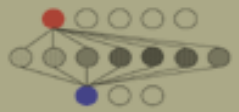


Reconhecimento de Padrões

Métodos, Técnicas e Ferramentas para Aprendizado e Classificação de Dados

Prof. Dr. rer.nat. Aldo von Wangenheim

The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR

Aula

2.1: Conceitos de Indução, Dedução e Técnicas estatísticas Exploratórias

The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR

Caracterísitcas: Reconhecimento de Padrões

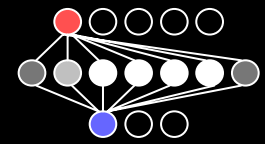
- Geralmente dados sob a forma de pares atributo-valor
  - atributo é definido pela posição do valor em um vetor de dados,
  - ex.: (1 2 4 11.01 1), representando: (num_imóveis, num_carros, num_filhos, idade_media_filhos, bom/mau pagador)
- Dados numéricos ou representações numéricas de dados simbólicos
  - ex: 1=fumante, 2=não fumante, 3=fumante eventual.
- Os relacionamentos entre os dados desconhecidos.
  - Se os padrões são pré-classificados em classes, essas geralmente são resultado da experiência de usuários humanos no domínio de aplicação.
- Objetivo quase sempre será o de reconhecer padrões
  - classificá-los, porém sem gerar explicações complexas sobre o porquê desta classificação.

Como vamos proceder para classificar padrões ?

- Já vimos: Se já sabemos quantas categorias possuímos e temos exemplos de instâncias de cada categoria: queremos criar um **classificador**
  - Vai utilizar nossos exemplos para classificar novos dados.
  - Exemplo: Banco tem base de casos de bons e maus pagadores de empréstimos e usa-a para classificar novos clientes
- Não vimos ainda: Se temos dados mas não sabemos como se organizam, temos de minerá-los: queremos criar um **agrupador**
  - O resultado desse agrupamento pode ser utilizado para classificações futuras.

2 formas de apresentação/uso dos padrões

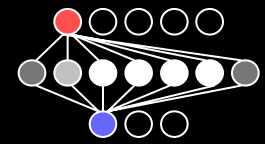
- Aprendizado Supervisionado
  - figura do professor, que apresenta os exemplos a serem aprendidos e controla a avaliação da qualidade do aprendizado
  - geração direta de classificadores
- Aprendizado Não-Supervisionado
  - estruturas existentes nos exemplos a serem aprendidos devem ser descobertas pelo aprendizado
  - dados precisam ser minerados



Forma de apresentação dos padrões

- Problemas Classificatórios:
  - conhecemos o domínio de aplicação bem o suficiente e sabemos de antemão quais as classes de padrões que vão existir.
  - objetivo do aprendizado é criar um classificador capaz de replicar este conhecimento e, eventualmente, refiná-lo.
    - Possuímos um conjunto de padrões pré-classificados que podemos utilizar como ponto de partida.
    - Ex.: Banco MaxiCred S.A., que possue coleção de descrições de clientes e de empréstimos que foram atendidos nos últimos anos e classificados como bom-pagador, médio-pagador e mau-pagador. O objetivo é integrar este conhecimento em um ssitema através de aprendizado de forma a futuramente poder utilizá-lo na classificação de novos clientes.

Aprendizado Supervisionado



Forma de apresentação dos padrões

- Problemas Exploratórios:
  - possuímos uma coleção de padrões representando situações ou configurações de dados
  - não possuímos nenhum julgamento *a priori* sobre estes dados:
  - não sabemos em quantas classes este conjunto se deixa dividir nem o significado de cada padrão.
- O nosso objetivo é analisar estes dados para descobrir uma divisão satisfatória em classes, de acordo com características dos mesmos, que ajudem a prover algum tipo de significado aos mesmos.

Aprendizado Não-Supervisionado

Processo de aprendizado/construção do classificador

- Processo de aprendizado:
  - novos conhecimentos são produzidos como conclusões do processo.
  - uma conclusão C pode ser sempre representada da seguinte forma: P1, ... , Pn -> C
    - onde P1, ... , Pn são o conjunto de premissas que justificam a conclusão.
    - podem ser conclusões lógicas, abstrações ou analogias.
  - Tipo de conhecimento gerado depende do modelo de inferência subjacente ao processo.
  - O modelo de inferência determina também a semântica de uma conclusão à qual um sistema de aprendizado chegou.

Processo de aprendizado

Aprendizado Sintético (indução)

Aprendizado Analítico (dedução)

Aprendizado por analogia

Processo de aprendizado:
Aprendizado analítico-> raciocínios dedutivos

- Vale:

"... *when you have eliminated the impossible, whatever remains, however improbable, must be the truth.*"

Sherlock Holmes

Processo de aprendizado: dedução

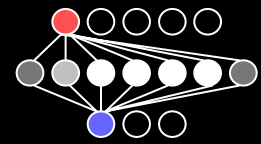
- Aprendizado analítico: raciocínios dedutivos.
  - Aqui, tudo o que se exige para que a conclusão seja verdadeira, é que todas as premissas também sejam verdadeiras.
    - Conhecemos as leis que amarram a conclusão às premissas.
  - Objeto de pesquisas em lógicas formais, como Lógica de Predicados, Lógica Modal, Lógica Temporal, etc.
  - Típicos em situações do tipo "Sherlock Holmes”:
    - possuímos conhecimento estruturado do domínio de aplicação de nosso problema e
    - aplicamos este conhecimento para refinar nossos conhecimentos sobre uma situação específica.

Processo de aprendizado: dedução

- Exemplos de raciocínios dedutivos
  - "Todo carro turbinado é veloz. Meu Lada possui um turbo. Logo meu Lada é veloz."
  - "Para matar a punhaladas é preciso ter acesso a um punhal. João é o único que possui um punhal. Logo João é o assassino."
- Raciocínio dedutivo: top-down
  - usa-se conhecimento abstrato, genérico (ex.: "Todo carro turbinado é veloz."), para provar a conclusão.
  - Dedução é utilizada em Raciocínio Baseado em Modelos
  - Sem uma Teoria do Domínio de Aplicação não há dedução.

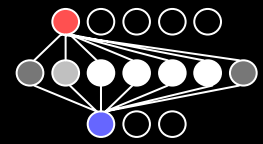
Processo de aprendizado: dedução

- Não existe raciocínio dedutivo sem que haja um **modelo** do mundo subjacente ao processo !



Processo de aprendizado
Aprendizado sintético -> indução

- Aprendizado sintético: raciocínios indutivos
  - permitimos que fatos observados **induzam** a **síntese** de conclusões abstratas sobre a natureza do observado.
  - Idéia básica: uma conclusão **C** é a descrição, abstrata, geral de fenômeno que obedece a certas leis, sintetizadas através de **C**.
- A indução
  - primeiro passo na descoberta de novos fenômenos naturais.
  - observa-se que sempre que determinada situação ocorre, determinado fenômeno é observável. O
    - bservação induz à conclusão de que o fenômeno está associado à situação, sem no entanto explicar a natureza desta associação.



Processo de aprendizado

- O aprendizado sintético é o mais utilizado em RP:
  - facilidade de aplicação a um domínio cujo objetivo é sempre encontrar regras gerais que nos permitam classificar um conjunto de padrões cujos interrelacionamentos muitas vezes não conhecemeos de antemão.
  - No aprendizado sintético pode se aprender fatos e ralacionamentos de natureza diversa:
    - Conceitos (superconceitos) através da generalização sobre exemplos conhecidos.
    - Fatos sobre as quais se conhece apenas o relacionamento entre entradas e saídas, mas não a relação entre causa e efeito como regra geral de formação do efeito.
    - Probabilidades para a ocorrência de determinadas situações.

Foco desta parte da Disciplina

Como é realizado computacionalmente o Raciocínio Indutivo ?

- Técnicas Clássicas de Mineração de Dados
  - Técnicas da Estatística Exploratória ou Estatística Multivariada
  - Baseiam-se em diferentes formas de interpretar o processo de Análise da Variância
- Técnicas de Aprendizado de Máquina
  - Técnicas conexionistas de Aprendizado Não-Supervisionado

Estatística Multivariada

Técnicas úteis para reconhecimento e descoberta de padrões em ambientes onde os fenômenos são descritos/baseados em uma grande variedade de dados são conhecidas como:

- Análise de Dados Exploratória (ADE) ou
- Estatística Exploratória.

O que é Análise de Dados Exploratória (ADE) ?

- Relacionada de forma próxima com o conceito de Mineração de Dados.
- Contrário dos testes de hipóteses tradicionais:
  - Projetados para verificar uma hipótese a priori acerca de relacionamentos entre variáveis
  - Ex.: "Existe uma correlação positiva entre a IDADE de uma pessoa e o NÍVEL_DE_VIOLÊNCIA dos filmes locados em uma locadora ?"
- Utilizada para a identificação de relacionamentos sistemáticos entre variáveis quando não existem expectativas a priori acerca da natureza destes relacionamentos ou estas são incompletas.
  - Tipicamente muitas variáveis diferentes são consideradas e comparadas.
  - Grande variedade de técnicas e modelos matemáticos com o objetivo de se encontrar padrões nestes dados.

O que é um conjunto de dados multivariado ?

- Até agora:
  - Um conjunto de dados multivariado era um conjunto de dados onde cada caso ou observação de um fenômeno era descrito por um conjunto de várias variáveis, sendo representado tipicamente por um padrão (n+1)-dimensional, onde n>2 é número de variáveis necessárias para descrever o fenômeno ou observação e a variável n+1 descreve a classe à qual este determinado padrão pertence.
- No mundo da Estatística:
  - faz-se uma diferenciação rigorosa se um fenômeno é baseado em uma, duas ou muitas variáveis (mono-, bi- ou multivariado)
  - pois as técnicas estatísticas utilizadas para cada um desses três casos variam muito e são tratadas separadamente
  - vamos trabalhar com mais de duas variáveis

Como aplicamos ADE ao RP ?

- 2 formas diferentes:

- **Indução**: As técnicas de ADE são técnicas de RP, já que são projetadas para detectar regularidades, correlações e fatores agrupadores ou diferenciadores em um conjunto de dados.
  - Realizar mineração de dados ou aplicar ADEs a um problema, já é uma forma de realizar reconhecimento de padrões;
  - Aplicação das técnicas de ADE em RP sob esta ótica dispensa maiores explicações, basta que aprendamos as técnicas.

- **Construção de Classificadores**: uso para a extração de informações com o objetivo de utilizá-las para a implementação de um classificador.
  - O desenvolvimento completo de uma solução envolve mais do que as técnicas de ADE em si;
  - Exige que se utilize uma técnica adicional para a implementação do classificador.

Filosofia: ADE + RP

Passos:

- Os resultados da aplicação de técnicas da estatística exploratória a um fenômeno utilizados para o desenvolvimento de classificadores
- Classificadores a posteriori utilizados para classificar novos dados produzidos pelo mesmo fenômeno anteriormente analisado.

Filosofia:

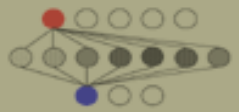
- Análise inicial com técnicas exploratórias provê dados para a elaboração de um mecanismo de classificação utilizando técnicas tradicionais, como k-NN:
  - através de informações sobre distribuições de dados ou variáveis-chave para classificação dos dados em classes.
- Onde não conhecemos a priori em quais e quantas classes os dados se permitem agrupar, pode-se determinar estas classes e utilizar esta informação para um posterior mecanismo de classificação.

Resumo da Filosofia:

1. **Exploração**: Primeiramente escolhemos uma técnica da ADE para gerar um conjunto de informações a partir de um conjunto inicial de dados gerado por um processo que desejamos dominar capaz de servir para utilização em um classificador e subseqüentemente ser utilizado para a classificação de novos casos gerados pelo mesmo processo que gerou os dados originais;
2. **Utilização**: Escolhemos uma técnica de RP adequada ao tipo de informação gerada pelo método de ADE utilizado e também adequada ao tipo de classificação que queremos obter para dados futuros e utilizamos a informação gerada pela ADE para alimentar ou implementar o classificador.

Técnicas clássicas de AED

- Análise de Agrupamentos (Cluster Analysis),
- Árvores de Classificação/Decisão (Classification Trees).
- Análise de Discriminantes (Discriminant Function Analysis),
- Análise de Séries Temporais (Time Series Analysis),
- Análise Fatorial (Factor Analysis),
- Análise de Correspondências (Correspondence Analysis),
- Escalonamento Multidimensional (Multidimensional Scaling),
- Análise Log-Linear (Log-linear Analysis),
- Correlação Canônica (Canonical Correlation),
- Regressão Parcialmente Linear e Não-Linear (Stepwise Linear and Nonlinear Regression),



2.2. Análise de Agrupamentos e Árvores de Decisão e Classificação

Parte I: Introdução e Unificação ou Agrupamento em Árvore

The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR

Análise de Agrupamentos

- Termo Análise de Agrupamentos primeiramente usado por (Tyron, 1939)
  - Variedade de algoritmos de classificação diferentes, todos voltados para uma questão importante em várias áreas da pesquisa: Como organizar dados observados em estruturas que façam sentido, ou como desenvolver taxonomias capazes de classificar dados observados em diferentes classes.
  - Importante é considerar inclusive, que essas classes devem ser classes que ocorrem "naturalmente" no conjunto de dados.
- Biólogos, por exemplo, têm de organizar dados observados em estruturas que "façam sentido", ou seja, desenvolver taxonomias:
  - Zoologistas confrontados com uma variedade de espécies de um determinado tipo têm de conseguir classificar os espécimes observados em grupos antes que tenha sido possível descrever-se esses animais em detalhes de formas a se destacar detalhadamente as diferenças entre espécies e subespécies.

Exemplo clássico: F

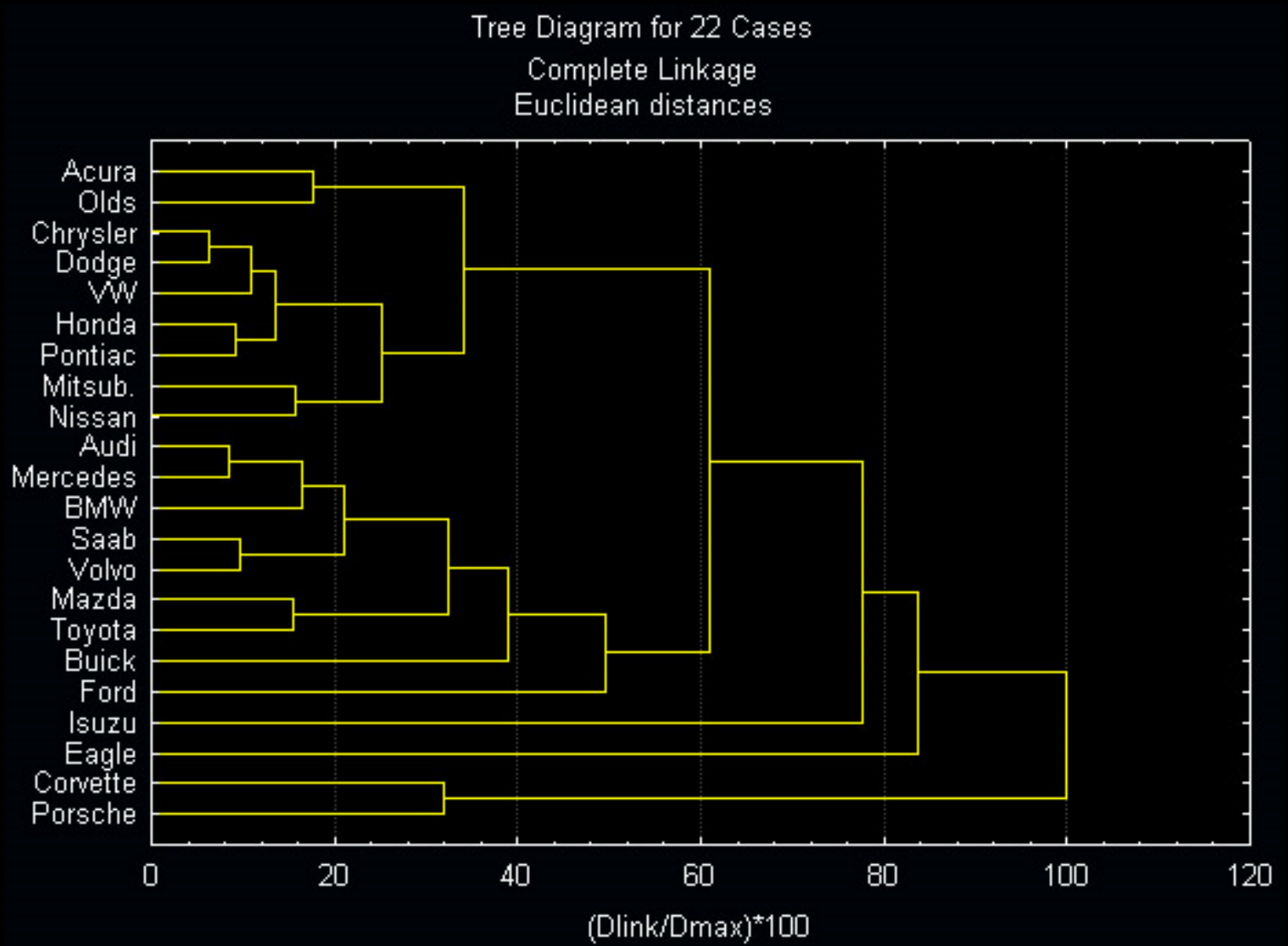
Em quantas espécies se cl

	CompSépalas	LargSépalas	CompPétalas	LargPétalas	TipoIris
1	5	3,3	1,4	0,2	SETOSA
2	6,4	2,8	5,6	2,2	VIRGINIC
3	6,5	2,8	4,6	1,5	VERSICOL
4	6,7	3,1	5,6	2,4	VIRGINIC
5	6,3	2,8	5,1	1,5	VIRGINIC
6	4,6	3,4	1,4	0,3	SETOSA
7	6,9	3,1	5,1	2,3	VIRGINIC
8	6,2	2,2	4,5	1,5	VERSICOL
9	5,9	3,2	4,8	1,8	VERSICOL
10	4,6	3,6	1	0,2	SETOSA
11	6,1	3	4,6	1,4	VERSICOL
12	6	2,7	5,1	1,6	VERSICOL
13	6,5	3	5,2	2	VIRGINIC
14	5,6	2,5	3,9	1,1	VERSICOL
15	6,5	3	5,5	1,8	VIRGINIC
16	5,8	2,7	5,1	1,9	VIRGINIC
17	6,8	3,2	5,9	2,3	VIRGINIC
18	5,1	3,3	1,7	0,5	SETOSA
19	5,7	2,8	4,5	1,3	VERSICOL
20	6,2	3,4	5,4	2,3	VIRGINIC
21	7,7	3,8	6,7	2,2	VIRGINIC
22	6,3	3,3	4,7	1,6	VERSICOL
23	6,7	3,3	5,7	2,5	VIRGINIC
24	7,6	3	6,6	2,1	VIRGINIC
25	4,9	2,5	4,5	1,7	VIRGINIC
26	5,5	3,5	1,3	0,2	SETOSA
27	6,7	3	5,2	2,3	VIRGINIC
28	7	3,2	4,7	1,4	VERSICOL
29	6,4	3,2	4,5	1,5	VERSICOL
30	6,1	2,8	4	1,3	VERSICOL
31	4,8	3,1	1,6	0,2	SETOSA
32	5,9	3	5,1	1,8	VIRGINIC

Análise de Agrupamentos

- A idéia aqui é a de um processo *data-driven*, ou seja, dirigido pelos dados observados de forma a agrupar esses dados segundo características comuns que ocorram neles.
- Processo deve levar em conta a possibilidade de se realizar uma organização hierárquica de grupos:
  - a cada nível de abstração maior, são também maiores as diferenças entre elementos contidos em cada grupo,
  - Ex.: espécies animais do mesmo gênero têm muito em comum entre si, mas espécies animais que possuem apenas o filo ou a ordem em comum possuem pouca similaridade.

Exemplo de hierarquia: Carros



Análise de Agrupamentos: Significância Estatística

- Discussão comum e importante na Estatística.
- Na ADE as discussões até o momento não mencionaram a questão da significância estatística ou de seu teste.
- O ponto aqui é que, utilizamos métodos de análise de agrupamentos quando não possuímos nenhuma hipótese a priori sobre a estrutura ou comportamento de nosso dados e necessitamos iniciar com alguma coisa.
  - Por que então não deixar um software descobrir quais regularidades são interessantes no conjunto dos dados ?
  - Por causa disso, testes de significância estatística ainda não são apropriados nesta altura do campeonato.

Análise de Agrupamentos: Áreas de Aplicação

- Agrupamento tem sido aplicado em uma enorme gama de áreas:
  - (Hartigan 1975) já provê uma visão geral ampla de vários estudos publicados acerca da utilização de técnicas de Análise de Agrupamentos.
- Na área médica, por exemplo, agrupamento de doenças por sintoma ou curas pode levar a taxonomias muito úteis.
  - Em psiquiatria p.ex. considera-se que o agrupamento de sintomas como paranóia, esquizofrenia e outros é essencial para a terapia adequada.
- Na arqueologia, também se tem tentado agrupar civilizações ou épocas de civilizações com base em ferramentas de pedra, objetos funerários, etc.
- De forma geral, toda vez que se faz necessário que se classifique uma "montanha" de dados desconhecidos em pilhas gerenciáveis, se utiliza métodos de agrupamento.

Análise de Agrupamentos: Métodos de Agrupamento

- Há 2 algoritmos de agrupamento de dados baseados em métodos estatísticos interessantes para efeitos de classificação de padrões:
  - Unificação ou Agrupamento em Árvore
  - Agrupamento por k-Médias
- Vamos analisar e discutir cada um dos dois e vamos, ao final, discutir como utilizar os resultados da aplicação destes métodos para o particionamento de um grupo de dados cujo comportamento intrínseco ainda não conhecemos na confecção de sistemas de reconhecimento de padrões que sejam capazes de automaticamente classificar novas observações em uma das classes "detectadas" por um destes dois métodos.

Unificação ou Agrupamento em Árvore

- A agrupamento em árvore (*Tree Clustering*) tem por objetivo a construção de taxonomias de vários níveis
  - Identificação de atributos comuns
- Ele é considerado um método de agrupamento aglomerativo hierárquico.

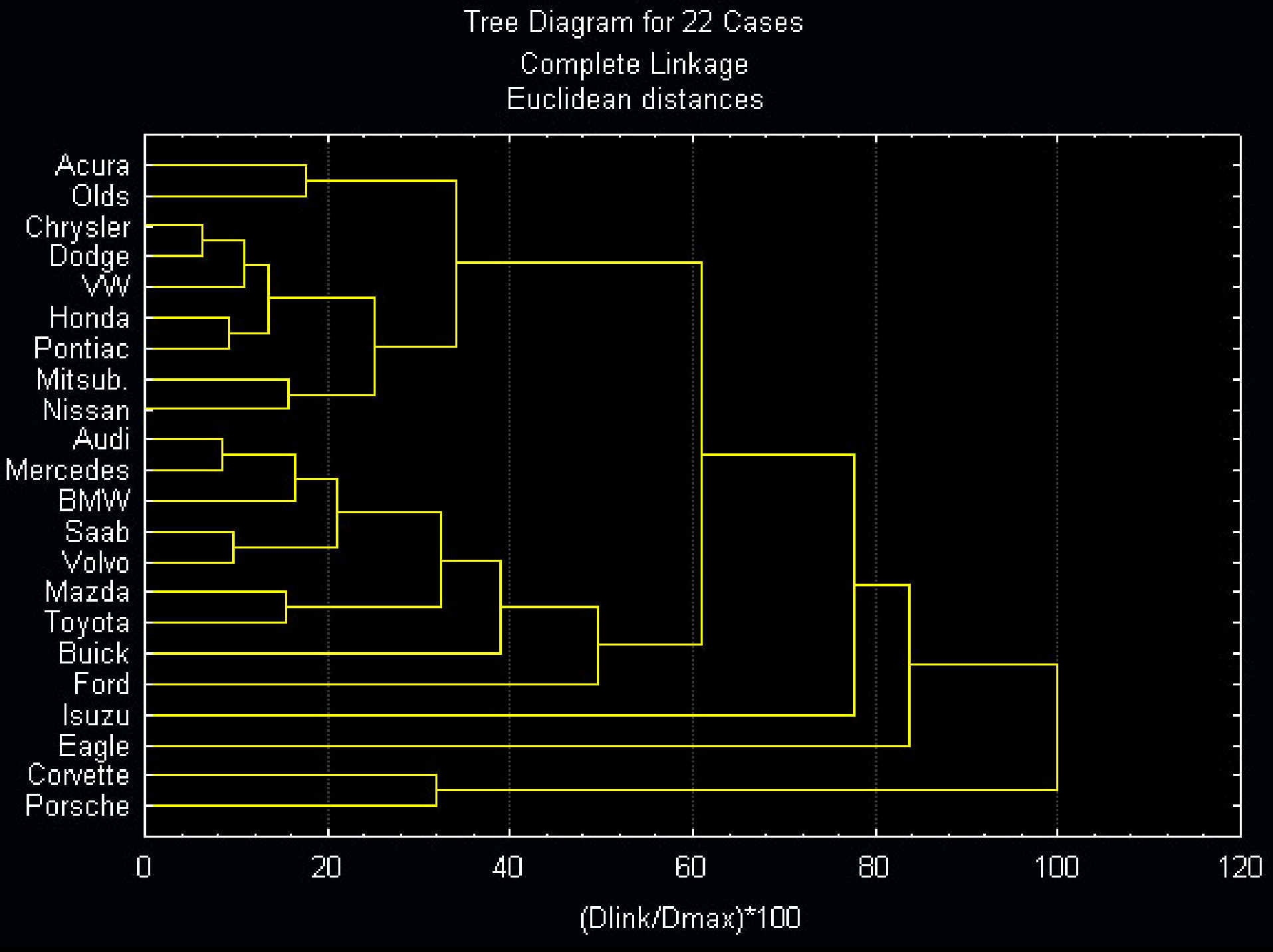
Ex.: Em quantas classes sensatas podemos dividir este conjunto de carros ?

Data: Cars.sta* (5v by 22c)

	Performance, fuel economy, and approximate price for various automobiles				
	1 Preço	2 Aceleração	3 Frenagem	4 Manutenção	5 Kilometragem
Acura	-0,521072363	0,477252671	-0,00657103855	0,381619066	2,07875356
Audi	0,865652474	0,208033216	0,31869537	-0,0913735792	-0,677061608
BMW	0,495859184	-0,801539742	0,192202878	-0,0913735792	-0,153805564
Buick	-0,613520685	1,68874022	0,933087475	-0,20962174	-0,153805564
Corvette	1,23544576	-1,8111127	-0,494470651	0,972859872	-0,677061608
Chrysler	-0,613520685	0,0734234878	0,427117506	-0,20962174	-0,153805564
Dodge	-0,705969008	-0,195795968	0,481328574	0,145122743	-0,153805564
Eagle	-0,613520685	1,21760617	-4,19889364	-0,20962174	-0,677061608
Ford	-0,705969008	-1,54189324	0,987298543	0,145122743	-1,7235737
Honda	-0,42862404	0,409947807	-0,00657103855	0,0268745821	0,369450479
Isuzu	-0,79841733	0,409947807	-0,0607821066	-4,23005922	1,0671252
Mazda	0,126065894	0,679167263	-0,133063531	0,499867227	-1,7235737
Mercedes	1,05054912	0,00611862399	0,119921454	-0,0913735792	-0,153805564
Mitsub.	-0,613520685	-1,00345433	0,0837807415	0,381619066	0,718287842
Nissan	-0,42862404	0,0734234878	-0,00657103855	0,263370905	0,997357732
Olds	-0,613520685	-0,734234878	0,40904715	0,381619066	2,11363729
Pontiac	-0,613520685	0,679167263	0,535539642	0,145122743	0,195031798
Porsche	3,4542055	-2,21494188	-0,295696735	0,618115388	-1,02589897
Saab	0,588307506	0,679167263	0,246413946	0,263370905	0,0206131169
Toyota	-0,0588307506	1,21760617	0,22834359	0,73636355	-0,851480289
VW	-0,705969008	-0,128491104	0,101851098	0,381619066	0,195031798
Volvo	0,218514217	0,611862399	0,13799181	-0,20962174	0,369450479

Unificação ou Agrupamento em Árvore

- O objetivo deste algoritmo é o de unificar objetos em classes ou grupos sucessivamente maiores através da utilização de alguma medida de similaridade ou de distância.
- O resultado deste enfoque é uma árvore hierárquica, como no exemplo do dendrograma visto anteriormente:

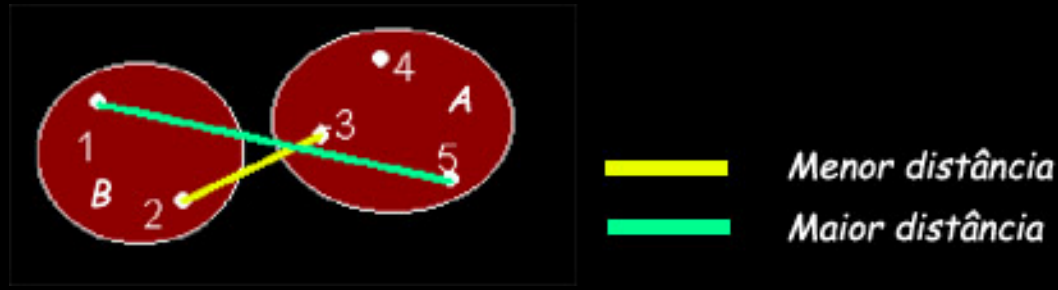

Tree Diagram for 22 Cases
Complete Linkage
Euclidean distances
Acura
Olds
Chrysler
Dodge
VW
Honda
Pontiac
Mitsub.
Nissan
Audi
Mercedes
BMW
Saab
Volvo
Mazda
Toyota
Buick
Ford
Isuzu
Eagle
Corvette
Porsche
0
20
40
60
80
100
120
(Dlink/Dmax)*100

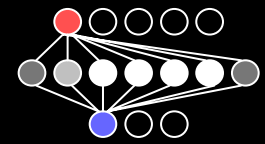
Processo de Unificação ou Agrupamento em Árvore

- Para construir a árvore utilizamos alguma medida de distância entre classes:
  - Chamamos esta distância de **distância de conexão** ou *linkage distance.*

Há três filosofias de análise da distância de conexão:

- Simples - consideramos a distância entre os vizinhos mais próximos como a distância entre agrupamentos.
  Neste caso no exemplo abaixo $d(A, B) = d(2,3)$
- Completa - consideramos a distância entre os vizinhos mais distantes como a distância entre agrupamentos.
  Neste caso no exemplo abaixo $d(A, B) = d(1,5)$
- Média - Consideramos a distância média segundo a fórmula adiante como a distância entre agrupamentos.





Processo de Unificação ou Agrupamento em Árvore

- A fórmula para cálculo da distância média $d_{\text{média}}$ é dada pela média das distâncias entre todos os pares de pontos:

$$\overline{d} = \frac{d(1,3) + d(1,4) + \ldots + d(2,5)}{6}$$

- Montagem da árvore de classificação ou dendrograma:
  - procedemos unindo sempre grupos apresentando a menor distância de acordo com uma das três regras anteriores.
  - Veja os exemplos adiante para esclarecer suas dúvidas:

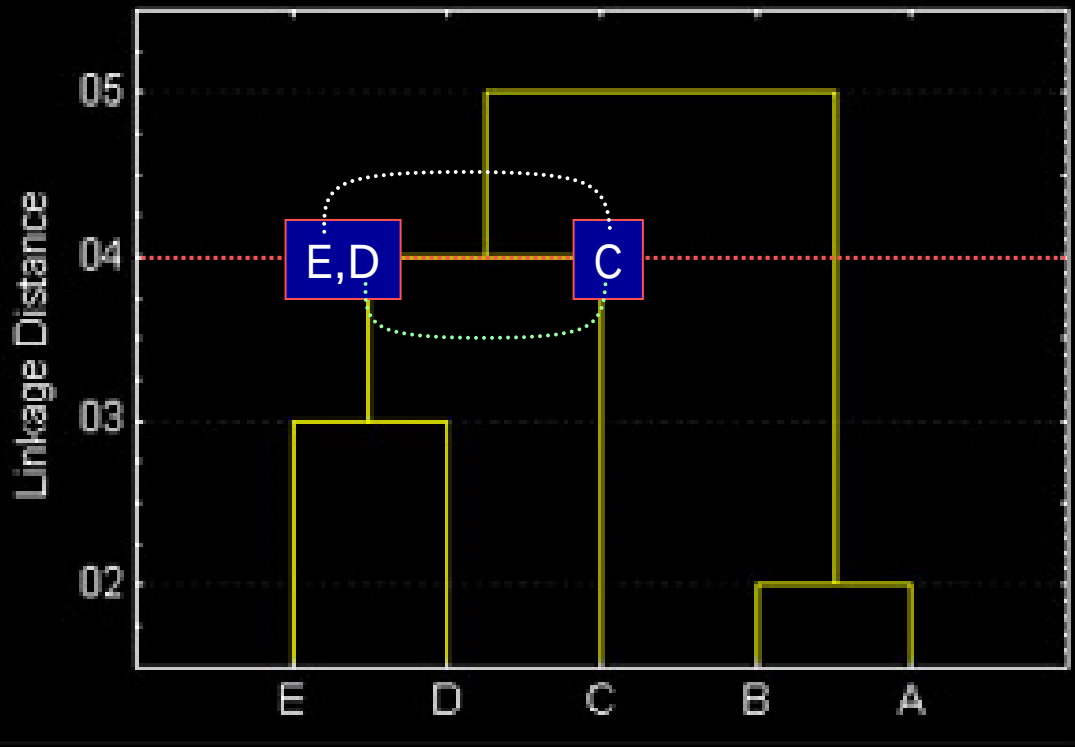
Processo de Unificação ou Agrupamento em Árvore

Simples - vizinho mais próximo

Matriz de distâncias

	A	B	C	D	E
A					
B	2				
C	6	5			
D	10	9	4		
E	9	8	5	3	

Dendograma

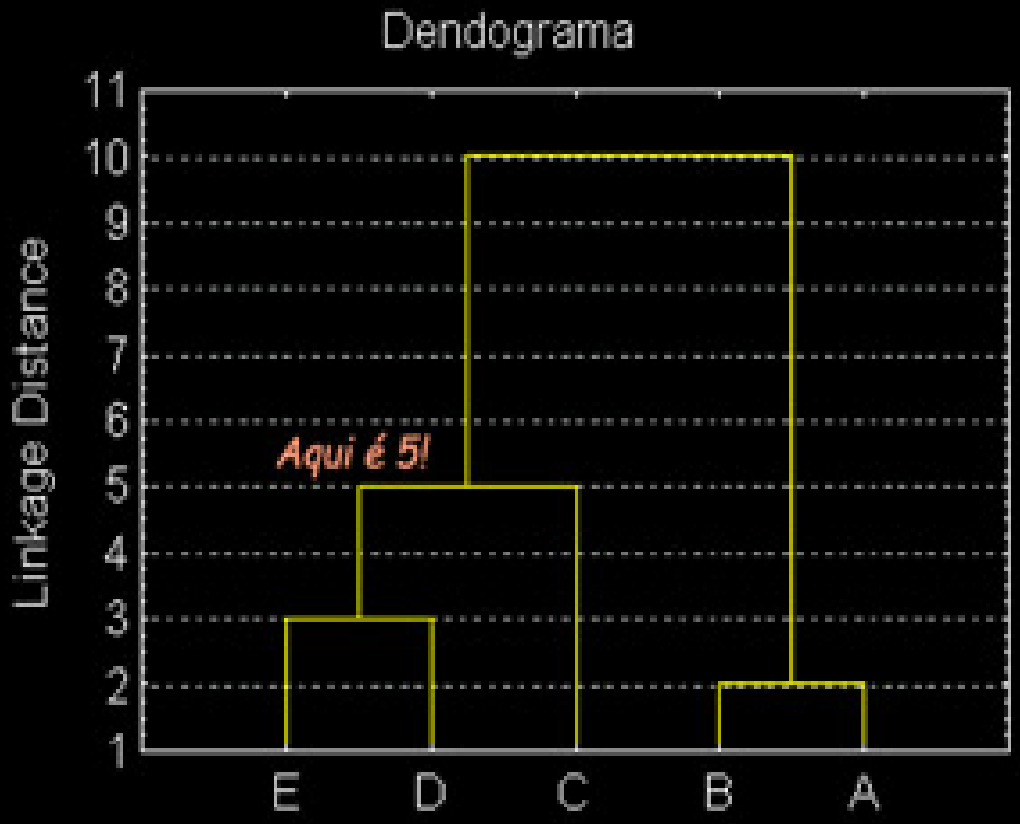


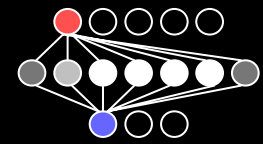
Processo de Unificação ou Agrupamento em Árvore

Completa (complete linkage) - vizinho mais distante

Matriz de distâncias

	A	B	C	D	E
A					
B	2				
C	6	5			
D	10	9	4		
E	9	8	5	3	





Processo de Unificação ou Agrupamento em Árvore

- Para utilizarmos o método para nos dar um determinado conjunto de classes:
  - percorremos a árvore a partir da raiz, até termos o número de classes que nos agrada mais.
  - No caso dos carros, se percorrermos a partir da raiz em direção às folhas e pararmos em Dlink/Dmax = 0.7, teremos 4 classes, dadas cada qual pelo seu ramo correspondente.
- Número “certo” de classes:
  - Quanto mais para a direita no diagrama de árvore, maiores as distâncias de conexão: Diversidades intra-agrupamentos cada vez maiores
  - Solução: Se gráfico mostra um platô claro, significa que muitos clusters foram formados a aproximadamente a mesma distância de conexão. Esta distância poderia ser um local de corte para a divisão dos dados em grupos ou classes.

Considerações: Estandardização dos Dados

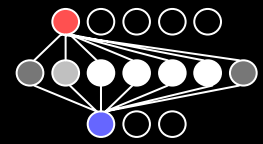
- O valor estandardizado de uma distribuição de dados é também chamado de *z-score* ou valor transformado.
- Para a maioria dos métodos estatísticos se utiliza a estandardização dos dados.
- A estandardização é diferente da normalização dos dados, onde se objetiva que cada variável se encontre em uma faixa de valores no intervalo [0,1].
- Na estandardização a faixa de valores pode variar e depende do desvio padrão e se utiliza a fórmula abaixo:

$$x_{estandardizado} = (x - \text{média})/\text{desvio-padrão}$$

calculada, como na normalização, de forma independente para cada variável.

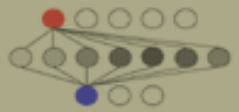
Considerações: Estandardização dos Dados

- Toma informações sobre a média e o desvio padrão de uma variável e produz um valor correspondente a cada valor original que especifica a posição deste valor original dentro da distribuição original de dados.
- O valor original, independentemente de pertencer a uma distribuição de dados que inclua valores negativos, é transformado em um valor com sinal, de tal forma que:
  - O sinal indique se o valor original , também chamado de score, está acima (+) ou abaixo (-) da média.
  - O módulo do valor estandardizado indique a distância entre entre a média e o valor original em termos de desvios-padrão.
- Dessa forma, um valor estandardizado de -2,4 representa um valor original que se encontra mais de dois desvios-padrão abaixo da média, representando um valor bem difícil de ocorrer de de probabilidade muito baixa.



Trabalho (entrega: próxima aula)

- Implemente o Método de Unificação ou Agrupamento em Árvore
  - O sistema deve ser capaz de ler um conjunto qualquer de dados em formato texto, por exemplo separado por tabs
  - Deve possuir interface gráfica que apresente o dendrograma gerado
  - Bole um “analisador de dendrograma” que encontre o local ótimo de corte, dados dois limites: maxClass e minClass
- Para testes tome um conjunto de quatro sets de dados, dentre estes:
  - (2) Os dados da flor do Gênero Iris e dos carros disponíveis na página
  - (1) Os dados de câncer cerebral (gliomas): http://www.inf.ufsc.br/~awangenh/RP/glioma-daumas-duport.xls
  - (1) Outro conjunto qualquer que você deverá procurar nos ”Links Úteis”.



2.2. Análise de Agrupamentos e Árvores de Decisão e Classificação

Parte II: Agrupamento por k-Médias

The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR

Prof. Dr. rer.nat. Aldo von Wangenheim

k-Médias (k-Means)

- Método não-hierárquico por repartição
  - muito diferente do método em Árvore
- Pressuposto: você já tem as hipóteses a respeito do número de conjuntos em seus casos ou variáveis
  - Você tenta formar exatamente *k* conjuntos que devem ser tão distintos quanto o possível.
  - Este é o tipo de pesquisa que pode ser feita pelo algoritmo de aglomeramento por k-Médias.
  - O método k-Médias produzirá exatamente k diferentes conjuntos com a maior distinção possivel entre eles.

Ilustrando: k-Médias

- Veja exemplo dos carros
  - Pesquisador pode ter um "pressentimento" da experiência de análises de mercado que os carros caem basicamente em três categorias diferentes no que diz respeito à relação custo-benefício.
    - Ele pode querer saber se esta intuição pode quantificada
  - **Hipótese**: análise de agrupamento por k-Médias das medidas da relação custo-benefício dada pelas variáveis descritoras dos carros produzirá os três conjuntos de marcas de carros como esperados.
    - Médias das diferentes medidas de relação custo-benefício (frenagem, manutenibilidade, etc) para cada conjunto representam uma maneira quantitativa de expressar a hipótese ou intuição do pesquisador.

Método: k-Médias

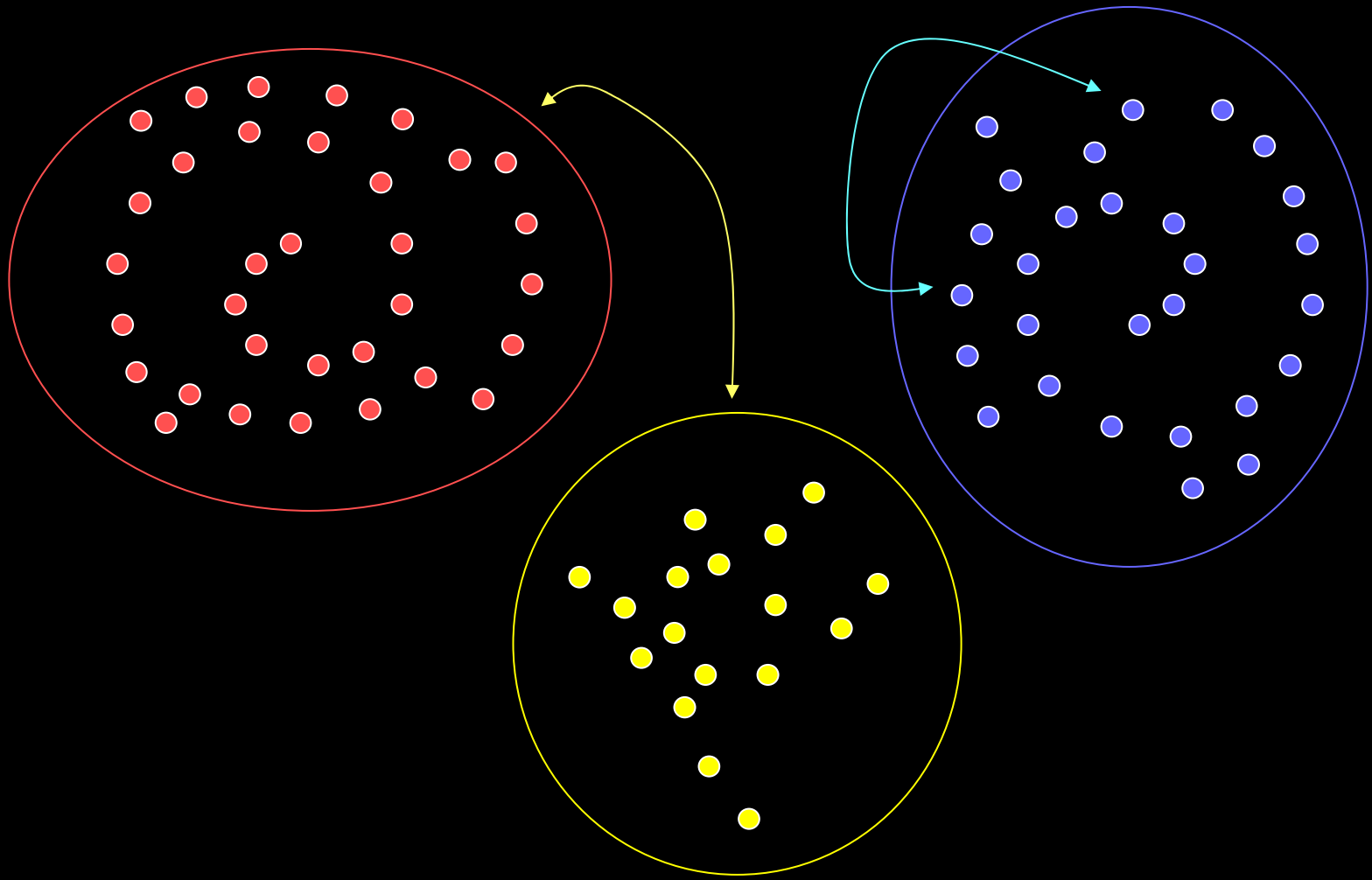
- Inicia-se com k-conjuntos aleatórios.
- Move-se os objetos entre estes conjuntos com o objetivo de:

  (1) minimizar o variabilidade dentro dos conjuntos e

  (2) maximizar o variabilidade entre conjuntos.
- Isto é semelhante à Análise de Variância ANOVA ao contrário
  - teste de significancia ANOVA avalia a variabilidade entre-grupos com a variabilidade intra-grupo: hipótese de que as medias dos grupos são diferentes para cada grupo.
  - em k-Médias o método move objetos entre grupos para ter resultados ANOVA mais significativos

Método: k-Médias

Maximizar: variâncias σ_g inter-grupos Minimizar: variâncias σ_i intra-grupos



Algoritmo Básico do Método das k-Médias

1. Padronize todos os dados

   Descreva cada variável em termos de distância de seu valor em desvios-padrão da sua média.

2. Fixe o número de agrupamentos desejado = k

3. Divida os casos aleatoriamente em k grupos

4. Calcule o centróide de cada grupo

5. Com os dados padronizados, calcule, para cada caso, a distância euclidiana em relação ao centróide de cada grupo;

6. Transfira o caso para o grupo cuja distância ao centróide é mínima

7. Repita (4), (5) e (6) até que nenhum caso seja mais transferido.

Características: k-Médias

- Algoritmo rápido
  - apesar de a média de cada grupo ser modificada por cada movimentação de um elemento e ter de ser recalculada para os dois grupos afetados, o processamento é gerenciável
- Pode ser considerado uma espécie de descida em gradiente
  - nenhum mínimo local descrito na literatura

Avaliação de resultados: k-Médias

- Duas perguntas:
  - Qual o melhor *k* ?
    - Como eu comparo diferentes agrupamentos gerados pelo método ?
  - O “melhor” *k* que eu achei é realmente “bom” ?
    - Posso rejeitar a hipótsese nula de que não existe um agrupamento “natural” dos dados em *k* classes ?
- Necessito de uma métrica:
  - Teste F de Fisher

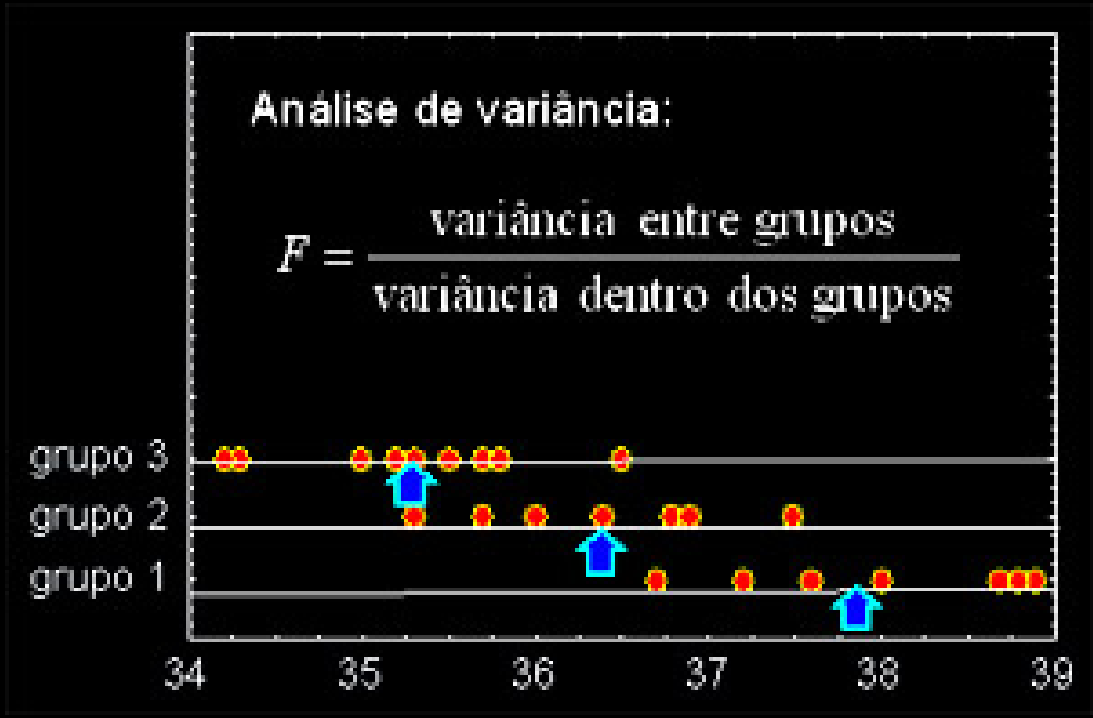
Avaliação de resultados: k-Médias

- Comparar:
  - a variância intracluster (que deverá ser pequena se a divisão em classes for adequada)
  - à variância inter-clusters (que deverá ser grande se a classificação em categorias for boa).
  - Isto significa que uma boa divisão de um conjunto de observações em grupos ou categorias é aquela onde os elementos de uma mesma categoria são o mais parecidos entre si (menor variância intra-cluster) e onde os elementos de grupos diferentes são o mais diferentes entre si possível (variância inter-cluster ou inter-grupos). Isto é dito verificar a robustez dos grupos de objetos ou categorias geradas.
- Realiza-se uma de variância padrão inter-grupos.
  - Para um dado k e para cada variável, uma medida de discriminação entre grupos é dada pelo cálculo das variâncias inter- e intra-grupos e pelo coeficiente de discriminação F

Avaliação de resultados: k-Médias

- Um **Teste-F** é qualquer teste estatístico que possui **distribuição-F** se a hipótese nula for verdadeira. Os principais são:
  - A hipótese de que as médias de múltiplas populações distribuídas normalmente, todas apresentando o mesmo desvio padrão, são iguais. Esta é provavelmente a mais conhecida hipótese testada través de um teste-F e é o problema básico na Análise de Variância (ANOVA).
  - A hipótese de que os desvios padrão de duas populações são iguais e que conseqüentemente elas são de origem comparável.
- Importante:
  - Se é a igualdade de variâncias ou de desvios-padrão que está sendo testada, o teste-F é extremamente pouco robusto para distribuições diferentes da normalidade. Se os dados demonstram desvios da normalidade, mesmo que pequenos, desaconselha-se o uso do teste-F.

Avaliação de resultados: k-Médias

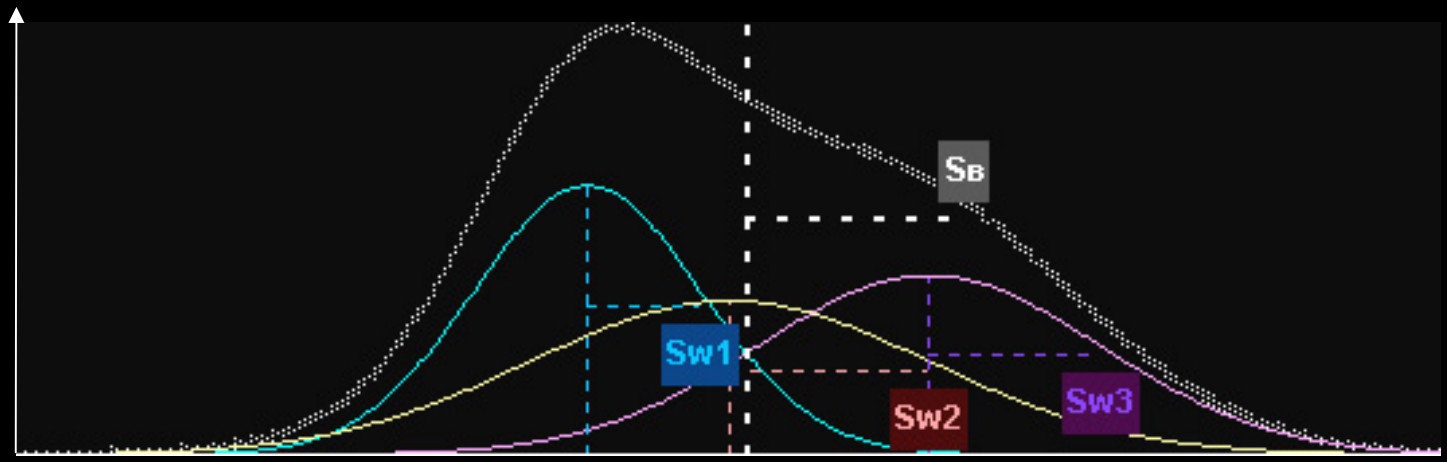


ou $F = \frac{S_B^2}{S_w^2}$

O Teste-F nos dá a taxa entre as variâncias entre-grupos (S_B) e intra-grupos (S_w).
A hipótese nula (“não há grupos”) é rejeitada quando o F calculado encontra-se acima de um valor crítico para um determinado par de graus de liberdade e um α de significância.

Entendendo o Significado do Valor de F calculado

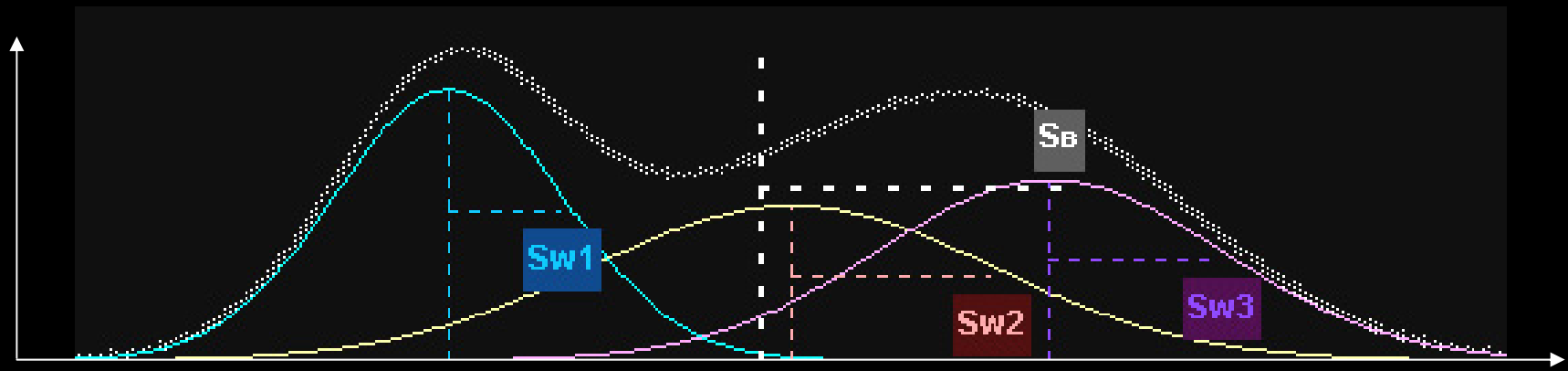
- Considere três amostras (Sw_1, Sw_2 e Sw_3) representadas em turquesa, amarelo e lilás na figura:



- Cada amostra tem sua média (linha vertical pontilhada) e medida de dispersão (Sw_1, Sw_2 e Sw_3), dada pelo seu desvio-padrão.
- Podemos imaginar que, tomando-se as três amostras conjuntamente, exista uma média geral, com sua respectiva medida de dispersão (S_B), dada pelo desvio-padrão global.

Entendendo o Significado do Valor de F calculado

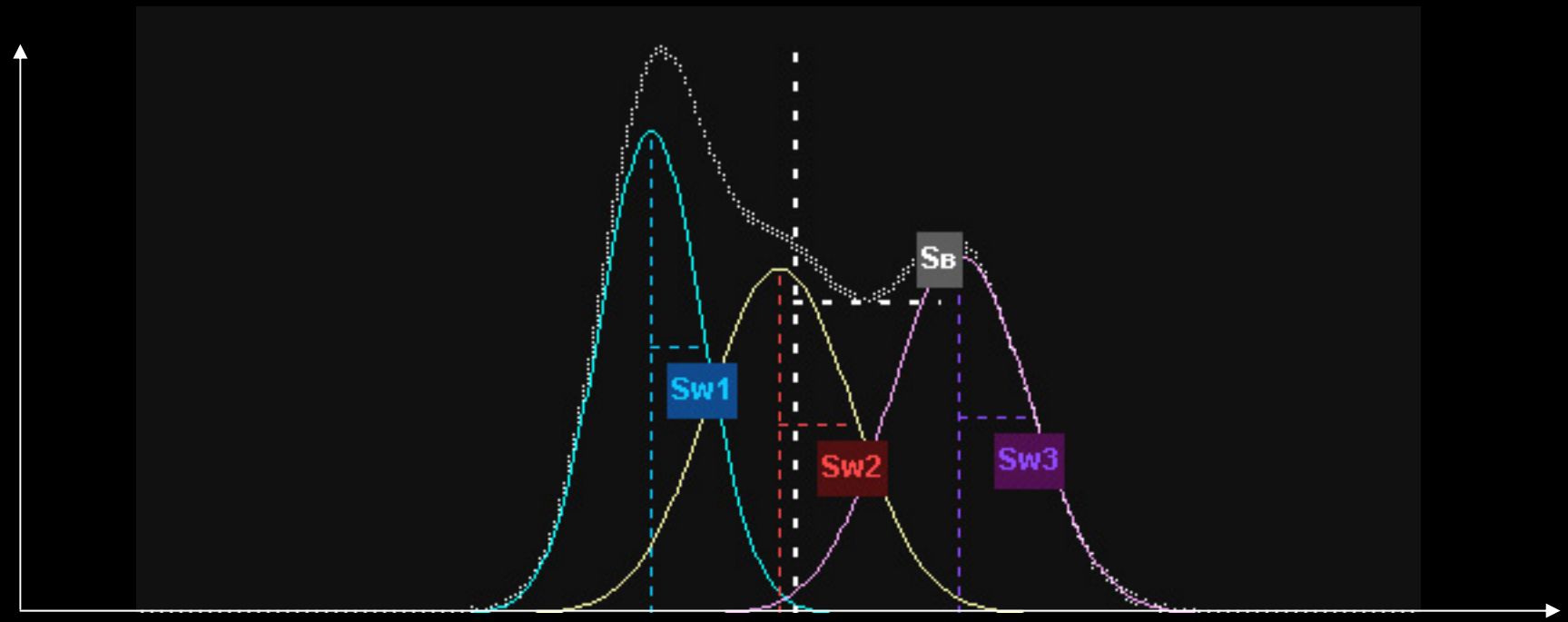
- Caso a dispersão S_w (dentro dos grupos) seja mantida, mas as médias de cada amostra sejam mais distantes entre si, aumenta-se a dispersão entre os grupos, S_B como mostra a figura:



- Como o teste F é dado por: $$F = \frac{S_B^2}{S_w^2}$$
- ... onde S_W^2 é obtido por uma composição de S_{w1}, S_{w2} e S_{w3}, se as médias são mais afastadas entre si, S_B e F (conseqüentemente) aumentam.

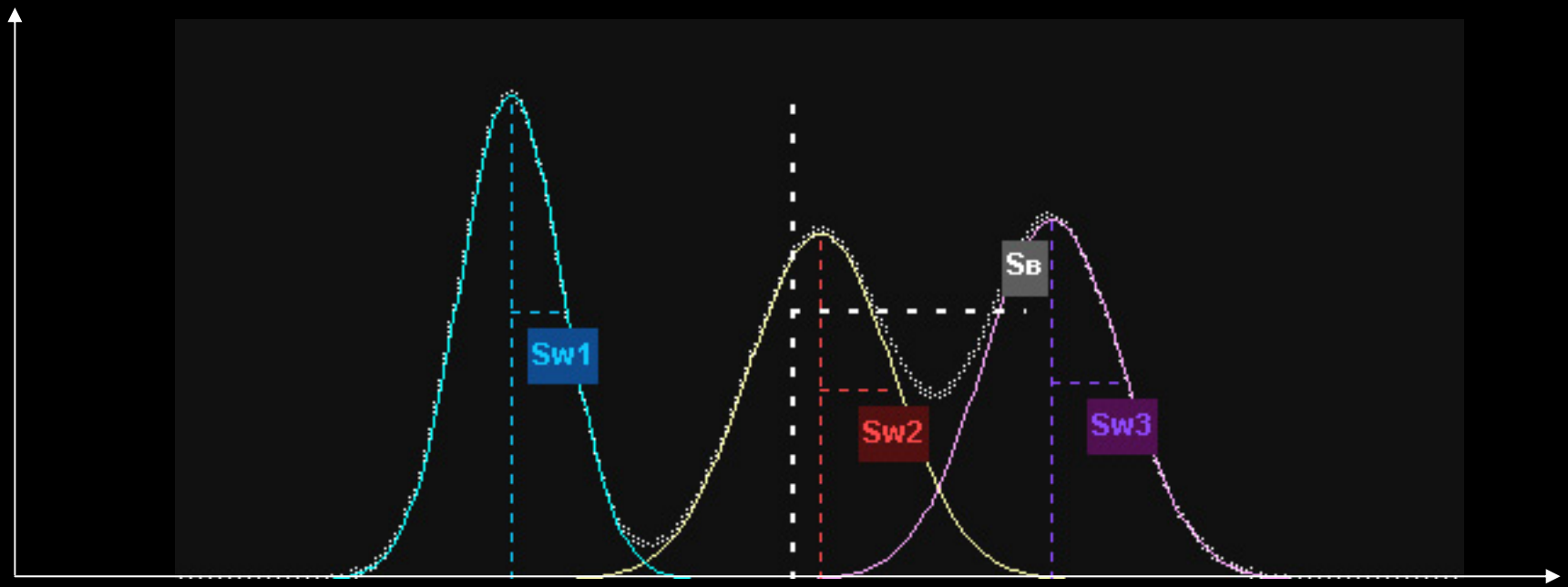
Entendendo o Significado do Valor de F calculado

- Mantendo-se as mesmas médias do primeiro exemplo, mas diminuindo-se a dispersão entre os grupos ($S_{w\#}$), o que sofre maior redução é o denominador, S_W^2, o que também leva a um aumento do valor de **F**:



Entendendo o Significado do Valor de F calculado

- Compare esta figura com as anteriores
  - Ggrosseiramente falando, quanto menos sobreposição entre as amostras, mais provavelmente o teste indicará significância estatística.
  - As distribuições abaixo estão bastante separadas e muito provavelmente indicam situações diferentes de um mesmo fenômeno.



A Distribuição-F

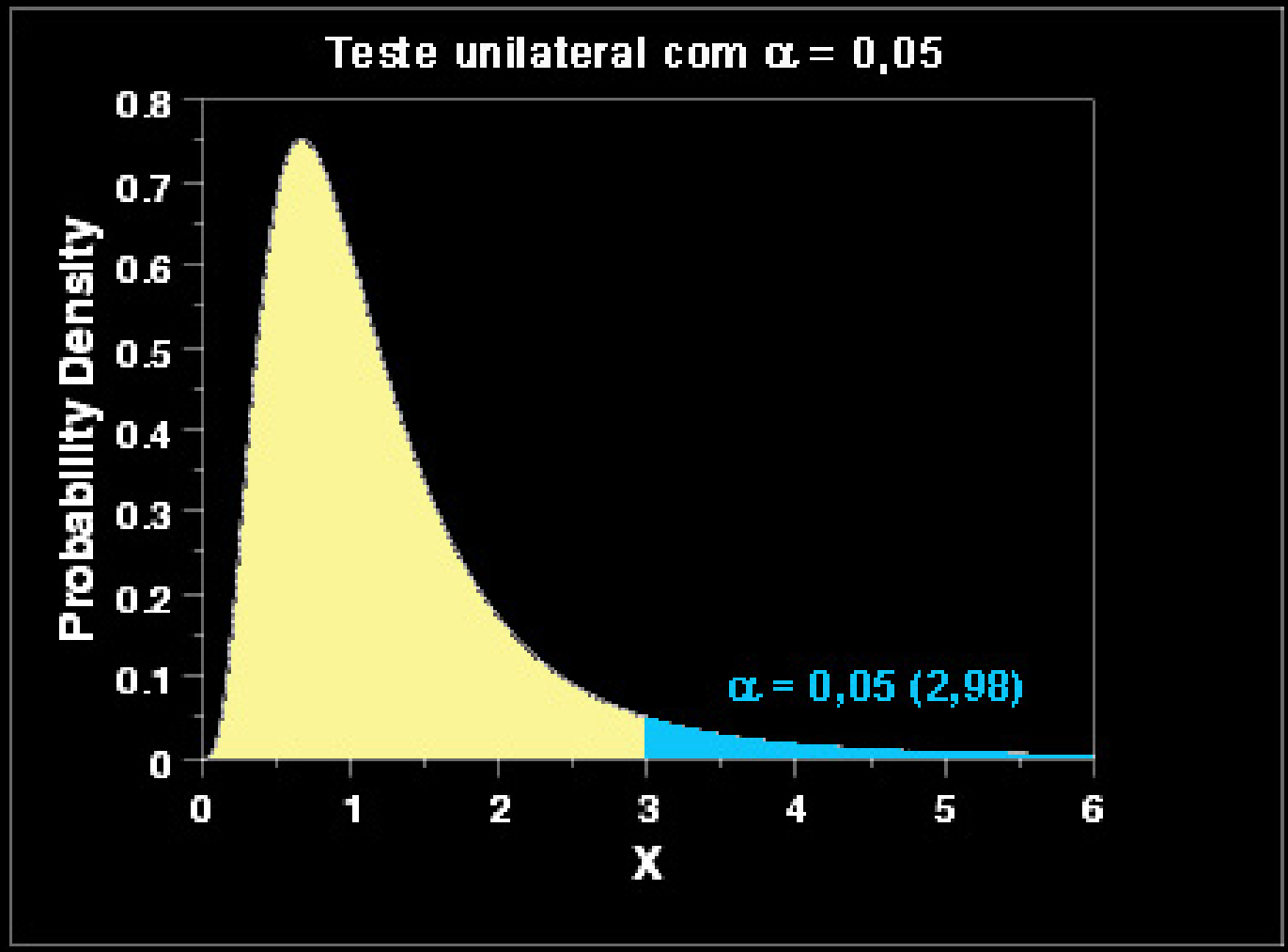
- Dada pela fórmula: $F(x, d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1} \; d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x \, \mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$

- Onde B é a Função-Beta:

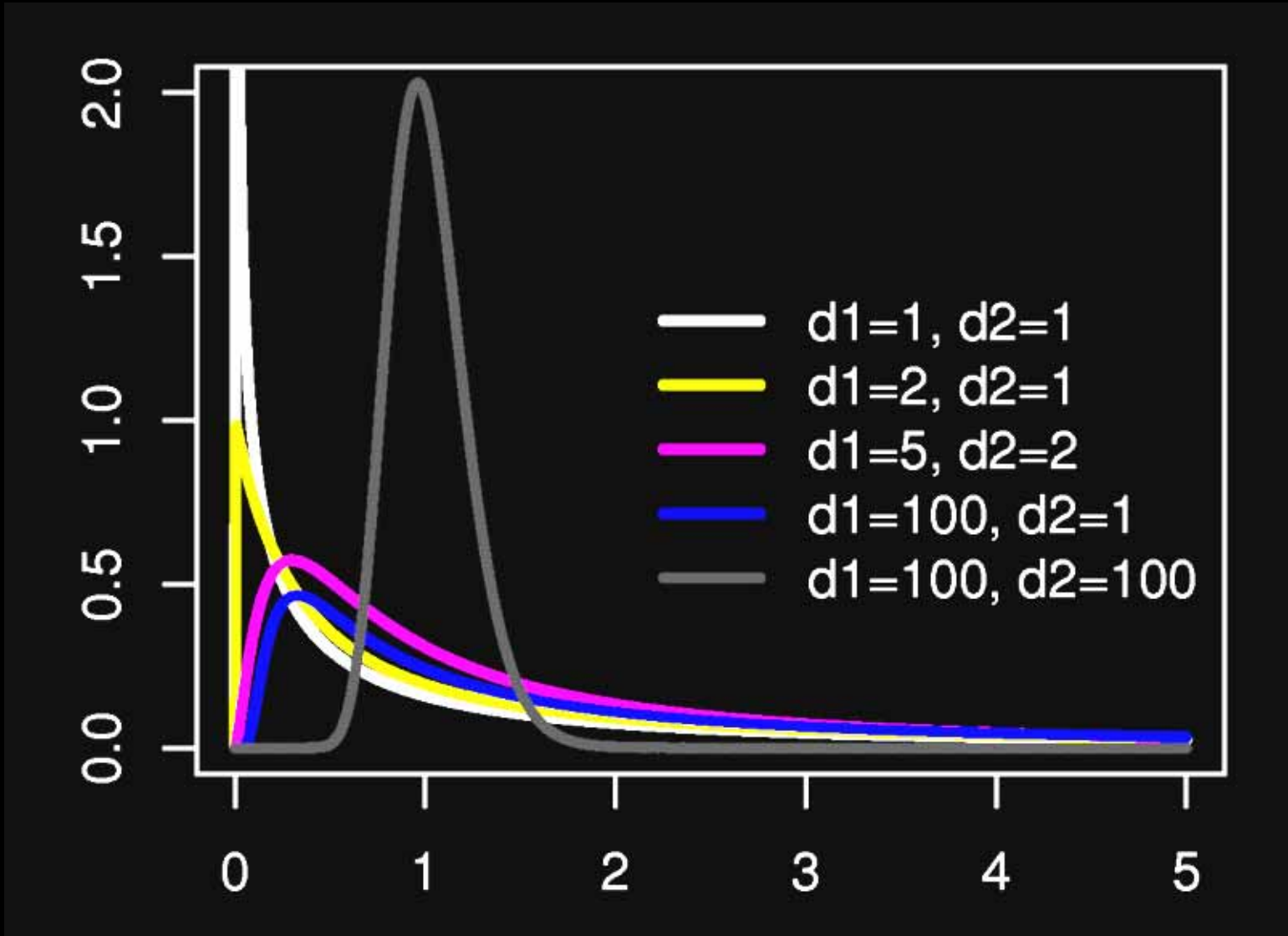
$$\mathrm{B}(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} \, dt$$

- $d_1 > 0$ e $d_2 > 0$ são os graus de liberdade e
- **x** é o grau de confiança expresso em desvios-padrão

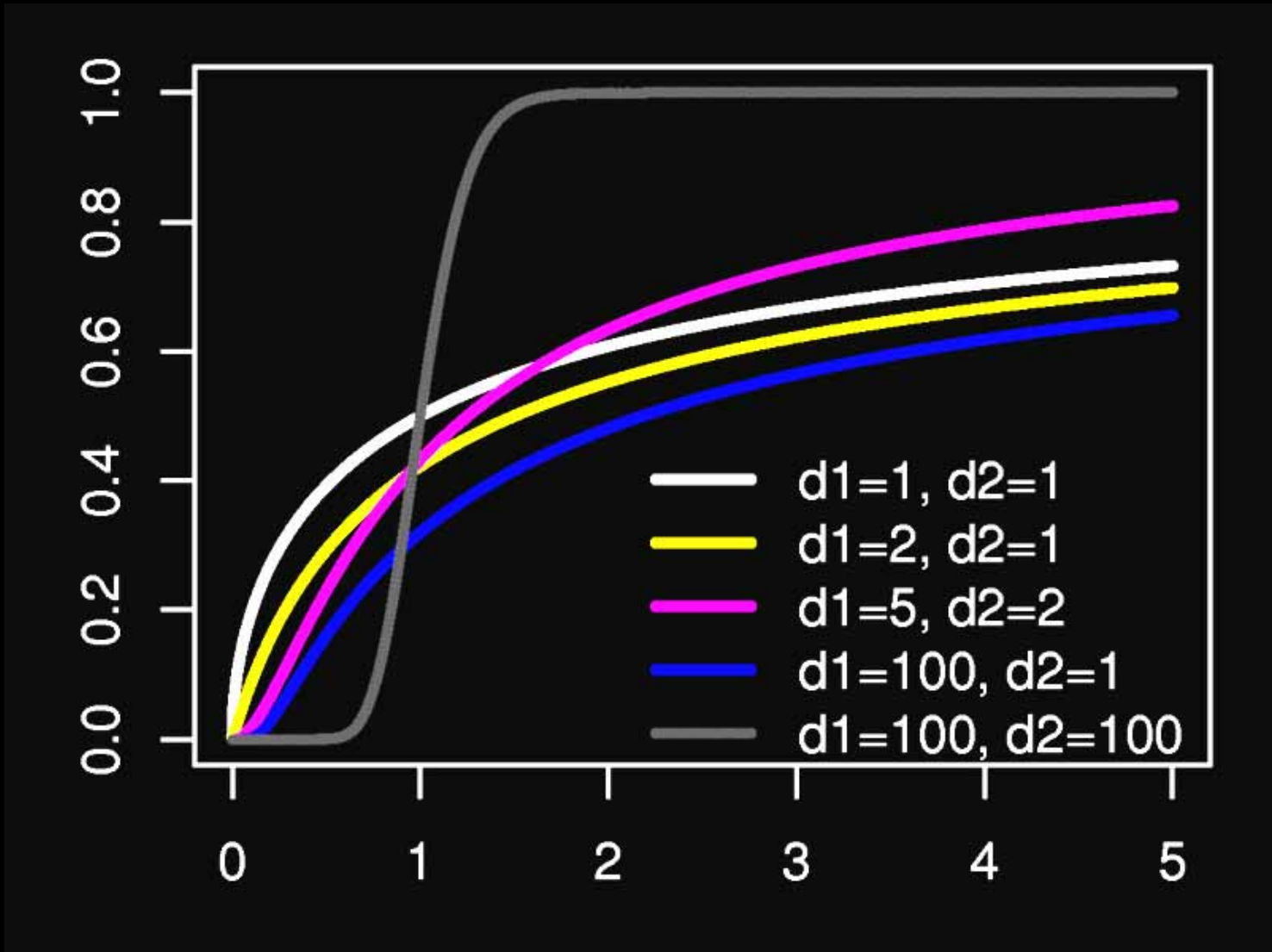
Exemplo de limite de aceitação de 95% com F (2,98, d_1 = 10, d_2 = 10)



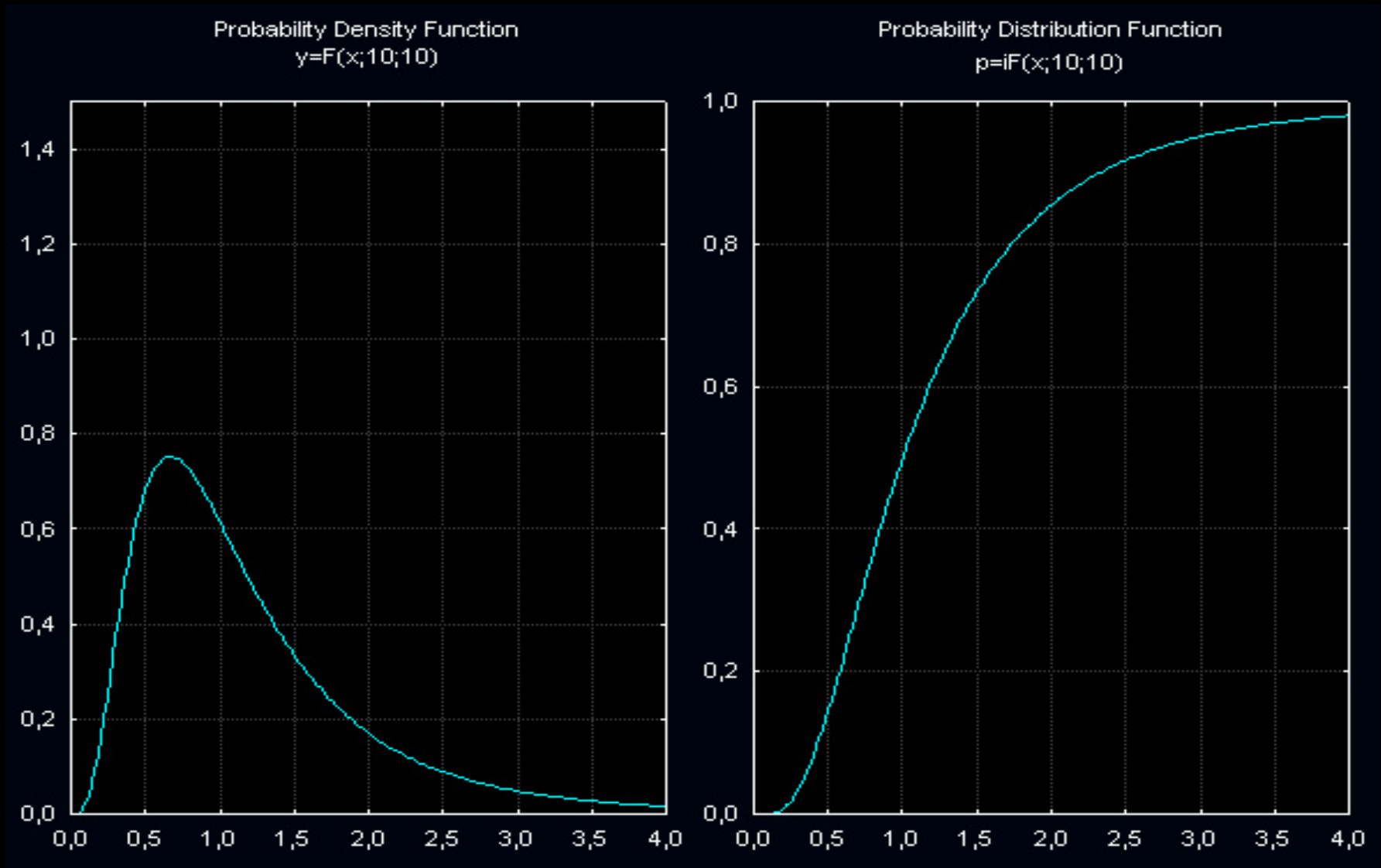
A função de densidade de probabilidade da Distribuição-F



A função de probabilidade acumulada da Distribuição-F



A distribuição de probabilidade da Distribuição-F para um determinado conjunto de graus de liberdade é dada pelos programas de estatística



Calculando a taxa-F (F_0) para o meu caso

- Utilizamos a taxa entre as medidas da variância:
  - Entre grupos e
  - Intra grupos
- Para o cálculo utilizamos a soma os quadrados dos erros como medida dessa variância
  - como essas somas variam com o tamanho da amostra, utilizamos as médias das variações
  - uma média dos quadrados é a soma dos quadrados dividida pelos seus graus de liberdade

Calculando a taxa-F (F_0) para o meu caso

- Cálculo da soma dos quadrados entre grupos:

número de grupos

número de elementos por grupo

média do grupo i

$$SQ_{entre} = n \cdot \sum_{i=1}^{k} (\overline{x_i} - \overline{x_{total}})^2$$

média geral

Obs.: Pressupõe grupos de mesmo tamanho.

Calculando a taxa-F (F_0) para o meu caso

- Cálculo da soma dos quadrados intra-grupos: passo 1: erro quadrado total

número de elementos por grupo

número de grupos

elemento j do grupo i

$$SQ_{total} = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \overline{x_{total}})^2$$

média geral

Calculando a taxa-F (F_0) para o meu caso

- Cálculo da soma dos quadrados intra-grupos: passo 2: erro quadrado interno

$$SQ_{dentro} = SQ_{total} - SQ_{entre}$$

Calculando a taxa-F (F_0) para o meu caso

- Cálculo das médias das soma dos quadrados entre-grupos:

$$\overline{SQ_{entre}} = \frac{SQ_{entre}}{k-1}$$

número de graus de liberdade
= grupos - 1 (d_1)

Calculando a taxa-F (F_0) para o meu caso

- Cálculo das médias das soma dos quadrados intra-grupos:

$$\overline{SQ_{dentro}} = \frac{SQ_{dentro}}{N-k}$$

número de graus de liberdade
= total de elemento - grupos (d_2)

Calculando a taxa-F (F_0) para o meu caso

- Alternativa para cálculo das médias das soma dos quadrados intra-grupos:

$$\overline{SQ_{dentro}} = \frac{SQ_{dentro}}{k(n-1)}$$

número de graus de liberdade
= grupos * (graus de liberdade do grupo)

Obs.: Pressupõe grupos de mesmo tamanho.

Calculando a taxa-F (F_0) para o meu caso

- Razão-F do experimento (F_0):

$$F_0 = \frac{\overline{SQ_{entre}}}{\overline{SQ_{dentro}}}$$

Se este valor for superior ao da Distribuição-F para o meu limiar de aceitação *x* e meus graus de liberdade d_1 e d_2, então posso rejeitar a hipótese H_0 de que não há grupos.

Avaliando a significância da razão-F (F_0) para o meu caso

- Duas formas:
  - Calculando o valor crítico da Distribuição-F para o meu limiar de aceitação (x) e meus graus de liberdade (d_1, d_2). Desvantagem: cálculo complicado.
  - Comparando F_0 calculado a uma tabela de valores críticos da função-F
    - existem muitas na Internet e todo bom livro de estatística possui uma.
    - http://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm

Upper critical values of the F distribution

for ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom

5% significance level

$$F_{.05}(\nu_1, \nu_2)$$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.882	240.543	241.882
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321

Usando os resultados de sua Análise de Agrupamentos

- O nosso objetivo será sempre o de implementar um classificador
  - para tanto temos de utilizar os resultados da Análise de Agrupamentos
  - associando as classes detectadas aos padrões e
  - criando um classificador
    - k-Nearest Neighbour
    - IBL
    - ....