MAE 0330

ANÁLISE MULTIVARIADA DE DADOS

Análise de Correspondência

Júlia M Pavan Soler

pavan@ime.usp.br

2° Sem/2012

Análise de Correspondência

	Variável Coluna					
u.a. / Variável Linha	1	2	…	j	…	J
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1J}
2	Y_{21}	Y_{22}		Y_{2j}		Y_{2J}
…	…	…	…	…		…
i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{iJ}
…	…	…	…	…		…
I	Y_{I1}	Y_{I2}		Y_{Ij}		Y_{IJ}

Identificar a estrutura dos dados multivariados com “Tabelas de Contingência”

Objetivos:

- Descrever graficamente os dados dispostos em tabelas de contingência
- Representar graficamente o padrão de associação entre variáveis $\Rightarrow$ os vetores linha e os vetores coluna da tabela são visualizados como pontos em um espaço vetorial

TÉCNICA GRÁFICA MULTIDIMENSIONAL (similar ao Escalonamento!!)

(essencialmente descritiva, não adota qualquer modelo estrutural, auxilia a análise inferencial)

Análise de Correspondência

			Ano			
Jornal	**1976**	**1977**	**1978**	**1979**	**1980**	**Total**
A	64	58	67	59	60	308
B	18	18	23	20	17	96
C	12	10	9	12	9	52
D	36	25	34	31	27	153
E	29	21	25	20	20	115
F	133	115	116	107	89	560
G	34	28	30	26	29	147
H	178	143	180	150	148	799
I	8	8	5	6	6	33
J	101	113	143	112	107	576
K	66	56	60	58	53	293
L	87	69	79	68	69	372
M	23	19	17	19	17	95
N	34	24	29	26	23	136
O	70	56	60	55	50	291
P	29	20	25	19	18	111
Q	46	40	38	38	33	195
R	123	122	149	122	112	628
S	79	68	70	61	57	335
T	130	109	148	110	100	597
U	22	17	19	15	16	89
Total	1322	1139	1326	1134	1060	5981

Ao longo de 5 anos, em cada ano, cerca de 1000 pessoas de uma cidade foram amostradas e questionadas sobre quais jornais, dentre 21, eles liam regularmente.

Como representar o hábito de leitura de jornais dos cidadãos e sua variação ao longo do tempo?

Análise de Correspondência

Distribuição de 5.387 estudantes escoceses de acordo com a cor dos olhos e dos cabelos (Fisher, 1940)

	Cor do cabelo					
Cor olhos	**Claro**	**Ruivo**	**Médio**	**Escuro**	**Preto**	**Total**
Claros	688	116	584	188	4	1580
Azul	326	38	241	110	3	718
Médio	343	84	909	412	26	1774
Escuro	98	48	403	681	85	1315
Total	1455	286	2137	1391	118	5387

Como descrever graficamente o padrão de associação entre as variáveis cor dos olhos e dos cabelos dos estudantes escoceses ?

Análise de Correspondência

Distribuição dos funcionários de uma empresa de acordo com o tabagismo.

	Hábito de Fumar				
Funcionário	**Não**	**Pouco**	**Médio**	**Muito**	**Total**
Nível 1	4	2	3	2	11
Nível 2	4	3	7	4	18
Nível 3	25	10	12	4	51
Nível 4	18	24	33	13	88
Nível 5	10	6	7	2	25
Total	61	45	62	25	193

Para aderir a uma campanha nacional anti-tabagismo, o gerente de Recursos Humanos de uma empresa deseja conhecer o hábito de fumar dos funcionários. Os dados acima foram coletados para esta finalidade.

A representação gráfica dos dados é, em geral, de fácil entendimento. Como representar o padrão de associação entre o nível do funcionário e o hábito de fumar em um gráfico ?

Análise de Correspondência

Influência da idade da adolescente no tipo de namoro (Everitt, 2004)

	Faixa Etária				
	< 16	16-17	17-18	18-19	19-20
Nenhum namorado	21	21	14	13	8
Namoro sem sexo	8	9	6	8	2
Namoro com sexo	2	3	4	10	10
Total	31	33	24	31	20

Como descrever graficamente o padrão de associação entre as variáveis faixa etária da adolescente e o tipo de namoro ?

Análise de Correspondência

(Everitt, 2004)

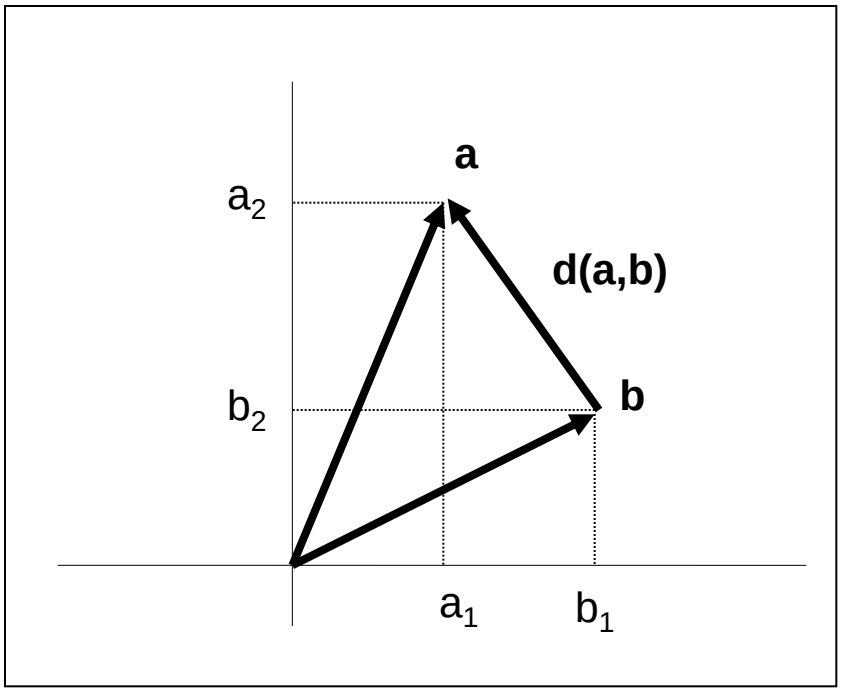
- Método de decomposição da estatística Qui-Quadrado, usada para testar independência em uma tabela de contingência, em componentes que correspondem a diferentes dimensões da heterogeneidade entre as variáveis coluna da tabela.

- Método que simultaneamente atribui uma escala às linhas e, separadamente, uma escala às colunas da tabela de tal forma a maximizar a correlação entre as duas escalas.

- Método de obtenção de coordenadas para representar as categorias de ambas as variáveis linha e coluna da tabela, de tal forma que o padrão de associação seja representado graficamente $\Rightarrow$ é um tipo de Escalonamento Multidimensional para uma medida de distância específica para dados categorizados, conhecida como distância Qui-Quadrado.

Análise de Correspondência

Notação



Distância Euclidiana entre pontos:

$$d^2(\mathbf{a},\mathbf{b}) = \|\mathbf{a}-\mathbf{b}\| = (b_1-a_1)^2 + (a_2-b_2)^2 = (\mathbf{a}-\mathbf{b})'(\mathbf{a}-\mathbf{b})$$

Escalas diferentes

Heterocedasticidade

Estrutura de Covariância

Distância Euclidiana Ponderada:

$$d^2(\mathbf{a}^*,\mathbf{b}^*) = \|\mathbf{a}^*-\mathbf{b}^*\| = \left(\frac{b_1}{s_1}-\frac{a_1}{s_1}\right)^2 + \left(\frac{a_2}{s_2}-\frac{b_2}{s_2}\right)^2 = (\mathbf{a}-\mathbf{b})'\mathbf{D}_{s_{jj}}^{-1}(\mathbf{a}-\mathbf{b})$$

Análise de Correspondência

Distância Euclidiana entre Vetores de Frequências

Distribuição da intenção de voto de eleitores

Ano	Partido 1	Partido 2	Partido 3	Partido 4	Partido 5	Total
2002	0,25	0,44	0,1	0,16	0,05	1
2006*						
O	1195	2290	545	771	199	5000
E	1270	2210	510	780	230	5000

* Resultados parciais de uma pesquisa eleitoral

O: freq. Observadas E: freq. Esperadas (sob as proporções de 2002)

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{75^2}{1270} + \frac{80^2}{2210} + \frac{35^2}{510} + \frac{9^2}{780} + \frac{31^2}{230} = 14{,}01$$

Estatística de um teste de aderência

$$\chi^2 = (\mathbf{O}_{5\times1} - \mathbf{E}_{5\times1})' \mathbf{D}^{-1}_{E_{5\times5}} (\mathbf{O}_{5\times1} - \mathbf{E}_{5\times1}) \Rightarrow$$

Distância ao quadrado entre as freqüências observadas e esperadas com pesos iguais ao inverso das freq. esperadas

A estatística χ^2 é uma medida de distância Euclidiana ao quadrado ponderada

Análise de Correspondência

Distância Euclidiana entre Vetores de Frequências

Distribuição da intenção de voto de eleitores

Ano	Partido 1	Partido 2	Partido 3	Partido 4	Partido 5	Total
2002	0,25	0,44	0,1	0,16	0,05	1
2006*						
O	1195	2290	545	771	199	5000
E	1270	2210	510	780	230	5000
$\mathbf{p} = (1/n)\mathbf{O}$	0,239	0,458	0,109	0,154	0,04	1
$\bar{\mathbf{p}} = (1/n)\mathbf{E}$	0,254	0,442	0,102	0,156	0,046	1

$\mathbf{O}$, $\mathbf{E}$ $\Rightarrow$ $\mathbf{p}$, $\bar{\mathbf{p}}$

$$\chi^2 = (\mathbf{O}-\mathbf{E})' \mathbf{D}_E^{-1} (\mathbf{O}-\mathbf{E})$$

$$\chi^2 = n(\mathbf{p}-\bar{\mathbf{p}})' \mathbf{D}_{\bar{\mathbf{p}}}^{-1} (\mathbf{p}-\bar{\mathbf{p}}) = n \sum_j \frac{(p_j - \bar{p}_j)^2}{\bar{p}_j}$$

A significância da estatística χ^2 depende do tamanho amostral $\Rightarrow$a distância Euclidiana ponderada entre o vetor de freq. relativas observadas e média é proporcional (a menos do fator *n*) a esta estatística!

Análise de Correspondência

Distância Euclidiana entre Vetores de Frequências

$$\chi^2 = (\mathbf{O}-\mathbf{E})' \mathbf{D}_E^{-1} (\mathbf{O}-\mathbf{E}) \quad \Rightarrow \quad \chi^2 = n(\mathbf{p}-\bar{\mathbf{p}})' \mathbf{D}_{\bar{\mathbf{p}}}^{-1} (\mathbf{p}-\bar{\mathbf{p}}) = n \sum_j \frac{(p_j - \bar{p}_j)^2}{\bar{p}_j}$$

- Análise de Correspondência considera a dispersão de vetores de frequência relativa (**p**) em um espaço multidimensional. O vetor **p** é denominado um perfil de frequências relativas.
- É calculada a distância Euclidiana entre o vetor de frequências relativas observadas de cada população $\mathbf{p_i}$ e o vetor de frequências relativas médias $\bar{\mathbf{p}}$, ponderada pelas frequências relativas médias.
- Ainda, a análise é flexível no sentido de dar pesos (“massas”) aos perfis **p** das populações sob estudo, como veremos a seguir.

Análise de Correspondência

Representação dos Perfis Linha da Tabela

	Variável Coluna					
Variável Linha	1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

Estatística Qui-Quadrado:

$$\chi^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(O_{11}-E_{11})^2}{E_{11}} + \frac{(O_{12}-E_{12})^2}{E_{12}} + \frac{(O_{IJ}-E_{IJ})^2}{E_{IJ}}; \qquad O_{ij} = n_{ij} \quad E_{ij} = \frac{n_{i.}n_{.j}}{n}$$

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(\frac{n_i}{n_{i.}}O_{ij} - \frac{n_i}{n_{i.}}E_{ij}\right)^2}{E_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{i.}\frac{(p_{ij}-\bar{p}_j)^2}{\bar{p}_j} = \sum_{i=1}^{I} n_{i.}\sum_{j=1}^{J}\frac{(p_{ij}-\bar{p}_j)^2}{\bar{p}_j} = \sum_{i=1}^{I} n_{i.} d_i^2$$

$$d_i^2 = (\mathbf{p_i} - \bar{\mathbf{p}})' \mathbf{D}_{\bar{\mathbf{p}}}^{-1} (\mathbf{p_i} - \bar{\mathbf{p}})$$

$\mathbf{p_i}$: perfil de freqüências relativas da linha i

Análise de Correspondência

Representação dos Perfis Linha da Tabela

	Variável Coluna					
Variável Linha	1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

$$\mathbf{p_i} = (p_{i1}, p_{i2}, ..., p_{iJ})'; \qquad p_{ij} = \frac{n_{ij}}{n_{i.}} \qquad j = 1,2,...,J \quad i = 1,2,...,I$$

$$\bar{\mathbf{p}} = \left(\frac{n_{.1}}{n}, \frac{n_{.2}}{n}, ..., \frac{n_{.J}}{n}\right)' \quad \text{Centróide (linha)}$$

$$d_i^2 = (\mathbf{p_i} - \bar{\mathbf{p}})' \mathbf{D}_{\bar{\mathbf{p}}}^{-1} (\mathbf{p_i} - \bar{\mathbf{p}}) = \sum_{j=1}^{J} \frac{(p_{ij} - \bar{p}_j)^2}{\bar{p}_j}$$

Distância Euclidiana ponderada ao quadrado do perfil de freqüências relativas da linha i ao centróide

Como representar tais perfis linha em um espaço multidimensional?

Análise de Correspondência

Variável Linha	Variável Coluna 1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

$$\mathbf{p_i} = \left(p_{i1}, p_{i2}, ..., p_{iJ}\right)' \qquad \bar{\mathbf{p}} = \left(\frac{n_{.1}}{n}, \frac{n_{.2}}{n}, ..., \frac{n_{.J}}{n}\right)' \qquad d_i^2 = \left(\mathbf{p_i} - \bar{\mathbf{p}}\right)' \mathbf{D}_{\bar{\mathbf{p}}}^{-1} \left(\mathbf{p_i} - \bar{\mathbf{p}}\right)$$

$$\chi^2 = \sum_{j=1}^{J}\sum_{i=1}^{I} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = \frac{\left(O_{11} - E_{11}\right)^2}{E_{11}} + \frac{\left(O_{12} - E_{12}\right)^2}{E_{12}} + \frac{\left(O_{IJ} - E_{IJ}\right)^2}{E_{IJ}};$$

$$in(I) = \chi^2 / n$$

Ponderação da estatística por n

in(I): medida de Inércia total do conjunto dos I perfis. Mede a variação dos perfis individuais $\mathbf{p_i}$ em torno do centróide $\bar{\mathbf{p}}$.

O objetivo da análise de Correspondência é encontrar um subespaço de "baixa" dimensão que melhor contenha os perfis $\mathbf{p_i}$

Análise de Correspondência

$$\mathbf{Y}_{I\times J} = \begin{pmatrix} \mathbf{p}_{1_{1\times J}} \\ \dots \\ \mathbf{p}_{I_{1\times J}} \end{pmatrix}$$

Matriz de dados (frequências relativas) com a soma de cada linha igual a uma constante c (c=1). O vetor centróide é dado por:

$$\bar{\mathbf{p}} = \left(\bar{p}_1,\dots,\bar{p}_J\right)'$$

Considere as seguintes matrizes:

Matriz de pesos: $D_{q\,J\times J} = diag\left(q_j = 1/\,\bar{p}_j = n/n_{.j}\right)$

Matriz de massas: $D_{w\,I\times I} = diag\left(w_i = n_{i.}/n\right)$ associada à marginal fixada

Então, os eixos principais (denotados por F) dos perfis linha $\mathbf{p_i}$ podem ser obtidos da decomposição espectral da matriz ***Y*** tal que, para k dimensões e com I>J, tem-se:

$$Y_{I\times J} = N_{I\times I}\; D_{\lambda\,I\times I}\; M'_{I\times J} \quad ; \qquad N'D_wN = M'D_qM = I \quad \Rightarrow \quad F_{I\times k} = N_{I\times I}D_{\lambda_{I\times 2}}$$

$\Rightarrow in(I) = \sum_{i=1}^{I}\lambda_i^2$: inércia total

$\Rightarrow \dfrac{\lambda^2{}_1 + \lambda^2{}_2 + \dots + \lambda^2{}_k}{\sum_i \lambda_i^2}$: proporção da inércia descrita pelos eixos

Análise de Correspondência

Nível do funcionário vs tabagismo

	Hábito de Fumar				
Funcionário	**Não**	**Pouco**	**Médio**	**Muito**	**Total**
Nível 1	4	2	3	2	11
Nível 2	4	3	7	4	18
Nível 3	25	10	12	4	51
Nível 4	18	24	33	13	88
Nível 5	10	6	7	2	25
Total	61	45	62	25	193

Expected Frequencies

	F0	F1	F2	F3
N1	3,48	2,56	3,53	1,42
N2	5,69	4,20	5,78	2,33
N3	16,12	11,89	16,38	6,61
N4	27,81	20,52	28,27	11,40
N5	7,90	5,83	8,03	3,24

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Chi-Square Distances

	F0	F1	F2	F3	Total
N1	0,079	0,124	0,081	0,232	0,516
N2	0,502	0,341	0,256	1,194	2,293
N3	4,893	0,301	1,173	1,028	7,395
N4	3,463	0,591	0,792	0,225	5,070
N5	0,557	0,005	0,132	0,474	1,168
Total	9,493	1,362	2,434	3,153	16,442

Estatística χ^2 = 16,442 (p=0,172)

0,232/16,442

Relative Inertias

	F0	F1	F2	F3	Total
N1	0,005	0,008	0,005	0,014	0,031
N2	0,031	0,021	0,016	0,073	0,139
N3	0,298	0,018	0,071	0,063	0,450
N4	0,211	0,036	0,048	0,014	0,308
N5	0,034	0,000	0,008	0,029	0,071
Total	0,577	0,083	0,148	0,192	1,000

Inércia total=16,442/193=0,08518

Análise de Correspondência

Distribuição de funcionários de acordo com o tabagismo

	Hábito de Fumar				
Funcionário	**Não**	**Pouco**	**Médio**	**Muito**	**Total**
Nível 1	4	2	3	2	11
Nível 2	4	3	7	4	18
Nível 3	25	10	12	4	51
Nível 4	18	24	33	13	88
Nível 5	10	6	7	2	25
Total	61	45	62	25	193

Perfis Linha da Tabela — Y_{5x4}

	F0	F1	F2	F3	Mass
N1	0,364	0,182	0,273	0,182	0,057
N2	0,222	0,167	0,389	0,222	0,093
N3	0,490	0,196	0,235	0,078	0,264
N4	0,205	0,273	0,375	0,148	0,456
N5	0,400	0,240	0,280	0,080	0,130
Mass	0,316	0,233	0,321	0,130	

$$\Rightarrow \quad Y_{5\times4} = N\ D_{\lambda}\ M' \qquad F_{(k=2)} = N_{5\times5} D_{\lambda_{5\times2}}$$

λ=(0,2734 0,1001 0,0203)

$$F_{(k=2)} = \begin{pmatrix} 0,066 & -0,194 \\ -0,259 & -0,243 \\ 0,381 & -0,011 \\ -0,233 & 0,058 \\ 0,201 & 0,078 \end{pmatrix}$$

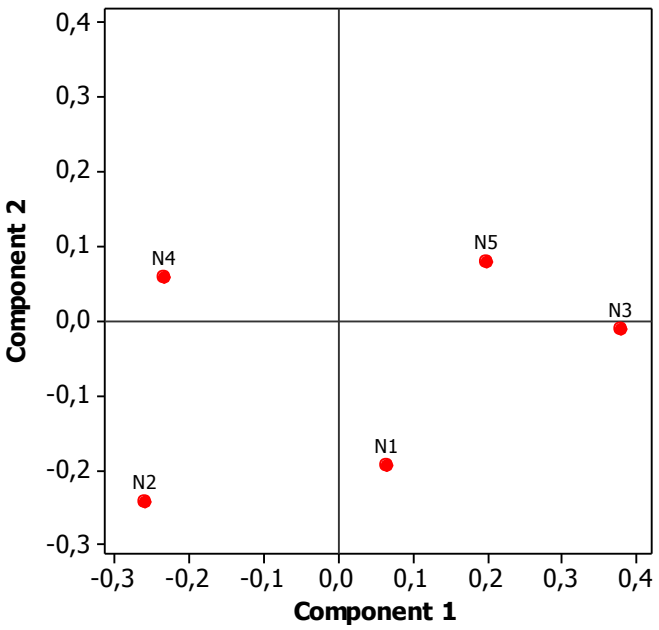
Nível1
Nível2
Nível3
Nível4
Nível5

Análise de Correspondência

Distribuição de funcionários de acordo com o tabagismo

	Hábito de Fumar				
Funcionário	**Não**	**Pouco**	**Médio**	**Muito**	**Total**
Nível 1	4	2	3	2	11
Nível 2	4	3	7	4	18
Nível 3	25	10	12	4	51
Nível 4	18	24	33	13	88
Nível 5	10	6	7	2	25
Total	61	45	62	25	193

Representação dos perfis linha



$$\Rightarrow in(eixo1) = \lambda_1^2 = (0{,}2734)^2 = 0{,}0748$$

$$\Rightarrow in(eixo2) = \lambda_2^2 = (0{,}1001)^2 = 0{,}01$$

0,0848/0,08518=0,995

$\Rightarrow$ 99,5% da inércia total dos dados está representada no plano

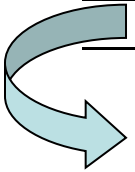
$\Rightarrow$ os funcionários níveis N5 e N3 são mais semelhantes em seu hábito de fumar. N2 e N4 estão mais distantes deste grupo, sendo mais semelhantes entre si. N1 ocupa uma posição intermediária entre estes grupos.

Análise de Correspondência

Representação dos Perfis Coluna da Tabela

Distribuição de funcionários de acordo com o tabagismo

	Hábito de Fumar				
Funcionário	**Não**	**Pouco**	**Médio**	**Muito**	**Total**
Nível 1	4	2	3	2	11
Nível 2	4	3	7	4	18
Nível 3	25	10	12	4	51
Nível 4	18	24	33	13	88
Nível 5	10	6	7	2	25
Total	61	45	62	25	193



Problema Dual:

Estudar o padrão de variação da variável hábito de fumar em função do nível funcional na empresa

$\Rightarrow$ Como representar os perfis das frequências relativas das colunas?

Análise de Correspondência

Representação dos Perfis Coluna da Tabela

	Variável Coluna					
Variável Linha	1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

$$p_{ij}^{c} = \frac{n_{ij}}{n_{.j}} \qquad i = 1,2,...,I$$

Estatística Qui-Quadrado:

$$\chi^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} = \frac{\left(O_{11}-E_{11}\right)^2}{E_{11}} + \frac{\left(O_{12}-E_{12}\right)^2}{E_{12}} + \frac{\left(O_{IJ}-E_{IJ}\right)^2}{E_{IJ}}; \qquad O_{ij} = n_{ij} \quad E_{ij} = \frac{n_{i.}n_{.j}}{n}$$

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{.j}\frac{\left(p_{ij}^{c}-\bar{p}_i\right)^2}{\bar{p}_i} = \sum_{j=1}^{J} n_{.j}\sum_{i=1}^{I}\frac{\left(p_{ij}^{c}-\bar{p}_i\right)^2}{\bar{p}_i} = \sum_{j=1}^{J} n_{.j} d_j^2$$

$$d_j^2 = \left(\mathbf{p}_j^{c}-\bar{\mathbf{p}}^{c}\right)' \mathbf{D}_{\bar{\mathbf{p}}^{c}}^{-1}\left(\mathbf{p}_j^{c}-\bar{\mathbf{p}}^{c}\right)$$

$\mathbf{p}_j^{c}$: perfil de freqüências relativas da coluna j

Análise de Correspondência

Representação dos Perfis Coluna da Tabela

	Variável Coluna					
Variável Linha	1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

$$\mathbf{p}_j = \left(p_{1j}, p_{2j}, ..., p_{Ij}\right)'; \qquad p_{ij}^c = \frac{n_{ij}}{n_{\cdot j}} \qquad i = 1,2,...,I$$

$$\overline{\mathbf{p}} = \overline{\mathbf{p}}^c = \left(\frac{n_{1\cdot}}{n}, \frac{n_{2\cdot}}{n}, ..., \frac{n_{I\cdot}}{n}\right)' \quad \text{Centróide (coluna)}$$

$$d_j^2 = \left(\mathbf{p}_j^c - \overline{\mathbf{p}}^c\right)' \mathbf{D}_{\overline{\mathbf{p}}^c}^{-1} \left(\mathbf{p}_j^c - \overline{\mathbf{p}}^c\right) = \sum_{i=1}^{I} \frac{\left(p_{ij}^c - \overline{p}_j^c\right)^2}{\overline{p}_j^c}$$

Distância Euclidiana ponderada ao quadrado do perfil de freqüências relativas da coluna j ao centróide

Como representar tais perfis coluna em um espaço multidimensional?

Análise de Correspondência

Representação dos Perfis Linha e Coluna da Tabela

	Variável Coluna					
Variável Linha	1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

linhas

$$\mathbf{Y}^{L}{}_{I\times J} = \begin{pmatrix} \mathbf{p}^{L}_{1_{1\times J}} \\ \dots \\ \mathbf{p}^{L}_{I_{1\times J}} \end{pmatrix} \Rightarrow$$

colunas

$$\mathbf{Y}^{C}{}_{J\times I} = \begin{pmatrix} \mathbf{p}^{C}_{1_{1\times I}} \\ \dots \\ \mathbf{p}^{C}_{J_{1\times I}} \end{pmatrix}$$

$$\bar{\mathbf{p}}^{C} = \left(\bar{p}^{C}_{1.},\dots,\bar{p}^{C}_{I.}\right)'$$

$$\bar{\mathbf{p}}^{L}_{1\times J} = \left(\bar{p}^{L}_{.1},\dots,\bar{p}^{L}_{.J}\right)'$$

$$D^{L}_{q\,J\times J} = diag\left(q^{L}_{j} = 1/\,\bar{p}^{L}_{.j}\right)$$

$$D^{L}_{w\,I\times I} = diag\left(w^{L}_{i} = n_{i.}/n\right)$$

$$D^{C}_{q\,I\times I} = diag\left(q^{C}_{i} = 1/\,\bar{p}^{C}_{i.}\right)$$: matriz de pesos

$$D^{C}_{w\,J\times J} = diag\left(w_{j} = n_{.j}/n\right)$$: matriz de massas

Análise de Correspondência

Representação dos Perfis Coluna da Tabela

$$\mathbf{Y}^{L}{}_{I\times J} = \begin{pmatrix} \mathbf{p}^{L}_{1_{1\times I}} \\ \dots \\ \mathbf{p}^{L}_{I_{1\times I}} \end{pmatrix} \quad \Rightarrow \quad \mathbf{Y}^{C}{}_{J\times I} = \begin{pmatrix} \mathbf{p}^{C}_{1_{1\times I}} \\ \dots \\ \mathbf{p}^{C}_{J_{1\times I}} \end{pmatrix} \qquad \bar{\mathbf{p}}^{C} = \left(\bar{p}^{C}_{1.},\dots,\bar{p}^{C}_{I.}\right)'$$

$$D^{C}_{q\,I\times I} = diag\left(q^{C}_{i} = 1/\,\bar{p}^{C}_{i.}\right) \qquad D^{C}_{w\,J\times J} = diag\left(w_{j} = n_{.j}/n\right)$$

Obter os eixos principais G dos perfis colunas $\mathbf{p}^{C}_{j} \Rightarrow$ obter a decomposição espectral da matriz $\boldsymbol{Y^{C}}$, tal que, para dimensões de ordem k, tem-se:

$$G_{(k)} = N^{C}_{(k)} D^{C}_{\lambda(k)} \quad \Rightarrow \quad Y^{C}{}_{J\times I} = N^{C}\; D^{C}_{\lambda}\; M^{C\,\prime} \quad ; \qquad N^{C\,\prime} D^{C}_{w} N^{C} = M^{C\,\prime} D^{C}_{q} M^{C} = I$$

$$\Rightarrow D^{C}_{\lambda} = D^{L}_{\lambda}$$

$\Rightarrow$ Os valores singulares da representação dos perfis de linha e coluna são os mesmos (a menos de autovalores nulos) $\Rightarrow$ o subespaço ótimo para a representação dos perfis linha e coluna é o mesmo !!

Análise de Correspondência

Distribuição de funcionários de acordo com o tabagismo

	Hábito de Fumar				
Funcionário	**Não**	**Pouco**	**Médio**	**Muito**	**Total**
Nível 1	4	2	3	2	11
Nível 2	4	3	7	4	18
Nível 3	25	10	12	4	51
Nível 4	18	24	33	13	88
Nível 5	10	6	7	2	25
Total	61	45	62	25	193

Perfis Coluna da Tabela — $Y^{C'}_{4x5}$

	F0	F1	F2	F3	Mass
N1	0,066	0,044	0,048	0,080	0,057
N2	0,066	0,067	0,113	0,160	0,093
N3	0,410	0,222	0,194	0,160	0,264
N4	0,295	0,533	0,532	0,520	0,456
N5	0,164	0,133	0,113	0,080	0,130
Mass	0,316	0,233	0,321	0,130	

$$\Rightarrow \quad Y^{C'}{}_{4\times5} = N^{C}\ D^{C}_{\lambda}\ M^{C'} \qquad G_{(k)} = N^{C}_{(k)} D_{\lambda(k)}$$

$$\lambda = (\ 0{,}2734 \quad 0{,}1001 \quad 0{,}0203\ )$$

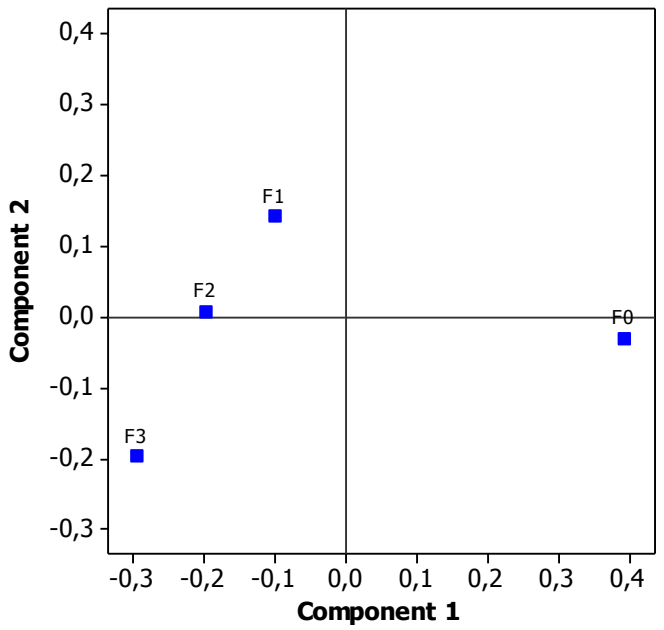
$$G_{(k=2)} = \begin{pmatrix} 0{,}393 & -0{,}031 \\ -0{,}1 & 0{,}141 \\ -0{,}196 & 0{,}007 \\ -0{,}294 & -0{,}198 \end{pmatrix} \begin{matrix} \leftarrow F0 \\ \leftarrow F1 \\ \leftarrow F2 \\ \leftarrow F3 \end{matrix}$$

Análise de Correspondência

Distribuição de funcionários de acordo com o tabagismo

	Hábito de Fumar				
Funcionário	**Não**	**Pouco**	**Médio**	**Muito**	**Total**
Nível 1	4	2	3	2	11
Nível 2	4	3	7	4	18
Nível 3	25	10	12	4	51
Nível 4	18	24	33	13	88
Nível 5	10	6	7	2	25
Total	61	45	62	25	193

Representação dos perfis coluna



$$\Rightarrow in(eixo1) = \lambda_1^2 = (0{,}2734)^2 = 0{,}0748$$

$$\Rightarrow in(eixo2) = \lambda_2^2 = (0{,}1001)^2 = 0{,}01$$

0,0848/0,08518=0,995

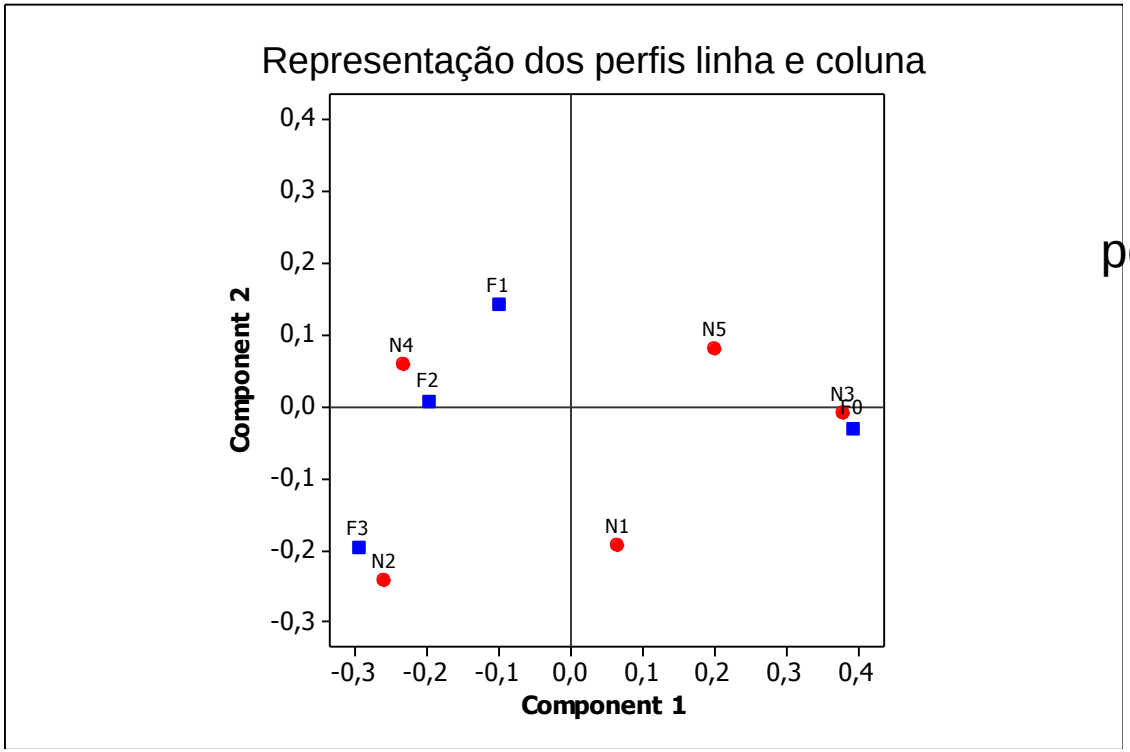
$\Rightarrow$ 99,5% da inércia total dos dados está representada no plano

$\Rightarrow$ disposição linear (C1) dos níveis de hábito de fumar. O grupo de não fumantes está bem distante dos demais

Análise de Correspondência

Distribuição de funcionários de acordo com o tabagismo

	Hábito de Fumar				
Funcionário	**Não**	**Pouco**	**Médio**	**Muito**	**Total**
Nível 1	4	2	3	2	11
Nível 2	4	3	7	4	18
Nível 3	25	10	12	4	51
Nível 4	18	24	33	13	88
Nível 5	10	6	7	2	25
Total	61	45	62	25	193



Representação conjunta dos perfis da frequência relativa das linhas e colunas da tabela

Análise de Correspondência via Escalonamento Multidimensional

Tabela de Contingência

	Variável Coluna					
Variável Linha	1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

Matriz de Distâncias Qui-Quadrado entre perfis (de proporções):

Entre os Perfis Linha

	1 2 ... I
1	
2	$d_{ij}^{2\,Linhas}$: distância entre as linhas i e j
...	
I	

Entre os Perfis Coluna

	1 2 ... J
1	
2	$d_{ij}^{2\,Colunas}$: distância entre as colunas i e j
...	
J	

Análise de Correspondência e Escalonamento Multidimensional

Variável Linha	Variável Coluna 1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

Perfis Linha

Variável Linha	Variável Coluna 1	…	J	Total
1	$p_{11}=n_{11}/n_{1.}$		$p_{1J}=n_{1J}/n_{1.}$	1
…	…	…	…	…
I	$p_{I1}=n_{I1}/n_{I.}$		$p_{IJ}=n_{IJ}/n_{I.}$	1

$$\Rightarrow p_{ij}^{L} = \frac{n_{ij}}{n_{i.}}$$

Perfis Coluna

Variável Linha	Variável Coluna 1	…	J
1	$p_{11}=n_{11}/n_{.1}$		$p_{1J}=n_{1J}/n_{.J}$
…	…	…	…
I	$p_{I1}=n_{I1}/n_{.1}$		$p_{IJ}=n_{IJ}/n_{.J}$
Total	1	…	1

$$\Rightarrow p_{ij}^{C} = \frac{n_{ij}}{n_{.j}}$$

Análise de Correspondência e Escalonamento Multidimensional

Variável Linha	Variável Coluna 1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

Distância Qui-Quadrado – Perfis Linha

$$p_{ij}^{L} = \frac{n_{ij}}{n_{i.}} \qquad i = 1,2,...,I$$

$$d_{ij}^{2\,Linhas} = \sum_{k=1}^{J} \frac{\left(p_{ik}^{c} - p_{jk}^{c}\right)^2}{p_{.k}}$$

Distância Qui-Quadrado – Perfis Coluna

$$p_{ij}^{c} = \frac{n_{ij}}{n_{.j}} \qquad j = 1,2,...,J$$

$$d_{ij}^{2\,Colunas} = \sum_{k=1}^{I} \frac{\left(p_{ki}^{L} - p_{kj}^{L}\right)^2}{p_{k.}}$$

$\Rightarrow$ Extrair as Coordenadas Principais das Matrizes de distâncias

D^{Linhas} e $D^{Colunas}$ $\Rightarrow$ resultados equivalentes à solução via dvs de Y^{L} e Y^{C}.

Análise de Correspondência e Escalonamento Multidimensional

Influência da idade da adolescente no tipo de namoro (Everitt, 2004)

	Faixa Etária				
	< 16	16-17	17-18	18-19	19-20
Nenhum namorado	21 (68)	21 (64)	14 (58)	13 (42)	8 (40)
Namoro sem sexo	8 (26)	9 (27)	6 (25)	8 (26)	2 (10)
Namoro com sexo	2 (6)	3 (9)	4 (17)	10 (32)	10 (50)
Total	31 (100%)	33 (100%)	24 (100%)	31 (100%)	20 (100%)

$d_{ij}^{Colunas}$

	< 16	**16-17**	**17-18**	**18-19**	**19-20**
<16	0,00	0,09	0,26	0,66	1,07
16-17		0,00	0,19	0,59	1,01
17-18			0,00	0,41	0,83
18-19				0,00	0,51
19-20					0,00

d_{ij}^{Linhas}

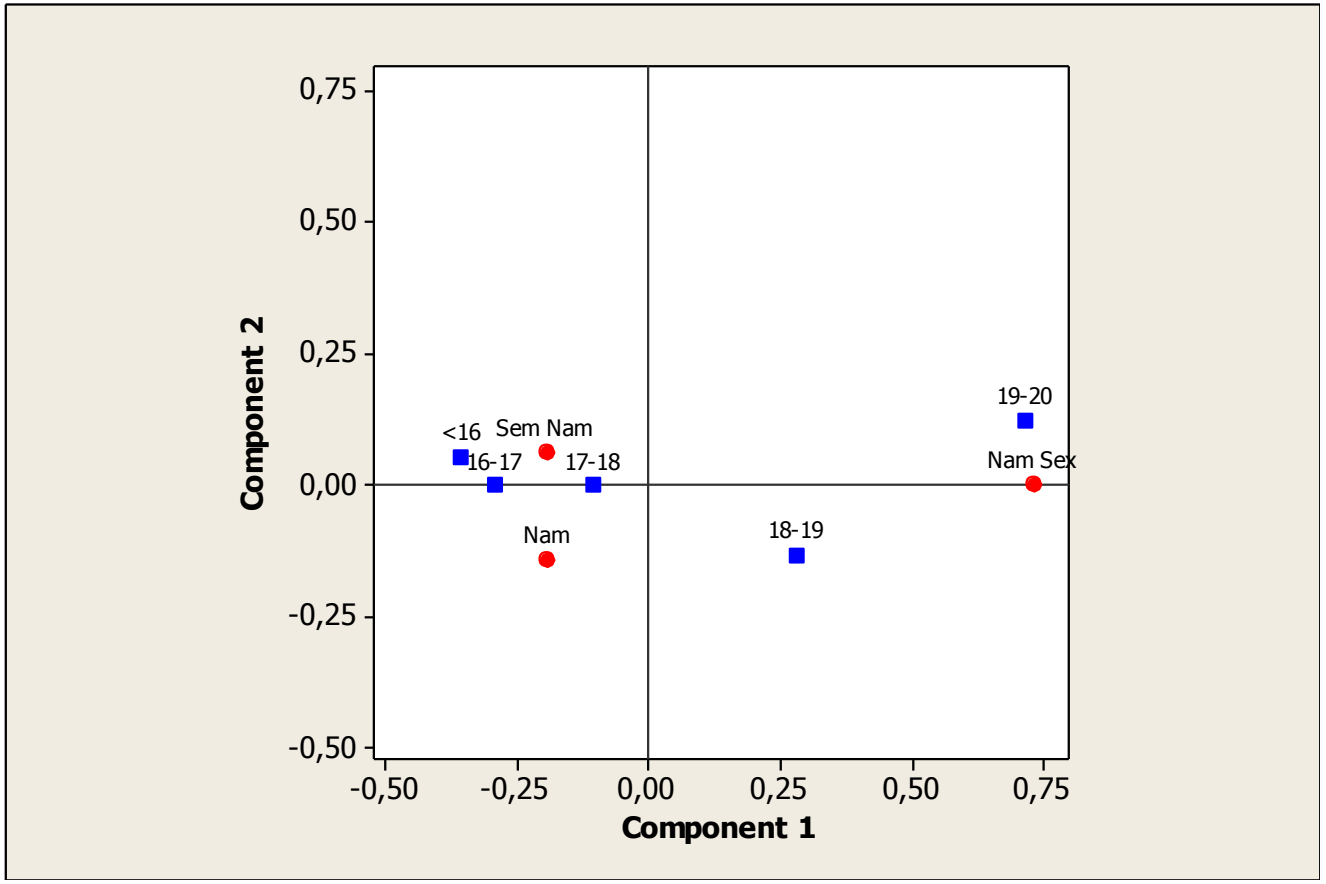
	Sem Nam	**Nam**	**NamSexo**
Sem Nam	0,00	0,21	0,93
Nam		0,00	0,93
NamSexo			0,00

Obter as coordenadas principais a partir das matrizes de distâncias Qui-Quadrado.

Análise de Correspondência e Escalonamento Multidimensional

Influência da idade da adolescente no tipo de namoro (Everitt, 2004)

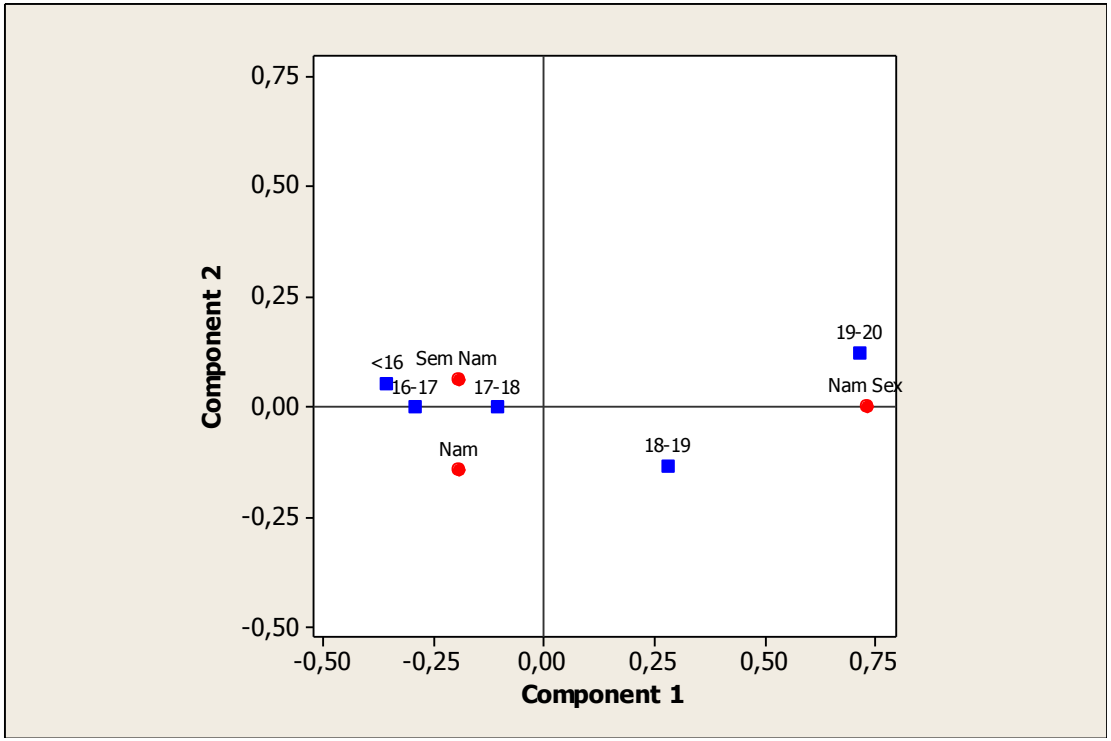
	CP1	CP2
Sem Nam	-0,1933	0,0610
Nam	-0,1924	-0,1425
Nam Sex	0,7322	0,0002
<16	-0,3547	0,0550
16-17	-0,2897	-0,0003
17-18	-0,1033	-0,0001
18-19	0,2806	-0,1342
19-20	0,7169	0,1234



Análise de Correspondência e Escalonamento Multidimensional

Influência da idade da adolescente no tipo de namoro (Everitt, 2004)

	CP1	CP2
Sem Nam	-0,1933	0,0610
Nam	-0,1924	-0,1425
Nam Sex	0,7322	0,0002
<16	-0,3547	0,0550
16-17	-0,2897	-0,0003
17-18	-0,1033	-0,0001
18-19	0,2806	-0,1342
19-20	0,7169	0,1234



$$d_{Euclid}(<16,16-17) = \sqrt{(-0,3547+0,2897)^2 + (0,055+0,0003)^2} = 0,09$$

$$d_{Qui-Quad}(<16,16-17) = \sqrt{\frac{(0,68-0,64)^2}{0,55} + \frac{(0,26-0,27)^2}{0,24} + \frac{(0,06-0,09)^2}{0,21}} = 0,09$$

Análise de Correspondência e Escalonamento Multidimensional

	Variável Coluna					
Variável Linha	1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

• Em um gráfico de coordenadas principais representando somente as categorias linha (ou coluna), as distâncias entre os pontos são distâncias Euclidianas.

• Mas, em um gráfico onde ambos os espaços (linha e coluna) estão representados simultaneamente, é preciso ter cuidado com a comparação entre categorias linha e coluna pois neste caso a medida de distância Euclidiana pode não ser válida $\Rightarrow$ uma melhor aproximação pode ser conseguida com a padronização das coordenadas principais (dividir os valores pela raiz quadrada da inércia do componente) $\Rightarrow$ coordenadas assimétricas (no Minitab)