Análise Multivariada de Dados - Aula 01

O quê é Análise Multivariada? O que é Análise de Conglomerados?

Kaique Matias de Andrade Roberto

Ciências Atuariais

HECSA - Escola de Negócios

FIAM-FAAM-FMU

Conteúdo

1. Conceitos que aprendemos em Aulas anteriores
2. O que é Análise Multivariada?
3. Exemplos de Dados Multivariados
4. Visão Geral dos Métodos Multivariados
5. Representação de Dados Multivariados
6. O que é Análise de Conglomerados?
7. Comentários Finais
8. Referências

Conceitos que aprendemos em Aulas anteriores

Conceitos que aprendemos em Aulas anteriores

- recapitulamos os tipos de variáveis;
- aprendemos como classificar variáveis;
- recapitulamos alguns tópicos de Estatística Descritiva;
- tivemos um primeiro contato com os conceitos de Amostragem;
- relembramos o conceito de Variável Aleatória;
- listamos as principais distribuições discretas e contínuas.

N_1 P_1 0 - 10 } " 0 - 14 "
E 0 - 4 }

N_2 P_2 0 - 9

APS 1

Software / linguagem:

R, Python (free)

O que é Análise Multivariada?

SPSS, STATA (pago)

XLSTATA (Excel)

O que é Análise Multivariada?

Definição 2.1

Análise multivariada se refere a todas as técnicas estatísticas que simultaneamente analisam múltiplas medidas sobre indivíduos ou objetos sob investigação.

O que é Análise Multivariada?

Assim, qualquer análise simultânea de mais do que duas variáveis pode ser considerada, a princípio, como multivariada.

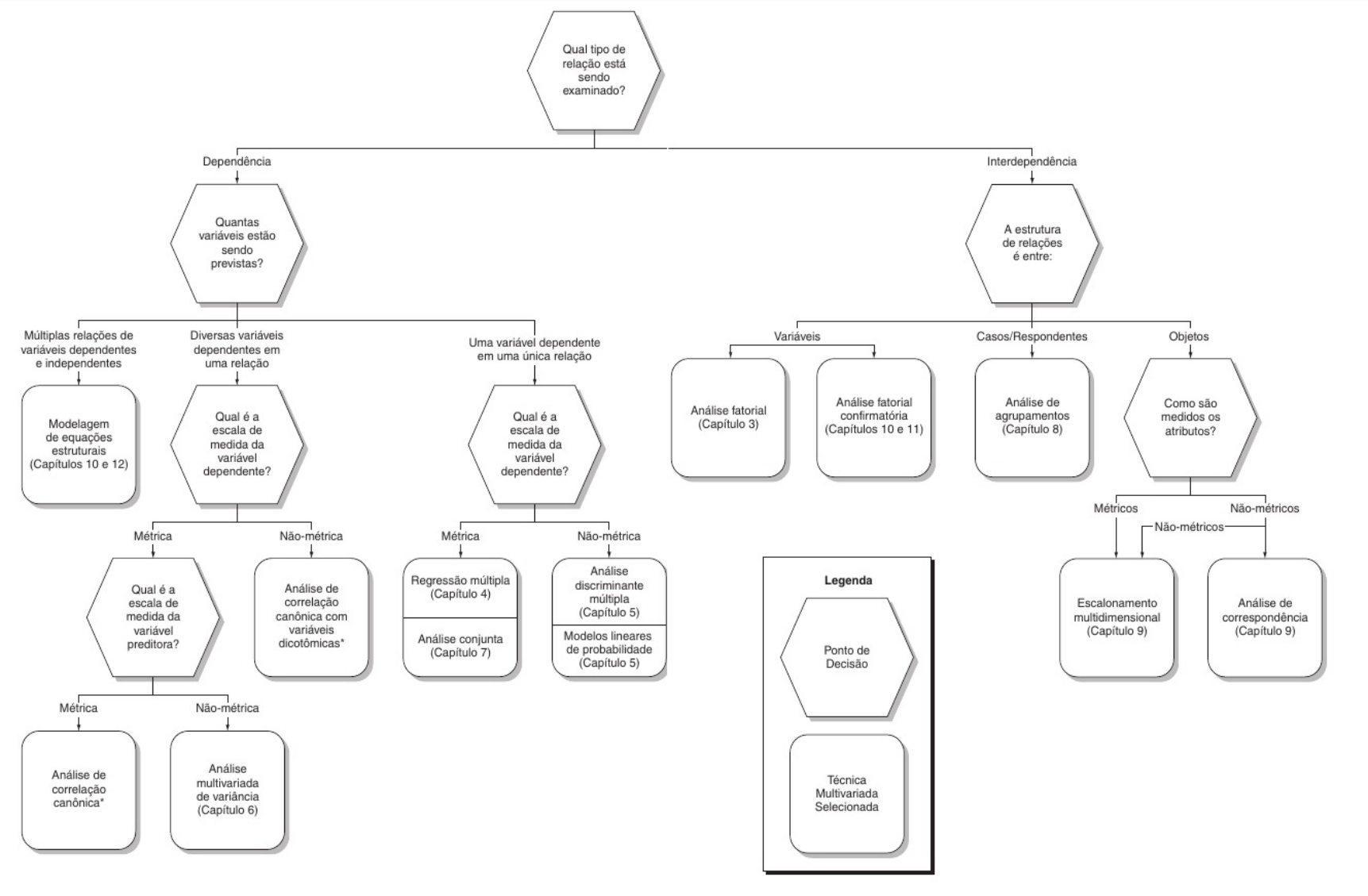
O que é Análise Multivariada?

Muitas técnicas multivariadas são extensões da análise univariada (análises de distribuições de uma única variável) e da análise bivariada (correlação, análise de variância e regressão simples usadas para analisar duas variáveis).

O que é Análise Multivariada?

Outras técnicas multivariadas, não obstante, são exclusivamente planejadas para lidar com aspectos multivariados, como a análise fatorial, que identifica a estrutura inerente a um conjunto de variáveis, ou a análise discriminante, que distingue entre grupos baseada em um conjunto de variáveis.

O que é Análise Multivariada?



Exemplos de Dados Multivariados

Exemplos de Dados Multivariados

Exemplo 3.1 (Pardais sobreviventes de tempestade)

Após uma forte tempestade em 01 de fevereiro de 1898, diversos pardais moribundos foram levados ao laboratório biológico de Hermon Bumpus na Universidade de Brown em Rhode Island.

Exemplos de Dados Multivariados

Subsequentemente cerca de metade dos pássaros morreram, e Bumpus viu isso como uma oportunidade de encontrar suporte para a teoria de seleção natural de Charles Darwin.

Exemplos de Dados Multivariados

Para esse fim, ele fez oito medidas morfológicas em cada pássaro, e também os pesou. Os resultados de cinco das medidas são mostrados na guia "pardais" planilha "dados-multivariados" para fêmeas somente.

Exemplos de Dados Multivariados

Dos dados que obteve, Bumpus (1898) concluiu que "os pássaros que morreram, morreram não por acidente, mas porque eles eram fisicamente desqualificados, e que os pássaros que sobreviveram, sobreviveram porque eles possuíam certas características físicas".

Exemplos de Dados Multivariados

Concluiu também que "o processo de eliminação seletiva é mais severo com indivíduos extremamente variáveis, não importando em qual direção a variação possa ocorrer".

Exemplos de Dados Multivariados

Em outras palavras, ele concluiu que é tão perigoso estar acima de um certo padrão de excelência orgânica como estar visivelmente abaixo do padrão.

Exemplos de Dados Multivariados

Isso queria dizer que ocorreu seleção estabilizadora, de modo que indivíduos com medidas próximas da média sobrevivem melhor do que indivíduos com medidas longe da média.

Exemplos de Dados Multivariados

De fato, o desenvolvimento dos métodos de análise multivariada havia recém-iniciado em 1898 quando Bumpus estava escrevendo. Apesar disso, seus métodos de análise foram sensíveis. Muitos autores têm reanalisado seus dados e, em geral, têm confirmado suas conclusões.

Exemplos de Dados Multivariados

Tomando os dados como um exemplo para ilustrar métodos multivariados, surgem muitas questões interessantes. Em particular:

Exemplos de Dados Multivariados

- Como estão relacionadas as várias variáveis? Por exemplo, um valor grande para uma das variáveis tende a ocorrer com valores grandes para as outras variáveis?

Exemplos de Dados Multivariados

- Os sobreviventes e os não-sobreviventes têm diferenças estatisticamente significantes para seus valores médios das variáveis?

Exemplos de Dados Multivariados

- Os sobreviventes e não-sobreviventes mostram quantidades similares de variação para as variáveis?

Exemplos de Dados Multivariados

- Se os sobreviventes e não-sobreviventes diferem em termos das distribuições das variáveis, então é possível construir alguma função dessas variáveis que separe os dois grupos? Então seria conveniente se valores grandes da função tendessem a ocorrer com os sobreviventes enquanto que a função seria então aparentemente um índice de ajuste darwiniano dos pardais.

Exemplos de Dados Multivariados

Exemplo 3.2 (Cães pré-históricos da Tailândia)

Escavações de locais pré-históricos no nordeste da Tailândia têm produzido uma coleção de ossos caninos cobrindo um período em tomo de 3500 a.C. até o presente. Entretanto, a origem dos cães pré-históricos não é certa.

Exemplos de Dados Multivariados

Podem descender dos jacais dourados (Canis aureus) ou do lobo, mas o lobo não é nativo da Tailândia. As fontes de origem mais próximas são a parte ocidental da China (Canis lupus chanco) ou o subcontinente indiano (Canis lupus pallides).

Exemplos de Dados Multivariados

Para tentar esclarecer os ancestrais dos cães pré-históricos, foram feitas medidas da mandíbula dos espécimens disponíveis.

Exemplos de Dados Multivariados

Estas foram então comparadas com as mesmas medidas feitas no chacal dourado, no lobo chinês e no lobo indiano.

Exemplos de Dados Multivariados

As comparações foram também estendidas para incluir o dingo, o qual tem suas origens na índia, o cuon (Cuon alpinus), o qual é indígena do sudeste da Ásia e os cães modernos de cidade da Tailândia.

Exemplos de Dados Multivariados

Na aba "caes-pre-historicos" da planilha "dados-multivariados" temos os valores médios para as seis medidas de mandíbulas para espécimens de todos os sete grupos.

Exemplos de Dados Multivariados

A questão principal aqui é o que as medidas sugerem sobre o relacionamento entre os grupos e, em particular, como os cães pré-históricos poderiam se relacionar com os outros grupos.

Exemplos de Dados Multivariados

Exemplo 3.3 (Empregos em países europeus)

Considere os dados na planilha "dados-multivariados". Eles mostram as porcentagens da força de trabalho em nove diferentes tipos de indústrias para 30 países europeus.

Exemplos de Dados Multivariados

Nesse caso, métodos multivariados podem ser úteis para isolar grupos de países com padrões similares de empregos, e, em geral, ajudar o entendimento dos relacionamentos entre os países.

Exemplos de Dados Multivariados

Diferenças entre países que são relacionados a grupos políticos (UE, a União Européia; AELC, a área européia de livre comércio; países do leste europeu e outros países) podem ser de particular interesse.

Visão Geral dos Métodos Multivariados

Visão Geral dos Métodos Multivariados

Os exemplos que acabamos de considerar são dados brutas típicos para métodos estatísticos multivariados.

Visão Geral dos Métodos Multivariados

Em todos os casos, existem várias variáveis de interesse e elas são claramente não-independentes umas das outras.

Visão Geral dos Métodos Multivariados

Nesse momento, é útil dar uma breve visão prévia do que está por vir nas próximas aulas.

Visão Geral dos Métodos Multivariados

Definição 4.1

A análise de componentes principais é elaborada para reduzir o número de variáveis que necessitam ser consideradas a um número menor de índices (chamados de componentes principais) os quais são combinações lineares das variáveis originais.

Exemplo 4.2

Muito da variação nas medidas do corpo dos pardais (X_1 a X_5) mostrada no Exemplo 3.1 está relacionada ao tamanho geral dos pássaros, e o total

$$I_1 = X_1 + X_2 + X_3 + X_4 + X_5$$

deve medir muito bem esse aspecto dos dados. Este índice é responsável por uma dimensão dos dados.

Visão Geral dos Métodos Multivariados

Outro índice é

$$I_2 = X_1 + X_2 + X_3 - X_4 - X_5$$

o qual é um contraste entre as três primeiras medidas e as duas últimas. Este reflete outra dimensão dos dados.

Visão Geral dos Métodos Multivariados

A análise de componentes principais fornece uma maneira **objetiva** de encontrar índices desse tipo de modo que a variação nos dados pode ser levada em consideração tão concisamente quanto possível.

Visão Geral dos Métodos Multivariados

Pode muito bem acontecer que dois ou mais componentes principais forneçam um bom resumo de todas as variáveis originais.

Visão Geral dos Métodos Multivariados

A consideração dos valores dos componentes principais ao invés dos valores das variáveis originais pode tornar muito mais fácil entender o que os dados têm a dizer.

Visão Geral dos Métodos Multivariados

Em poucas palavras, a análise de componentes principais é um meio de simplificar dados pela redução do número de variáveis.

Definição 4.3

A **análise de fatores** também tem como objetivo estudar a variação em uma quantidade de variáveis originais usando um número menor de variáveis índices ou fatores.

Visão Geral dos Métodos Multivariados

Assume-se que cada variável original possa ser expressa como uma combinação linear desses fatores, mais um termo residual que reflete o quanto a variável é independente das outras variáveis.

Exemplo 4.4

Por exemplo, um modelo de dois fatores para os dados dos pardais assume que

$$
\begin{aligned}
X_1 &= a_{11}F_1 + a_{12}F_2 + e_1 \\
X_2 &= a_{21}F_1 + a_{22}F_2 + e_2 \\
X_3 &= a_{31}F_1 + a_{32}F_2 + e_3 \\
X_4 &= a_{41}F_1 + a_{42}F_2 + e_4 \\
X_5 &= a_{51}F_1 + a_{52}F_2 + e_5
\end{aligned}
$$

em que os valores a_{ij} são constantes, F_1 e F_2 são fatores e e_i representa a variação em X_i que é independente da variação nas outras variáveis X.

Aqui F_1 pode ser o fator tamanho. Nesse caso, os coeficientes $a_{11}, a_{21}, a_{31}, a_{41}, a_{51}$ seriam todos positivos, refletindo o fato de que alguns pássaros tendem a ser grandes e alguns pássaros tendem a ser pequenos em todas as medidas do corpo.

O segundo fator F_2 poderia então medir um aspecto da forma dos pássaros, com alguns coeficientes positivos e alguns negativos.

Se esse modelo de dois fatores ajustar bem os dados, então ele forneceria uma descrição relativamente direta do relacionamento entre as cinco medidas do corpo que estão sendo consideradas.

Visão Geral dos Métodos Multivariados

Definição 4.5

A **análise de função discriminante** refere-se à possibilidade de separar diferentes grupos com base nas medidas disponíveis.

Exemplo 4.6

Isso pode ser usado, por exemplo, para ver quão bem pardais sobreviventes e não-sobreviventes podem ser separados usando suas medidas do corpo (Exemplo 3.1), ou como crânios de diferentes épocas podem ser separados, novamente usando medidas de tamanho (Exemplo 3.2).

Visão Geral dos Métodos Multivariados

Assim como a análise de componentes principais, a análise de função discriminante é baseada na ideia de encontrar combinações lineares convenientes das variáveis originais para atingir o objetivo desejado.

- Matriz
- Base
- Diagonalização
- Combinação linear
- vetor
- Matriz Transposta
- Álgebra Linear

É tudo novidade

Combinação Linear das variáveis $X_1, \ldots, X_n$:

$$\alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n$$

com $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$

$X_1 + X_2$, $2X_1 - 3X_2$,

$(X_1)^2 - X_2$ não é linear

$\frac{X_1}{X_2}$ não é linear

cluster

Definição 4.7

A **análise de agrupamentos/conglomerados** diz respeito à identificação de grupos de objetos similares.

Exemplo 4.8

Não há muito sentido em fazer esse tipo de análise com dados como os do Exemplos 3.1 pois os grupos já são conhecidos (sobreviventes / não-sobreviventes).

Visão Geral dos Métodos Multivariados

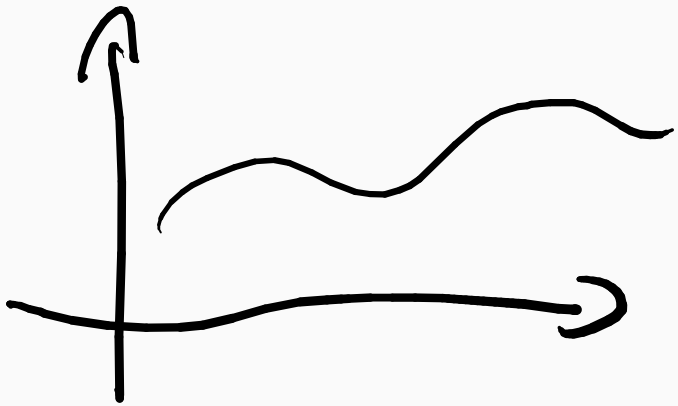
No entanto, no Exemplo 3.2 o principal ponto de interesse está na similaridade entre cães pré-históricos tailandeses e outros animais.

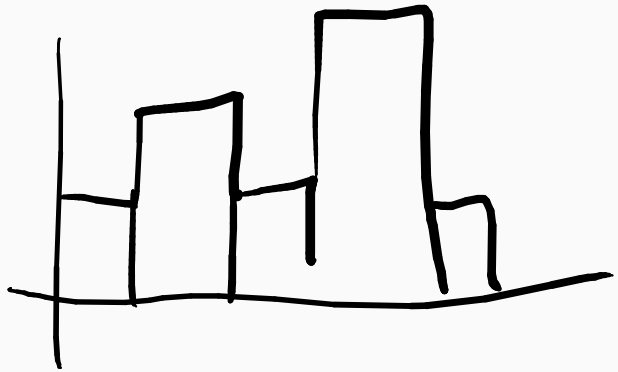
Visão Geral dos Métodos Multivariados

Da mesma forma, no Exemplo 3.3 os países europeus podem possivelmente ser agrupados em termos de suas similaridades no padrão de empregos.

Representação de Dados Multivariados

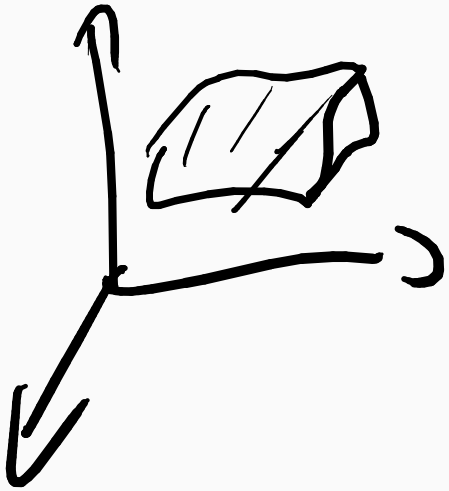
Representação de Dados Multivariados





Gráficos precisam ser apresentados em duas dimensões, ou sobre papel ou na tela de um computador.

Representação de Dados Multivariados



É consideravelmente mais complicado mostrar uma variável representada contra outras duas, mas ainda possível.

Representação de Dados Multivariados

Não é possível mostrar uma variável representada contra outras três ao mesmo tempo em alguma extensão da representação tridimensional.

Representação de Dados Multivariados

Portanto, é um problema de magnitude maior o de mostrar de uma maneira simples os relacionamentos que existem entre objetos individuais em um conjunto multivariado de dados onde estes objetos são descritos por quatro ou mais variáveis cada um.

Representação de Dados Multivariados

Em síntese, podemos dizer que não existe um método para representação de dados em muitas variáveis ao mesmo tempo que seja completamente satisfatório em situações nas quais não é desejável reduzir essas variáveis a duas ou três variáveis índices.

Perfil Bivariado ou Análise Draftsman

	X_1	X_2	X_3
X_1	histograma X_1	correl X_1 e X_2	correl X_1 e X_3
X_2	Dispersão X_2 e X_1	histograma X_2	correl X_2 e X_3
X_3	Dispersão X_3 e X_1	Dispersão X_3 e X_2	histograma X_3

Representação de Dados Multivariados

Mesmo assim, vamos entender como usar gráficos univariados/bivariados para obter insights sobre dados multivariados, usando os dados da planinha "dados-multivariados".

SUGESTÃO

Uma Análise Estatística bem-executada está fortemente apoiada em um procedimento consistente de coleta de dados. Como sugestão, faça o Exercício 1.4 para aprofundamento no tema.

O que é Análise de Conglomerados?

O que é Análise de Conglomerados?

A **análise de agrupamentos/conglomerados** representa um conjunto de técnicas exploratórias que podem ser aplicadas quando há a intenção de se verificar a existência de comportamentos semelhantes entre observações (indivíduos, empresas, municípios, países, etc) em relação a determinadas variáveis e o objetivo de se criarem grupos, ou clusters, em que prevaleça a homogeneidade interna.

O que é Análise de Conglomerados?

Além disso, a inclusão de novas observações ou variáveis também pode fazer com que haja um rearranjo completo das observações nos grupos, tornando *necessário* a reaplicação da modelagem.

O que é Análise de Conglomerados?

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.

O que é Análise de Conglomerados?

Outra forma de aplicação seria, 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.

O que é Análise de Conglomerados?

Entretanto, independentemente da natureza do objetivo, a análise de agrupamentos continuará *exploratória*.

O que é Análise de Conglomerados?

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.

O que é Análise de Conglomerados?

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.

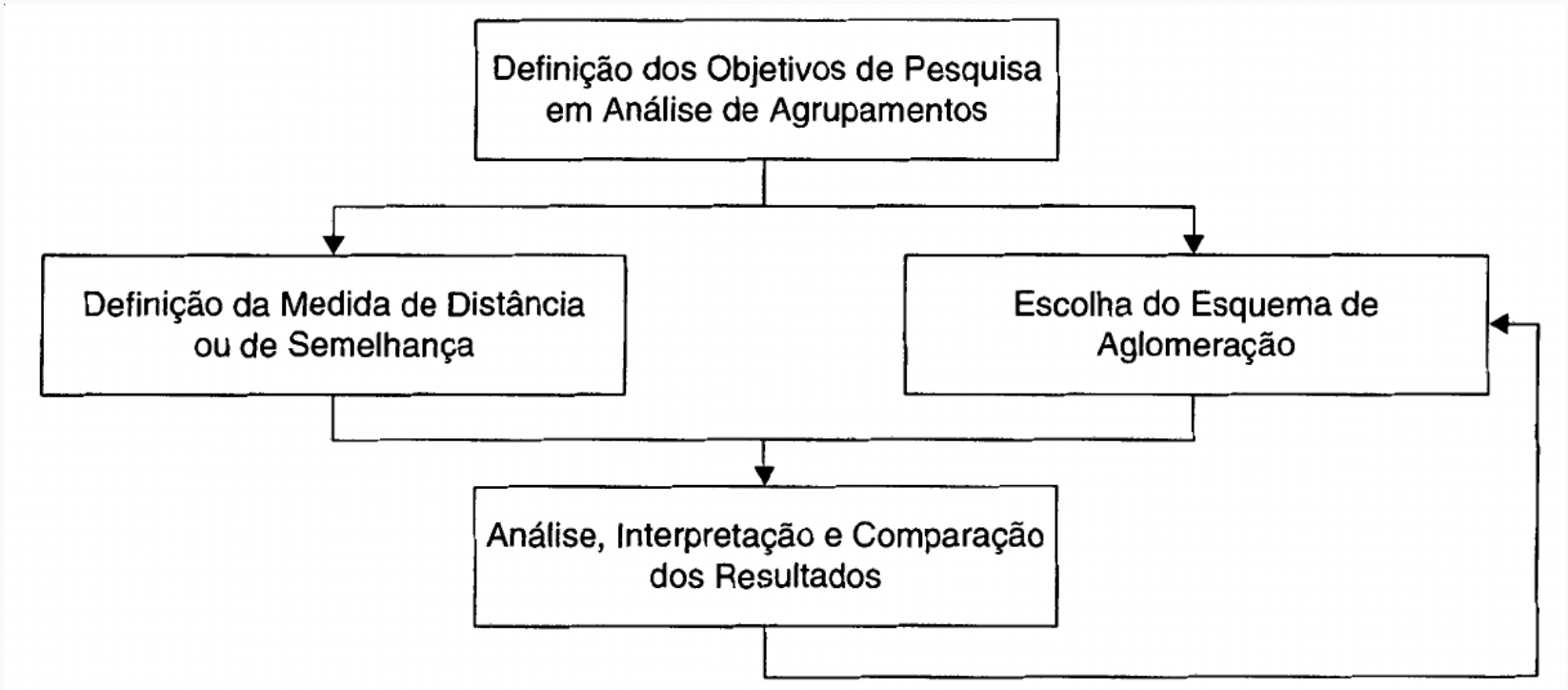
O que é Análise de Conglomerados?

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.

O que é Análise de Conglomerados?

A Figura a seguir apresenta a lógica a partir da qual a análise de agrupamentos pode ser elaborada.

O que é Análise de Conglomerados?



O que é Análise de Conglomerados?

Exemplo 6.1

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$).

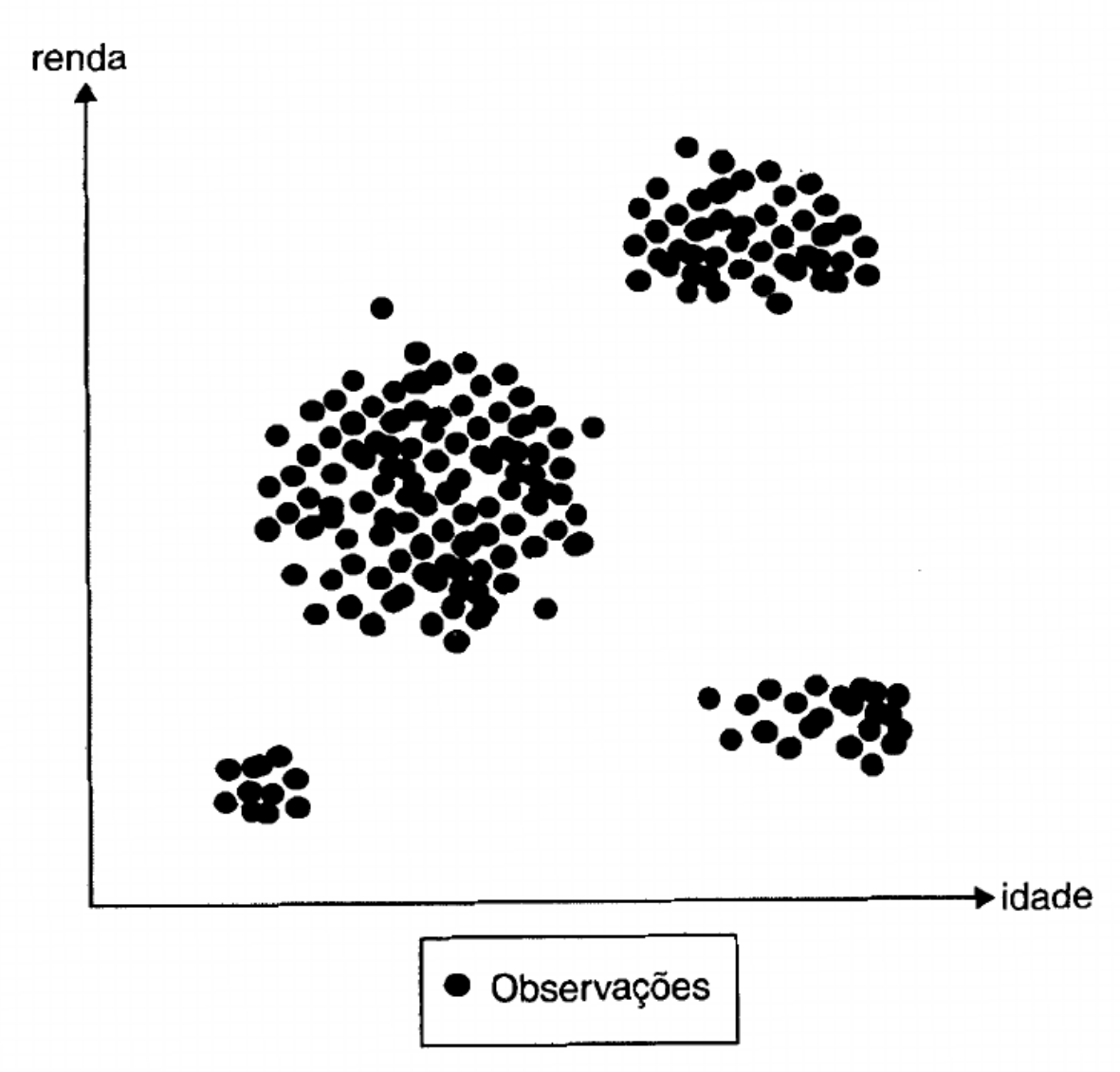
O que é Análise de Conglomerados?

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.

O que é Análise de Conglomerados?

Após a coleta dos dados, o pesquisador elaborou um gráfico de dispersão, como o apresentado na Figura a seguir.

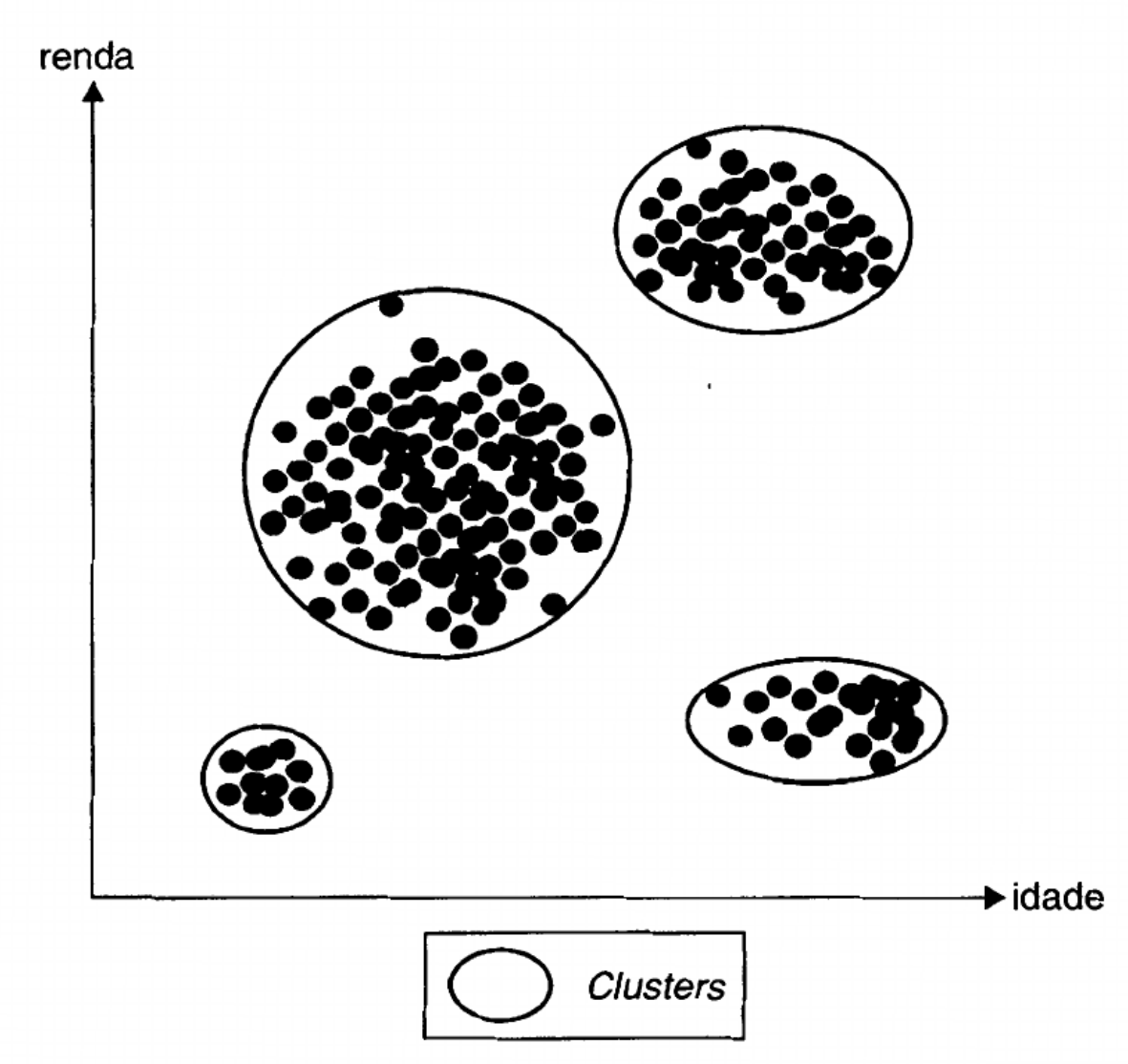
O que é Análise de Conglomerados?



O que é Análise de Conglomerados?

Com base no gráfico da Figura anterior, o pesquisador identificou quatro clusters, destacando-os em novo gráfico.

O que é Análise de Conglomerados?



O que é Análise de Conglomerados?

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.

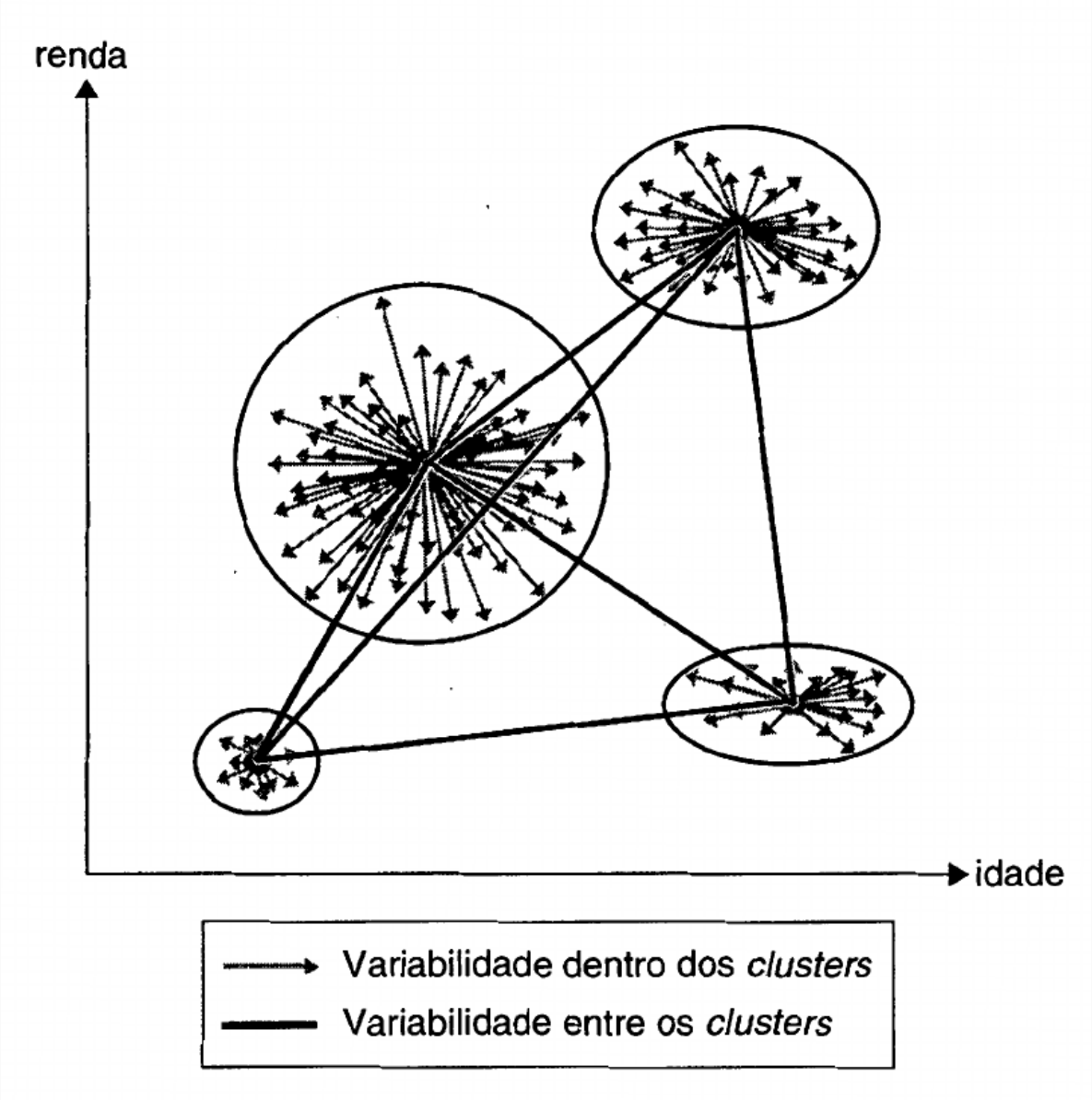
O que é Análise de Conglomerados?

Com isso, a expectativa é poder embasar de maneira clara e consciente, sua decisão a respeito da alocação dos indivíduos nesses quatro novos programas sociais.

O que é Análise de Conglomerados?

A fim de ilustrar essa questão, o pesquisador elaborou o gráfico da Figura a seguir.

O que é Análise de Conglomerados?



O que é Análise de Conglomerados?

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.

O que é Análise de Conglomerados?

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).

O que é Análise de Conglomerados?

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.

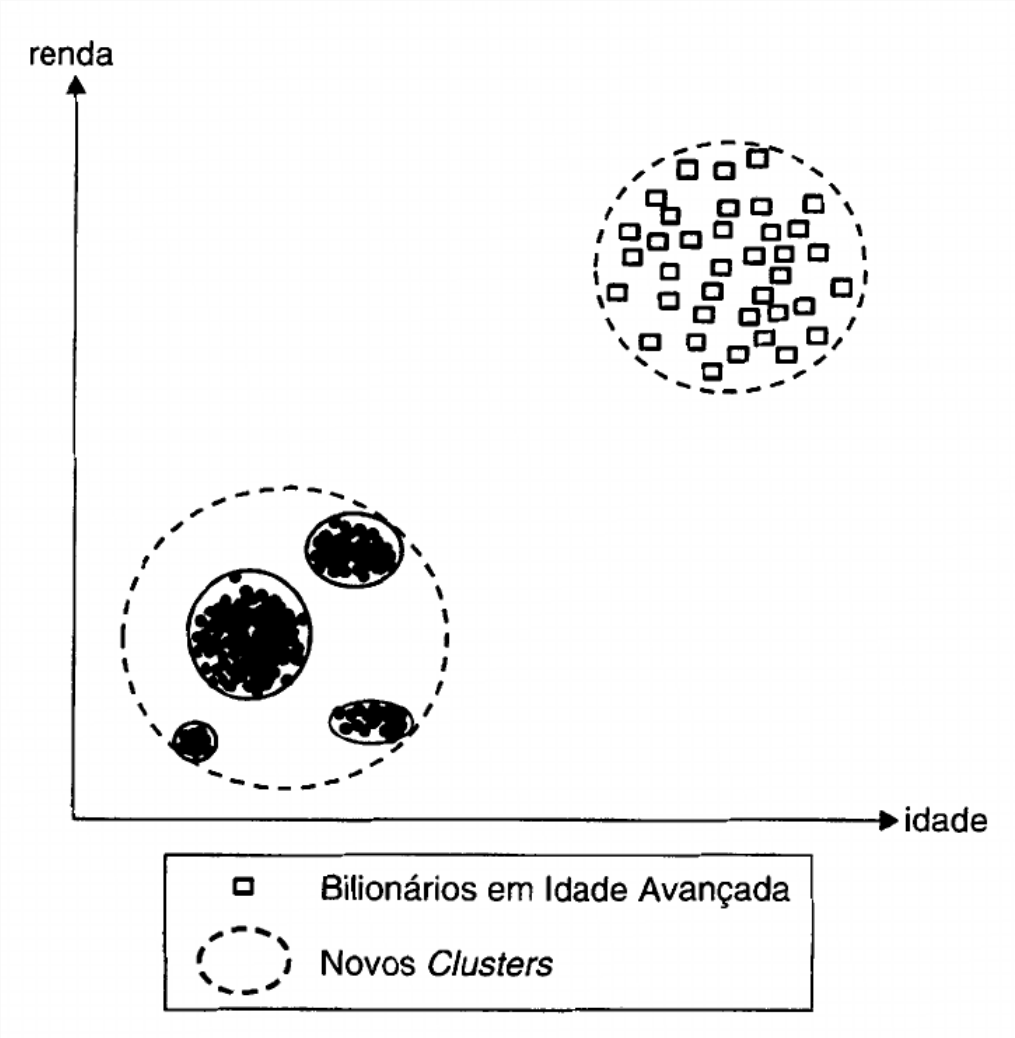
O que é Análise de Conglomerados?

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 que é Análise de Conglomerados?

O novo gráfico de dispersão encontra-se na Figura a seguir.

O que é Análise de Conglomerados?



O que é Análise de Conglomerados?

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.

O que é Análise de Conglomerados?

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.

Comentários Finais

Comentários Finais

Em resumo, na aula de hoje nós:

- vimos alguns objetivos da Análise Multivariada;
- tivemos uma visão geral dos métodos multivariados;
- lidamos com o problema da representação de dados multivariados;
- começamos a exposição da Análise de Conglomerados.

Comentários Finais

Nas próximas aulas focaremos na Análise de Conglomerados. Em particular, na próxima aula nós iremos lidar com:

- medidas de distância;
- medidas de similaridade.

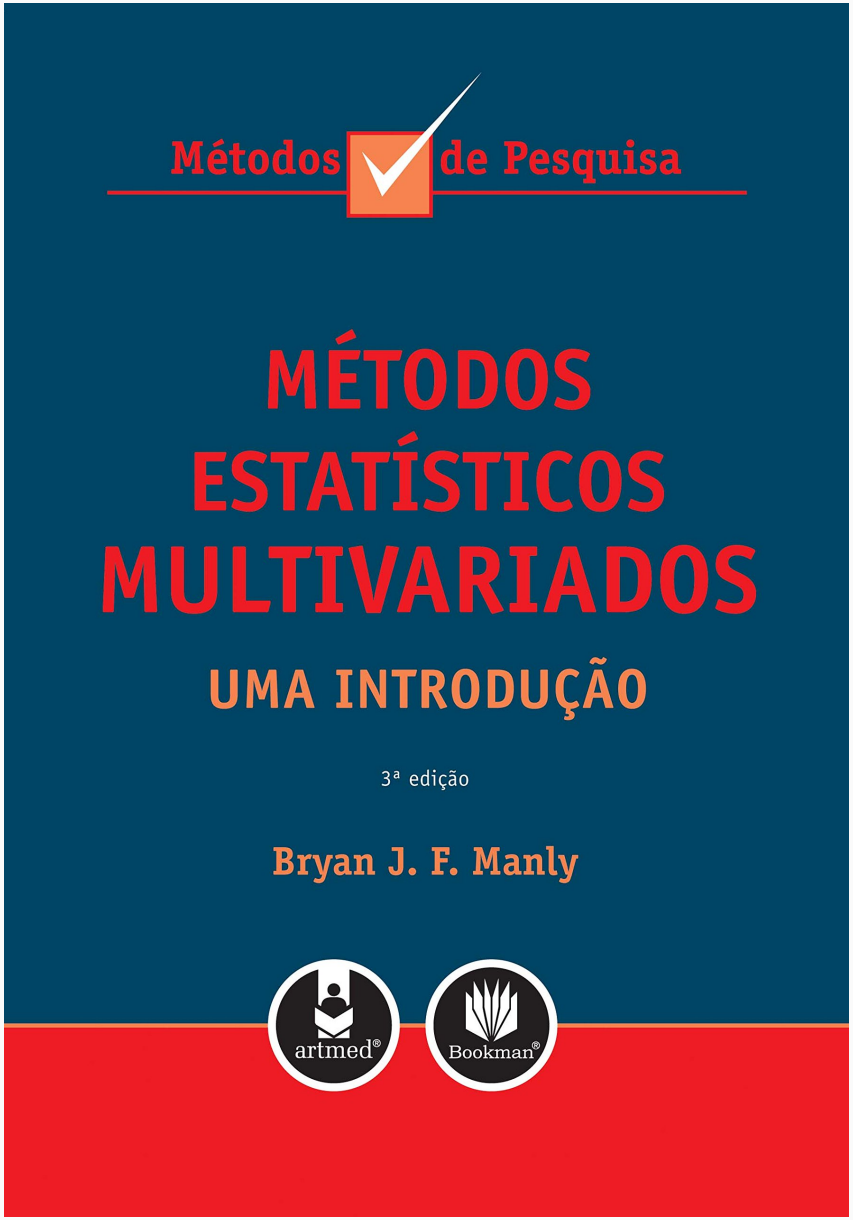
Comentários Finais

ATIVIDADE PARA ENTREGAR (E COMPOR A NOTA N1)

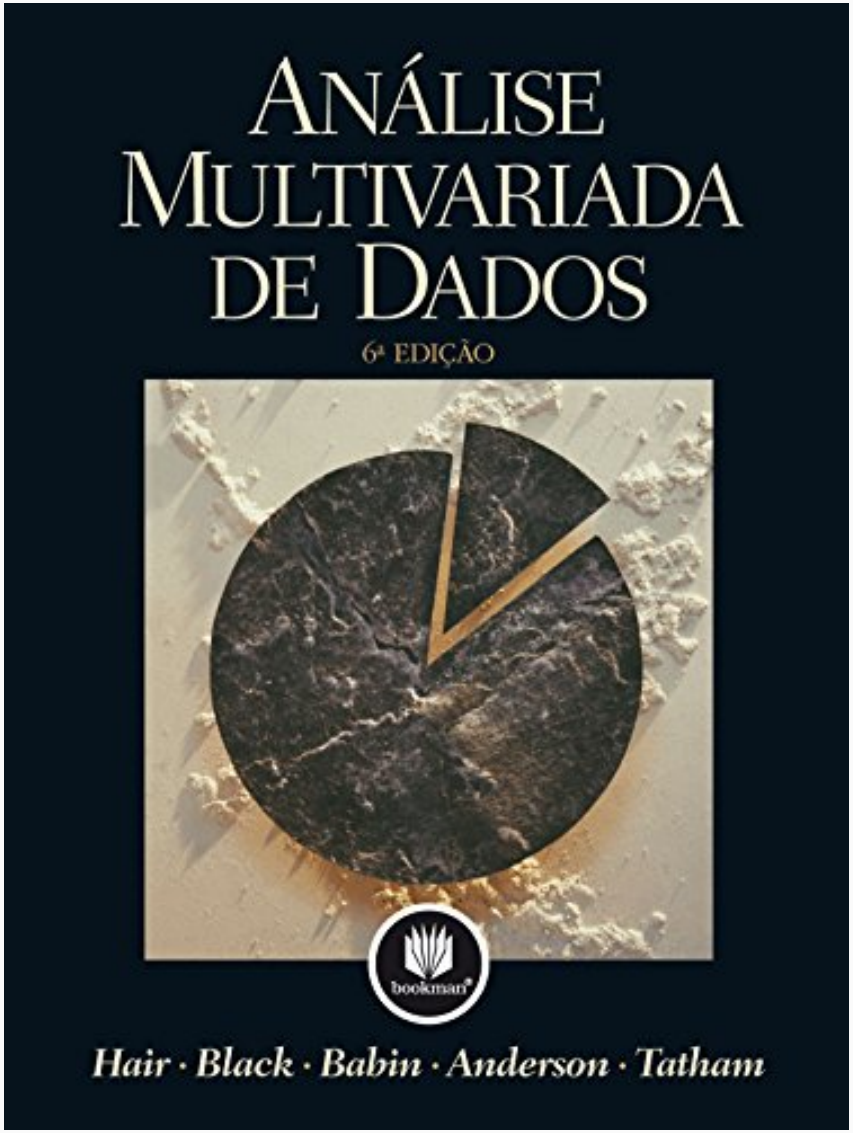
Resolva em grupos de até 4 integrantes os Exercícios 1.1, 1.2, 1.5.

Referências

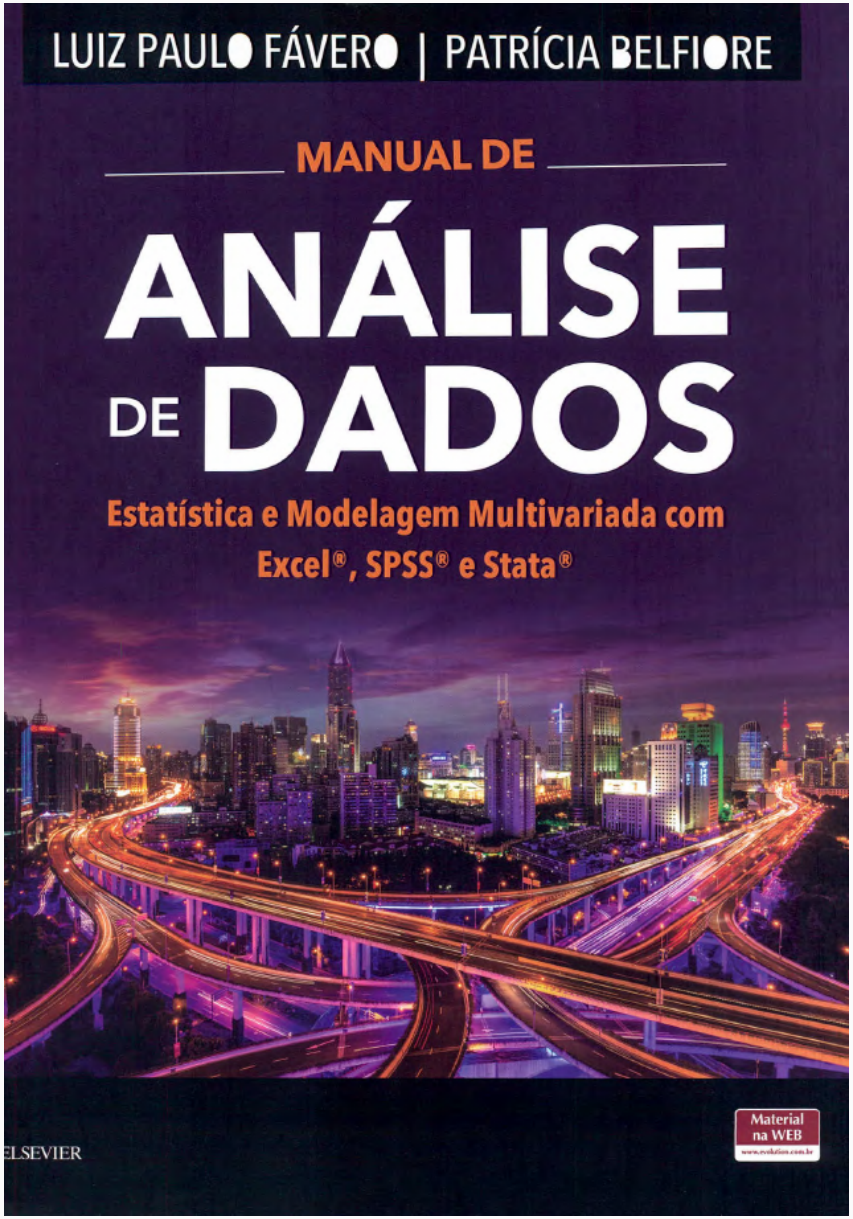
Referências



Referências



Referências



Bons Estudos!