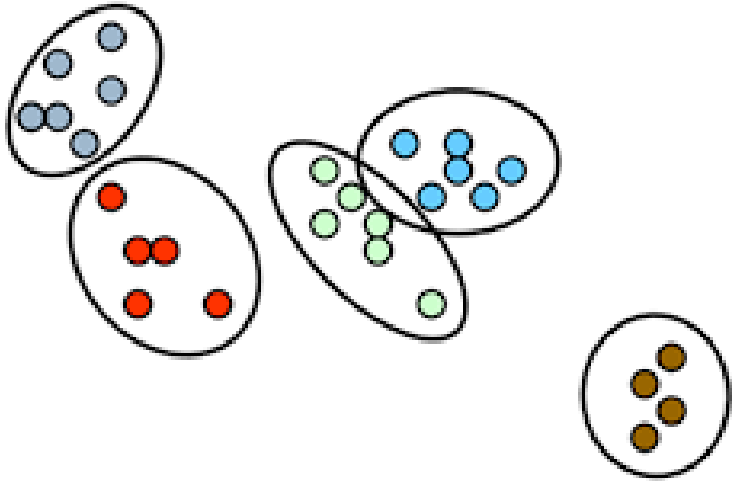
# Curso de Estatística Multivariada
# Método k means de agrupamento



Brasília - DF
Dezembro de 2018

# Sumário

# Lista I - Método K-means

## Técnicas de agrupamentos não hierárquicas

Os métodos não hierárquicos são métodos que tem em como objetivo encontrar diretamente uma partição de n elementos em k grupos (clusters), de modo que a partição satisfaça dois requisitos básicos: coesão interna (ou semelhança interna) e isolamento (ou separação) dos clusters formados. Para se buscar a melhor partição de ordem k, algum critério de qualidade de partição deve ser empregado. È impossível, computacionalmente, criar todas as partições possíveis de ordem k. Desse modo, são necessários passos que investiguem algumas das partições possíveis com o objetivo de encontrar uma quase ótima.

Os métodos não hierárquicos divergem dos hierárquicos em vários aspectos. Primeiramente, requerem que o usuário tenha especificado previamente o número de clusters k desejado, ao contrário das técnicas aglomerativas. Em cada estágio do agrupamento, os novos grupos podem ser formados através da divisão ou junção de grupos já combinados em passos anteriores. Isso significa que, se em algum passo do algoritmo dois elementos estiverem sido colocados num mesmo conglomerado, não necessariamente estarão juntos na partição final. Como consequência, não é mais possível a construção de dendrogramas. Geralmente, os algoritmos computacionais utilizados nos métodos não hierárquicos são do tipo interativo e, em comparação com os métodos hierárquicos, têm uma maior capacidade de análise de conjuntos de dados de maior porte, ou seja, com um grande número de observações.

## Introdução

A análise de grupamentos, também conhecida como análise de conglomerados, classificação ou cluster, tem como objetivo dividir os elementos da amostra, ou população em grupos de forma que os elementos pertencentes a um mesmo grupo sejam similares entre si com respeito ás variáveis (características) que neles foram medidas, e os elementos em grupos diferentes sejam heterogêneos em relação a estas mesmas características. (MINGOTI,2007 pág. 155)

Em outras palavras, a analise de conglomerados (*cluster analysis*) em uma amostra ou população é uma técnica exploratória de dados que visa formar grupos os mais distintos possíveis, tal que os elementos que os constituem sejam semelhantes, segundo as características observadas.

A definição acima, apresenta dois problemas a serem enfrentados: como medir a semelhança entre os indivíduos analisados; e o número de grupos a serem formados.

O grau de semelhança ou diferença entre dois indivíduos são determinados por medidas de similaridade ou de dissimilaridade, respectivamente. Já o número de grupos a serem formados podes ser definido *à priori* pelo pesquisador ou com auxílio de medidas geradas no decorrer da analise.

Este texto procura apresentar de forma introdutória e prática os conceitos de medidas de similaridade e dissimilaridade, critérios para determinar o número de clusters e métodos de formação de cluster.

## Medidas de Similaridade e Dissimilaridade

Em Bussab *et al.* (1990) é apresentada uma amostra de n=6 indivíduos foram coletados o peso (Kg) e altura (cm), conforme a tabela 1.

Tabela 1: Peso (kg) e altura (cm) de seis indivíduo

| Indivíduo | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| Peso | 79 | 75 | 70 | 63 | 71 | 60 |
| Altura | 180 | 175 | 170 | 167 | 180 | 165 |

Como estão sendo observadas p=2 variáveis, esse dados colocados na Figura (1) mostra que pelo critério de menor distância entre os pontos, forma-se dois conglomerados $C_1$ e $C_2$. A distância foi utilizada como uma medida de dissimilaridade, pois quanto menor a distância mais semelhantes são os indivíduos comparados. Todas as distâncias, com base na altura e pesos, entre esse indivíduos estão na matriz de abaixo:

$$D_{6\times 6} = \begin{bmatrix} & A & B & C & D & E & F \\ A & 0 & & & & & \\ B & 6,04 & 0 & & & & \\ C & 13,45 & 7,07 & 0 & & & \\ D & 20,61 & 14,42 & 7,61 & 0 & & \\ E & 8,00 & 6,40 & 10,04 & 15,26 & 0 & \\ F & 24,20 & 18,02 & 11,18 & 3,60 & 18,60 & 0 \end{bmatrix}$$

Ao analisar $D_{6\times 6}$, temos {D} e{F} são os indivíduos mais próximo, com distância euclidiana igual a $d(C,F) = 3,60$, que formam o conglomerado $C_1 = \{D,F\}$ . Ao continuar com o critério da menor distância temos: {B} une-se a {A}; depois {B} une-se a {E}; e finalmente {B} une-se a {C} formando o conglomerado $C_2 = \{A,B,C,E\}$.

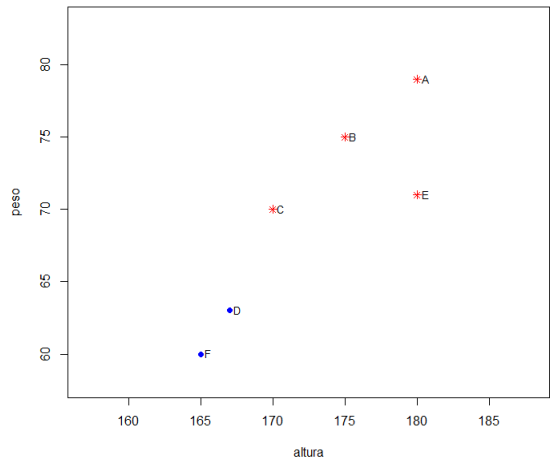



Figura 1: Peso e altura de seis indivíduos.

Para os dados da Tabela 1, fica claro que a há dois conglomerados 'naturais' induzidos pelas variáveis peso e altura e que a distância euclidiana foi uma medida de dissimilaridade eficiente, no sentido de permitir a formação dos conglomerados.

Neste texto será utilizada a notação apresentada em Mingotti(2007), assim dado um conjunto de dados constituídos de n elementos amostrais, tendo-se medido p-variáveis aleatórias de cada um deles, com objetivo de agrupar esses elementos em g grupos. Para cada elemento amostral j, tem-se, portanto, o vetor de medidas $X_j$ definido por:

$$X_j = [X_{1j}, \ldots, X_{pj}]'$$

onde $X_{ij}$ representa o valor observado da variável i medida no elemento j. Para que se possa proceder ao agrupamento de elementos, é necessário que se decida a priori a medida de similaridade ou dissimilaridade que será utilizada.

## Medidas de Dissimilaridade

As medidas de dissimilaridade são aquelas que quanto maior os seus valores menos semelhantes são os indivíduos comparados, a distância Euclidiana é a medida de dissimilaridade mais popular.

Distância euclidiana entre dois elementos $X_l$ e $X_k$, $l \neq k$ , é definida por:

$$d\left(X_l, X_k\right) = \sqrt[2]{\sum_{i=1}^{p}\left(X_{il} - X_{ik}\right)^2}$$

ou seja, os dois elementos são comparados em cada variável pertencente ao vetor de observações. Seja os dados dos indivíduos {D} e {F} da tabela 1, assim tem-se $X'_D = [63; 167]$ e $X'_F = [60; 165]$, logo

$$d\left(X_D, X_F\right) = \sqrt[2]{\left(63-60\right)^2 + \left(167-165\right)^2} = 3,60$$

Há outras medidas de distâncias que se pode citar: Distância de Mahalanobis, Distância Euclidiana Média e Distância de MinKowsky, da qual a distância Euclidiana é um caso especial. As distâncias entre os elementos amostrais são armazenadas na matriz $D_{nxn}$, chamada de matriz de distâncias. As medidas distância só podem ser aplicadas em variáveis quantitativas

## Medidas de Similaridade

As medidas de similaridade são aquelas que quanto maior os seus valores mais semelhantes são os indivíduos comparados.

Em geral, quando a análise envolve variáveis qualitativas, a similaridade entre os indivíduos baseia-se na presença ou na ausência das características observadas. A presença ou ausência de uma característica pode ser representada por uma variável binária do tipo 0 ou 1. Elementos parecidos têm mais itens similares que dissimilares.

Seja os dados que compara 3 indivíduos, quanto a presença os não de 5 características, conforme Tabela 2.

Tabela 2: Presença (1) ou ausência (0) de 5 características em 3 indivíduos

| 2*Indivíduo | Caracteristica | | | | |
|---|---|---|---|---|---|
| | X1 | X2 | X3 | X4 | X5 |
| A | 1 | 1 | 1 | 0 | 0 |
| B | 1 | 1 | 0 | 0 | 1 |
| C | 1 | 1 | 1 | 1 | 0 |

Nessa situação em que envolve variáveis qualitativas é razoável assumir que os os indivíduos que apresentam mais características em comum, são mais semelhantes. Assim, pelos dados pode-se afirmar que {C} é mais semelhante a {A} do que a {B} .

Uma média de similaridade é o Coeficiente de Concordância, que é definido como a razão entre a soma do número de características presentes e ausentes simultaneamente pelo total de características analisadas, quanto maior essa razão, maior a similaridade entre os indivíduos. Entre {A} é {C} existem três características presente e uma característica ausente simultaneamente e entre {B} é {C} existem 2 características presente e nenhuma ausente simultaneamente, assim

$$S\left(A, C\right) = \frac{4}{5} = 0,8 \text{ e } S\left(B, C\right) = \frac{2}{5} = 0,3$$

Uma outra medida de similaridade é o Coeficiente de Concordância Positiva, que é a razão entre o número de característica presentes simultaneamente nos indivíduos e número de características analisadas Entre {A} é {C} existem três características presente simultaneamente e entre {B} é {C} existem 2

características presente simultaneamente, assim

$$S(A,C) = \frac{3}{5} = 0,6 \text{ e } S(B,C) = \frac{2}{5} = 0,3$$

As medidas de similaridade entre os elementos amostrais são armazenadas na matriz $S_{nxn}$, chamada de matriz similaridade.

## Uso simultâneo de medidas de similaridade e de dissimilaridade

Há situações em que os indivíduos são caracterizados tanto por variáveis quantitativas como por variáveis qualitativas, nesta situação devemos trabalha ou somente com medidas de similaridade ou somente com medidas de dissimilaridades.

É importante ressaltar que qualquer medida de distância usada para as variáveis quantitativas pode ser transformada num coeficiente de similaridade. Suponho que dois elementos A e B sendo comparados e que a distância entre eles seja d(A,B). Então o coeficiente de similaridade entre A e B será definido por:

$$s(A,B) = 1 - d^*(A,B)$$

onde

$$d^* = \frac{(d(A,B) - min(D))}{(max(D) - min(D))}$$

sendo $min(D)$ e $max(D)$ o menor e a maior distância observados na matriz de distância $D_{nxn}$, sem considerar os seus elementos da diagonal principal. O coeficiente de similaridade entre um elemento e ele mesmo é definido como sendo 1, isto é, $d(A,A) = 1$. Suponha, por exemplo, que haja n = 4 elementos amostrais e que a matriz de distâncias entre eles seja igual a:

$$D_{4\times4} = \begin{bmatrix} 0 & 8 & 3 & 6 \\ 8 & 0 & 5 & 4 \\ 3 & 5 & 0 & 2 \\ 6 & 4 & 2 & 0 \end{bmatrix}$$

Então $min(D) = 2$ e $max(D) = 8$ . A distância entre os elementos 1 e 3 será transformada no coeficiente de similaridade $s(1,3)$, dado por:

$$s(1,3) = 1 - \frac{3-2}{8-2} = 1 - \frac{1}{6} = \frac{5}{6} = 0,83$$

Assim, a matriz de distâncias $D_{4x4}$ pode ser transformada na matriz de similaridade $S_{4x4}$, definida por:

$$S_{4\times4} = \begin{bmatrix} 1 & 0 & 0,83 & 0,33 \\ 0 & 1 & 0,50 & 0,66 \\ 0,83 & 0,50 & 1 & 1 \\ 0,33 & 0,66 & 1 & 1 \end{bmatrix}$$

Assim, na situação em que são observadas p-variáveis quantitativas e q-variáveis qualitativas nos mesmos elementos amostrais, tem-se basicamente três alternativas:

1. Transformar as q-variáveis qualitativas em quantitativas através de atribuição de valores numéricos às várias categorias. Existem situações na quais há uma ordenação natural no grau de importância das categorias e a atribuição de valores cria um conjunto de variáveis ordinais. No caso de classes nominais, a atribuição de valores é arbitrária. Após a transformação das variáveis, para comparação dos elementos amostrais nas p+q variáveis, utiliza-se as medidas de distância.

2. Transformar as p-variáveis quantitativas em variáveis qualitativas através da categorização de seus valores por algum critério. A partir dessa transformação utiliza-se os coeficientes de concordância apresentados para comparar os elementos amostrais nas p+q variáveis. Essa é a alternativa de uso menos comum em problemas práticos devido a perda de informação que se tem ao categorizar variáveis contínuas.

3. Construir medidas de semelhanças mistas e utilizá-las para a comparação dos elementos amostrais. A proposta mais simples é construir uma combinação linear das medidas de comparação para variáveis quantitativas e qualitativas, isto é, se temos dois elementos amostrais A e B, o coeficiente de semelhança de A e B será definido como

$$c(A,B) = \omega_p c_p(A,B) + \omega_q c_q(A,B)$$

onde os coeficientes $c_q(\cdot)$ e $c_q(\cdot)$ são coeficientes de similaridade (ou dissimilaridade) entre a A e B nas p-variáveis quantitativas e nas q-variáveis qualitativas, respectivamente, e $\omega_p$ e $\omega_q$ são os pesos de ponderação para as comparações quantitativa e qualitativa. Para que o coeficiente de combinado faça sentido, os coeficientes $c_q(\cdot)$ e $c_q(\cdot)$ precisam ter a mesma direção e precisam estar definidos no mesmo intervalo de variação. Assim, se os coeficientes de similaridade são usados na comparação dos elementos amostrais nas variáveis qualitativas e a distância Euclidiana é usada para comparar os elementos amostrais na variáveis quantitativas, os valores de distância devem ser transformados para o intervalo [0,1]. A dificuldade que se tem na implementação desse coeficiente combinado é a determinação dos pesos e $\omega_p$ e $\omega_q$. Uma possibilidade é escolher esses valores como função do número de variáveis quantitativas e qualitativas, por exemplo

$$\omega_p = \frac{p}{p+q} \text{ e } \omega_q = \frac{q}{p+q}$$

## Método das K-Médias

O método das k-médias (Hartingan; Wong, 1979) é provavelmente um dos mais conhecidos e utilizados em problemas práticos. Basicamente cada elemento amostral é alocado àquele cluster cujo centroide (vetor de média amostral) é o mais próximo do vetor de valores observados para o respectivo elemento. Originalmente, o método é composto por 4 passos:

1. Primeiramente escolhe-se k centroides, chamados de sementes ou protótipos, para se iniciar o processo de partição;

2. Cada elemento do conjunto de dados é, então, comparado com cada centroide inicial, através de uma medida de distância que, em geral, é a distância Euclidiana. O elemento é alocado ao grupo cuja distância é menor;

3. Depois de aplicar o passo 2 para cada um dos n elementos amostrais, recalcula-se os valores dos centroides para cada novo grupo formado, e repete-se o passo 2, considerando-se os centroides destes novos grupos;

4. Os passos 2 e 3 devem ser repetidos até que todos os elementos amostrais estejam bem alocados em seus grupos, isto é, até que nenhuma realocação de elementos seja necessária.

Computacionalmente, os softwares estatísticos podem utilizar formas diferentes de implementação do método de k-medias. A escolha das sementes iniciais de agrupamentos influencia no agrupamento final. Portanto, cuidados são necessários na escolha das sementes. Algumas sugestões são apresentadas a seguir.

1. Técnicas hierárquicas aglomerativas

   Primeiramente utiliza-se algum dos métodos de agrupamento das técnicas hierárquicas aglomerativas para se obter os g=k grupos iniciais. A partir, daí, calcula-se o vetor de médias de cada grupo formado, sendo esses vetores de médias as sementes iniciais usadas no método k-médias.

2. Escolha aleatória

   As k sementes iniciais são escolhidas aleatoriamente dentro do conjunto de dados a ser analisado. Um procedimento amostral pode ser utilizado para isso é o de amostragem aleatória simples sem reposição. Essa estratégia de escolha não é eficiente, embora seja de execução simples. Uma forma de melhorar sua eficiência é selecionar m amostras de k sementes, M >1. Desse modo, o procedimento de amostragem aleatória simples é repetido m vezes e, no final, calcula-se o vetor das médias selecionadas para cada grupo. Estes vetores constituem centroides de inicialização do procedimento das k-Médias.

3. Escolha via uma variável aleatória

   Neste procedimento, escolhe-se a variável aleatória de maior variância dentre as p componentes do vetor aleatório X em consideração. Desse modo, a variável por si só já induz uma certa participação natural dos dados. Divide-se o domínio da variável em k intervalos, por exemplo. A semente inicial será o centroide de cada intervalo.

4. Observação dos valores discrepantes do conjunto de dados

   Através de uma análise estatística, busca-se k elementos discrepantes no conglomerado de dados. Cada um desses elementos constituirá a semente de um conglomerado inicial. A discrepância nesse neste caso é em relação às p-variáveis observadas conjuntamente.

5. Escolha prefixada

   As sementes são escolhidas arbitrariamente pelo pesquisador. É um método não muito recomendável devido ao alto grau de subjetividade. No entanto, podem ser usado em casos nos quais o pesquisador tem um grande grau de conhecimento do problema estudado ou esteja buscando validar uma solução já existente.

6. Os k primeiros valores dos bancos de dados

   A maioria dos softwares estatístico usa, como default para a escolha das sementes iniciais, as k primeiras observações do banco de dados, a menos que o pesquisador diretamente quais as sementes desejam utilizar para inicialização do algoritmo.

   Esse procedimento pode trazer bons resultados quando os k primeiros elementos amostrais são discrepantes entres si e não é recomendável quando esses são semelhantes entre si.

   Assim, caberá ao usuário verificar o que está ocorrendo com seu banco de dados e rearranjá-lo de modo a satisfazer o requisito para obtenção de melhores resultados com este critério de escolha de sementes.

# Método K-means no R

O método K-means pode ser aplicado a um conjunto de dados no R por meio da função `kmeans()`, que integra a biblioteca *stats*. Está função retorna uma *lista* de resultados que , entre outros, está um vetor com a identificação do cluster ao pertence cada observação. A chamada dessa função tem a forma básica:

```
kmeans(x, centers)
```

onde

1. `x` - matriz numérica de dados, ou um objeto que pode ser coagido para tal matriz (como um vetor numérico ou um `data frame` com todas as colunas numéricas).

2. `Centros` - o número de clusters, digamos k, ou um conjunto de centros de clusters iniciais (distintos). Se for um número, um conjunto aleatório de linhas (distintas) em x é escolhido como os centros iniciais.

A *lista* retornada pela função `kmeans()` tem os seguintes componente:

1. `cluster` - Um vetor de inteiros (de 1: k) indicando o cluster ao qual cada ponto é alocado.
2. `centers` - Uma matriz como os centroides de cada cluster.
3. `totss` - A soma dos quadrados total.
4. `withinss` - Vetor com a soma dos quadrados dentro de cada cluster.
5. `tot.withinss` - Soma total de quadrados dentro do cluster, isto é, `sum(withinss)`.
6. `betweenss` A soma dos quadrados entre grupos, isto é, `totss-tot.withinss`.
7. `size` - O número de pontos em cada cluster.
8. `iter` - O número de iterações (externas).
9. `indiferente` - integer: indicador de um possível problema no algoritmo - para especialistas

## Exercício

**Texto Base - Iris data set**

O conjunto de dados da flor Iris é um conjunto de dados multivariado utilizado pelo estatístico e biólogo britânico Ronald Fisher em seu artigo de 1936, *O uso de múltiplas medições em problemas taxonômicos como exemplo de análise discriminante linear*. É chamado às vezes base de dados Iris de Andersen, pois foi obtido por Edgar Andersen para quantificar as variações morfológicas das flores íris de três espécies relacionadas. Duas das três espécies foram coletadas na Península de Gaspé "todas do mesmo pasto, e colhidas no mesmo dia e medidas ao mesmo tempo pela mesma pessoa com o mesmo aparelho". O conjunto de dados consiste em 50 amostras de cada uma das três espécies de Iris (**Iris setosa** , **Iris virginica** e **Iris versicolor**). Quatro características foram medidas a partir de cada amostra: o comprimento e a largura das sépalas e pétalas , em centímetros. Essa base de dados já está incorporada no R, é o `data frame` iris.

**Questão 1** - Execute e comente o programa abaixo:

```
dados <- iris
cor(dados[,-5])
set.seed(12345)
f <- kmeans(dados[,-5],3)
table(dados$Species, f$cluster)
wss <- NULL
for (i in 1:15) wss[i] <- sum(kmeans(dados[,-5], centers=i)$withinss)
plot(1:15, wss, type="l", xlab="Numero de Cluster",
     ylab="Soma dos Quadrados Dentro", axes=F)
points(1:15, wss, pch=16, col="red")
axis(1, 1:15, las=2)
axis(2,seq(0,700,by=100))
fit <- prcomp(dados[,-5])$x
plot(fit,col=f$cluster,pch='.')
text(fit[,1], fit[,2], substr(dados[,5],1,2), col=f$cluster, cex=0.9)
```

**Texto Base - Sementes**

Foram examinado grãos pertencentes a três variedades diferentes de trigo: Kama, Rosa e canadense, 70 elementos cada, selecionados aleatoriamente para o experimento. A visualização de alta qualidade da estrutura interna do núcleo foi detectada usando uma técnica de raios X. Essa técnica não é destrutiva e consideravelmente mais barata do que outras técnicas de imagem mais sofisticadas, como a microscopia de varredura ou a tecnologia a laser. As imagens foram gravadas em placas KODAK de raios X de 13x18 cm. Os dados analizados são sete parâmetros, todos eram valores reais, geométricos dos grãos de trigo:

i - área

ii - perímetro

iii - compactação

iv - comprimento do núcleo

v - largura do núcleo

vi - coeficiente de assimetria

vii - comprimento do sulco do núcleo.

**Questão 2** - Faça uma análise de cluster com o data frame `semente`, que além dos dados numéricos acima, contém a variável `tipo`, com os seguintes passos:

i - Padronize as variáveis para que tenha média 0 (zero) e desvio padrão 1(um).

ii - Defina o número de cluster usando o gráfico do número de cluster e a soma dos quadrados dentro.

iii - Faça uma tabela de frequência, na qual as linhas representam o tipo de semente e as colunas as classificação dada pelo procedimento k-means.

**Análise de Cluster em Redes Sociais**

O conjunto de dados a ser usado representa uma amostra aleatória de 30.000 estudantes do ensino médio dos EUA que tiveram perfis em uma conhecida rede social de 2006 a 2009. Das 500 principais palavras que aparecem em todas as páginas, 36 palavras foram escolhidas para representar cinco categorias de interesses, a saber: atividades extracurriculares, moda, religião, romance e comportamento antissocial. Esas palavras incluem termos como futebol, sexy, beijo, bíblia, compras, morte e drogas. O conjunto de dados **snsdata.csv** indica, para cada pessoa, quantas vezes cada palavra apareceu no seu perfil desta rede social.

**Questão 3** - Aplique o método k-means aos dados em **snsdata.csv**, com os seguintes passos:

i - Crie o data frame `dado` a partir do arquivo **snsdata.csv**.

ii - Observe as seis primeiras linhas do data frame `dado`.

iii - Crie o data frame `dado2` que contenha apenas as variáveis que armazenar a frequência das palavras analisadas ocorrem em cada perfil.

iv - Defina o número de cluster usando o gráfico do número de cluster e a soma dos quadrados dentro.

v - Faça uma tabela de frequência com as palavras de cada cluster.

REFERÊNCIAS

[1] FÁVERO, L. P.; BELFIORE, P.; SILVA, F. L.; CHAN, B. L. Análise de dados: modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier, 2009.

[2] HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. Análise multivariada de dados. Porto Alegre: Bookman, 2005.