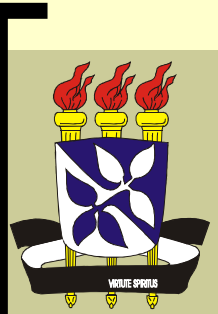


UNIVERSIDADE FEDERAL DA BAHIA
ESCOLA POLITÉCNICA
MESTRADO EM ENGENHARIA AMBIENTAL E URBANA

Apresentação do Curso – Introdução

Professora:

Cira Souza Pitombo

Disciplina:

Aplicações de técnicas de análise de dados

A DISCIPLINA

- **Objetivos**
  - **Fornecer um conjunto de técnicas estatísticas que permitem analisar problemas com um número elevado de variáveis.**
  - **Mostrar aplicações de técnicas diversas em diferentes trabalhos científicos.**
    - **Utilizar *software* estatístico.**
  - **Estimular os alunos a tratar seus diferentes bancos de dados (estudos de caso)**

A DISCIPLINA

- **METODOLOGIA**
- **Aulas expositivas**
- **Estudos de caso**
- **Aulas práticas – banco de dados individuais (alunos)**
- **Trabalho prático de análise de dados a cada estudo de caso**

A DISCIPLINA

AVALIAÇÃO

- **A avaliação de rendimento dos alunos será feita através de 01 trabalho que resultará em paper (nota peso 6) e entrega de atividades práticas (nota peso 4)**

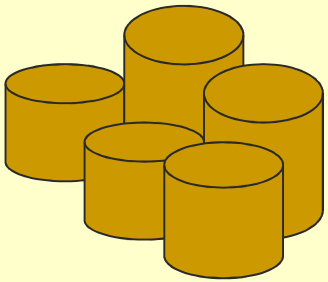
OBJETIVOS DA AULA

- Explicar o que é análise multivariada e quando sua aplicação é adequada
- Definir e discutir as técnicas específicas incluídas na análise multivariada
- Determinar qual técnica multivariada é apropriada a um problema específico de pesquisa
- Discutir a natureza das escalas de medida e sua relação com técnicas multivariadas.
- Descrever os aspectos conceituais e estatísticos inerentes à análise multivariada

O QUE É ANÁLISE MULTIVARIADA?

- Métodos estatísticos que simultaneamente analisam múltiplas medidas sobre cada indivíduo ou objeto sob investigação
- **Qualquer análise simultânea de mais de duas variáveis**
- Muitas técnicas multivariadas são extensões da análise univariada (análise de distribuição de uma única variável) e da análise bivariada (correlação, análise de variância, regressão simples).

O QUE É ANÁLISE MULTIVARIADA?



GRANDES BASES DE DADOS

Análise de Cluster

Árvore de Decisão

Y $X_1, X_2, X_3, \ldots X_N$

Análise Fatorial

Regressão Múltipla

Regressão Logística

CONHECIMENTO

VARIÁVEIS

Escalas de medidas

- **Não métricos (qualitativos) - categóricas**
- **Métricos (quantitavos) – numéricas (contínuas ou discretas)**
- **Dados não métricos : atributos, características, ou propriedades categóricas que identificam ou descrevem um objeto**
- **Dados métricos: refletem quantidade relativa**

VARIÁVEIS/ Banco de dados

Para exemplificar, considere um processo que utilize a base de dados cadastrais dos clientes de um Banco. A *unidade observacional* é o cliente. Para cada cliente, temos diversas características como Nome, Sexo, Estado Civil, Renda, etc. Cada uma dessas características é chamada tecnicamente de *variável*. Um cadastro, ou base de dados cadastrais é formado de casos ou registros (que no caso são os clientes) e variáveis. Para cada cliente temos o resultado das variáveis. Por exemplo, o primeiro cliente da base de dados tem o *nome* Fulano de Tal, o *Sexo* Masculino, o *Estado Civil* Solteiro, a *Renda* R$2.500,00 por mês, etc.

Nome	Sexo	Estado Civil	Renda Mensal
João	M	C	1.200,00
Maria	F	S	1.350,00
Antonio	M	C	*

VARIÁVEIS/ Banco de dados

Numérica – Contínua, discreta
Categórica

Nome	Idade	Estado civil	Genero	Renda	CEP	CIC	No de dependentes
João	52	1	1	1200	4180040	8009765	3
José	32	1	1	9000	4052000	5678900	2
Patrícia	48	1	2	5000	4254003	12356780	2
Maria	29	2	2	8000	4810204	79956420	1

VARIÁVEIS

Zona	Nome	Domicílios	Famílias	População	Matrículas	Empregos	Automóveis
1	Sé	825	1.175	3.120	4.890	76.594	428
2	Parque Dom Pedro	1.559	2.283	5.859	4.299	47.285	558
3	Praça João Mendes	3.983	4.747	12.275	4.424	21.368	1.295
4	Ladeira da Memória	7.165	9.850	19.956	4.222	33.178	4.516
5	República	5.014	6.491	12.040	4.932	73.136	3.090
6	Santa Efigênia	4.162	8.066	17.670	712	52.023	2.738
7	Luz	3.957	4.138	14.161	14.033	35.379	1.934
8	Brás	3.757	4.043	13.622	7.015	36.130	1.877
9	Independência	4.244	5.194	15.138	2.287	25.842	2.990
10	Cambuci	4.525	4.280	16.951	3.334	12.328	4.758

VARIÁVEIS

INDIVIDUO	SIT_FAMI	IDADE	SEXO	SE_ESTUD	GRAU_INS	USA_VTRA	RENDA_FA
1	1	38	1	1	3	1	437,00
2	2	37	2	1	3	2	437,00
3	3	15	1	3	3	2	437,00
4	3	12	2	3	3	2	437,00
5	3	10	1	3	3	2	437,00
6	1	30	1	1	6	2	1019,00
7	2	25	2	1	4	2	1019,00
8	4	60	2	1	3	2	1019,00
12	1	54	1	1	7	2	0,00

TRANSALVADOR — SALVADOR

RELATÓRIO ESTATÍSTICO DE TRÂNSITO

SALVADOR - 2010 ATUALIZADO EM: 12/01/10

SEAC - *Setor de Estatística e Acidentologia*

Quadro 1

QUADRO DE ACIDENTES - 2000 à 2010

ACIDENTES		Meses												Total / Ano	Média Mensal
		Jan*	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez*		
NÚMERO DE ACIDENTES	2000	2245	2266	2692	2167	2288	2532	1797	2620	2554	2333	2065	670	26229	2186
	2001	2199	1972	2447	2304	2305	1632	1439	2228	2257	2300	2113	2348	25544	2129
	2002	1698	1495	2293	2011	2212	1966	2065	2299	2345	2412	2054	2466	25316	2110
	2003	1835	1756	1867	1725	2246	1837	1725	2260	2007	2388	2382	2431	24459	2038
	2004	2056	2034	2240	2071	2035	1832	2159	2205	2118	2384	2146	2277	25557	2130
	2005	2034	1874	2429	2321	2357	2003	1884	2394	2188	2351	2313	2447	26595	2216
	2006	2322	2020	2425	2047	2233	1966	1616	1907	1919	2034	1904	2182	24575	2048
	2007	1792	1800	2268	2034	2252	2163	2109	2340	2374	2362	2424	2552	26470	2206
	2008	2242	2113	2557	2131	1920	1003	2421	3095	3156	3709	3765	3277	31389	2616
	2009	2883	2910	3339	3076	3072	2378	3018	3600	3382	3608	3619	3423	38308	3192
	2010	715												715	

VARIÁVEIS

Municípios e classes de tamanho da população	População					
	Total	Sexo (%)		Situação do domicílio (%)		Proporção de pessoas naturais dos municípios (%) (1)
		Masculino	Feminino	Urbana	Rural	
Acre	**557 882**	**50,4**	**49,6**	**66,5**	**33,5**	**71,3**
Até 5 000	10 190	52,3	47,7	35,6	64,4	84,0
De 5 001 até 10 000	47 214	53,3	46,7	33,6	66,4	70,6
De 10 001 até 20 000	97 443	52,2	47,8	46,3	53,7	66,8
De 20 001 até 50 000	82 535	51,4	48,6	50,1	49,9	91,6
De 50 001 até 100 000	67 441	50,3	49,7	57,8	42,2	81,8
De 100 001 até 500 000	253 059	48,7	51,3	89,4	10,6	63,1

TIPOS DE TÉCNICAS MULTIVARIADAS

(1) Análise de componentes principais e fatorial

(2) Regressão Múltipla

(3) Análise discriminante múltipla

(4) Análise Multivariada de variância e covariância

(5) Análise conjunta

(6) Correlação Canônica

(7) Análise de Agrupamentos

(8) Escalonamento Multidimensional

TIPOS DE TÉCNICAS MULTIVARIADAS

(1) Análise de componentes principais e fatorial

(2) Regressão Múltipla

(3) Análise discriminante múltipla

(4) Análise Multivariada de variância e covariância

(5) Análise conjunta

(6) Correlação Canônica

(7) Análise de Agrupamentos

(8) Escalonamento Multidimensional

TIPOS DE TÉCNICAS MULTIVARIADAS

TÉCNICAS EMERGENTES

(9) Análise de correpondências

(10) Modelos lineares de probabilidade – *logit* e *probit*

(11) Modelagem de equações simultâneas/estruturais

(12) Mineração de dados

TIPOS DE TÉCNICAS MULTIVARIADAS

ANÁLISE FATORIAL

Abordagem estatística que pode ser usada para analisar inter-relações entre um grande número de variáveis e explicar essas variáveis em termos de suas dimensões inerentes comuns (fatores)

Objetivo: encontrar um meio de condensar a informação contida em um número de variáveis originais em um conjunto menor de variáveis (fatores) com uma perda mínima de informações

TIPOS DE TÉCNICAS MULTIVARIADAS

ANÁLISE FATORIAL

Em geral, a análise fatorial aborda o problema de analisar a estrutura das interrelações (correlações) entre um grande número de variáveis, definindo um conjunto de dimensões latentes, chamada de fatores.

Ao resumir os dados, a análise fatorial obtém dimensões latentes que, quando interpretadas e compreendidas, descrevem os dados em um número muito menor de conceitos do que as variáveis individuais originais.

Pode desempenhar um papel único na aplicação de outras técnicas multivariadas – novas variáveis não mais altamente correlacionadas

TIPOS DE TÉCNICAS MULTIVARIADAS

ANÁLISE FATORIAL

Difere das técnicas de dependência – Regressão múltipla, análise discriminante, etc – nas quais uma ou mais variáveis são explicitamente consideradas como variáveis dependentes e todas as outras são variáveis independentes.

A Análise Fatorial é uma técnica de interdependência na qual todas as variáveis são simultaneamente consideradas, cada uma relacionada com todas as outras.

TIPOS DE TÉCNICAS MULTIVARIADAS

ANÁLISE FATORIAL

$$y_1 = a_{11} x_1 + a_{12} x_2 + \ldots + a_{1p} x_p$$

$$y_2 = a_{21} x_1 + a_{22} x_2 + \ldots + a_{2p} x_p$$

$$\ldots$$

$$y_p = a_{p1} x_1 + a_{p2} x_2 + \ldots + a_{pp} x_p$$

$X_1, X_2, \ldots, X_p$ são as variáveis medidas e correlacionadas entre si;

$y_1, y_2, \ldots, y_p$ são as variáveis não correlacionadas que designam as componentes principais;

a_{ij} =>i=1,...,p; j=1,...,p, são os pesos de cada variável original na equação geral, que definem as novas variáveis y_p

TIPOS DE TÉCNICAS MULTIVARIADAS

ANÁLISE FATORIAL

ID	A	B	C	D	E	F	G	H	I	J	L	M	N	O
1	1200	1,2	78	99	50000	5	1889	70	20	170	88	254	30	36
2	1000	1,89	98	98	30000	5	1990	78	25	169	87	239	28	37
3	2100	1,3	76	97	40000	8	1829	79	32	160	88	267	22	36
4	1450	2,4	52	99	50000	7	1877	67	31	178	89	288	27	36
5	1300	1,4	54	92	20000	6	1890	72	33	184	87	231	26	38

$$Fator1 = \alpha A + \beta B + \delta E$$

$$Fator2 = \lambda C + \alpha_1 F + \beta_1 B$$

$$Fator3 = \delta O + \beta_2 I + \alpha_2 N$$

85% da variância dos dados explicada

TIPOS DE TÉCNICAS MULTIVARIADAS

REGRESSÃO MÚLTIPLA

É o método de análise apropriado quando o problema de pesquisa envolve uma única variável dependente numérica considerada relacionada a duas ou mais variáveis independentes numéricas.

Sempre que o pesquisador estiver interessado em prever a quantia ou magnitude da variável dependente

Variável dependente – Despesas mensais com jantares fora de casa.
Variáveis independentes – renda familiar; tamanho da família; idade do chefe da família

TIPOS DE TÉCNICAS MULTIVARIADAS

REGRESSÃO MÚLTIPLA

ID	A	B	C	D	E	F	G	H	I	J	L	M	N	O
1	1200	1,2	78	99	50000	5	1889	70	20	170	88	254	30	36
2	1000	1,89	98	98	30000	5	1990	78	25	169	87	239	28	37
3	2100	1,3	76	97	40000	8	1829	79	32	160	88	267	22	36
4	1450	2,4	52	99	50000	7	1877	67	31	178	89	288	27	36
5	1300	1,4	54	92	20000	6	1890	72	33	184	87	231	26	38

$$y = \alpha B + \beta C + \delta D + \partial E + \gamma F + \eta G + \kappa H + \lambda I + \mu J + \theta L + \rho M + \sigma N + \omega O$$

TIPOS DE TÉCNICAS MULTIVARIADAS

REGRESSÃO LOGÍSTICA

Variável dependente não métrica com apenas dois grupos – variáveis *dummy* (dicotômicas)

A regressão logística não depende de suposições rígidas e é muito mais robusta quando pressupostos como normalidade multivariada e de iguais matrizes de variância co-variância nos grupos não são satisfeitos

É similar à regressão linear múltipla, com testes estatísticos diretos, habilidade de incorporar efeitos não-lineares e uma vasta gama de diagnósticos

TIPOS DE TÉCNICAS MULTIVARIADAS

ANÁLISE DE AGRUPAMENTOS - *CLUSTER*

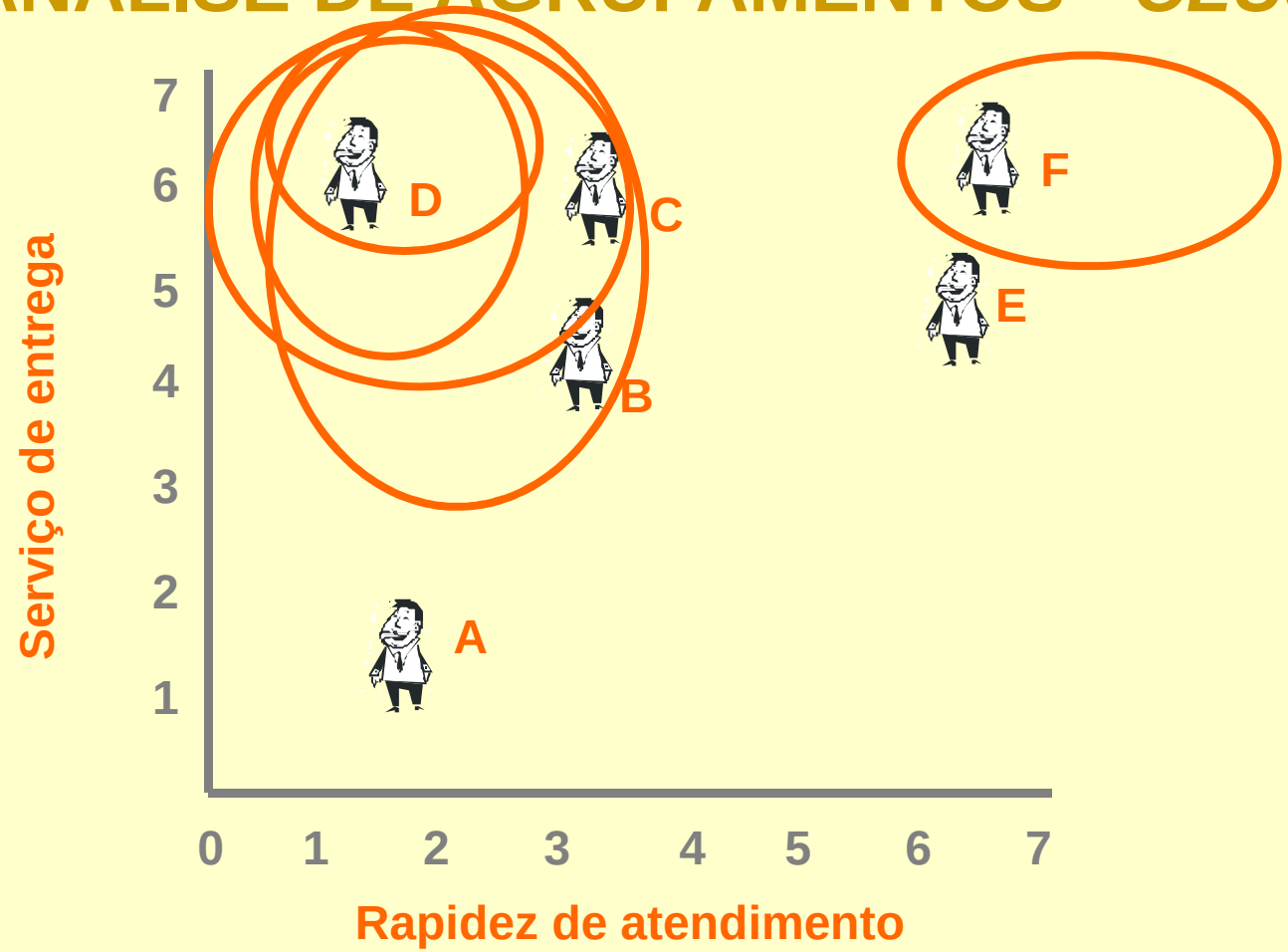
Técnica analítica para desenvolver subgrupos significativos de indivíduos ou objetos.

O objetivo é classificar uma amostra de entidades em um pequeno número de grupos mutuamente excludentes, com base nas similaridades entre as entidades

Na análise de agrupamentos os grupos não são predefinidos. A técnica é usada para identificar os grupos.

TIPOS DE TÉCNICAS MULTIVARIADAS

ANÁLISE DE AGRUPAMENTOS - *CLUSTER*



TIPOS DE TÉCNICAS MULTIVARIADAS

ÁRVORE DE DECISÃO

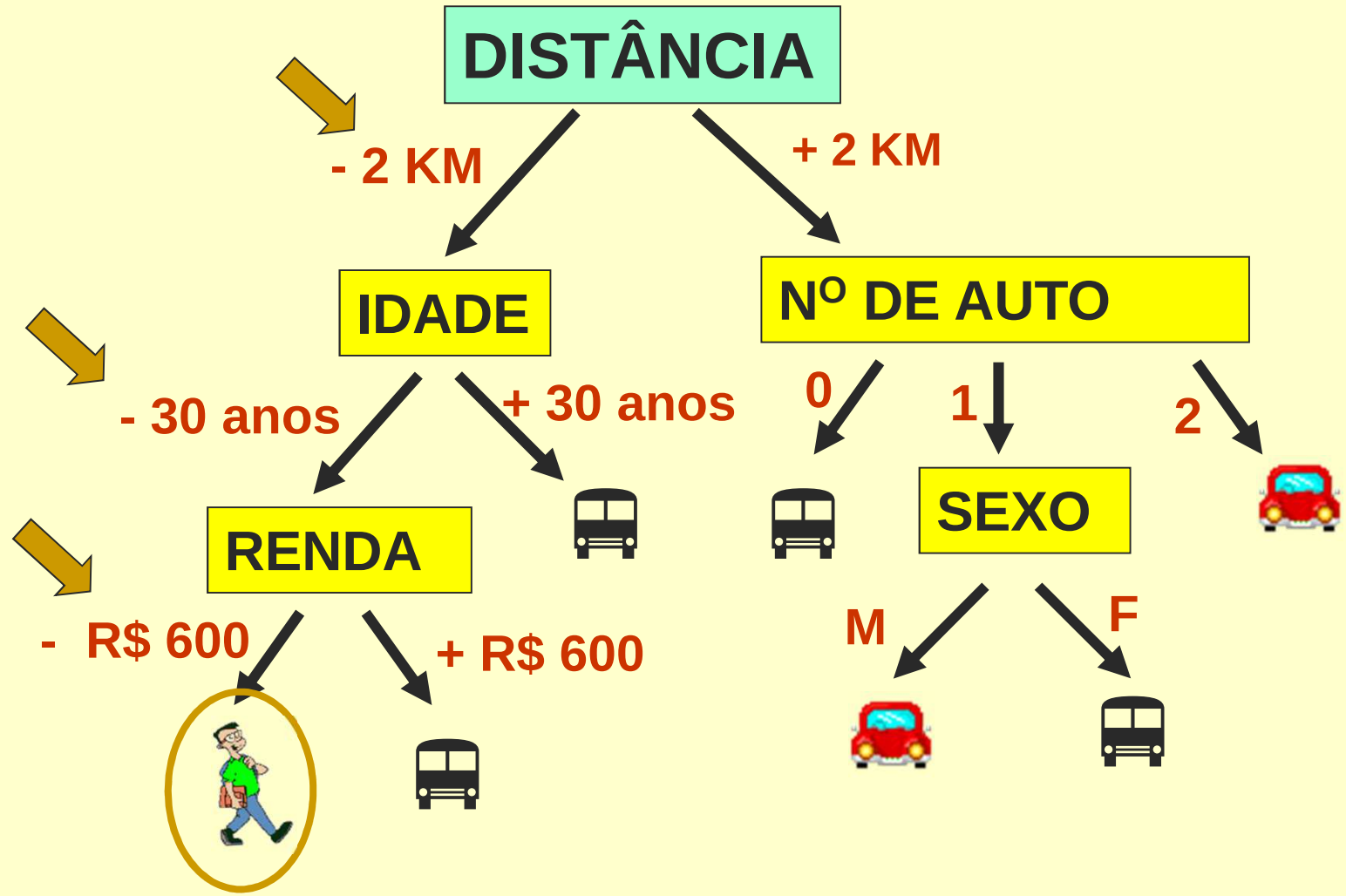
- Forma simples de representação de relações existentes em um conjunto de dados
- Variável dependente ⇨ Variáveis independentes
- Divisão seqüencial do conjunto de dados, considerando os valores das variáveis
- C4.5, CHAID, CART

TIPOS DE TÉCNICAS MULTIVARIADAS

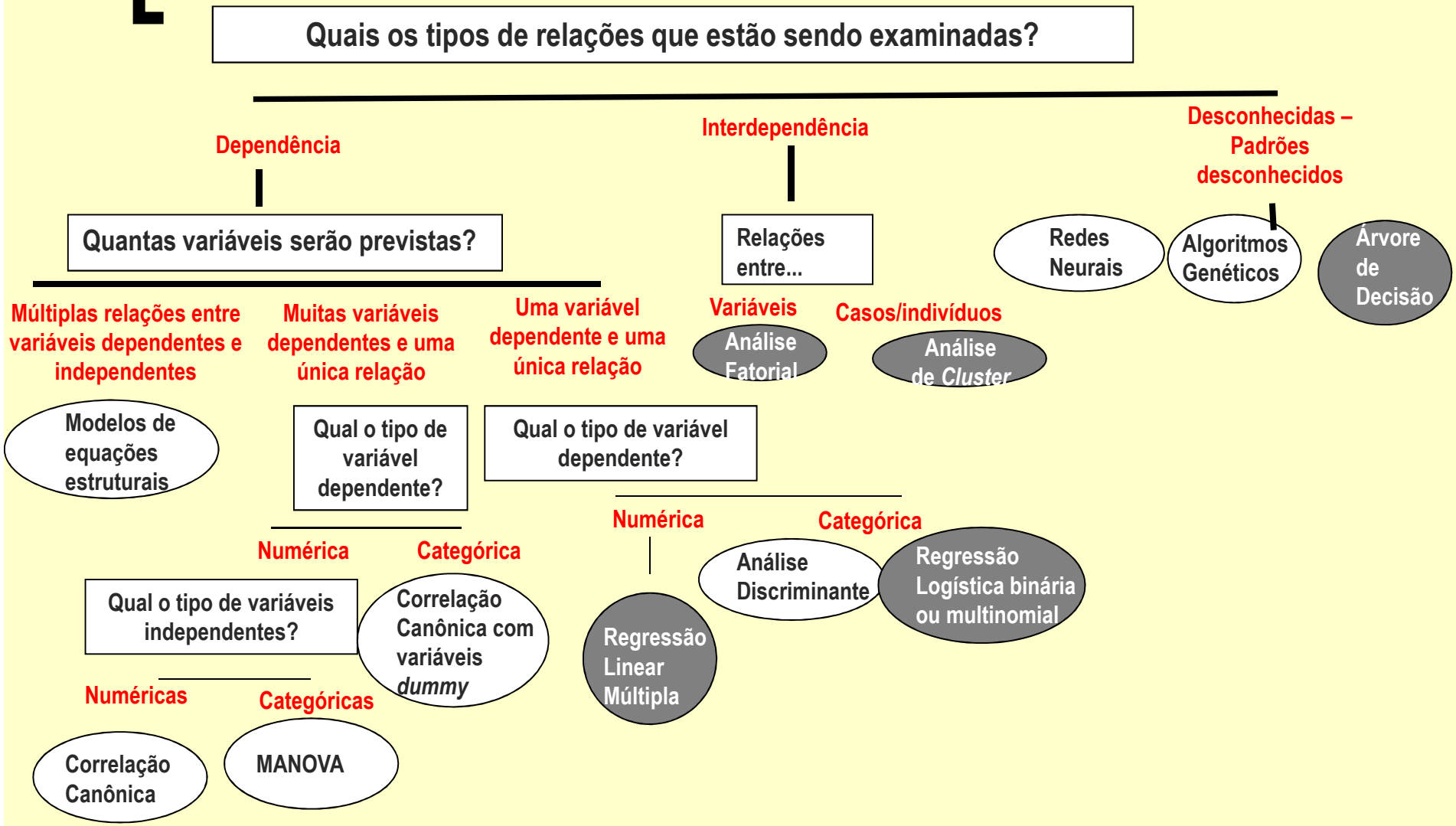
ÁRVORE DE DECISÃO

TIPOS DE TÉCNICAS MULTIVARIADAS

ÁRVORE DE DECISÃO



TIPOS DE TÉCNICAS MULTIVARIADAS



ATIVIDADE 30.11.2012

REVISAR CONCEITOS:

1. **VARIÂNCIA/CO-VARIÂNCIA**
2. **MÉDIA E DESVIO PADRÃO**
3. **DISTRIBUIÇÃO NORMAL**
4. **DISTRIBUIÇÃO BINOMIAL**
5. **COEFICIENTE DE CORRELAÇÃO**

TRAZER BANCO DE DADOS

DESCREVER O BANCO DE DADOS E OBJETIVOS

INSTALAR O SPSS