ANÁLISE DE AGRUPAMENTOS

(*CLUSTER ANALYSIS*)

Material preparado pelo Prof. Dr. César Gonçalves de Lima – FZEA/USP

DEFINIÇÕES INICIAIS

Objetivo: Reunir objetos ou indivíduos em grupos (*clusters*) com base em diversas características medidas neles.

Busca classificar um objeto ou indivíduo no grupo que apresenta características mais similares às suas, considerando um critério de seleção previamente escolhido (CORRAR, 2017).

O grupo escolhido deve apresentar um alto grau de homogeneidade interna e alta heterogeneidade externa.

Quando a classificação é bem-sucedida, os objetos ou indivíduos de um mesmo grupo estarão juntos quando *plotados* geometricamente e os diferentes grupos estarão distantes uns dos outros.

O processo de agrupamento (*clustering*) é **diferente** do método de análise discriminante, que envolve o conhecimento prévio de um número de grupos e tem o objetivo operacional de designar/classificar novos objetos/indivíduos a um desses grupos conhecidos.

A **análise de agrupamentos** é uma técnica mais primitiva em que nenhuma suposição é feita sobre o número ou sobre a estrutura dos grupos e está baseada exclusivamente em medidas de similaridade ou de distâncias (dissimilaridade) entre os objetos/indivíduos.

Na literatura aparece também com os nomes: análise de conglomerados, análise de *clusters*, *Q-analysis*, *typology*, *classification analysis* e *numerical taxonomy*.

Esta variedade de nomes deve-se, em parte, ao uso da técnica em diversas áreas de pesquisa.

Segundo Corrar (2007) existem três questões básicas em análise de agrupamentos:

- Como medir a semelhança entre objetos?
- Supondo que se possa medir a semelhança relativa entre objetos, como juntar os objetos semelhantes em *clusters*/grupos?
- Após ter efetuado o agrupamento, como descrever os *clusters* e como saber se eles são "reais" ou se são resultados de um simples artifício estatístico?

Vamos buscar uma medida de distância adequada e escolher bem as variáveis (características) que serão medidas/avaliadas nos objetos e que irão caracterizá-los.

Os agrupamentos também podem fornecer um meio informal para identificar *outliers* multivariados que, geralmente, representam observações identificadas como atípicas ou verdadeiras "anomalias" e que não são representativas da população em estudo.

Os *outliers* podem distorcer a estrutura final dos grupos derivados do estudo.

Sugestões para identificação de *outliers*

- Quando as variáveis são medidas na mesma escala, construir um gráfico com os perfis das respostas de todos os indivíduos.
- Usar a Distância de Mahalanobis e identificar os indivíduos mais distantes do centro do grupo.

A Análise de Agrupamento visa à obtenção de um esquema que possibilite reunir unidades amostrais (tratamentos, genótipos, objetos, indivíduos, entidades *etc*.) sobre as quais são medidas p variáveis, em um número de grupos escolhido pelo analista, de tal modo que exista grande homogeneidade dentro de cada grupo e uma grande heterogeneidade entre os grupos.

O agrupamento é feito com base em **medidas de similaridades** ou de **distâncias** (dissimilaridades).

Sugestão: Rever os slides sobre Distâncias Multivariadas.

Problema recorrente: Técnicas baseadas em diferentes medidas de similaridades/distâncias nem sempre levam aos mesmos resultados.

Ideia

- Antes de realizar a Análise de Agrupamento podemos diminuir o número de variáveis utilizando a análise fatorial (AF) ou a análise de componentes principais (ACP) e utilizar os **escores** dos primeiros fatores ou componentes para criar os grupos.

Para comprovar que os grupos formados têm características diferentes, podemos usar a MANOVA e comparar os vetores de médias dos grupos formados, esperando que esses vetores não sejam considerados iguais.

DISTÂNCIAS, MEDIDAS DE SIMILARIDADE E DE DISSIMILARIDADE

Problema: Escolher um critério para medir (avaliar) a distância entre dois indivíduos ou para quantificar o quanto eles são parecidos.

Geralmente, os algoritmos utilizados na Análise de Agrupamentos estão baseados em **distâncias** ou **medidas de dissimilaridade**.

Medidas de Dissimilaridade: quanto menor o valor observado dessas medidas, mais parecidos são os indivíduos.

Exemplo: distância euclidiana, distância de Mahalanobis *etc*.

Importante: Quando as p variáveis não são avaliadas na mesma escala de medida ou as suas variâncias são muito diferentes, costuma-se trabalhar com os **dados padronizados.**

O Minitab disponibiliza o uso das distâncias Euclidiana, Manhattan e de Pearson, cujas fórmulas de cálculo são as seguintes:

Distância	Fórmula
Euclidiana	$d(i,k) = \sqrt{\sum_{j=1}^{p}\left(x_{ij} - x_{kj}\right)^2}$
Manhattan	$d(i,k) = \sum_{j=1}^{p}\left\|x_{ij} - x_{kj}\right\|$
Pearson	$d(i,k) = \sqrt{\sum_{j=1}^{p}\left(x_{ij} - x_{kj}\right)^2 / v_j}$

Em que $d(i,k)$ é a distância entre os indivíduos (linhas) i e k; $j = 1, \ldots, p$ são as p colunas (variáveis) da matriz de dados e v_j é a variância da variável (coluna) j.

MÉTODOS DE AGRUPAMENTO DE INDIVÍDUOS

No processo de agrupamento de n indivíduos precisamos calcular $n(n-1)/2$ distâncias, o que torna impraticável um exame visual das suas estimativas para agrupar os indivíduos.

Exemplo: Com $n = 50$ indivíduos, precisaríamos calcular

$$50(49)/2 = 1225$$

distâncias entre pares de objetos!

Os métodos de agrupamento são numerosos e o pesquisador deverá decidir qual é o mais indicado ao seu trabalho, lembrando que o uso de métodos diferentes pode levar a diferentes padrões de agrupamento.

Os métodos de agrupamento podem ser classificados como:

a) **Métodos Hierárquicos**: são utilizados em análises exploratórias com o intuito de identificar possíveis agrupamentos. A história de agrupamento pode ser resumida em um diagrama de árvore, também chamado de dendrograma (ou dendograma).

b) **Métodos de Otimização ou não-hierárquicos**: os grupos são formados pela otimização de um critério de agrupamento. O número de grupos deve ser especificado previamente pelo pesquisador.

MÉTODOS HIERÁRQUICOS DE AGRUPAMENTO

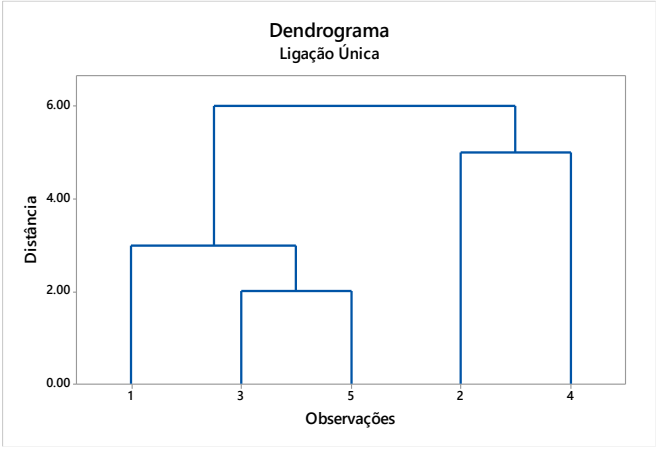
Os indivíduos são classificados nos grupos em diferentes etapas, de modo hierárquico (ordenado), produzindo uma árvore de classificação.

Os métodos hierárquicos podem ainda ser divididos em:

- **Métodos aglomerativos (mais comuns!)**: O processo é iniciado com n agrupamentos. Os dois indivíduos mais similares são fun-didos e formam o primeiro grupo; este processo vai se repetindo, com os grupos iniciais sendo fundidos com outros de acordo com suas similaridades. Eventualmente, com a diminuição das similaridades, todos os subgrupos são fundidos em um único agrupamento.

- **Métodos divisivos**: Partem de um único grupo e por divisões sucessivas vão sendo divididos (separados) em 2, 3, … subgrupos, de tal modo que os indivíduos em um subgrupo estejam longe dos indivíduos do outro subgrupo. Esses indivíduos são novamente separados em outros subgrupos e o processo continua até que cada objeto forme um grupo.

Os resultados dos dois métodos podem ser mostrados na forma de um diagrama bidimensional conhecido como **dendrograma** ou **diagrama de árvore** que ilustra as fusões (ou divisões) que são feitas em cada um dos níveis sucessivos do processo.



Em um dos seus eixos indicamos o nível de dissimilaridade (distância) utilizado e no outro, os indivíduos aparecem em uma ordem relacionada à história do agrupamento.

ALGUNS MÉTODOS HIERÁRQUICOS AGLOMERATIVOS (MÉTODOS DE LIGAÇÃO)

a) MÉTODO DA LIGAÇÃO ÚNICA (*single linkage*) OU DO VIZINHO MAIS PRÓXIMO

Os grupos serão formados a partir dos indivíduos, fundindo os vizinhos mais próximos, ou seja, aqueles que têm a **menor distância**. O algoritmo do método resume-se em:

1) Calcular a matriz (simétrica) $\mathbf{D} = \{d_{ik}\}$, $n \times n$, de distâncias (ou similaridades) entre os pares de indivíduos.

2) Encontrar a menor distância na matriz **D** e ligar os objetos correspondentes, digamos, U e V, para formar o grupo (UV).

3) Atualizar as entradas na matriz **D**, apagando as linhas e colunas correspondentes aos objetos U e V e adicionando uma nova linha e uma nova coluna para indicar as distâncias entre o grupo (UV) e os grupos/objetos restantes. A distância entre o grupo (UV) e outro grupo W é calculada como:

$$d_{(\mathrm{UV})\mathrm{W}} = min\{d_{\mathrm{UW}}, d_{\mathrm{VW}}\} \qquad (9.13)$$

onde d_{UW} e d_{VW} são as distâncias entre os vizinhos mais próximos dos grupos (U e W) e (V e W), respectivamente.

4) Repetir os passos 2 e 3 até que os n objetos formem um único grupo.

Exemplo 1. Para ilustrar o **algoritmo de ligação única** vamos considerar as distâncias euclidianas hipotéticas entre os pares de cinco objetos apresentadas a seguir e realizar o processo de agrupamento, passo-a-passo.

Com as distâncias guardadas na matriz **D** apresentada a seguir, iniciamos o processo identificando a menor distância.

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	**2**	8	0

1) Juntar os itens 3 e 5, porque $min(d_{ik}) = d_{35} = 2$.

2) Calcular as distâncias entre o grupo (35) e os demais objetos (1, 2 e 4):

$$d_{(35)1} = min\{d_{31}, d_{51}\} = min\{3, 11\} = 3$$

$$d_{(35)2} = min\{d_{32}, d_{52}\} = min\{7, 10\} = 7$$

$$d_{(35)4} = min\{d_{34}, d_{54}\} = min\{9,\ 8\} = 8$$

⇒ A nova matriz **D** fica: $\mathbf{D} =$

$$\begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

3) Juntar os grupos (35) e 1 porque $min(d_{ik}) = d_{(35)1} = 3$.

4) Calcular as distâncias entre o grupo (135) e os demais objetos 2 e 4:

$$d_{(1,35)2} = min\{d_{(1)2}, d_{(35)2}\} = min\{9, 7\} = 7$$

$$d_{(1,35)4} = min\{d_{(1)4}, d_{(35)4}\} = min\{6, 8\} = 6$$

⇒ A nova matriz **D** fica: $\mathbf{D} = \begin{matrix} \\ (135) \\ 2 \\ 4 \end{matrix} \begin{matrix} (135) & 2 & 4 \\ \end{matrix} \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & 5 & 0 \end{bmatrix}$

5) Juntar os grupos 2 e 4 porque $min(d_{ik}) = d_{24} = 5$. Neste ponto nós já os dois grupos (135) e (24) e a menor distância entre vizinhos é:

$$d_{(135)(24)} = min\{d_{(135)2}, d_{(135)4}\} = min\{7, 6\} = 6$$

A matriz de distâncias final é igual a: $\mathbf{D} = \begin{matrix} (135) \\ (24) \end{matrix} \begin{bmatrix} 0 & \\ 6 & 0 \end{bmatrix}$

6) Finalmente juntamos os grupos (135) e (24) em um único grupo.

Para realizar esta análise no MINITAB precisamos, primeiramente, carregar o arquivo **Cluster – Ex1**, com as distâncias e juntar as informações numa matriz, usando:

MTB > copy c1-c5 M1

Depois escolher **Stat > Multivariado > Agrupamento de observações > Matriz de distância ou variáveis** escolher a matriz **M1**; em **Método de ligação** escolher **Único**; em **Medida de distâncias** escolher **Euclidiano**, marcar **Exibir dendrograma** e em **Personalizar** escolher a opção **Distância**.

A sequência dos agrupamentos aparece em **Agrupados reunidos** e o nome dos novos grupos aparece em **Novo agrupado.** O tamanho dos agrupamentos aparece na última coluna.

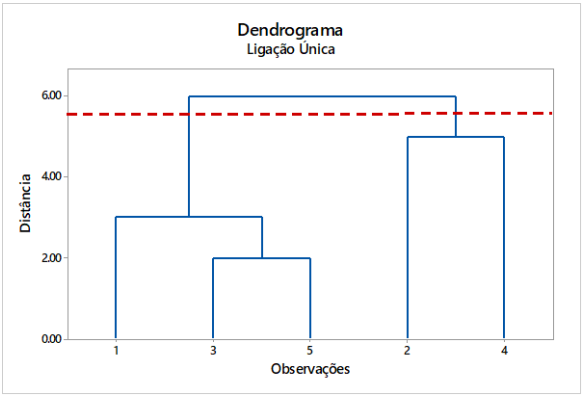
Análise de Agrupamentos de Observações: M1

Ligação Única

Passos de Amalgamação

Passo	Número de agrupados	Nível de similaridade	Nível de distância	Agrupados reunidos		Novo agrupado	Número de obs. no novo agrupado
1	4	81.8182	2	3	5	3	2
2	3	72.7273	3	1	3	1	3
3	2	54.5455	5	2	4	2	2
4	1	45.4545	6	1	2	1	5

Passo 1: agrupa (3) com (5) → (**3**) [35]
Passo 2: agrupa (1) com (3) → (**1**) [1,35]
Passo 3: agrupa (2) com (4) → (**2**) [24]
Passo 4: agrupa (1) com (2) criando um único grupo (**1**).

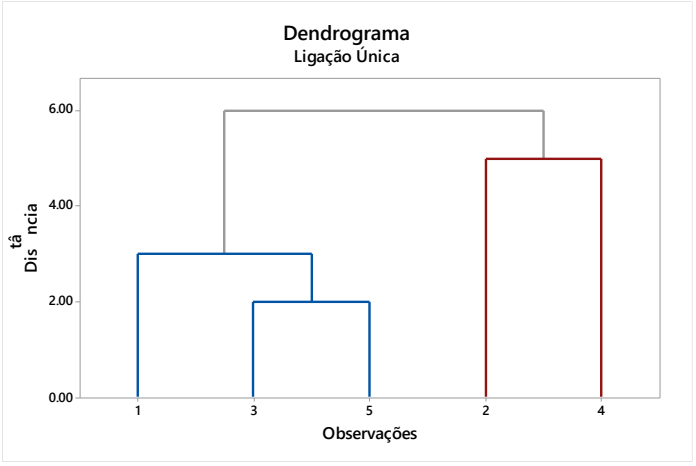


O dendrograma mostra a sequência de agrupamentos das cinco observações.

A linha de corte vermelha foi traçada depois de escolhermos 2 agrupamentos.

Repetindo o processo de análise, em **Número de agrupamentos** digitamos 2 e em **Armazenamento > Coluna de indicadores de agrupados** digitamos **Grupo**.

O dendrograma indicará a formação dos dois grupos com cores diferentes.



A coluna **Grupo** da planilha do Minitab indica o grupo ao qual cada observação foi alocada pelo método do vizinho mais próximo.

Cluster - Ex1.MTW ***

↓	C1	C2	C3	C4	C5	C6
	X1	X2	X3	X4	X5	Grupo
1	0	9	3	6	11	1
2	9	0	7	5	10	2
3	3	7	0	9	2	1
4	6	5	9	0	8	2
5	11	10	2	8	0	1

Perceba que o processo de agrupamento foi realizado sem conhecer os objetos/indivíduos!

Só precisamos da matriz de distâncias entre eles...

B) MÉTODO DA LIGAÇÃO COMPLETA (*complete linkage*) OU DO VIZINHO MAIS DISTANTE

Em cada estágio, a distância entre grupos é determinada pela **maior distância** entre dois elementos, um de cada grupo. Isto assegura que todos os itens em um grupo estão dentro de alguma distância máxima (ou similaridade mínima) de cada outro.

O algoritmo aglomerativo geral inicia procurando a menor distância em $\mathbf{D} = \{d_{ik}\}$ e juntando os objetos correspondentes, tais como U e V, num grupo (UV). No passo seguinte, as distâncias entre (UV) e qualquer outro grupo W são calculadas por:

$$d_{(\mathrm{UV})\mathrm{W}} = max\{d_{\mathrm{UW}}, d_{\mathrm{VW}}\} \qquad (9.14)$$

onde d_{UW} e d_{VW} são as distâncias entre os vizinhos mais distantes dos grupos (U e W) e (V e W), respectivamente.

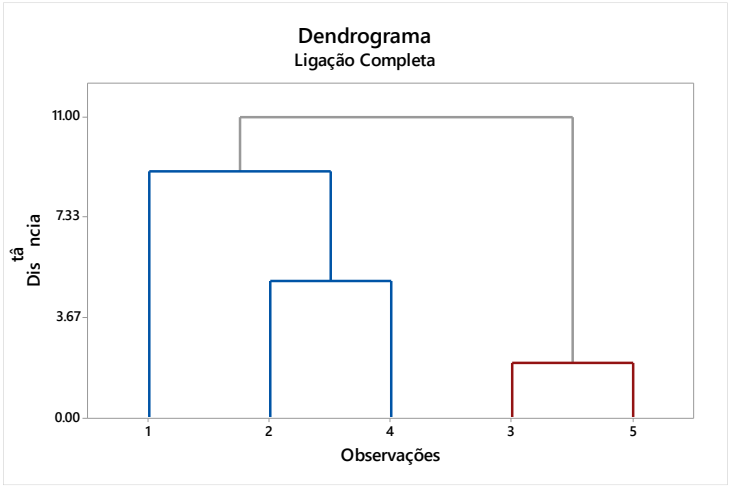
No Exemplo 1, escolhendo em **Método de ligação** a opção **Completa**, tem-se os seguintes passos de agrupamento:

Análise de Agrupamentos de Observações: M1

Ligação Completa

Passos de Amalgamação

Passo	Número de agrupados	Nível de similaridade	Nível de distância	Agrupados reunidos		Novo agrupado	Número de obs. no novo agrupado
1	4	81.8182	2	3	5	3	2
2	3	54.5455	5	2	4	2	2
3	2	18.1818	9	1	2	1	3
4	1	0.0000	11	1	3	1	5

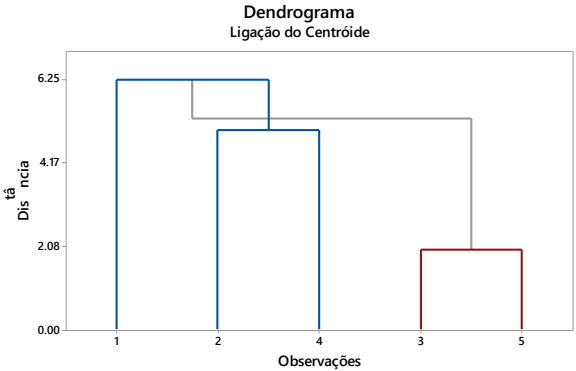


O dendrograma representa o processo de agrupamento das observações.

C) MÉTODO DO CENTRÓIDE

Cada grupo é substituído pelo valor médio das medidas dos indivíduos que pertencem ao grupo, que é chamado **centróide** do grupo.

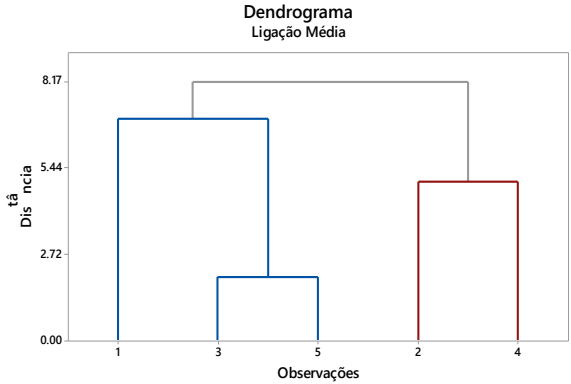
Em cada passo, calculamos a matriz de distâncias entre os centróides e juntamos os indivíduos que apresentam a menor distância com os respectivos centróides.



D) MÉTODO DA LIGAÇÃO MÉDIA (*average linkage*)

Trata as distâncias entre dois grupos como a **média das distâncias de todos os pares de itens** onde cada membro de um par pertence a um grupo.

Iniciamos o processo procurando na matriz de distâncias $\mathbf{D} = \{d_{ik}\}$ os objetos mais próximos, por exemplo, os objetos U e V. Esses objetos são juntados no grupo (UV).

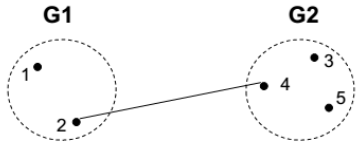


No próximo passo, as distâncias entre (UV) e outro grupo W são calculadas por:

$$d_{(\mathrm{UV})\mathrm{W}} = \frac{\sum_i \sum_k d_{ik}}{n_{(UV)} n_W} \tag{9.15}$$

onde d_{ik} é a distância entre o objeto i no grupo (UV) e o objeto k no grupo W, e $n_{(UV)}$ e n_W correspondem aos números de objetos existentes nos grupos (UV) e W, respectivamente.

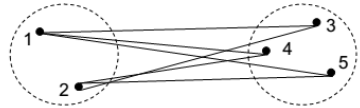
A Figura 1 apresenta um resumo de como são calculadas as distâncias nos principais métodos de agrupamento hierárquico.



Distância entre Grupos

Ligação Simples

$$d(G_1, G_2) = d_{24}$$



Ligação Média

$$d(G_1, G_2) = \frac{d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25}}{6}$$

2x3

Ligação Completa

$$d(G_1, G_2) = d_{15}$$

Figura 1. Cálculo de distância em alguns métodos hierárquicos de agrupamento

E) MÉTODO DE WARD

O método de Ward não calcula distâncias entre grupos, mas forma grupos maximizando a homogeneidade dentro deles ou minimizando o total das somas de quadrados dentro de grupos, que também é conhecida como soma de quadrados de erros – SQE (inglês: ESS $error\ sums\ squares$).

Inicialmente, cada grupo é formado por um único item ($SQE = 0$). Em cada passo do algoritmo, são formados grupos de tal modo que a solução resultante tenha a menor SQE.

Em cada passo são consideradas as uniões de todos os possíveis pares de grupos. Os dois grupos cuja combinação resulta em um menor aumento de SQE (mínima perda de informação) são juntados.

Na outra extremidade, quando todos os grupos forem combinados em um único grupo de n itens, o valor de SQE é calculado por

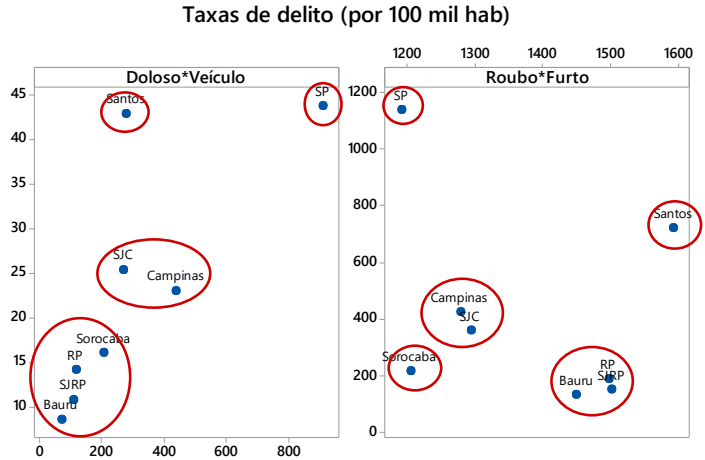
$$SQE = \sum_{j=1}^{n} (\boldsymbol{x}_j - \overline{\boldsymbol{x}})' (\boldsymbol{x}_j - \overline{\boldsymbol{x}})$$

onde $\boldsymbol{x}_j$ é o vetor multivariado de medidas associado com o j-ésimo item e $\overline{\boldsymbol{x}}$ é a média de todos os itens.

Os resultados do método de Ward podem ser apresentados em um dendrograma, indicando os valores de SQE em um dos eixos.

Exemplo 2. Agrupar as delegacias seccionais do Departamento de Polícia Judiciária de São Paulo Interior (Deinter) com base nas taxas de delitos (por 100 mil habitantes) em 2002, utilizando o método do vizinho mais próximo.

DEINTER	Homicídio doloso	Furto	Roubo	Roubo e furto de veículos
SJRP	10.85	1500.80	149.35	108.38
RP	14.13	1496.07	187.99	116.66
Bauru	8.62	1448.79	130.97	69.98
Campinas	23.04	1277.33	424.87	435.75
Sorocaba	16.04	1204.02	214.36	207.06
SP	43.74	1190.94	1139.52	909.21
SJC	25.39	1292.91	358.39	268.24
Santos	42.86	1590.66	721.90	275.89



Dúvidas: Quantos agrupamentos são possíveis? Quais os integrantes de cada grupo?

Carregar a planilha **Cluster - Delitos**. Escolher **Stat > Multivariado > Agrupamento de observações**; em **Matriz de distância ou variáveis** escolher as variáveis **Doloso, Furto, Roubo e Veículo**; em **Método de ligação** escolher **Único**; marcar **Exibir dendrograma**; em **Personalizar > Rótulo de casos**, escolher a coluna **C1: DEINTER** e em **Rótulo do eixo y** escolher **Distância.** Resultando em:

Distância Euclideana, Ligação Única

Passos de Amalgamação

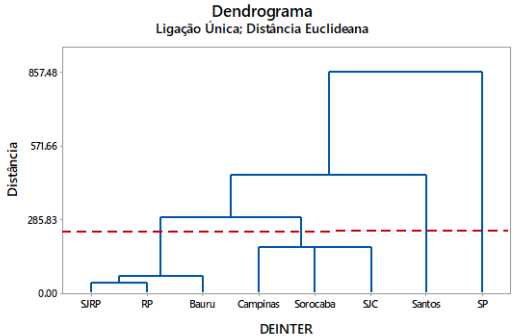
Passo	Número de agrupados	Nível de similaridade	Nível de distância	Agrupados reunidos		Novo agrupado	Número de obs. no novo agrupado
1	7	97.01	39.93	1	2	1	2
2	6	94.97	67.25	1	3	1	3
3	5	86.53	180.21	5	7	5	2
4	4	86.48	180.91	4	5	4	3
5	3	78.11	292.77	1	4	1	6
6	2	65.55	460.82	1	8	1	7
7	1	35.89	857.48	1	6	1	8

Partição Final

	Número de observações	Dentro da soma de quadrados do agrupado	Distância média do centróide	Distância máxima do centróide
Agrupado1	8	1551609	366.725	964.602

Avaliando o dendrograma ao lado, optamos por criar quatro agrupamentos:

- {SJRP, RP, Bauru},
- {Campinas, Sorocaba, SJC}
- {Santos}
- {SP}



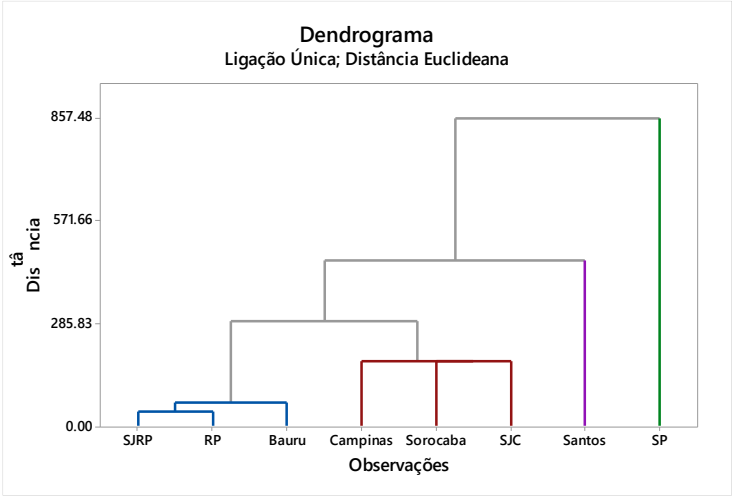


Figura 2. Dendrograma dos itens do Exemplo 2 (DEINTER) utilizando o método da ligação único (vizinho mais próximo).

Para criar uma coluna com os componentes de cada agrupamento, repetimos a análise, em **Número de agrupamentos** indicamos **4** e em **Armazenamento > Coluna de indicadores agrupados** indicamos a coluna **Grupo**.

Um gráfico de dispersão dos dados de **Furto** e **Doloso** serve para visualizar os agrupamentos obtidos.

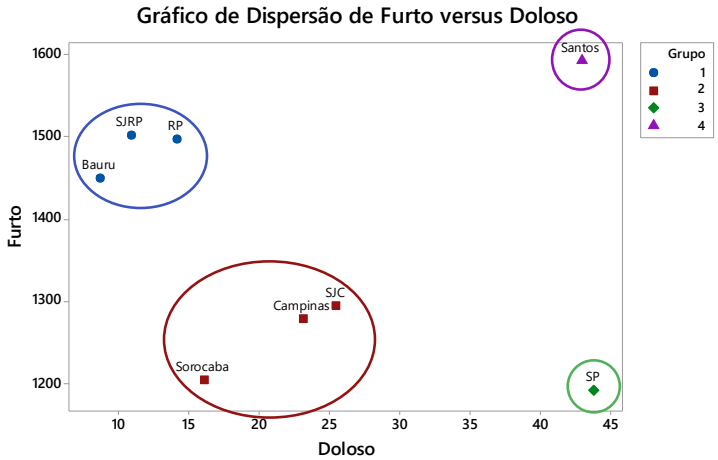


Figura 3. Dispersão do número de furtos e de crimes dolosos com indicação do agrupamento (DEINTER).

ALGUMAS CARACTERÍSTICAS DOS MÉTODOS HIERÁRQUICOS

- **Vantagem**: Não requerem o conhecimento *a priori* do número de grupos ou da partição inicial.
- **Desvantagem**: uma vez que um item foi designado a um grupo, ele não pode ser realocado em outro grupo.
- O método de ligação completa (*complete linkage*) é menos afetado pela presença de *outliers* do que o método de ligação única (*single linkage*),
- A técnica do vizinho mais distante (*complete linkage*) tende a identificar grupos mais compactos, com itens bastante similares.
- O método de Ward tende a encontrar grupos mais compactos e de tamanhos semelhantes.

- Os métodos hierárquicos são usados muitas vezes de forma exploratória e a solução resultante é submetida a um método não hierárquico para refinar ainda mais a solução.
- Os métodos hierárquicos e não hierárquicos podem ser vistos como complementares ao invés de competidores.

COMO ESCOLHER O NÚMERO DE GRUPOS? COMO AVALIAR A QUALIDADE DOS AGRUPAMENTOS

Após obtermos um agrupamento, precisamos avaliar a solução e escolher o número de grupos presentes nos dados.

Existem algumas estatísticas que podem auxiliar nessa avaliação, mas elas não são calculadas pelo Minitab.

- **RMSSTD (*Root Mean Square Standard Deviation*):** é o desvio padrão ponderado de todas as variáveis que formam cada um dos grupos
- **RS (*R squared*).** É calculado dividindo-se a soma de quadrados *entre* grupos (SQB) pela soma de quadrados total corrigida pela média (SQT). Valores próximos de zero indicam pouca diferença entre grupos e valores próximos de um, diferenças máximas entre grupos.
- **Pseudo F e T^2.** Testa a igualdade dos vetores de médias dos agrupamentos em cada passo do algoritmo. Maiores valores dessas estatísticas indicam maior heterogeneidade entre os grupos.
- ***SPR* (*Semipartial R squared*).** Avalia a perda de homogeneidade devida à combinação de dois grupos para formar um novo.

- **Distância entre grupos.** O cálculo dessa distância depende do método de agrupamento utilizado. Por exemplo: no método centróide, a distância corresponde à distância euclidiana entre os centróides dos dois grupos que são juntados. Um valor alto da distância entre dois grupos indica que eles são muito distintos.

Importante:

i) Devemos olhar para um grande salto no valor da estatística usada. Se os valores da estatística forem colocados num gráfico, deveremos procurar por um *cotovelo*.

ii) Devemos avaliar também o dendrograma e os itens que compõem cada um dos grupos.

Exemplo 3. Dados relativos a 21 países fornecido pela ONU (2002), disponível no site www.undp.org/hdro. As variáveis são os índices de Expectativa de vida (ExpecVida), Educação (Educ), Renda (PIB) e Estabilidade política e de segurança (EstabSeg). Os dados estão disponíveis na planilha **Cluster – Mingoti (ONU).**

Para agrupar os países em 4 *clusters* usando o **método de Ward** escolhemos **Stat > Multivariado > Agrupamento de observações**; em **Matriz de distância ou variáveis** escolher as quatro variáveis; em **Método de ligação** escolher **Ward**; em **Número de agrupamentos** digitar **4**; marcar **Exibir dendrograma**; clicar em **Personalizar**, em **Rótulos de casos** escolher **País** e marcar **Distância**; em **Armazenamento > Coluna de indicadores de agrupados** escrever **Grupo**.

Os passos de amalgamação podem ser visualizados no dendrograma, em que já estão indicados os quatro grupos de países.

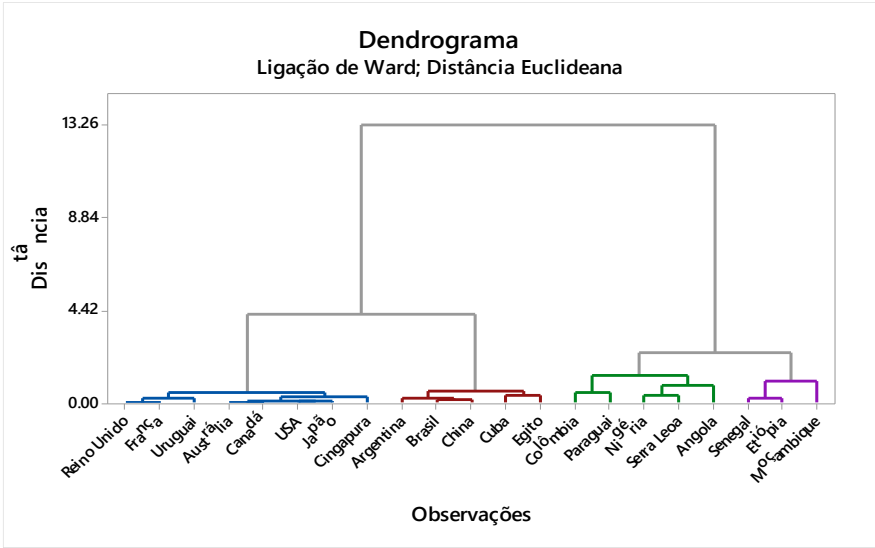


Figura 4. Dendrograma dos dados do Exemplo 3 – Ligação de Ward

Partição Final

	Número de observações	Dentro da soma de quadrados do agrupado	Distância média do centróide	Distância máxima do centróide
Agrupado1	8	0.15692	0.123194	0.240124
Agrupado2	5	0.25524	0.211137	0.295689
Agrupado3	5	1.24016	0.461494	0.676541
Agrupado4	3	0.48820	0.379292	0.552751

(1) Na partição final percebe-se que o Agrupado 1 é mais homogêneo que os demais, pois apresenta a menor SQ do Agrupado.

Centróides do grupo

Variável	Agrupado1	Agrupado2	Agrupado3	Agrupado4	Centróide global
ExpecVida	0.884	0.766	0.506	0.340	0.688
Educ	0.954	0.814	0.590	0.363	0.750
PIB	0.907	0.674	0.494	0.377	0.678
EstabSeg	1.185	0.338	-1.366	-0.343	0.158

(2) Percebem-se diferenças numéricas entre os componentes dos centróides dos Agrupados.

Distâncias Entre Centróides do Grupo

	Agrupado1	Agrupado2	Agrupado3	Agrupado4
Agrupado1	0.00000	0.89740	2.63697	1.80606
Agrupado2	0.89740	0.00000	1.74751	0.96809
Agrupado3	2.63697	1.74751	0.00000	1.06703
Agrupado4	1.80606	0.96809	1.06703	0.00000

3) Note que os Grupos 1 e 3 estão mais **distantes** e os Grupos 1 e 2, mais **próximos**

Vamos usar a MANOVA para verificar se os centróides dos quatro grupos são iguais:

$$H_0: \begin{bmatrix}\mu_{11}\\ \mu_{21}\\ \mu_{31}\\ \mu_{41}\end{bmatrix} = \begin{bmatrix}\mu_{12}\\ \mu_{22}\\ \mu_{32}\\ \mu_{42}\end{bmatrix} = \begin{bmatrix}\mu_{13}\\ \mu_{23}\\ \mu_{33}\\ \mu_{43}\end{bmatrix} = \begin{bmatrix}\mu_{14}\\ \mu_{24}\\ \mu_{34}\\ \mu_{44}\end{bmatrix}$$

Em que μ_{ij} é a média da i-ésima variável no grupo j.

País	ExpecVida	Educ	PIB	EstabSeg	Grupo
Austrália	0.90	0.99	0.93	1.26	1
Canadá	0.90	0.98	0.94	1.24	1
Cingapura	0.88	0.87	0.91	1.41	1
França	0.89	0.97	0.92	1.04	1
Japão	0.93	0.93	0.93	1.20	1
Reino Unido	0.88	0.99	0.91	1.10	1
Uruguai	0.82	0.92	0.75	1.05	1
USA	0.87	0.98	0.97	1.18	1
Argentina	0.81	0.92	0.80	0.55	2
Brasil	0.71	0.83	0.72	0.47	2
China	0.76	0.80	0.61	0.39	2
Cuba	0.85	0.90	0.64	0.07	2
Egito	0.70	0.62	0.60	0.21	2
Angola	0.34	0.36	0.51	-1.98	3
Colômbia	0.77	0.85	0.69	-1.36	3
Nigéria	0.44	0.58	0.37	-1.36	3
Paraguai	0.75	0.83	0.63	-0.87	3
Serra Leoa	0.23	0.33	0.27	-1.26	3
Etiópia	0.31	0.35	0.32	-0.55	4
Moçambique	0.24	0.37	0.36	0.20	4
Senegal	0.47	0.37	0.45	-0.68	4

Em **Estat > ANOVA > MANOVA generalizada** em **Respostas** escolhemos as 4 variáveis e em **Modelo** incluímos **Grupo.**

Testes de MANOVA para Grupo

Critério	Estatística de teste	F aproximadamente	GL Núm.	GL Denom.	P
Wilks	0.01322	12.846	12	37	0.000
Lawley-Hotelling	23.15694	24.443	12	38	0.000
Pillai	1.82883	6.246	12	48	0.000
Roy	21.34484				

s = 3 m = 0.0 n = 6.0

Como valor-p $\cong 0$ rejeitamos H_0 e concluímos que os centróides dos quatro grupos criados não são iguais entre si em todas as características, ou seja, foram criados grupos de países que têm características diferentes.

Podemos continuar a análise rodando ANOVA´s individuais para cada variável, buscando identificar as possíveis diferenças entre os grupos.

9.4. MÉTODOS NÃO HIERÁRQUICOS DE AGRUPAMENTO

- Os grupos são formados iterativamente pela otimização de um critério de agrupamento.
- Requerem o conhecimento *a priori* do número de grupos: os centróides ou a partição inicial têm que ser identificados antes do uso dessa técnica de agrupamento.
- Os métodos não hierárquicos são mais sensíveis à partição inicial. O processo pode chegar a soluções diferentes se for iniciado com partições diferentes.
- Esses métodos têm um baixo desempenho quando usamos partições iniciais escolhidas sem qualquer critério.

O desempenho de um método não hierárquico é superior quando utiliza os resultados de um método hierárquico para formar a partição inicial.

- A existência de um *outlier* multivariado pode produzir, no mínimo, um grupo com itens muito dispersos.

A escolha por uma técnica de agrupamento hierárquico ou não hierárquico depende do objetivo do estudo e das propriedades já conhecidas dos vários algoritmos de agrupamento.

MÉTODO DAS K-MÉDIAS ($k-means\ method$)

É um método iterativo que tem como função de classificação a distância do objeto ao centro de cada grupo (centróide).

O método k-médias minimiza a soma de todas as distâncias euclidianas entre os objetos e os centróides.

O algoritmo de agrupamento consiste nas seguintes etapas:

1) Com base na partição inicial em k grupos calculamos seus k centróides.
2) Calcular a distância de cada objeto aos k centróides e movê-lo para o grupo mais próximo.
3) Se ocorreu alguma inclusão/exclusão de objetos em algum grupo, recalcular os centróides.
4) Repetir as etapas 2 e 3 até que os valores dos centróides não mudem em duas iterações sucessivas.

Vamos aplicar o método não hierárquico k-médias para analisar os dados do Exemplo 3, usando os agrupamentos já formados pelo método de Ward como partição inicial.

Escolher **Estat > Multivariada > Agrupamento de k-Médias** e em **Variáveis** escolher as 4 variáveis; em **Número de agrupamentos** digitar **4**; em **Coluna da partição inicial** escolher a coluna **Grupo** e em **Armazenamento > Coluna de indicadores de agrupados** digitar **KMedias**.

Análise de Agrupamentos de K-médias: ExpecVida; Educ; ... ; EstabSeg

Método

Número de agrupados	4
Variáveis padronizadas	Não

Partição Final

	Número de observações	Dentro da soma de quadrados do agrupado	Distância média do centróide	Distância máxima do centróide
Agrupado1	8	0.157	0.123	0.240
Agrupado2	5	0.255	0.211	0.296
Agrupado3	5	1.240	0.461	0.677
Agrupado4	3	0.488	0.379	0.553

(1) Mostra o número de países de cada grupo, a SQ dentro do grupo que mede a sua variabilidade *etc*.

Centróides do grupo

Variável	Agrupado1	Agrupado2	Agrupado3	Agrupado4	Centróide global
ExpecVida	0.8838	0.7660	0.5060	0.3400	0.6881
Educ	0.9537	0.8140	0.5900	0.3633	0.7495
PIB	0.9075	0.6740	0.4940	0.3767	0.6776
EstabSeg	1.1850	0.3380	-1.3660	-0.3433	0.1576

(2) Mostra os centróides de cada grupo e o centróide global.

Distâncias Entre Centróides do Grupo

	Agrupado1	Agrupado2	Agrupado3	Agrupado4
Agrupado1	0.0000	0.8974	2.6370	1.8061
Agrupado2	0.8974	0.0000	1.7475	0.9681
Agrupado3	2.6370	1.7475	0.0000	1.0670
Agrupado4	1.8061	0.9681	1.0670	0.0000

(3) Mostra as distâncias entre os centróides dos 4 grupos: os mais distantes são os grupos 1 e 3.

Com o comando:

MTB > table Grupo KMedias

Você poderá verificar que o método K-Médias não alterou a classificação inicial indicada pelo método hierárquico de Ward: todos os países dos grupos iniciais permaneceram nos grupos finais.

Linhas: Grupo Colunas: KMedias

	1	2	3	4	Todos
1	8	0	0	0	8
2	0	5	0	0	5
3	0	0	5	0	5
4	0	0	0	3	3
Todos	8	5	5	3	21

Conteúdo da Célula
Contagem

Se você usar uma partição inicial resultante de outro método hierárquico (vizinho mais próximo, por exemplo), irão ocorrer alterações nos grupos finais. (Sugestão: Tente iniciar o processo com outra solução inicial)

AGRUPAMENTO DE VARIÁVEIS

Objetivo: Juntar variáveis em agrupamentos que compartilham características comuns.

As variáveis são agrupadas com aquelas que são semelhantes (correlacionadas entre si), indicando a presença de construtos.

Para juntar as variáveis o Minitab calcula medidas de similaridade e de dissimilaridade (distância). Essas medidas podem ser resumidas em um dendograma.

Vamos aproveitar os dados do Exemplo 3, disponíveis na planilha **Cluster – Mingoti (ONU)**, para realizar o agrupamento dos índices de Expectativa de vida (ExpecVida), Educação (Educ), Renda (PIB) e Estabilidade política e de segurança (EstabSeg).

Em **Stat > Multivariado > Agrupar variáveis > Matriz de distância ou variáveis** escolher as 4 variáveis **ExpecVida-EstabSeg**; em **Método de ligação** escolher **Único** e em **Medida de distância** escolher **Correlação**; marcar **Exibir dendograma** e em **Personalizar** escolher **Similaridade**.

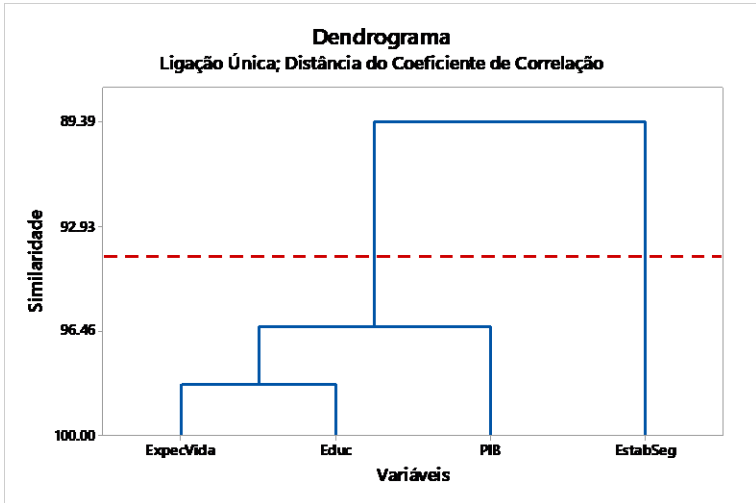
O quadro seguinte mostra os passos para realizar os agrupamentos dos índices.

Distância do Coeficiente de Correlação, Ligação Única
Passos de Amalgamação

Passo	Número de agrupados	Nível de similaridade	Nível de distância	Agrupados reunidos		Novo agrupado	Número de obs. no novo agrupado
1	3	98.2785	0.034431	1	2	1	2
2	2	96.2970	0.074060	1	3	1	3
3	1	89.3885	0.212230	1	4	1	4

O dendograma indica a formação de dois agrupamentos: um com os índices ExpecVida, Educ e PIB e outro com o índice EstabSeg.

Para confirmar esta indicação, podemos usar a Análise de itens para comparar a consistência interna dos dados com os quatro índices e sem a coluna EstabSeg.


Dendrograma
Ligação Única; Distância do Coeficiente de Correlação
Similaridade
89.39
92.93
96.46
100.00
ExpecVida
Educ
PIB
EstabSeg
Variáveis

PROBLEMA: Como realizar a análise de agrupamentos quando as características medidas nos itens são dicotômicas?

Os itens são muitas vezes comparados com base na **presença** ou **ausência** de certas características.

A **presença** ou **ausência** de uma característica pode ser descrita matematicamente introduzindo uma **variável binária**, que assume o valor 1 se a característica está presente e valor 0 se a característica está ausente.

Por exemplo, para $p = 5$ variáveis binárias, as respostas para dois itens i e k podem ser arranjados da seguinte maneira:

Item	X_1	X_2	X_3	X_4	X_5
i	1	0	0	1	1
k	1	1	0	1	0

Quadro com o resumo (empates e desempates) para os itens i e k

		Item k		
		1	0	Totais
Item i	1	a	b	$a+b$
	0	c	d	$c+d$
Totais		$a+c$	$b+d$	$p=a+b+c+d$

O quadrado da distância euclidiana não é muito utilizado nestas situações, porque pondera os empates 1-1 e 0-0 igualmente.

Às vezes o empate 1-1 é uma indicação mais forte de similaridade do que o empate 0-0. Por exemplo, em grupos de pessoas, a evidência que duas pessoas leem livros de pensadores gregos é uma evidência mais forte de similaridade do que a ausência desta característica.

Tabela 1. Alguns coeficientes de similaridade para agrupamentos de itens utilizados em melhoramento genético.

Coeficiente	Lógica
$\frac{a+d}{p}$	Pesos iguais para os empates 1-1 e 0-0
$\frac{2(a+d)}{2(a+d)+(b+c)}$	Peso duplo para os empates 1-1 e 0-0
$\frac{a+d}{a+d+2(b+c)}$	Peso duplo para os desempates
a/p	Nenhum empate 0-0 no numerador
$\frac{a}{a+b+c}$	Os empates 0-0 são tratados como irrelevantes.

Coeficiente	Lógica
$\frac{2a}{2a+b+c}$	Nenhum empate 0-0 no numerador ou denominador. Peso duplo para os empates 1-1
$\frac{a}{a+2(b+c)}$	Nenhum empate 0-0 no numerador ou denominador. Peso duplo para os desempates.
$\frac{a}{b+c}$	Razão do número de empates pelo número de desempates, com os empates 0-0 excluídos.

Infelizmente, o Minitab não disponibiliza o cálculo desses coeficientes de similaridade para agrupamentos quando as características medidas nos itens são dicotômicas

Bibliografia consultada

BUSSAB, W.O.; MIAZAK, E.S.; ANDRADE, D.F. **Introdução à Análise de Agrupamentos.** 9º Simpósio Brasileiro de Probabilidade e Estatística. São Paulo: IME – USP, 1990.

CORRAR, J. L.; PAULO, E.; DIAS FILHO, J. M. **Análise Multivariada: para cursos de Administração, Ciências Contábeis e Economia**. FIPECAFI, 1ª ed. – 5ª re-impr. – São Paulo: Atlas, 2014.

HAIR, J. F. JR.; ANDERSON, R.E.; TATHAM, R.L; BLACK, W.C. **Análise Multivariada de Dados**; tradução Adonai Schlup Sant'Anna. 6 ed. Porto Alegre: Bookman, 2009.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. Prentice Hall, Englewood Cliffs, 1998.

LATTIN, J.M.; CARROLL, J.D.; GREEN, P.E. **Análise de dados multivariados** - São Paulo: CENGAGE Learning, 2011.

MANLY, B.F. **Métodos estatísticos multivariados: uma introdução.** Tradução de Sara Ianda Correa Carmona. Bookman. Porto Alegre. 2008.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: Editora UFMG, 2005.

No Minitab:

https://support.minitab.com/pt-br/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-observations/before-you-start/overview/