pertencentes ao grupo M , que também possui um vetor com dimensão k $\bar{\mathbf{x}}_M (\bar{x}_{M1}, \bar{x}_{M2}, \dots, \bar{x}_{Mk})$, de modo que:

$$d_{iM} = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}}_M)' \cdot \mathbf{S}_M^{-1} \cdot (\mathbf{x}_i - \bar{\mathbf{x}}_M)} \quad (9.30)$$

em que \mathbf{S}_M representa a matriz de covariâncias das m observações.

3. Todas as observações com distâncias de Mahalanobis menores que determinado limiar são adicionadas ao grupo *M* de observações. Esse limiar é definido como um percentil corrigido da distribuição χ^2 (85% no padrão do Stata).

Os passos 2 e 3 devem ser reaplicados até que não existam mais modificações no grupo *M*, que possuirá apenas observações consideradas não *outliers*. Portanto, as excluídas do grupo serão consideradas **outliers multivariados**.

Weber (2012) codifica o algoritmo proposto no trabalho de Billor, Hadi e Velleman (2000) no Stata, criando o comando **bacon**. Na sequência, apresentamos um exemplo em que é utilizado esse comando, cuja principal vantagem é ser computacionalmente muito rápido, mesmo quando aplicado a grandes bancos de dados.

B. Exemplo: O Comando **bacon** no Stata

Antes da elaboração específica deste procedimento no Stata, devemos instalar o comando **bacon**, digitando **findit bacon** e clicando no link **st0197 from <http://www.stata-journal.com/software/sj10-3>**. Na sequência, devemos clicar em **click here to install**. Por fim, retornando à tela de comandos do Stata, podemos digitar **ssc install moremata** e **mata: mata mlib index**. Feito isso, temos condições de aplicar o comando **bacon**.

Para o uso do comando, utilizaremos o arquivo **Bacon.dta**, que apresenta dados de 20.000 engenheiros sobre renda média familiar (R\$), idade (anos) e tempo de formado (anos). Inicialmente, podemos digitar o comando **desc**, que possibilita a análise das características do banco de dados. A Figura 9.76 apresenta esse primeiro *output*.

```
. desc
```

obs:	20,000			
vars:	3			
size:	200,000 (99.6% of memory free)			

variable name	storage type	display format	value label	variable label
renda	float	%9.0g		renda média familiar (R\$)
idade	byte	%8.0g		idade (anos)
tformado	byte	%8.0g		tempo de formado (anos)

Sorted by:

Figura 9.76 Descrição do banco de dados **Bacon.dta**.

Na sequência, podemos digitar o seguinte comando, que identifica, com base no algoritmo apresentado, as observações consideradas *outliers* multivariados:

bacon renda idade tformado, generate(outbacon)

em que o termo **generate(outbacon)** faz com que seja gerada uma nova variável *dummy* no banco de dados, denominada *outbacon*, que apresenta valores iguais a 0 para observações não consideradas *outliers*, e valores iguais a 1 para as consideradas como tal. Esse *output* encontra-se na Figura 9.77.

```
. bacon renda idade tformado, generate(outbacon)
```

Total number of observations:	20000
BACON outliers (p = 0.15):	4
Non-outliers remaining:	19996

Figura 9.77 Aplicação do comando **bacon** no Stata.

Por meio dessa figura, é possível verificarmos que quatro observações são classificadas como *outliers* multivariados. Além disso, o Stata considera 85% o padrão de percentil da distribuição χ^2 , utilizado como limiar de separação entre observações tidas como *outliers* e não *outliers*, conforme discutido anteriormente e destacado por Weber (2012). Essa é a razão de, nos *outputs*, aparecer o termo **BACON outliers (p = 0.15)**. Esse valor poderá ser alterado em função de algum critério estabelecido pelo pesquisador, porém, ressalta-se que o padrão **percentile(0.15)** é bastante adequado para a obtenção de respostas consistentes.

A partir do comando a seguir, que gera o *output* da Figura 9.78, podemos investigar quais as observações classificadas como *outliers*:

list if outbacon == 1

```
. list if outbacon==1
```

	renda	idade	tformado	outbacon
1935.	30869.93	30	15	1
2468.	34773.54	42	17	1
14128.	41191.15	50	21	1
16833.	32924.19	31	16	1

Figura 9.78 Observações classificadas como *outliers* multivariados.

Mesmo que estejamos trabalhando com três variáveis, podemos elaborar gráficos de dispersão bidimensionais, que permitem identificar as posições das observações consideradas *outliers* em relação às demais. Para tanto, vamos digitar os seguintes comandos, que geram os referidos gráficos para cada par de variáveis:

```
scatter renda idade, ml(outbacon) note("0 = não outlier, 1 = outlier")
```

```
scatter renda tformado, ml(outbacon) note("0 = não outlier, 1 = outlier")
```

```
scatter idade tformado, ml(outbacon) note("0 = não outlier, 1 = outlier")
```

Os três gráficos encontram-se nas Figuras 9.79, 9.80 e 9.81.

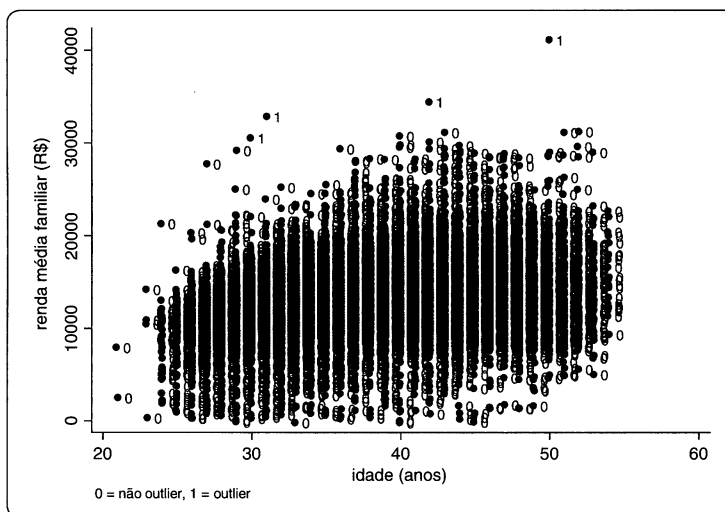


Figura 9.79 Variáveis *renda* e *idade* – Posição relativa das observações.

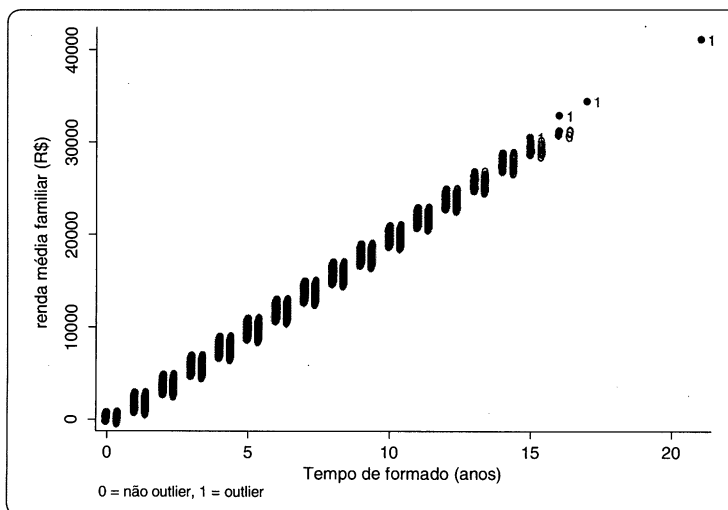


Figura 9.80 Variáveis *renda* e *tformado* – Posição relativa das observações.

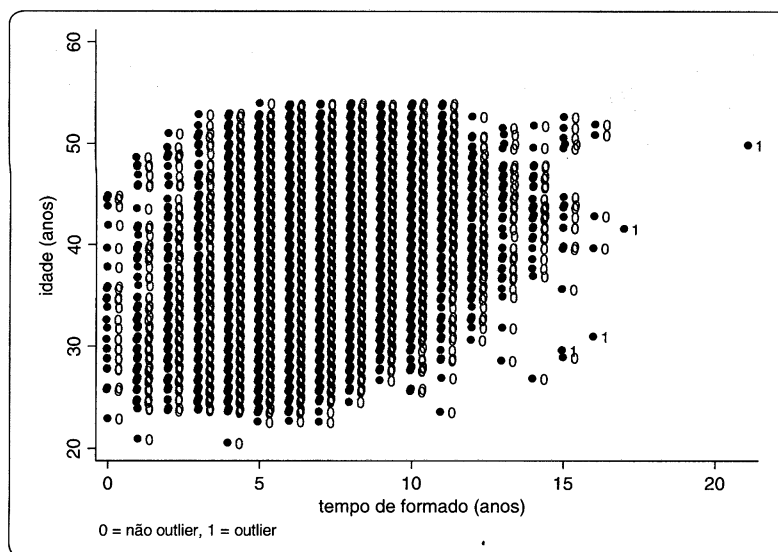


Figura 9.81 Variáveis *idade* e *tformado* – Posição relativa das observações.

Embora os *outliers* tenham sido identificados, é importante mencionar que a decisão sobre o que fazer com essas observações pertence totalmente ao pesquisador, que deverá tomá-la em função de seus objetivos de pesquisa. Conforme discutimos ao longo do capítulo, a exclusão desses *outliers* da base pode representar uma opção a ser considerada. Porém, o estudo sobre as razões que os tornaram multivariadamente discrepantes também pode gerar muitos frutos interessantes de pesquisa.