CAPÍTULO 11

# Análise de Correspondência Simples e Múltipla

*O mundo recompensa com mais frequência as aparências do mérito do que o próprio mérito.*
**François de La Rochefoucauld**

Ao final deste capítulo, você terá condições de:

- Estabelecer as circunstâncias a partir das quais as técnicas de análise de correspondência podem ser utilizadas.
- Saber diferenciar a análise de correspondência simples da análise de correspondência múltipla.
- Entender como os bancos de dados devem ser dispostos para a elaboração das técnicas.
- Saber interpretar os resultados do teste $\chi^2$.
- Compreender os conceitos de frequências absolutas e relativas e de resíduos em tabelas de contingência.
- Saber calcular e interpretar as inércias principais parciais e totais.
- Gerar coordenadas das categorias das variáveis e construir mapas perceptuais.
- Entender as diferenças entre o método da matriz binária e o método da matriz de Burt para a elaboração da análise de correspondência múltipla.
- Elaborar as técnicas de análise de correspondência simples e múltipla de maneira algébrica e por meio do IBM SPSS Statistics Software® e do Stata Statistical Software® e interpretar seus resultados.

## 11.1. INTRODUÇÃO

As técnicas exploratórias de análise de correspondência simples e múltipla são muito úteis quando há a intenção de se trabalhar com variáveis que apresentam dados categóricos, como as variáveis qualitativas, e deseja-se investigar a **associação** entre as variáveis e entre suas categorias.

Imagine que um pesquisador tenha interesse em estudar a **relação de interdependência** entre duas variáveis categóricas, por exemplo, comportamento de consumo, descrito pela preferência por determinados tipos de estabelecimento varejista, e faixa de idade dos consumidores. Nessa situação, a **análise de correspondência simples** pode ser utilizada, uma vez que é uma técnica bivariada que permite investigar a associação entre duas, e somente duas, variáveis categóricas.

Em outra situação, pode-se investigar a relação entre o país de origem, o setor de atuação e a faixa de lucratividade de empresas de capital aberto. Nesse caso, a **análise de correspondência múltipla** pode ser utilizada, já que se trata de uma técnica multivariada que possibilita a investigação da existência de associação entre mais de duas variáveis categóricas.

Segundo Greenacre (2008), as técnicas de análise de correspondência são métodos de representação de linhas e colunas de tabelas cruzadas de dados como **coordenadas** em um gráfico, chamado **mapa perceptual**, a partir do qual se podem interpretar as similaridades e diferenças de comportamento entre variáveis e entre categorias. Portanto, essas técnicas têm como principal objetivo avaliar a significância dessas similaridades, determinar coordenadas das categorias com base na distribuição dos dados em tabelas cruzadas e, a partir dessas coordenadas, construir **mapas perceptuais**, que nada mais são que **diagramas de dispersão** que representam as categorias das variáveis na forma de pontos em relação a eixos de coordenadas ortogonais. São, portanto, mapas de categorias.

Embora a origem teórica dessas técnicas regrida à primeira metade do século XX, com o seminal trabalho de Hirschfeld (1935), foi o matemático e linguista francês Jean-Paul Benzécri que deu um impulso realmente significativo

às aplicações modernas da análise de correspondência, a partir da década de 1960, com estudos realizados na Universidade de Rennes e, posteriormente, na Universidade de Paris. Anos mais tarde, o holandês Jan de Leeuw e o japonês Chikio Hayashi também fizeram importantes contribuições para o desenvolvimento teórico e prático das técnicas. Em 1984, Greenacre publica uma importante obra (*Theory and Applications of Correspondence Analysis*), que acaba por contribuir para uma ampla difusão das técnicas de análise de correspondência em diversas partes do mundo.

As técnicas de análise de correspondência simples e múltipla permitem considerar todo e qualquer tipo de categoria de variáveis, sem que o pesquisador precise fazer uso do **incorreto procedimento de ponderação arbitrária**, infelizmente ainda tão praticado em ambientes acadêmicos e organizacionais. Variáveis em **escala Likert**, por exemplo, sofrem constantemente com esse tipo de manipulação, visto que, com frequência, pesquisadores atribuem pesos arbitrários a cada uma das possíveis categorias. As técnicas de análise de correspondência são bastante úteis para que o pesquisador perceba a incoerência desse tipo de prática!

Conforme discutido nos dois capítulos anteriores, a análise de correspondência deve ser definida com base na teoria subjacente e na experiência do pesquisador, de modo que seja possível aplicá-la de forma correta e analisar os resultados obtidos.

Neste capítulo, trataremos das técnicas de análise de correspondência simples e múltipla, com os seguintes objetivos: (1) introduzir os conceitos; (2) apresentar, de maneira algébrica e prática, o passo a passo da modelagem; (3) interpretar os resultados obtidos; e (4) propiciar a aplicação das técnicas em SPSS e Stata. Seguindo a lógica dos dois capítulos anteriores, será inicialmente elaborada a solução algébrica de um exemplo vinculada à apresentação dos conceitos. Somente após a introdução dos conceitos serão apresentados os procedimentos para a elaboração das técnicas em SPSS e Stata.

## 11.2. ANÁLISE DE CORRESPONDÊNCIA SIMPLES

A análise de correspondência simples, também conhecida por **Anacor**, é uma técnica de análise bivariada por meio da qual é estudada a associação entre duas variáveis categóricas e entre suas categorias, bem como a intensidade dessa associação, a partir de uma tabela cruzada de dados, conhecida por **tabela de contingência**, em que são dispostas em cada célula as **frequências absolutas observadas** para cada par de categorias das duas variáveis. A tabela de contingência também é chamada de **tabela de correspondência**, **tabela de classificação cruzada** ou ***cross-tabulation***.

Nas seções seguintes, apresentaremos o desenvolvimento teórico da técnica, bem como a elaboração de um exemplo prático. Enquanto nas seções 11.2.1 a 11.2.4 serão apresentados os principais conceitos, a seção 11.2.5 é destinada à resolução de um exemplo prático por meio de solução algébrica a partir de um banco de dados.

### 11.2.1. Notação

Imaginemos um banco de dados que apresenta apenas e tão somente duas variáveis categóricas, em que a primeira possui $I$ categorias, e a segunda, $J$ categorias. Logo, a partir desse banco de dados, é possível definir uma tabela de contingência $\mathbf{X_o}$ (*cross-tabulation*) que apresenta as frequências absolutas observadas das categorias das duas variáveis, em que determinada célula $ij$ contém certa quantidade $n_{ij}$ ($i = 1, ..., I$ e $j = 1, ..., J$) de observações. A quantidade total de observações $N$ do banco de dados pode, portanto, ser expressa por:

$$N = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \tag{11.1}$$

A representação geral de uma tabela de contingência é:

**Tabela 11.1** Representação geral de uma tabela de contingência (frequências absolutas observadas).

| | 1 | 2 | ... | $J$ |
|---|---|---|---|---|
| 1 | $n_{11}$ | $n_{12}$ | ... | $n_{1J}$ |
| 2 | $n_{21}$ | $n_{22}$ | ... | $n_{2J}$ |
| ⋮ | ⋮ | ⋮ | ... | ⋮ |
| $I$ | $n_{I1}$ | $n_{I2}$ | ... | $n_{IJ}$ |

Na forma matricial, a tabela pode ser representada da seguinte maneira:

$$\mathbf{X_o}=\begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1J} \\ n_{21} & n_{22} & \cdots & n_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ n_{I1} & n_{I2} & \cdots & n_{IJ} \end{pmatrix} \tag{11.2}$$

Como um dos principais objetivos da análise de correspondência simples é estudar a existência de associação estatisticamente significante a determinado nível de significância entre duas variáveis categóricas e entre as categorias de cada uma, devemos partir para o estudo do teste $\chi^2$ e dos resíduos em tabelas de contingência.

### 11.2.2. Associação entre duas variáveis categóricas e entre suas categorias: teste $\chi^2$ e análise dos resíduos

Uma vez que a matriz $\mathbf{X_o}$ da expressão (11.2) apresenta as frequências absolutas observadas para cada combinação de categorias das duas variáveis, podemos definir a expressão de uma matriz $\mathbf{X_e}$ que oferece as **frequências absolutas esperadas** em cada célula. Para tanto, à Tabela 11.1 podem ser acrescentados os valores totais das frequências absolutas observadas em cada linha e coluna, conforme mostra a Tabela 11.2.

**Tabela 11.2** Tabela de contingência com valores totais por linha e coluna.

| | 1 | 2 | ... | $J$ | **Total** |
|---|---|---|---|---|---|
| 1 | $n_{11}$ | $n_{12}$ | ... | $n_{1J}$ | $\sum l_1$ |
| 2 | $n_{21}$ | $n_{22}$ | | $n_{2J}$ | $\sum l_2$ |
| ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ |
| $I$ | $n_{I1}$ | $n_{I2}$ | | $n_{IJ}$ | $\sum l_I$ |
| **Total** | $\sum c_1$ | $\sum c_2$ | ... | $\sum c_J$ | $N$ |

Obviamente, sabemos que:

$$\sum c_1+\sum c_2+\ldots+\sum c_J=\sum l_1+\sum l_2+\ldots+\sum l_I=N \tag{11.3}$$

Logo, a tabela que apresenta as frequências absolutas esperadas de cada célula pode ser definida de acordo com o apresentado na Tabela 11.3.

**Tabela 11.3** Tabela com frequências absolutas esperadas em cada célula.

| | 1 | 2 | ... | $J$ |
|---|---|---|---|---|
| 1 | $\left(\frac{\sum c_1 \cdot \sum l_1}{N}\right)$ | $\left(\frac{\sum c_2 \cdot \sum l_1}{N}\right)$ | ... | $\left(\frac{\sum c_J \cdot \sum l_1}{N}\right)$ |
| 2 | $\left(\frac{\sum c_1 \cdot \sum l_2}{N}\right)$ | $\left(\frac{\sum c_2 \cdot \sum l_2}{N}\right)$ | | $\left(\frac{\sum c_J \cdot \sum l_2}{N}\right)$ |
| ⋮ | ⋮ | | | ⋮ |
| $I$ | $\left(\frac{\sum c_1 \cdot \sum l_I}{N}\right)$ | $\left(\frac{\sum c_2 \cdot \sum l_I}{N}\right)$ | | $\left(\frac{\sum c_J \cdot \sum l_I}{N}\right)$ |

Na forma matricial, essa tabela pode ser escrita como:

$$\mathbf{X_e} = \begin{pmatrix} \left(\frac{\sum c_1 \cdot \sum l_1}{N}\right) & \left(\frac{\sum c_2 \cdot \sum l_1}{N}\right) & \cdots & \left(\frac{\sum c_J \cdot \sum l_1}{N}\right) \\ \left(\frac{\sum c_1 \cdot \sum l_2}{N}\right) & \left(\frac{\sum c_2 \cdot \sum l_2}{N}\right) & \cdots & \left(\frac{\sum c_J \cdot \sum l_2}{N}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{\sum c_1 \cdot \sum l_I}{N}\right) & \left(\frac{\sum c_2 \cdot \sum l_I}{N}\right) & \cdots & \left(\frac{\sum c_J \cdot \sum l_I}{N}\right) \end{pmatrix} \tag{11.4}$$

Portanto, podemos definir uma **matriz de resíduos, E**, cujos valores se referem às diferenças, para cada célula, entre as frequências absolutas observadas e esperadas. Logo, temos que:

$$\mathbf{E} = \begin{pmatrix} n_{11} - \left(\frac{\sum c_1 \cdot \sum l_1}{N}\right) & n_{12} - \left(\frac{\sum c_2 \cdot \sum l_1}{N}\right) & \cdots & n_{1J} - \left(\frac{\sum c_J \cdot \sum l_1}{N}\right) \\ n_{21} - \left(\frac{\sum c_1 \cdot \sum l_2}{N}\right) & n_{22} - \left(\frac{\sum c_2 \cdot \sum l_2}{N}\right) & \cdots & n_{2J} - \left(\frac{\sum c_J \cdot \sum l_2}{N}\right) \\ \vdots & \vdots & \ddots & \vdots \\ n_{I1} - \left(\frac{\sum c_1 \cdot \sum l_I}{N}\right) & n_{I2} - \left(\frac{\sum c_2 \cdot \sum l_I}{N}\right) & \cdots & n_{IJ} - \left(\frac{\sum c_J \cdot \sum l_I}{N}\right) \end{pmatrix} \tag{11.5}$$

E, com base nas matrizes $\mathbf{X_e}$ e $\mathbf{E}$, podemos definir a estatística $\chi^2$ conforme segue, de maneira análoga ao exposto na expressão (3.1) do Capítulo 3:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left[n_{ij} - \left(\frac{\sum c_j \cdot \sum l_i}{N}\right)\right]^2}{\left(\frac{\sum c_j \cdot \sum l_i}{N}\right)} \tag{11.6}$$

com $(I-1) \times (J-1)$ graus de liberdade, conforme estudamos no Capítulo 3.

Em outras palavras, a estatística $\chi^2$ corresponde à somatória, para todas as células, dos valores correspondentes à razão entre o resíduo ao quadrado e a frequência esperada em cada célula. Sendo assim, para dado número de graus de liberdade e determinado nível de significância, se o valor total da estatística $\chi^2$ for maior que seu valor crítico, poderemos afirmar que existe associação estatisticamente significante entre as duas variáveis categóricas, ou seja, a distribuição das frequências das categorias de uma variável segundo as categorias da outra não será aleatória, e, portanto, haverá um padrão de dependência entre essas variáveis. Podemos, portanto, definir as hipóteses nula e alternativa do teste $\chi^2$ referente a essa estatística da seguinte maneira:

$H_0$: as duas variáveis categóricas se associam de forma aleatória.

$H_1$: a associação entre as duas variáveis categóricas não se dá de forma aleatória.

É importante mencionar que a estatística $\chi^2$ aumenta à medida que cresce o tamanho da amostra ($N$), o que pode prejudicar a análise da associação existente em tabelas de contingência. Para que tal problema seja superado, segundo Beh (2004), a análise de correspondência faz uso da **inércia principal total** de uma tabela de contingência para descrever o nível de associação entre duas variáveis categóricas, expressa por:

$$I_T = \frac{\chi^2}{N} \tag{11.7}$$

Ainda segundo Beh (2004), a decomposição da inércia principal total de uma tabela de contingência pode auxiliar o pesquisador na identificação de fontes importantes de informação que possam ajudar a descrever a associação entre duas variáveis categóricas e, como consequência, propiciar a construção de mapas perceptuais. O tipo mais comum de decomposição inercial corresponde à **determinação de autovalores**, a ser abordada na próxima seção.

Antes disso, porém, precisamos elaborar um estudo mais aprofundado das relações entre as duas variáveis, com foco em suas categorias, fazendo uso dos **resíduos padronizados** e dos **resíduos padronizados ajustados**. Enquanto o teste $\chi^2$ permite avaliar se a distribuição das frequências das categorias de uma variável segundo as categorias da outra é aleatória ou se há um padrão de dependência entre as duas, a análise dos resíduos padronizados ajustados, segundo Batista, Escuder e Pereira (2004), revela os padrões característicos de cada categoria de uma variável segundo o excesso ou a falta de ocorrências de sua combinação com cada categoria da outra variável. Vamos, então, introduzir seus conceitos.

Seguindo Barnett e Lewis (1994), podemos definir os resíduos padronizados em uma tabela de contingência dividindo-se em cada célula o valor do resíduo calculado pela raiz quadrada da respectiva frequência absoluta esperada. Sendo assim, temos, para determinada célula $ij$ ($i = 1, ..., I$ e $j = 1, ..., J$), que:

$$e_{\text{padronizado}_{ij}} = \frac{n_{ij} - ne_{ij}}{\sqrt{ne_{ij}}} \tag{11.8}$$

em que $n_{ij}$ e $ne_{ij}$ se referem, respectivamente, às frequências absolutas observadas e às frequências absolutas esperadas. Portanto, com base na Tabela 11.3 e na expressão (11.4), podemos definir uma **matriz de resíduos padronizados**, $\mathbf{E}_{\text{padronizado}}$, da seguinte forma:

$$\mathbf{E}_{\text{padronizado}} = \begin{pmatrix} \dfrac{n_{11} - \left(\dfrac{\sum c_1 \cdot \sum l_1}{N}\right)}{\sqrt{\left(\dfrac{\sum c_1 \cdot \sum l_1}{N}\right)}} & \dfrac{n_{12} - \left(\dfrac{\sum c_2 \cdot \sum l_1}{N}\right)}{\sqrt{\left(\dfrac{\sum c_2 \cdot \sum l_1}{N}\right)}} & \cdots & \dfrac{n_{1J} - \left(\dfrac{\sum c_J \cdot \sum l_1}{N}\right)}{\sqrt{\left(\dfrac{\sum c_J \cdot \sum l_1}{N}\right)}} \\ \dfrac{n_{21} - \left(\dfrac{\sum c_1 \cdot \sum l_2}{N}\right)}{\sqrt{\left(\dfrac{\sum c_1 \cdot \sum l_2}{N}\right)}} & \dfrac{n_{22} - \left(\dfrac{\sum c_2 \cdot \sum l_2}{N}\right)}{\sqrt{\left(\dfrac{\sum c_2 \cdot \sum l_2}{N}\right)}} & \cdots & \dfrac{n_{2J} - \left(\dfrac{\sum c_J \cdot \sum l_2}{N}\right)}{\sqrt{\left(\dfrac{\sum c_J \cdot \sum l_2}{N}\right)}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{n_{I1} - \left(\dfrac{\sum c_1 \cdot \sum l_I}{N}\right)}{\sqrt{\left(\dfrac{\sum c_1 \cdot \sum l_I}{N}\right)}} & \dfrac{n_{I2} - \left(\dfrac{\sum c_2 \cdot \sum l_I}{N}\right)}{\sqrt{\left(\dfrac{\sum c_2 \cdot \sum l_I}{N}\right)}} & \cdots & \dfrac{n_{IJ} - \left(\dfrac{\sum c_J \cdot \sum l_I}{N}\right)}{\sqrt{\left(\dfrac{\sum c_J \cdot \sum l_I}{N}\right)}} \end{pmatrix} \tag{11.9}$$

A partir dos resíduos padronizados, podemos calcular os resíduos padronizados ajustados propostos por Haberman (1973), cuja expressão geral, para cada célula $ij$ ($i = 1, ..., I$ e $j = 1, ..., J$), é dada por:

$$e_{\text{padronizado ajustado}_{ij}} = \frac{e_{\text{padronizado}_{ij}}}{\sqrt{\left(1 - \dfrac{\sum c_j}{N}\right) \cdot \left(1 - \dfrac{\sum l_i}{N}\right)}} \tag{11.10}$$

e, analogamente, podemos definir uma **matriz de resíduos padronizados ajustados**, $\mathbf{E}_{\text{padronizado ajustado}}$, da seguinte maneira:

$$\mathbf{E}_{\text{padronizado ajustado}} = \begin{pmatrix} \frac{e_{\text{padronizado}_{11}}}{\sqrt{\left(1-\frac{\sum c_1}{N}\right)\cdot\left(1-\frac{\sum l_1}{N}\right)}} & \frac{e_{\text{padronizado}_{12}}}{\sqrt{\left(1-\frac{\sum c_2}{N}\right)\cdot\left(1-\frac{\sum l_1}{N}\right)}} & \cdots & \frac{e_{\text{padronizado}_{1J}}}{\sqrt{\left(1-\frac{\sum c_J}{N}\right)\cdot\left(1-\frac{\sum l_1}{N}\right)}} \\ \frac{e_{\text{padronizado}_{21}}}{\sqrt{\left(1-\frac{\sum c_1}{N}\right)\cdot\left(1-\frac{\sum l_2}{N}\right)}} & \frac{e_{\text{padronizado}_{22}}}{\sqrt{\left(1-\frac{\sum c_2}{N}\right)\cdot\left(1-\frac{\sum l_2}{N}\right)}} & \cdots & \frac{e_{\text{padronizado}_{2J}}}{\sqrt{\left(1-\frac{\sum c_J}{N}\right)\cdot\left(1-\frac{\sum l_2}{N}\right)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{e_{\text{padronizado}_{I1}}}{\sqrt{\left(1-\frac{\sum c_1}{N}\right)\cdot\left(1-\frac{\sum l_I}{N}\right)}} & \frac{e_{\text{padronizado}_{I2}}}{\sqrt{\left(1-\frac{\sum c_2}{N}\right)\cdot\left(1-\frac{\sum l_I}{N}\right)}} & \cdots & \frac{e_{\text{padronizado}_{IJ}}}{\sqrt{\left(1-\frac{\sum c_J}{N}\right)\cdot\left(1-\frac{\sum l_I}{N}\right)}} \end{pmatrix} \quad (11.11)$$

Segundo Batista, Escuder e Pereira (2004, tanto para o estudo da associação entre as variáveis (teste $\chi^2$) quanto para o dos padrões característicos de cada categoria de uma variável segundo o excesso ou a falta de ocorrências de sua combinação com cada categoria da outra variável (análise dos resíduos padronizados ajustados), é comum adotar, como veremos mais adiante, o nível de significância de 5% para o excesso de ocorrências em determinada célula, que corresponde a um resíduo padronizado ajustado com valor positivo superior a 1,96 (distribuição normal padrão, conforme mostra a Tabela E do apêndice do livro). Nesse sentido, caso determinada célula apresente um resíduo padronizado ajustado com valor superior a 1,96, poderemos caracterizar a associação entre as duas categorias correspondentes a ela (cada uma proveniente de uma variável).

Sendo assim, tão importante quanto avaliar a existência de associação estatisticamente significante entre duas variáveis categóricas é estudar a relação de dependência entre cada par de categorias, o que, inclusive, facilitará a análise do mapa perceptual a ser construído, como veremos no final da seção 11.2.5.

Elaboradas as análises, podemos, de fato, partir para o estudo da decomposição inercial, a fim de que sejam definidas as coordenadas de cada categoria de cada variável e, consequentemente, construído o mapa perceptual.

### 11.2.3. Decomposição inercial: a determinação de autovalores

Tradicionalmente, o método de decomposição de autovalores é conhecido por **método Eckart-Young**, em que são gerados $m$ autovalores, sendo $m = \text{mín}(I-1, J-1)$. Se, por exemplo, determinada base de dados oferecer uma tabela de contingência com dimensões (3 × 3), serão calculados $m = 2$ autovalores que, na análise de correspondência, também são chamados de **inércias principais parciais**.

Inicialmente, vamos definir uma matriz de proporções **P**, também conhecida por **matriz de frequências relativas observadas**, cujos valores são calculados com base na matriz $\mathbf{X_o}$, conforme mostra a Tabela 11.4.

**Tabela 11.4** Tabela com frequências relativas observadas em cada célula.

| | 1 | 2 | ... | $J$ | **Total** |
|---|---|---|---|---|---|
| 1 | $\frac{n_{11}}{N}$ | $\frac{n_{12}}{N}$ | | $\frac{n_{1J}}{N}$ | $\frac{\sum l_1}{N}$ |
| 2 | $\frac{n_{21}}{N}$ | $\frac{n_{22}}{N}$ | ... | $\frac{n_{2J}}{N}$ | $\frac{\sum l_2}{N}$ |
| ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ |
| $I$ | $\frac{n_{I1}}{N}$ | $\frac{n_{I2}}{N}$ | | $\frac{n_{IJ}}{N}$ | $\frac{\sum l_I}{N}$ |
| **Total** | $\frac{\sum c_1}{N}$ | $\frac{\sum c_2}{N}$ | ... | $\frac{\sum c_J}{N}$ | **1,00** |

Na forma matricial, essa tabela pode ser representada por:

$$\mathbf{P}=\frac{1}{N}\cdot\mathbf{X}_o=\begin{pmatrix}\frac{n_{11}}{N} & \frac{n_{12}}{N} & \cdots & \frac{n_{1J}}{N}\\ \frac{n_{21}}{N} & \frac{n_{22}}{N} & \cdots & \frac{n_{2J}}{N}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{n_{I1}}{N} & \frac{n_{I2}}{N} & \cdots & \frac{n_{IJ}}{N}\end{pmatrix} \tag{11.12}$$

Com base na tabela de frequências relativas observadas (matriz **P**), podemos definir o conceito de **massa**, que representa uma medida de influência ou preponderância de determinada categoria em relação às demais, com base em sua frequência observada. Sendo assim, podemos determinar as massas das categorias da variável disposta em linha e, da mesma forma, das categorias da variável disposta em coluna na tabela de contingência. As Tabelas 11.5 e 11.6 apresentam essas massas, com destaque para as **massas médias** de cada categoria em linha ou em coluna.

**Tabela 11.5** Massas – *Column profiles*.

| | 1 | 2 | ... | $J$ | **Massa** |
|---|---|---|---|---|---|
| 1 | $\left(\frac{n_{11}}{\sum c_1}\right)$ | $\left(\frac{n_{12}}{\sum c_2}\right)$ | | $\left(\frac{n_{1J}}{\sum c_J}\right)$ | $\frac{\sum l_1}{N}$ |
| 2 | $\left(\frac{n_{21}}{\sum c_1}\right)$ | $\left(\frac{n_{22}}{\sum c_2}\right)$ | ... | $\left(\frac{n_{2J}}{\sum c_J}\right)$ | $\frac{\sum l_2}{N}$ |
| ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ |
| $I$ | $\left(\frac{n_{I1}}{\sum c_1}\right)$ | $\left(\frac{n_{I2}}{\sum c_2}\right)$ | | $\left(\frac{n_{IJ}}{\sum c_1}\right)$ | $\frac{\sum l_I}{N}$ |
| **Total** | **1,000** | **1,000** | **...** | **1,000** | |

**Tabela 11.6** Massas – *Row profiles*.

| | 1 | 2 | ... | $J$ | **Total** |
|---|---|---|---|---|---|
| 1 | $\left(\frac{n_{11}}{\sum l_1}\right)$ | $\left(\frac{n_{12}}{\sum l_1}\right)$ | | $\left(\frac{n_{1J}}{\sum l_1}\right)$ | **1,000** |
| 2 | $\left(\frac{n_{21}}{\sum l_2}\right)$ | $\left(\frac{n_{22}}{\sum l_2}\right)$ | ... | $\left(\frac{n_{2J}}{\sum l_2}\right)$ | **1,000** |
| ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ |
| $I$ | $\left(\frac{n_{I1}}{\sum l_I}\right)$ | $\left(\frac{n_{I2}}{\sum l_I}\right)$ | | $\left(\frac{n_{IJ}}{\sum l_I}\right)$ | **1,000** |
| **Massa** | $\frac{\sum c_1}{N}$ | $\frac{\sum c_2}{N}$ | **...** | $\frac{\sum c_J}{N}$ | |

Com base nos valores das massas médias em linha e em coluna, podemos definir duas matrizes diagonais, $\mathbf{D}_l$ e $\mathbf{D}_c$, que contêm, respectivamente, esses valores em suas diagonais principais. Sendo assim, temos que:

$$\mathbf{D}_l=\begin{pmatrix}\frac{\sum l_1}{N} & 0 & \cdots & 0\\ 0 & \frac{\sum l_2}{N} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \frac{\sum l_I}{N}\end{pmatrix} \tag{11.13}$$

e

$$\mathbf{D}_c = \begin{pmatrix} \frac{\sum c_1}{N} & 0 & \cdots & 0 \\ 0 & \frac{\sum c_2}{N} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\sum c_J}{N} \end{pmatrix} \tag{11.14}$$

Note que, enquanto os valores da diagonal principal da matriz $\mathbf{D}_l$ são oriundos da Tabela 11.5 (*column profiles*), os valores da diagonal principal da matriz $\mathbf{D}_c$ são provenientes da Tabela 11.6 (*row profiles*).

Segundo Johnson e Wichern (2007), a decomposição inercial para a elaboração da análise de correspondência consiste em calcular os autovalores de uma matriz $\mathbf{W} = \mathbf{A'A}$, em que $\mathbf{A}$ pode ser definida da seguinte forma:

$$\mathbf{A} = \mathbf{D}_l^{-1/2} \cdot (\mathbf{P} - lc') \cdot \mathbf{D}_c^{-1/2} \tag{11.15}$$

sendo:

$$\mathbf{P} - lc' = \begin{pmatrix} \left(\frac{n_{11}}{N} - \frac{\sum l_1}{N} \cdot \frac{\sum c_1}{N}\right) & \left(\frac{n_{12}}{N} - \frac{\sum l_1}{N} \cdot \frac{\sum c_2}{N}\right) & \cdots & \left(\frac{n_{1J}}{N} - \frac{\sum l_1}{N} \cdot \frac{\sum c_J}{N}\right) \\ \left(\frac{n_{21}}{N} - \frac{\sum l_2}{N} \cdot \frac{\sum c_1}{N}\right) & \left(\frac{n_{22}}{N} - \frac{\sum l_2}{N} \cdot \frac{\sum c_2}{N}\right) & \cdots & \left(\frac{n_{2J}}{N} - \frac{\sum l_2}{N} \cdot \frac{\sum c_J}{N}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{n_{I1}}{N} - \frac{\sum l_I}{N} \cdot \frac{\sum c_1}{N}\right) & \left(\frac{n_{I2}}{N} - \frac{\sum l_I}{N} \cdot \frac{\sum c_2}{N}\right) & \cdots & \left(\frac{n_{IJ}}{N} - \frac{\sum l_I}{N} \cdot \frac{\sum c_J}{N}\right) \end{pmatrix} \tag{11.16}$$

Pode-se provar que os valores das células da matriz $\mathbf{A}$ são iguais aos valores das respectivas células da matriz $\mathbf{E}_{\text{padronizado}}$ divididos pela raiz quadrada do tamanho da amostra ($\sqrt{N}$).

Se, por exemplo, $\mathbf{A}$ for uma matriz (3 × 3), $\mathbf{W}$ também será uma matriz (3 × 3) com a seguinte expressão:

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{pmatrix} \tag{11.17}$$

da qual podem ser calculados os autovalores ($\lambda^2$) da decomposição inercial, por meio da solução da seguinte equação:

$$\det(\lambda^2 \cdot \mathbf{I} - \mathbf{W}) = \begin{vmatrix} \lambda^2 - w_{11} & -w_{12} & -w_{13} \\ -w_{21} & \lambda^2 - w_{22} & -w_{23} \\ -w_{31} & -w_{32} & \lambda^2 - w_{33} \end{vmatrix} = 0 \tag{11.18}$$

em que $\mathbf{I}$ é a matriz identidade.

Genericamente, para uma tabela inicial de contingência de dimensões ($I \times J$), os $m$ autovalores obtidos obedecem à seguinte lógica:

$$\lambda_0^2 = 1 \geq \lambda_1^2 \geq \ldots \geq \lambda_m^2 \geq 0, \text{ em que } m = \text{mín}(I-1, J-1).$$

Além disso, a inércia principal total, já definida por meio da expressão (11.7), pode ser também escrita com base nos autovalores obtidos, conforme segue:

$$I_T = \frac{\chi^2}{N} = \sum_{k=1}^{m=\text{mín}(I-1,J-1)} \lambda_k^2, k = 1,2,\ldots,m \tag{11.19}$$

Em outras palavras, a decomposição inercial em determinada tabela de contingência, representada pelas diferenças entre as frequências absolutas observadas e esperadas, pode ser decomposta em $m$ componentes, que se referem aos valores das inércias principais parciais de cada dimensão e que nada mais são que o quadrado dos **valores singulares** $\lambda_k$ de cada dimensão. Como a análise de correspondência tem, como um de seus principais objetivos, propiciar ao pesquisador a construção de mapas perceptuais que mostram a relação entre as categorias das variáveis dispostas em linha e em coluna na tabela de contingência, cada componente da inércia principal total será utilizado para que se identifique como determinada linha ou coluna contribui para a construção de cada eixo (**dimensão**) do referido mapa.

Dessa forma, precisamos definir como são calculadas as coordenadas (também chamadas de ***scores***) das categorias de cada variável no mapa perceptual, com base nos conceitos estudados até o presente momento.

### 11.2.4. Definição das coordenadas (*scores*) das categorias no mapa perceptual

Seguindo a mesma lógica proposta por Johnson e Wichern (2007), vamos chamar a matriz diagonal de autovalores da matriz $\mathbf{W} = \mathbf{A'A}$ de $\Lambda^2$, em que:

$$\Lambda^2 = \begin{pmatrix} \lambda_1^2 & 0 & \cdots & 0 \\ 0 & \lambda_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m^2 \end{pmatrix} \tag{11.20}$$

sendo que cada $\lambda_k^2$ se refere à inércia principal parcial da k-ésima dimensão, e $\lambda_k$, ao respectivo valor singular. Logo, definidos os autovalores da matriz $\mathbf{W}$, podemos chegar aos autovetores da mesma matriz, que chamaremos de:

$$\mathbf{V} = \begin{pmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_J \end{pmatrix}$$

e

$$\mathbf{U} = \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_I \end{pmatrix}$$

Johnson e Wichern (2007) provam ainda que a relação entre os autovetores se dá por meio das seguintes expressões:

$$\mathbf{v}_k' = \mathbf{u}_k' \cdot [\mathbf{D}_l^{-1/2} \cdot (\mathbf{P} - lc') \cdot \mathbf{D}_c^{-1/2}] \cdot \lambda_k^{-1} \tag{11.21}$$

e

$$\mathbf{u}_k = [\mathbf{D}_l^{-1/2} \cdot (\mathbf{P} - lc') \cdot \mathbf{D}_c^{-1/2}] \cdot \mathbf{v}_k \cdot \lambda_k^{-1} \tag{11.22}$$

Além disso, Johnson e Wichern (2007) ainda demonstram que:

$$\mathbf{v}_k' \cdot \mathbf{D}_c^{1/2} \cdot \mathbf{1}_J = 0 \tag{11.23}$$

e

$$\mathbf{u}_k' \cdot \mathbf{D}_l^{1/2} \cdot \mathbf{1}_I = 0 \tag{11.24}$$

em que $\mathbf{1}_I$ e $\mathbf{1}_J$ representam, respectivamente, vetores de dimensões $I \times 1$ e $J \times 1$ com valores iguais a 1, respeitadas as seguintes condições:

$$(\mathbf{D}_c^{1/2} \cdot \mathbf{v}_k)' \cdot \mathbf{D}_c^{-1} \cdot (\mathbf{D}_c^{1/2} \cdot \mathbf{v}_k) = \mathbf{v}_k' \cdot \mathbf{v}_k = 1 \tag{11.25}$$

e

$$(\mathbf{D}_l^{1/2} \cdot \mathbf{u}_k)' \cdot \mathbf{D}_l^{-1} \cdot (\mathbf{D}_l^{1/2} \cdot \mathbf{u}_k) = \mathbf{u}_k' \cdot \mathbf{u}_k = 1 \tag{11.26}$$

Definidos a matriz diagonal de autovalores $\Lambda^2$ e os autovetores $\mathbf{U}$ e $\mathbf{V}$, as coordenadas (abcissa e ordenada) de cada categoria das variáveis podem ser calculadas com base nas seguintes expressões:

- **Variável em linha na tabela de contingência:**
  - Coordenadas da primeira dimensão (abcissas):

$$\mathbf{X}_l = \begin{pmatrix} \mathbf{x}_{l1} \\ \vdots \\ \mathbf{x}_{lI} \end{pmatrix} = \mathbf{D}_l^{-1} \cdot (\mathbf{D}_l^{1/2} \cdot \mathbf{U}) \cdot \mathbf{\Lambda} = \sqrt{\lambda_1} \cdot \mathbf{D}_l^{-1/2} \cdot \mathbf{u}_1 \tag{11.27}$$

  - Coordenadas da segunda dimensão (ordenadas):

$$\mathbf{Y}_l = \begin{pmatrix} \mathbf{y}_{l1} \\ \vdots \\ \mathbf{y}_{lI} \end{pmatrix} = \mathbf{D}_l^{-1} \cdot (\mathbf{D}_l^{1/2} \cdot \mathbf{U}) \cdot \mathbf{\Lambda} = \sqrt{\lambda_2} \cdot \mathbf{D}_l^{-1/2} \cdot \mathbf{u}_2 \tag{11.28}$$

  - Coordenadas da k-ésima dimensão:

$$\mathbf{Z}_l = \begin{pmatrix} \mathbf{z}_{l1} \\ \vdots \\ \mathbf{z}_{lI} \end{pmatrix} = \mathbf{D}_l^{-1} \cdot (\mathbf{D}_l^{1/2} \cdot \mathbf{U}) \cdot \mathbf{\Lambda} = \sqrt{\lambda_k} \cdot \mathbf{D}_l^{-1/2} \cdot \mathbf{u}_k \tag{11.29}$$

- **Variável em coluna na tabela de contingência:**
  - Coordenadas da primeira dimensão (abcissas):

$$\mathbf{X}_c = \begin{pmatrix} \mathbf{x}_{c1} \\ \vdots \\ \mathbf{x}_{cJ} \end{pmatrix} = \mathbf{D}_c^{-1} \cdot (\mathbf{D}_c^{1/2} \cdot \mathbf{V}) \cdot \mathbf{\Lambda} = \sqrt{\lambda_1} \cdot \mathbf{D}_c^{-1/2} \cdot \mathbf{v}_1 \tag{11.30}$$

  - Coordenadas da segunda dimensão (ordenadas):

$$\mathbf{Y}_c = \begin{pmatrix} \mathbf{y}_{c1} \\ \vdots \\ \mathbf{y}_{cJ} \end{pmatrix} = \mathbf{D}_c^{-1} \cdot (\mathbf{D}_c^{1/2} \cdot \mathbf{V}) \cdot \mathbf{\Lambda} = \sqrt{\lambda_2} \cdot \mathbf{D}_c^{-1/2} \cdot \mathbf{v}_2 \tag{11.31}$$

  - Coordenadas da k-ésima dimensão:

$$\mathbf{Z}_c = \begin{pmatrix} \mathbf{z}_{c1} \\ \vdots \\ \mathbf{z}_{cJ} \end{pmatrix} = \mathbf{D}_c^{-1} \cdot (\mathbf{D}_c^{1/2} \cdot \mathbf{V}) \cdot \mathbf{\Lambda} = \sqrt{\lambda_k} \cdot \mathbf{D}_c^{-1/2} \cdot \mathbf{v}_k \tag{11.32}$$

É importante ressaltar que as coordenadas da variável em linha também podem ser obtidas por meio das coordenadas da variável em coluna e vice-versa. Assim, caso o pesquisador tenha apenas as coordenadas das categorias de uma das variáveis, porém possua as massas de cada uma das categorias da outra, além dos valores singulares, poderá calcular as coordenadas das categorias desta última variável. Conforme comentam Fávero *et al.* (2009), as coordenadas das categorias da variável em linha para uma específica dimensão podem ser obtidas multiplicando-se a matriz de massas (*row profiles*) pelo vetor de coordenadas das categorias da variável em coluna e dividindo-se os valores obtidos pelo valor singular daquela determinada dimensão. Analogamente, as coordenadas das categorias da variável em coluna, também para dada dimensão, podem ser obtidas multiplicando-se a matriz de massas (*column profiles*) pelo vetor de coordenadas das categorias da variável em linha e dividindo-se também os valores obtidos pelo valor singular daquela dimensão.

Assim, temos que:

$$\mathbf{X}_l = \begin{pmatrix} \mathbf{x}_{l1} \\ \vdots \\ \mathbf{x}_{lI} \end{pmatrix} = \begin{pmatrix} \left(\frac{n_{11}}{\sum l_1}\right) & \left(\frac{n_{12}}{\sum l_1}\right) & \cdots & \left(\frac{n_{1J}}{\sum l_1}\right) \\ \left(\frac{n_{21}}{\sum l_2}\right) & \left(\frac{n_{22}}{\sum l_2}\right) & \cdots & \left(\frac{n_{2J}}{\sum l_2}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{n_{I1}}{\sum l_I}\right) & \left(\frac{n_{I2}}{\sum l_I}\right) & \cdots & \left(\frac{n_{IJ}}{\sum l_I}\right) \end{pmatrix} \cdot \begin{pmatrix} \mathbf{x}_{c1} \\ \vdots \\ \mathbf{x}_{cJ} \end{pmatrix} \cdot \lambda_1^{-1} \tag{11.33}$$

e

$$\mathbf{X}_c = \begin{pmatrix} \mathbf{x}_{c1} \\ \vdots \\ \mathbf{x}_{cJ} \end{pmatrix} = \begin{pmatrix} \left(\frac{n_{11}}{\sum c_1}\right) & \left(\frac{n_{12}}{\sum c_2}\right) & \cdots & \left(\frac{n_{1J}}{\sum c_J}\right) \\ \left(\frac{n_{21}}{\sum c_1}\right) & \left(\frac{n_{22}}{\sum c_2}\right) & \cdots & \left(\frac{n_{2J}}{\sum c_J}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{n_{I1}}{\sum c_1}\right) & \left(\frac{n_{I2}}{\sum c_2}\right) & \cdots & \left(\frac{n_{IJ}}{\sum c_J}\right) \end{pmatrix}' \cdot \begin{pmatrix} \mathbf{x}_{l1} \\ \vdots \\ \mathbf{x}_{lI} \end{pmatrix} \cdot \lambda_1^{-1} \tag{11.34}$$

Com base nas expressões (11.33) e (11.34), podem ser definidas, de forma análoga, as expressões das coordenadas das demais dimensões, sempre levando-se em considerando os respectivos valores singulares.

Por fim, podemos verificar que as coordenadas (*scores*) se relacionam com os valores singulares obtidos pode meio das seguintes expressões:

$$\lambda_1 = \sum_{i=1}^{I} \left[ (\mathbf{x}_{li})^2 \cdot \left( \frac{\sum l_i}{N} \right) \right] = \sum_{j=1}^{J} \left[ (\mathbf{x}_{cj})^2 \cdot \left( \frac{\sum c_j}{N} \right) \right] \tag{11.35}$$

$$\lambda_2 = \sum_{i=1}^{I} \left[ (\mathbf{y}_{li})^2 \cdot \left( \frac{\sum l_i}{N} \right) \right] = \sum_{j=1}^{J} \left[ (\mathbf{y}_{cj})^2 \cdot \left( \frac{\sum c_j}{N} \right) \right] \tag{11.36}$$

$$\lambda_k = \sum_{i=1}^{I} \left[ (\mathbf{z}_{li})^2 \cdot \left( \frac{\sum l_i}{N} \right) \right] = \sum_{j=1}^{J} \left[ (\mathbf{z}_{cj})^2 \cdot \left( \frac{\sum c_j}{N} \right) \right] \tag{11.37}$$

As coordenadas **X** e **Y** obtidas por meio das expressões (11.27) a (11.32) são utilizadas para construir um mapa perceptual conhecido como **mapa simétrico**, em que os pontos que representam as linhas e colunas das categorias das variáveis possuem a mesma escala, também conhecida por **normalização simétrica**. Caso o pesquisador deseje, por outro lado, privilegiar exclusivamente a visualização das massas em linha ou das massas em coluna de determinada tabela de contingência para a construção do mapa perceptual, poderá abrir mão da normalização simétrica e optar, respectivamente, por aquelas conhecidas como **principal linha** e **principal coluna**. Nesses casos, o cálculo das coordenadas é elaborado por expressões apresentadas no Quadro 11.1.

**Quadro 11.1** Expressões para determinação das abcissas e ordenadas em mapas perceptuais.

| Normalização | Expressão para as Abcissas | Expressão para as Ordenadas |
|---|---|---|
| Simétrica | $\mathbf{X} = \mathbf{D}_l^{-1} \cdot (\mathbf{D}_l^{1/2} \cdot \mathbf{U}) \cdot \mathbf{\Lambda}$ | $\mathbf{Y} = \mathbf{D}_c^{-1} \cdot (\mathbf{D}_c^{1/2} \cdot \mathbf{V}) \cdot \mathbf{\Lambda}$ |
| Principal Linha | $\mathbf{X} = \mathbf{D}_l^{-1} \cdot (\mathbf{D}_l^{1/2} \cdot \mathbf{U}) \cdot \mathbf{\Lambda}$ | $\mathbf{Y} = \mathbf{D}_c^{-1} \cdot (\mathbf{D}_c^{1/2} \cdot \mathbf{V})$ |
| Principal Coluna | $\mathbf{X} = \mathbf{D}_l^{-1} \cdot (\mathbf{D}_l^{1/2} \cdot \mathbf{U})$ | $\mathbf{Y} = \mathbf{D}_c^{-1} \cdot (\mathbf{D}_c^{1/2} \cdot \mathbf{V}) \cdot \mathbf{\Lambda}$ |

Enquanto, no perfil **linha**, apenas o cálculo das abcissas leva em consideração a matriz de valores singulares, no perfil **coluna**, essa matriz é utilizada apenas para o cálculo das ordenadas.

Com base na determinação das coordenadas de cada categoria, pode ser construído um mapa perceptual com $m$ dimensões. Embora essa possibilidade seja matematicamente possível, apenas as duas primeiras dimensões ($m$ = 2) são geralmente utilizadas para a elaboração da análise gráfica, o que gera um mapa perceptual conhecido por ***biplot***.

Na próxima seção, utilizaremos os conceitos apresentados para a elaboração analítica de um exemplo prático.

## 11.2.5. Exemplo prático de análise de correspondência simples (Anacor)

Imagine que o mesmo professor tenha agora o interesse em estudar se o perfil de investidor de seus alunos relaciona-se com o tipo de aplicação financeira realizada, ou seja, se existe associação estatisticamente significante, a determinado nível de significância, entre os perfis dos investidores e a forma como são alocados seus recursos financeiros.

Nesse sentido, o professor elaborou uma pesquisa com 100 alunos da escola onde leciona, solicitando que cada um declarasse em que tipo de aplicação financeira possuía a maior parte de seus recursos. Três possibilidades surgiram como resposta: **Poupança**, **CDB** e **Ações**. Na sequência, com base na estratificação do fator principal gerado a partir de uma análise fatorial por componentes principais aplicada anteriormente a diversas variáveis, os mesmos estudantes foram classificados pelo professor em três tipos de perfil de investidor: **Conservador**, **Moderado** ou **Agressivo**. Parte do banco de dados elaborado, que possui apenas essas duas variáveis categóricas, encontra-se na Tabela 11.7.

**Tabela 11.7** Exemplo: Perfil do investidor e tipo de aplicação financeira.

| Estudante | Perfil do Investidor | Tipo de Aplicação Financeira |
|---|---|---|
| Gabriela | Conservador | Poupança |
| Luiz Felipe | Conservador | Poupança |
| | ⋮ | |
| Renata | Conservador | CDB |
| Guilherme | Conservador | Ações |
| | ⋮ | |
| Kamal | Moderado | Poupança |
| Rodolfo | Moderado | CDB |
| | ⋮ | |
| Raquel | Moderado | CDB |
| Anna Luiza | Moderado | Ações |
| | ⋮ | |
| Nuno | Agressivo | Poupança |
| Bráulio | Agressivo | CDB |
| | ⋮ | |
| Estela | Agressivo | Ações |

O banco de dados completo pode ser acessado por meio do arquivo **Perfil_Investidor × Aplicação.xls**. Por meio dele, é possível definir a tabela de contingência de nosso exemplo, que possui dimensão 3 × 3 e oferece as frequências absolutas observadas para cada par perfil do investidor × tipo de aplicação (Tabela 11.8).

**Tabela 11.8** Tabela de contingência com frequências absolutas observadas.

| Aplicação / Perfil | Poupança | CDB | Ações | **Total** |
|---|---|---|---|---|
| Conservador | 8 | 4 | 5 | $\Sigma l_1 = 17$ |
| Moderado | 5 | 16 | 4 | $\Sigma l_2 = 25$ |
| Agressivo | 2 | 20 | 36 | $\Sigma l_3 = 58$ |
| **Total** | $\Sigma c_1 = 15$ | $\Sigma c_2 = 40$ | $\Sigma c_3 = 45$ | $N = 100$ |

Na forma matricial, a tabela de contingência com frequências absolutas observadas pode ser escrita, com base na expressão (11.2), da seguinte forma:

$$\mathbf{X}_o = \begin{pmatrix} 8 & 4 & 5 \\ 5 & 16 & 4 \\ 2 & 20 & 36 \end{pmatrix}$$

Por meio da Tabela 11.8 (ou da matriz $\mathbf{X}_o$), podemos verificar que há mais investidores com o perfil *Agressivo* que *Moderado* ou *Conservador*. Em relação ao tipo de aplicação financeira, verificamos que há uma quantidade maior de investidores com recursos alocados em *Ações* e em *CDB* que em *Poupança*. Entretanto, essa análise preliminar é apenas univariada, ou seja, leva em consideração a distribuição de frequências para cada variável isoladamente, sem uma análise de classificação cruzada. Nosso objetivo, portanto, é estudar se as categorias do perfil do investidor associam-se de forma estatisticamente significante com as categorias do tipo de aplicação financeira em uma perspectiva bivariada.

Conforme discutimos na seção 11.2.2, precisamos, portanto, investigar inicialmente se as categorias das duas variáveis associam-se de forma aleatória ou se existe uma relação de dependência entre elas. A fim de que seja calculada a estatística $\chi^2$, devemos definir as frequências absolutas esperadas e os resíduos de cada uma das células da tabela de classificação cruzada. Enquanto a Tabela 11.9 apresenta as frequências absolutas esperadas, a Tabela 11.10 mostra os resíduos.

**Tabela 11.9** Frequências absolutas esperadas.

| Aplicação / Perfil | Poupança | CDB | Ações |
|---|---|---|---|
| Conservador | $\left(\frac{15\times17}{100}\right)=2{,}55$ | $\left(\frac{40\times17}{100}\right)=6{,}80$ | $\left(\frac{45\times17}{100}\right)=7{,}65$ |
| Moderado | $\left(\frac{15\times25}{100}\right)=3{,}75$ | $\left(\frac{40\times25}{100}\right)=10{,}00$ | $\left(\frac{45\times25}{100}\right)=11{,}25$ |
| Agressivo | $\left(\frac{15\times58}{100}\right)=8{,}70$ | $\left(\frac{40\times58}{100}\right)=23{,}20$ | $\left(\frac{45\times58}{100}\right)=26{,}10$ |

**Tabela 11.10** Resíduos – Diferenças entre frequências absolutas observadas e esperadas.

| Aplicação / Perfil | Poupança | CDB | Ações |
|---|---|---|---|
| Conservador | 5,45 | -2,80 | -2,65 |
| Moderado | 1,25 | 6,00 | -7,25 |
| Agressivo | -6,70 | -3,20 | 9,90 |

Analogamente, na forma matricial, temos, com base nas expressões (11.4) e (11.5), que:

$$\mathbf{X_e} = \begin{pmatrix} 2{,}55 & 6{,}80 & 7{,}65 \\ 3{,}75 & 10{,}00 & 11{,}25 \\ 8{,}70 & 23{,}20 & 26{,}10 \end{pmatrix}$$

e

$$\mathbf{E} = \begin{pmatrix} 5{,}45 & -2{,}80 & -2{,}65 \\ 1{,}25 & 6{,}00 & -7{,}25 \\ -6{,}70 & -3{,}20 & 9{,}90 \end{pmatrix}$$

Obviamente, podemos verificar que a somatória dos resíduos é igual a 0 para cada linha e para cada coluna da matriz **E**.

Com base na expressão (11.6), podemos elaborar a Tabela 11.11, cuja somatória dos valores de cada célula fornece o valor da estatística $\chi^2$.

**Tabela 11.11** Valores de $\chi^2$ por célula.

| Aplicação / Perfil | Poupança | CDB | Ações |
|---|---|---|---|
| Conservador | $\frac{(5{,}45)^2}{2{,}55} = 11{,}65$ | $\frac{(-2{,}80)^2}{6{,}80} = 1{,}15$ | $\frac{(-2{,}65)^2}{7{,}65} = 0{,}92$ |
| Moderado | $\frac{(1{,}25)^2}{3{,}75} = 0{,}42$ | $\frac{(6{,}00)^2}{10{,}00} = 3{,}60$ | $\frac{(-7{,}25)^2}{11{,}25} = 4{,}67$ |
| Agressivo | $\frac{(-6{,}70)^2}{8{,}70} = 5{,}16$ | $\frac{(-3{,}20)^2}{23{,}20} = 0{,}44$ | $\frac{(9{,}90)^2}{26{,}10} = 3{,}76$ |

Assim, temos que:

$$\chi^2_{4\,g.l.} = \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{\left[n_{ij} - \left(\frac{\sum c_j \cdot \sum l_i}{100}\right)\right]^2}{\left(\frac{\sum c_j \cdot \sum l_i}{100}\right)} = \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{(resíduos_{ij})^2}{(frequências\ esperadas_{ij})} = 31{,}76$$

Para 4 graus de liberdade, já que $(I - 1) \times (J - 1) = (3 - 1) \times (3 - 1) = 4$, temos, por meio da Tabela D do apêndice do livro, que $\chi^2_c = 9{,}488$ ($\chi^2_c$ crítico para 4 graus de liberdade e para o nível de significância de 5%). Dessa forma, como o $\chi^2$ calculado $\chi^2_{cal} = 31{,}76 > \chi^2_c = 9{,}488$, podemos rejeitar a hipótese nula de que as duas variáveis categóricas se associam de forma aleatória, ou seja, existe associação estatisticamente significante, ao nível de significância de 5%, entre o perfil do investidor e o tipo de aplicação financeira.

Softwares como o SPSS e o Stata não oferecem o $\chi^2_c$ para os graus de liberdade definidos e determinado nível de significância. Todavia, oferecem o nível de significância do $\chi^2_{cal}$ para esses graus de liberdade. Portanto, em vez de analisarmos se $\chi^2_{cal} > \chi^2_c$, devemos verificar se o nível de significância do $\chi^2_{cal}$ é menor que 0,05 (5%) a fim de darmos continuidade à análise de correspondência. Assim:

Se *valor-P* (ou *P-value* ou *Sig.* $\chi^2_{cal}$ ou *Prob.* $\chi^2_{cal}$) < 0,05, a associação entre as duas variáveis categóricas não se dá de forma aleatória.

O nível de significância do $\chi^2_{cal}$ pode ser obtido no Excel por meio do comando **Fórmulas → Inserir Função → DIST.QUI**, que abrirá uma caixa de diálogo conforme mostra a Figura 11.1.

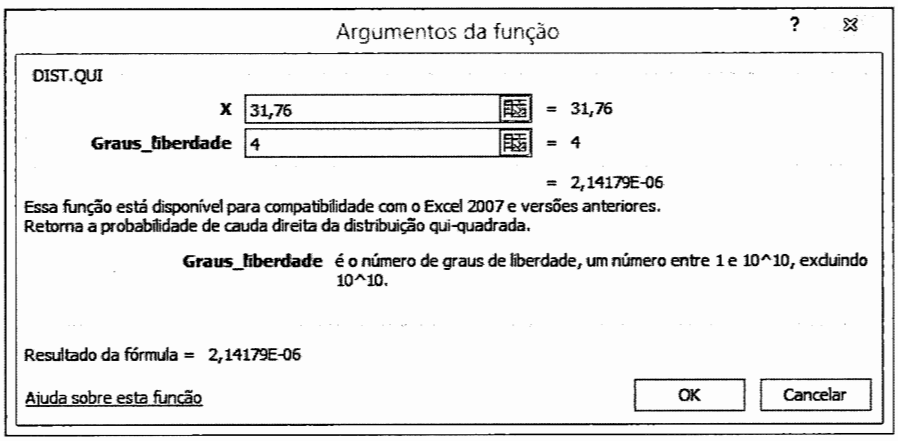



**Figura 11.1** Obtenção do nível de significância de $\chi^2$ (comando **Inserir Função**).

Conforme podemos observar por meio da Figura 11.1, o *valor-P* da estatística $\chi^2_{cal}$ é consideravelmente menor que 0,05 (*valor-P* $\chi^2_{cal} = 2{,}14 \times 10^{-6}$), ou seja, perfil do investidor e tipo de aplicação financeira não se combinam aleatoriamente.

Conforme discutimos na seção 11.2.2, embora o resultado do teste $\chi^2$ tenha mostrado a existência de um padrão de dependência entre o perfil do investidor e o tipo de aplicação financeira, é a análise dos resíduos padronizados ajustados que revelará os padrões característicos de cada categoria do perfil do investidor segundo o excesso ou a falta de ocorrências de sua combinação com cada categoria do tipo de aplicação financeira.

Logo, com base na expressão (11.8), podemos elaborar a Tabela 11.12, que apresenta o cálculo do resíduo padronizado em cada célula.

**Tabela 11.12** Resíduos padronizados.

| Perfil \ Aplicação | Poupança | CDB | Ações |
|---|---|---|---|
| Conservador | $\frac{8-2{,}6}{\sqrt{2{,}6}}=3{,}4$ | $\frac{4-6{,}8}{\sqrt{6{,}8}}=-1{,}1$ | $\frac{5-7{,}7}{\sqrt{7{,}7}}=-1{,}0$ |
| Moderado | $\frac{5-3{,}8}{\sqrt{3{,}8}}=0{,}6$ | $\frac{16-10}{\sqrt{10}}=1{,}9$ | $\frac{4-11{,}3}{\sqrt{11{,}3}}=-2{,}2$ |
| Agressivo | $\frac{2-8{,}7}{\sqrt{8{,}7}}=-2{,}3$ | $\frac{20-23{,}2}{\sqrt{23{,}2}}=-0{,}7$ | $\frac{36-26{,}1}{\sqrt{26{,}1}}=1{,}9$ |

Na forma matricial, a tabela de resíduos padronizados pode ser escrita, com base na expressão (11.9), da seguinte forma:

$$\mathbf{E}_{\text{padronizado}} = \begin{pmatrix} 3{,}4 & -1{,}1 & -1{,}0 \\ 0{,}6 & 1{,}9 & -2{,}2 \\ -2{,}3 & -0{,}7 & 1{,}9 \end{pmatrix}$$

Sendo assim, podemos elaborar a Tabela 11.13, que apresenta os resíduos padronizados ajustados. O valor de cada célula é calculado com base na expressão (11.10).

**Tabela 11.13** Resíduos padronizados ajustados.

| Perfil \ Aplicação | Poupança | CDB | Ações |
|---|---|---|---|
| Conservador | $\frac{3,4}{\sqrt{\left(1-\frac{15}{100}\right)\cdot\left(1-\frac{17}{100}\right)}}=4,1$ | $\frac{-1,1}{\sqrt{\left(1-\frac{40}{100}\right)\cdot\left(1-\frac{17}{100}\right)}}=-1,5$ | $\frac{-1,0}{\sqrt{\left(1-\frac{45}{100}\right)\cdot\left(1-\frac{17}{100}\right)}}=-1,4$ |
| Moderado | $\frac{0,6}{\sqrt{\left(1-\frac{15}{100}\right)\cdot\left(1-\frac{25}{100}\right)}}=0,8$ | $\frac{1,9}{\sqrt{\left(1-\frac{40}{100}\right)\cdot\left(1-\frac{25}{100}\right)}}=2,8$ | $\frac{-2,2}{\sqrt{\left(1-\frac{45}{100}\right)\cdot\left(1-\frac{25}{100}\right)}}=-3,4$ |
| Agressivo | $\frac{-2,3}{\sqrt{\left(1-\frac{15}{100}\right)\cdot\left(1-\frac{58}{100}\right)}}=-3,8$ | $\frac{-0,7}{\sqrt{\left(1-\frac{40}{100}\right)\cdot\left(1-\frac{58}{100}\right)}}=-1,3$ | $\frac{1,9}{\sqrt{\left(1-\frac{45}{100}\right)\cdot\left(1-\frac{58}{100}\right)}}=4,0$ |

A tabela de resíduos padronizados pode ser escrita matricialmente, com base na expressão (11.11), da seguinte forma:

$$\mathbf{E}_{\text{padronizado ajustado}}=\begin{pmatrix} 4,1 & -1,5 & -1,4 \\ 0,8 & 2,8 & -3,4 \\ -3,8 & -1,3 & 4,0 \end{pmatrix}$$

Note, na Tabela 11.13, que os resíduos padronizados ajustados com valores positivos superiores a 1,96 estão em destaque e correspondem ao excesso de ocorrências em cada célula, ao nível de significância de 5%, conforme discutimos ao final da seção 11.2.2. Podemos afirmar, portanto, que a análise dos resíduos padronizados ajustados permite caracterizar que o perfil *Conservador* se associa ao tipo de aplicação *Poupança*, o perfil *Moderado*, ao tipo de aplicação *CDB*, e o perfil *Agressivo*, ao tipo de aplicação *Ações*.

Visto que o perfil do investidor e o tipo de aplicação financeira não se associam de forma aleatória (teste $\chi^2$), e estudadas as relações entre cada par de categorias (resíduos padronizados ajustados), daremos sequência à análise de correspondência simples, com o objetivo de definir as coordenadas de cada uma das categorias para que, por meio delas, seja construído o mapa perceptual. Precisamos, dessa forma, calcular os autovalores (inércias principais parciais) e autovetores da matriz **W**, definida na seção 11.2.3 por meio da expressão (11.17). Conforme já discutimos, a partir dos quais, serão calculadas as coordenadas das categorias de ambas as variáveis.

Devemos inicialmente definir a matriz de frequências relativas observadas **P**, fazendo uso da expressão (11.12). Assim, temos que:

$$\mathbf{P}=\frac{1}{100}\cdot\mathbf{X}_o=\begin{pmatrix} 0,080 & 0,040 & 0,050 \\ 0,050 & 0,160 & 0,040 \\ 0,020 & 0,200 & 0,360 \end{pmatrix}$$

Por meio da matriz **P**, podemos elaborar as Tabelas 11.14 e 11.15, que apresentam as massas das categorias do perfil do investidor e do tipo de aplicação financeira, chamadas, respectivamente, de *column profiles* e *row profiles*.

**Tabela 11.14** Massas – *Column profiles*.

| Perfil \ Aplicação | Poupança | CDB | Ações | **Massa** |
|---|---|---|---|---|
| Conservador | 0,533 | 0,100 | 0,111 | $\frac{\sum l_1}{N}$ = **0,170** |
| Moderado | 0,333 | 0,400 | 0,089 | $\frac{\sum l_2}{N}$ = **0,250** |
| Agressivo | 0,133 | 0,500 | 0,800 | $\frac{\sum l_3}{N}$ = **0,580** |
| **Total** | **1,000** | **1,000** | **1,000** | |

**Tabela 11.15** Massas – *Row profiles*.

| Aplicação / Perfil | Poupança | CDB | Ações | **Total** |
|---|---|---|---|---|
| Conservador | 0,471 | 0,235 | 0,294 | **1,000** |
| Moderado | 0,200 | 0,640 | 0,160 | **1,000** |
| Agressivo | 0,034 | 0,345 | 0,621 | **1,000** |
| **Massa** | $\frac{\sum c_1}{N} = \mathbf{0,150}$ | $\frac{\sum c_2}{N} = \mathbf{0,400}$ | $\frac{\sum c_3}{N} = \mathbf{0,450}$ | |

As massas apresentadas nas Tabelas 11.14 e 11.15 influenciam diretamente o cálculo das coordenadas de cada uma das categorias das variáveis, uma vez que, por meio delas, é definida a matriz **W** e, consequentemente, seus autovalores e autovetores. É a partir das massas e da configuração de suas proporções em linha e em coluna, portanto, que o mapa perceptual da análise de correspondência começa a tomar forma. Vejamos de que maneira, tomando como exemplo a Tabela 11.15 (*row profiles*).

Inicialmente, vamos elaborar um gráfico que apresenta os percentuais em linha para cada categoria de perfil do investidor (Figura 11.2), do qual se pode analisar a alocação de recursos em cada uma das aplicações financeiras para dado perfil. Em outras palavras, essa visualização de frequências relativas permite elaborar uma comparação mais precisa de como são alocados os recursos financeiros para cada perfil de investidor.

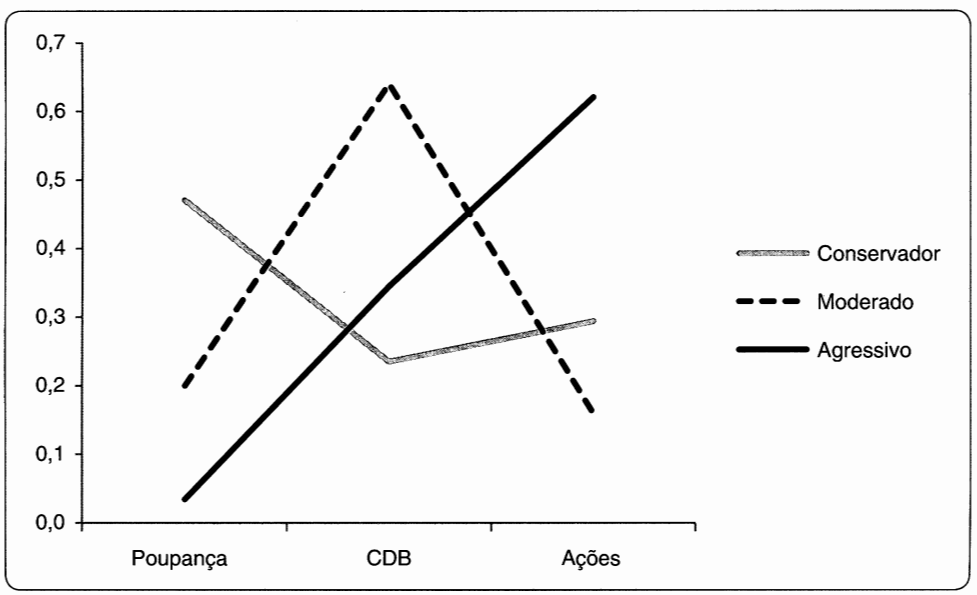



**Figura 11.2** Frequências relativas observadas de aplicação financeira por perfil do investidor (*row profiles*).

O gráfico da Figura 11.2 apresenta, em seu eixo horizontal, os tipos de aplicação financeira e, em seu eixo vertical, os percentuais de cada tipo de aplicação por perfil de investidor. Seguindo a lógica proposta por Greenacre (2008), vamos, na sequência, construir um gráfico tridimensional, em que cada eixo corresponde aos três tipos de aplicação financeira, conforme mostra a Figura 11.3. Dessa forma, plotamos nesse gráfico as coordenadas (0,471; 0,235; 0,294) para a categoria *Conservador*, (0,200; 0,640; 0,160), para a categoria *Moderado*, e (0,034; 0,345; 0,621), para a categoria *Agressivo*. Além disso, também plotamos as coordenadas (0,150; 0,400; 0,450) para a massa média do perfil do investidor.

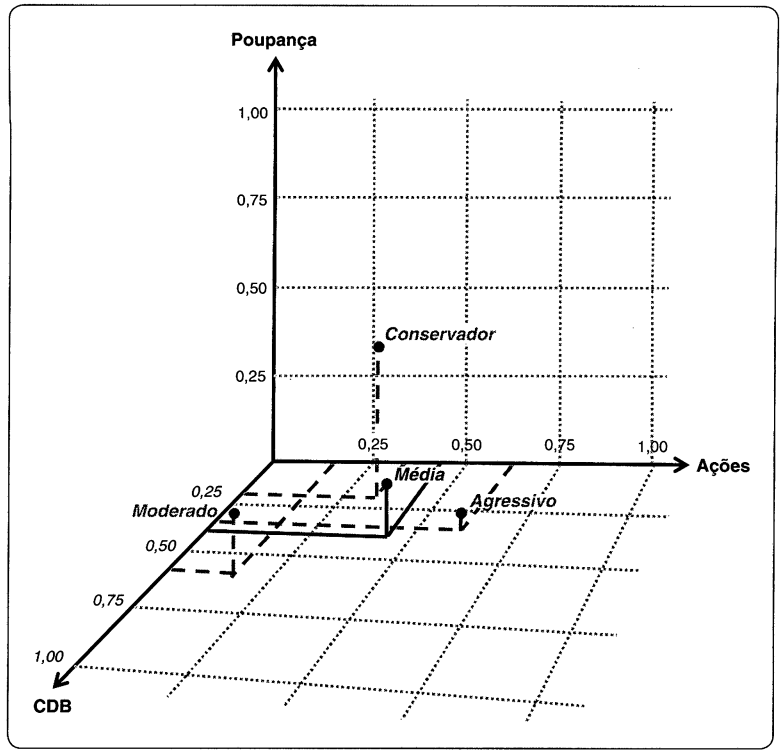



**Figura 11.3** Representação tridimensional das posições do perfil do investidor em relação aos tipos de aplicação financeira.

Ainda de acordo com Greenacre (2008), sobre a Figura 11.3 vamos construir um triângulo equilátero cujos vértices são as coordenadas (1; 0; 0), (0; 1, 0) e (0; 0; 1), ou seja, estão situados sobre cada um dos eixos e representam perfis concentrados somente em um tipo de aplicação financeira. Por exemplo, o vértice com coordenada (1; 0; 0) corresponde a um perfil de investidor que apresenta apenas aplicações financeiras em poupança. Já o vértice com coordenada (0; 0; 1) corresponde a outro perfil que possui apenas aplicações financeiras em ações. Essa nova representação gráfica, conhecida por **sistema triangular de coordenadas**, encontra-se na Figura 11.4.

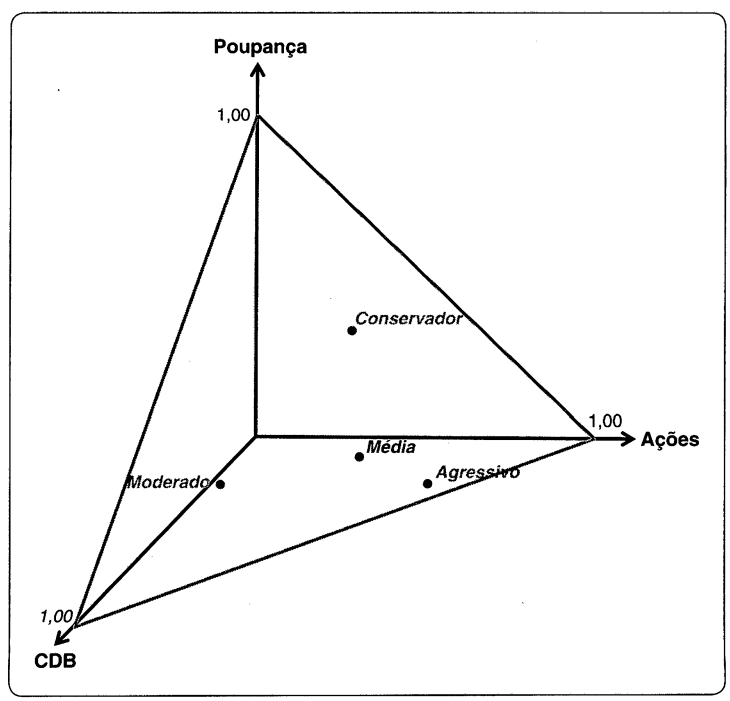



**Figura 11.4** Sistema triangular de coordenadas para o *row profile*.

O sistema triangular de coordenadas possibilita que projetemos os pontos referentes a cada uma das categorias do perfil do investidor sobre o triângulo equilátero, o que facilita a visualização de suas posições relativas. Isso gera o gráfico da Figura 11.5.

Por meio desse gráfico, temos condições de estudar a posição relativa de cada perfil de investidor em relação ao tipo de aplicação financeira. Assim, podemos verificar que, enquanto o perfil *Conservador* é o que mais se aproxima da aplicação *Poupança*, o *Moderado* é o que mais se aproxima da aplicação *CDB*. Por fim, o perfil *Agressivo* é o que mais se aproxima do vértice correspondente à aplicação *Ações*. O mais importante é que a posição relativa de cada ponto correspondente a cada perfil do investidor obedece à proporção de frequências relativas observadas (massas), apresentadas na Tabela 11.15 (*row profiles*).

Nesse sentido, tomemos, por exemplo, a categoria *Conservador*, cujas coordenadas são (0,471; 0,235; 0,294). Observe, por meio da Figura 11.6, que a posição relativa dessa categoria no sistema triangular de coordenadas obedece a essa proporção quando de sua projeção para cada um dos eixos respectivos às categorias *Poupança*, *CDB* e *Ações*, uma vez que linhas paralelas a esses eixos confluem para determinar a posição exata do ponto referente à categoria *Conservador*. Obviamente, a mesma lógica pode ser aplicada às categorias *Moderado* e *Agressivo*.

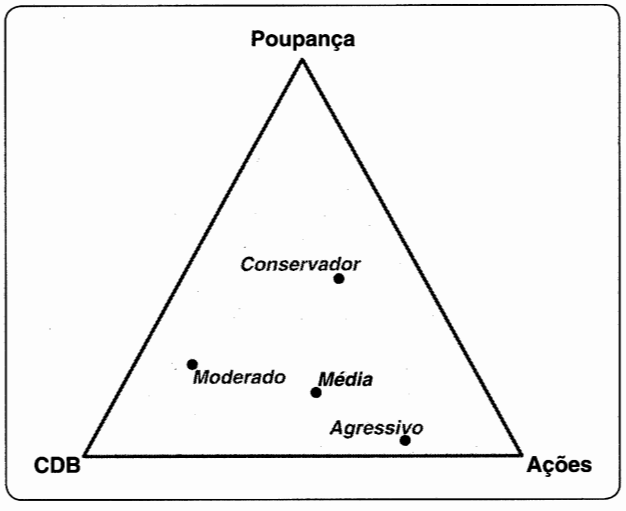



**Figura 11.5** Projeção das categorias do perfil do investidor no sistema triangular de coordenadas.

Segundo Greenacre (2008), na realidade, qualquer combinação de duas das três coordenadas dos perfis é suficiente para posicioná-los no sistema triangular de coordenadas, para uma variável com três categorias, sendo a terceira coordenada desnecessária, uma vez que a soma em linha das frequências relativas observadas será sempre igual a 1.

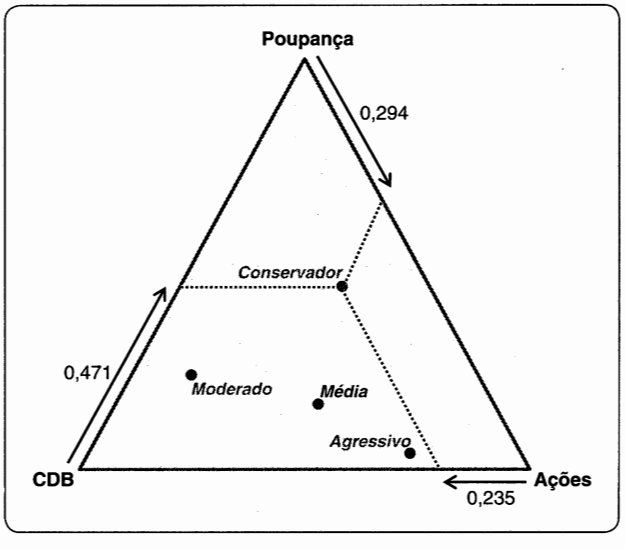



**Figura 11.6** Posição relativa da categoria *Conservador* no sistema triangular de coordenadas.

O sistema triangular de coordenadas somente pode ser utilizado para variáveis com três categorias. Como a dimensionalidade de um sistema de coordenadas é sempre igual ao número de categorias das variáveis menos 1, podemos comprovar, para nosso exemplo, que estamos lidando com um mapa, de fato, bidimensional (*biplot*).

Podemos, portanto, elaborar o gráfico do sistema triangular de coordenadas dando ênfase para o ponto com coordenadas (0,150; 0,400; 0,450), que corresponde à massa média do perfil do investidor. Esse gráfico encontra-se na Figura 11.7a.

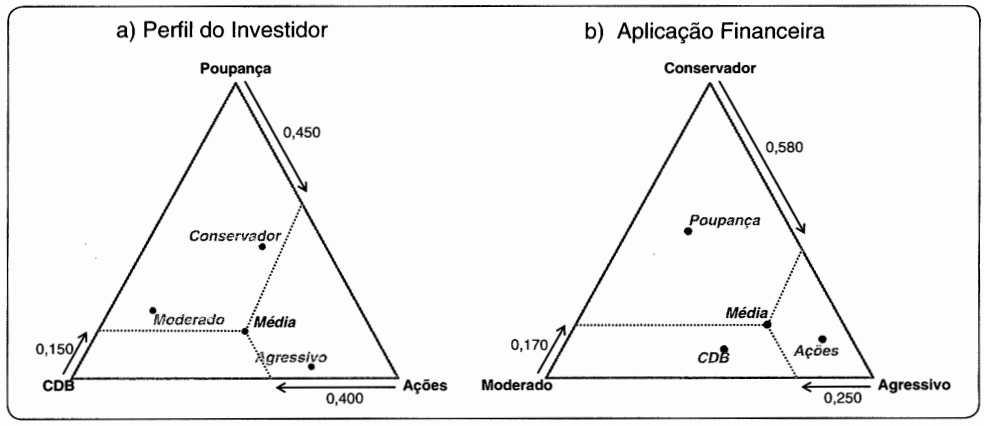



**Figura 11.7** Posições relativas das massas médias no sistema triangular de coordenadas.

Analogamente, podemos fazer uso das massas apresentadas na Tabela 11.14 (*column profiles*) para elaborar o gráfico da Figura 11.7b, em que cada vértice corresponde agora a cada uma das categorias do perfil de investidor, sendo plotadas as coordenadas (0,533; 0,333; 0,133) para a categoria *Poupança*, (0,100; 0,400; 0,500), para a categoria *CDB*, e (0,111; 0,089; 0,800), para a categoria *Ações*. No gráfico da Figura 11.7b, é dada ênfase para o ponto com coordenadas (0,170; 0,250; 0,580), que corresponde à massa média do tipo de aplicação financeira.

Dessa maneira, podemos verificar como as proporções das massas em linha e em coluna definem as posições relativas de cada categoria no mapa perceptual. Resta-nos, portanto, definir os eixos do mapa a fim de que o percentual da inércia principal parcial da primeira dimensão seja maximizado.

Para tanto, conforme discutimos ao final da seção 11.2.3, devemos definir uma matriz $\mathbf{W}$ e, a partir dela, calcular dois autovalores ($\lambda_1^2$ e $\lambda_2^2$) por meio do método Eckart-Young, correspondentes às duas inércias principais parciais das duas dimensões do mapa perceptual.

Nesse sentido, precisamos definir as duas matrizes diagonais, $\mathbf{D}_l$ e $\mathbf{D}_c$, que contêm, respectivamente, os valores das massas médias do tipo de aplicação financeira e do perfil do investidor em suas diagonais principais, em concordância com as expressões (11.13) e (11.14).

$$\mathbf{D}_l = \begin{pmatrix} 0{,}170 & 0 & 0 \\ 0 & 0{,}250 & 0 \\ 0 & 0 & 0{,}580 \end{pmatrix}$$

e

$$\mathbf{D}_c = \begin{pmatrix} 0{,}150 & 0 & 0 \\ 0 & 0{,}400 & 0 \\ 0 & 0 & 0{,}450 \end{pmatrix}$$

Note que, enquanto os valores da diagonal principal da matriz $\mathbf{D}_c$ são oriundos da Tabela 11.15 (*row profiles*), que também geraram o gráfico da Figura 11.7a, os valores da diagonal principal da matriz $\mathbf{D}_l$ são provenientes da Tabela 11.14 (*column profiles*), que também serviram de base para que fosse construído o gráfico da Figura 11.7b.

Ainda de acordo com o discutido na seção 11.2.3, a decomposição inercial para a elaboração da análise de correspondência consiste em calcular os autovalores de uma matriz $\mathbf{W} = \mathbf{A'A}$, em que $\mathbf{A}$ é definida de acordo com a expressão (11.15), reproduzida novamente a seguir:

$$\mathbf{A} = \mathbf{D}_l^{-1/2} \cdot (\mathbf{P} - lc') \cdot \mathbf{D}_c^{-1/2}$$

Precisamos, portanto, calcular os valores das células da matriz $\mathbf{P} - lc'$, com base na expressão (11.16). Logo, temos que:

$$\mathbf{P}-lc' = \begin{pmatrix} (0{,}080-0{,}170\times0{,}150) & (0{,}040-0{,}170\times0{,}400) & (0{,}050-0{,}170\times0{,}450) \\ (0{,}050-0{,}250\times0{,}150) & (0{,}160-0{,}250\times0{,}400) & (0{,}040-0{,}250\times0{,}450) \\ (0{,}020-0{,}580\times0{,}150) & (0{,}200-0{,}580\times0{,}400) & (0{,}360-0{,}580\times0{,}450) \end{pmatrix}$$

$$\mathbf{P}-lc' = \begin{pmatrix} 0{,}055 & -0{,}028 & -0{,}027 \\ 0{,}013 & 0{,}060 & -0{,}073 \\ -0{,}067 & -0{,}032 & 0{,}099 \end{pmatrix}$$

Note que as somatórias dos valores para cada linha e cada coluna da matriz $\mathbf{P} - lc'$ são, obviamente, sempre iguais a 0. Obtida a matriz, podemos chegar à matriz **A**:

$$\mathbf{A} = \begin{pmatrix} (0{,}170)^{-\frac{1}{2}} & 0 & 0 \\ 0 & (0{,}250)^{-\frac{1}{2}} & 0 \\ 0 & 0 & (0{,}580)^{-\frac{1}{2}} \end{pmatrix} \cdot \begin{pmatrix} 0{,}055 & -0{,}028 & -0{,}027 \\ 0{,}013 & 0{,}060 & -0{,}073 \\ -0{,}067 & -0{,}032 & 0{,}099 \end{pmatrix} \cdot \begin{pmatrix} (0{,}150)^{-\frac{1}{2}} & 0 & 0 \\ 0 & (0{,}400)^{-\frac{1}{2}} & 0 \\ 0 & 0 & (0{,}450)^{-\frac{1}{2}} \end{pmatrix}$$

$$\mathbf{A} = \begin{pmatrix} 0{,}341 & -0{,}107 & -0{,}096 \\ 0{,}065 & 0{,}190 & -0{,}216 \\ -0{,}227 & -0{,}066 & 0{,}194 \end{pmatrix}$$

Conforme mencionamos na seção 11.2.3, podemos realmente comprovar que os valores das células da matriz **A** são iguais aos das respectivas células da matriz $\mathbf{E}_{\text{padronizado}}$ divididos pela raiz quadrada do tamanho da amostra ($\sqrt{N} = 10$).

A matriz **W** pode ser obtida da seguinte maneira:

$$\mathbf{W} = \mathbf{A}'\mathbf{A} = \begin{pmatrix} 0{,}341 & 0{,}065 & -0{,}227 \\ -0{,}107 & 0{,}190 & -0{,}066 \\ -0{,}096 & -0{,}216 & 0{,}194 \end{pmatrix} \cdot \begin{pmatrix} 0{,}341 & -0{,}107 & -0{,}096 \\ 0{,}065 & 0{,}190 & -0{,}216 \\ -0{,}227 & -0{,}066 & 0{,}194 \end{pmatrix}$$

$$\mathbf{W} = \begin{pmatrix} 0{,}172 & -0{,}009 & -0{,}091 \\ -0{,}009 & 0{,}052 & -0{,}044 \\ -0{,}091 & -0{,}044 & 0{,}093 \end{pmatrix}$$

Os cálculos para obtenção das frequências absolutas esperadas (matriz $\mathbf{X}_e$), dos resíduos (matriz **E**), da estatística $\chi^2$, dos resíduos padronizados (matriz $\mathbf{E}_{\text{padronizado}}$), das massas e matrizes diagonais $\mathbf{D}_l$ e $\mathbf{D}_c$, da matriz **A** e da matriz **W** também podem ser verificados por meio do arquivo **Perfil_Investidor × Aplicação CálculoMatrizes.xls**.

Com base na expressão (11.18), podemos obter os autovalores da matriz **W**, de modo que:

$$\begin{vmatrix} \lambda^2 - 0{,}172 & 0{,}009 & 0{,}091 \\ 0{,}009 & \lambda^2 - 0{,}052 & 0{,}044 \\ 0{,}091 & 0{,}044 & \lambda^2 - 0{,}093 \end{vmatrix} = 0$$

de onde chegamos aos seguintes autovalores:

$$\begin{cases} \lambda_1^2 = 0{,}233 \\ \lambda_2^2 = 0{,}084 \end{cases}$$

valores das inércias principais parciais das duas dimensões que definem a matriz $\mathbf{\Lambda}^2$, de acordo com a expressão (11.20):

$$\mathbf{\Lambda}^2 = \begin{pmatrix} 0{,}233 & 0 \\ 0 & 0{,}084 \end{pmatrix}$$

Logo, a inércia principal total é $I_T = \lambda_1^2 + \lambda_2^2 = 0{,}318$. Por meio da expressão (11.7), também podemos verificar que:

$$I_T = \frac{\chi^2}{N} = \frac{31{,}76}{100} = 0{,}318$$

Os valores singulares de cada dimensão são, portanto, iguais a:

$$\begin{cases} \lambda_1 = 0{,}483 \\ \lambda_2 = 0{,}291 \end{cases}$$

A Tabela 11.16 apresenta a decomposição inercial para as duas dimensões.

**Tabela 11.16** Decomposição inercial para as duas dimensões.

| Dimensão | Valor Singular ($\lambda$) | Inércia Principal Parcial ($\lambda^2$) | Percentual da Inércia Principal Total |
|---|---|---|---|
| 1 | 0,483 | 0,233 | 73,42% |
| 2 | 0,291 | 0,084 | 26,58% |
| **Total** | | **0,318** | **100,00%** |

Por meio da análise da Tabela 11.16, podemos afirmar que as dimensões 1 e 2 explicam, respectivamente, 73,42% (0,233 / 0,318) e 26,58% (0,084 / 0,318) da inércia principal total. Na análise de correspondência, como os valores singulares da primeira dimensão são maximizados, serão sempre maiores que os da segunda dimensão, e assim sucessivamente, quando houver um número maior de dimensões. Portanto, o percentual da inércia principal total correspondente à primeira dimensão será sempre maior que o obtido para as dimensões subsequentes.

É importante mencionar que, quanto maior a inércia principal total, maior será a associação entre as categorias dispostas em linha e em coluna, o que afetará a disposição dos pontos no sistema triangular de coordenadas. De forma ilustrativa, imaginemos, para efeitos didáticos, três situações provenientes de três diferentes tabelas de contingência, conforme mostra a Figura 11.8.

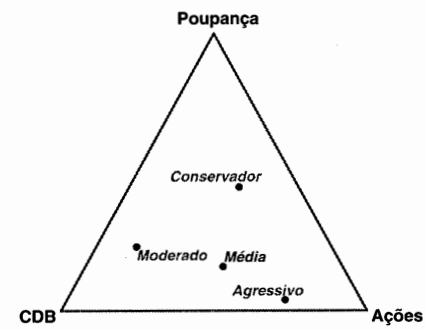
a) Dados do Nosso Exemplo

Tabela de Contingência:

| | Poupança | CDB | Ações |
|---|---|---|---|
| Conservador | 8 | 4 | 5 |
| Moderado | 5 | 16 | 4 |
| Agressivo | 2 | 20 | 36 |

$\chi^2 = 31{,}76$

Inércia Principal Total = 0,3176



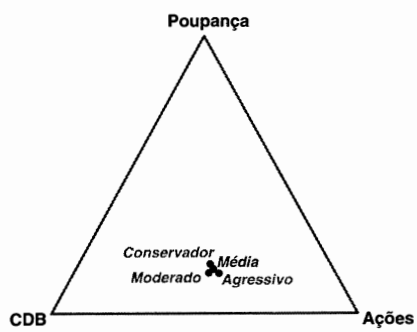

b) Inexistência de Associação

Tabela de Contingência:

| | Poupança | CDB | Ações |
|---|---|---|---|
| Conservador | 5 | 13 | 15 |
| Moderado | 5 | 14 | 15 |
| Agressivo | 5 | 13 | 15 |

$\chi^2 = 0{,}03$

Inércia Principal Total = 0,0003




c) Máxima Associação

Tabela de Contingência:

| | Poupança | CDB | Ações |
|---|---|---|---|
| Conservador | 15 | 0 | 0 |
| Moderado | 0 | 40 | 0 |
| Agressivo | 0 | 0 | 45 |

$\chi^2 = 200{,}00$

Inércia Principal Total = 2,000

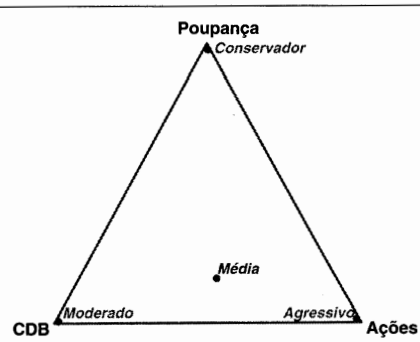



**Figura 11.8** Tabelas de contingência, inércias principais totais e o sistema triangular de coordenadas.

Por meio da Figura 11.8, podemos verificar que, quanto maior a inércia principal total, maior a associação entre as duas variáveis categóricas. Enquanto a Figura 11.8a mostra exatamente os dados do nosso exemplo, com foco em *row profiles* (exatamente igual à Figura 11.5), as Figuras 11.8b e 11.8c mostram situações opostas entre si, com inexistência de associação e associação máxima, respectivamente. Portanto, podemos afirmar que, quanto maior a inércia principal total (e, obviamente, o $\chi^2$), maior será a dispersão dos pontos no mapa perceptual e mais visível será a associação entre as variáveis cujas categorias são representadas por esses pontos. Note que a soma de cada coluna em cada uma das três situações não é alterada, o que faz as massas médias do perfil do investidor serem sempre iguais nas três situações.

Seguindo a lógica apresentada na seção 11.2.4, podemos, portanto, partir para o cálculo das coordenadas (*scores*) das categorias das duas variáveis em análise para os dados do nosso exemplo. Dessa forma, para calcularmos os autovetores da matriz $\mathbf{W}$ com base nos autovalores $\lambda_1^2$ e $\lambda_2^2$, devemos resolver o sistema de equações para cada uma das dimensões. Sendo assim, temos que:

- Primeira Dimensão ($\lambda_1^2 = 0{,}233$):

$$\begin{cases} 0{,}061 \cdot v_1 + 0{,}009 \cdot v_2 + 0{,}091 \cdot v_3 = 0 \\ 0{,}009 \cdot v_1 + 0{,}181 \cdot v_2 + 0{,}044 \cdot v_3 = 0 \\ 0{,}091 \cdot v_1 + 0{,}044 \cdot v_2 + 0{,}140 \cdot v_3 = 0 \end{cases}$$

De onde vem que:

$$\mathbf{v_1} = \begin{pmatrix} 0{,}822 \\ 0{,}093 \\ -0{,}562 \end{pmatrix}$$

Logo, por meio da expressão (11.22), podemos escrever que:

$$\mathbf{u_1} = \begin{pmatrix} \left\{ \dfrac{[0{,}341 \times 0{,}822] + [(-0{,}107) \times 0{,}093] + [(-0{,}096) \times (-0{,}562)]}{0{,}483} \right\} \\ \left\{ \dfrac{[0{,}065 \times 0{,}822] + [0{,}190 \times 0{,}093] + [(-0{,}216) \times (-0{,}562)]}{0{,}483} \right\} \\ \left\{ \dfrac{[(-0{,}227) \times 0{,}822] + [(-0{,}066) \times 0{,}093] + [0{,}194 \times (-0{,}562)]}{0{,}483} \right\} \end{pmatrix}$$

$$\mathbf{u_1} = \begin{pmatrix} 0{,}672 \\ 0{,}398 \\ -0{,}625 \end{pmatrix}$$

- Segunda Dimensão ($\lambda_2^2 = 0{,}084$):

$$\begin{cases} -0{,}088 \cdot v_1 + 0{,}009 \cdot v_2 + 0{,}091 \cdot v_3 = 0 \\ 0{,}009 \cdot v_1 + 0{,}032 \cdot v_2 + 0{,}044 \cdot v_3 = 0 \\ 0{,}091 \cdot v_1 + 0{,}044 \cdot v_2 - 0{,}009 \cdot v_3 = 0 \end{cases}$$

De onde vem que:

$$\mathbf{v_2} = \begin{pmatrix} 0{,}418 \\ -0{,}769 \\ 0{,}484 \end{pmatrix}$$

Analogamente, temos que:

$$\mathbf{u}_2 = \begin{pmatrix} \left\{\dfrac{[0{,}341\times 0{,}418]+[(-0{,}107)\times(-0{,}769)]+[(-0{,}096)\times 0{,}484]}{0{,}291}\right\} \\ \left\{\dfrac{[0{,}065\times 0{,}418]+[0{,}190\times(-0{,}769)]+[(-0{,}216)\times 0{,}484]}{0{,}291}\right\} \\ \left\{\dfrac{[(-0{,}227)\times 0{,}418]+[(-0{,}066)\times(-0{,}769]+[0{,}194\times 0{,}484]}{0{,}291}\right\} \end{pmatrix}$$

$$\mathbf{u}_2 = \begin{pmatrix} 0{,}616 \\ -0{,}769 \\ 0{,}172 \end{pmatrix}$$

Não serão aqui apresentados os cálculos, porém pode-se facilmente verificar, com base nos autovetores calculados, que as expressões (11.21) a (11.26) são satisfeitas.

Definidos a matriz diagonal de autovalores $\mathbf{\Lambda^2}$ e os autovetores $\mathbf{U}$ e $\mathbf{V}$, as coordenadas das abcissas e das ordenadas de cada uma das categorias da variável em linha e da variável em coluna na tabela de contingência podem ser calculadas por meio das expressões (11.27), (11.28), (11.30) e (11.31), de acordo como segue:

- **Variável em linha na tabela de contingência (perfil do investidor):**
  - Coordenadas das abcissas:

$$\mathbf{X}_l = \sqrt{0{,}483}\cdot\begin{pmatrix} (0{,}170)^{-\frac{1}{2}} & 0 & 0 \\ 0 & (0{,}250)^{-\frac{1}{2}} & 0 \\ 0 & 0 & (0{,}580)^{-\frac{1}{2}} \end{pmatrix}\cdot\begin{pmatrix} 0{,}672 \\ 0{,}398 \\ -0{,}625 \end{pmatrix}$$

$$\mathbf{X}_l = \begin{pmatrix} 1{,}132 \\ 0{,}553 \\ -0{,}570 \end{pmatrix}$$

que são as coordenadas, no mapa perceptual, das abcissas das categorias *Conservador*, *Moderado* e *Agressivo* do perfil do investidor.

  - Coordenadas das ordenadas:

$$\mathbf{Y}_l = \sqrt{0{,}291}\cdot\begin{pmatrix} (0{,}170)^{-\frac{1}{2}} & 0 & 0 \\ 0 & (0{,}250)^{-\frac{1}{2}} & 0 \\ 0 & 0 & (0{,}580)^{-\frac{1}{2}} \end{pmatrix}\cdot\begin{pmatrix} 0{,}616 \\ -0{,}769 \\ 0{,}172 \end{pmatrix}$$

$$\mathbf{Y}_l = \begin{pmatrix} 0{,}805 \\ -0{,}829 \\ 0{,}122 \end{pmatrix}$$

que são as coordenadas, no mapa perceptual, das ordenadas das categorias *Conservador*, *Moderado* e *Agressivo* do perfil do investidor.

- **Variável em coluna na tabela de contingência (tipo de aplicação financeira):**
  - Coordenadas das abcissas:

$$\mathbf{X}_c = \sqrt{0{,}483}\cdot\begin{pmatrix} (0{,}150)^{-\frac{1}{2}} & 0 & 0 \\ 0 & (0{,}400)^{-\frac{1}{2}} & 0 \\ 0 & 0 & (0{,}450)^{-\frac{1}{2}} \end{pmatrix}\cdot\begin{pmatrix} 0{,}822 \\ 0{,}093 \\ -0{,}562 \end{pmatrix}$$

$$\mathbf{X}_c = \begin{pmatrix} 1{,}475 \\ 0{,}102 \\ -0{,}582 \end{pmatrix}$$

que são as coordenadas, no mapa perceptual, das abcissas das categorias *Poupança*, *CDB* e *Ações* do tipo de aplicação financeira.

- Coordenadas das ordenadas:

$$\mathbf{Y}_c = \sqrt{0{,}291} \cdot \begin{pmatrix} (0{,}150)^{-\frac{1}{2}} & 0 & 0 \\ 0 & (0{,}400)^{-\frac{1}{2}} & 0 \\ 0 & 0 & (0{,}450)^{-\frac{1}{2}} \end{pmatrix} \cdot \begin{pmatrix} 0{,}418 \\ -0{,}769 \\ 0{,}484 \end{pmatrix}$$

$$\mathbf{Y}_c = \begin{pmatrix} 0{,}582 \\ -0{,}655 \\ 0{,}389 \end{pmatrix}$$

que são as coordenadas, no mapa perceptual, das ordenadas das categorias *Poupança*, *CDB* e *Ações* do tipo de aplicação financeira.

A Tabela 11.17, a seguir, apresenta as coordenadas das categorias das duas variáveis de forma consolidada.

**Tabela 11.17** Coordenadas (*scores*) das categorias das variáveis.

| Variável | Categoria | Coordenadas da 1ª Dimensão (Abcissas) | Coordenadas da 2ª Dimensão (Ordenadas) |
|---|---|---|---|
| Perfil do Investidor | Conservador | $\mathbf{x}_{l1} = 1{,}132$ | $\mathbf{y}_{l1} = 0{,}805$ |
| | Moderado | $\mathbf{x}_{l2} = 0{,}553$ | $\mathbf{y}_{l2} = -0{,}829$ |
| | Agressivo | $\mathbf{x}_{l3} = -0{,}570$ | $\mathbf{y}_{l3} = 0{,}122$ |
| Tipo de Aplicação Financeira | Poupança | $\mathbf{x}_{c1} = 1{,}475$ | $\mathbf{y}_{c1} = 0{,}582$ |
| | CDB | $\mathbf{x}_{c2} = 0{,}102$ | $\mathbf{y}_{c2} = -0{,}655$ |
| | Ações | $\mathbf{x}_{c3} = -0{,}582$ | $\mathbf{y}_{c3} = 0{,}389$ |

Conforme discutimos na seção 11.2.4 quando da apresentação das expressões (11.33) e (11.34), as coordenadas das categorias da variável em linha podem ser calculadas a partir das coordenadas das categorias da variável em coluna para determinada dimensão e vice-versa. Para tanto, devemos multiplicar a matriz de massas pelo vetor de coordenadas de uma variável e dividir pelo correspondente valor singular da dimensão em análise, para que sejam obtidas as coordenadas das categorias da outra variável. Vejamos dois exemplos, fazendo uso das expressões (11.33) e (11.34):

$$\mathbf{x}_{l1} = \frac{[0{,}471 \times 1{,}475] + [0{,}235 \times 0{,}102] + [0{,}294 \times (-0{,}582)]}{0{,}483} = 1{,}132$$

$$\mathbf{y}_{c2} = \frac{[0{,}100 \times 0{,}805] + [0{,}400 \times (-0{,}829)] + [0{,}500 \times 0{,}122]}{0{,}291} = -0{,}655$$

Finalmente, com base nas expressões (11.35) e (11.36), temos condições, por meio das coordenadas e das massas em linha e em coluna apresentadas nas Tabelas 11.14 e 11.15, de calcular, apenas para efeitos de verificação, os valores singulares obtidos anteriormente. Sendo assim, temos que:

$$\lambda_1 = [(1{,}132)^2 \times 0{,}170] + [(0{,}553)^2 \times 0{,}250] + [(-0{,}570)^2 \times 0{,}580] = 0{,}483$$

$$\lambda_1 = [(1{,}475)^2 \times 0{,}150] + [(0{,}102)^2 \times 0{,}400] + [(-0{,}582)^2 \times 0{,}450] = 0{,}483$$

e

$$\lambda_2 = [(0{,}805)^2 \times 0{,}170] + [(-0{,}829)^2 \times 0{,}250] + [(0{,}122)^2 \times 0{,}580] = 0{,}291$$

$$\lambda_2 = [(0{,}582)^2 \times 0{,}150] + [(-0{,}655)^2 \times 0{,}400] + [(0{,}389)^2 \times 0{,}450] = 0{,}291$$

Logo, com base nas coordenadas calculadas (*scores*), temos, enfim, condições de construir o mapa perceptual, a principal contribuição da análise de correspondência. A Figura 11.9 apresenta o mapa construído por meio das coordenadas consolidadas na Tabela 11.17.

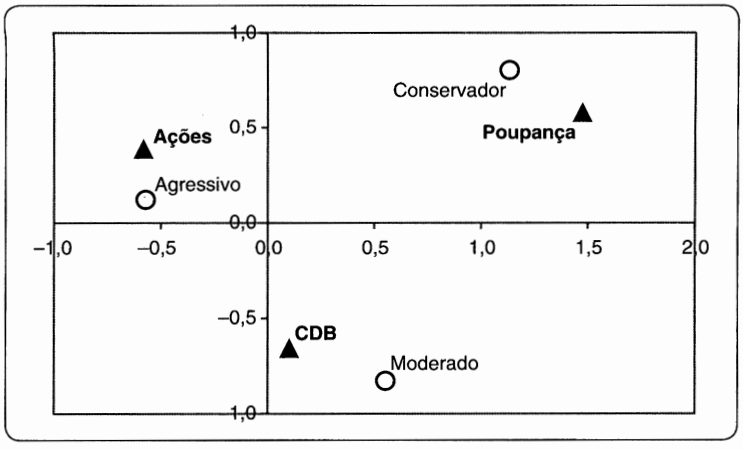



**Figura 11.9** Mapa perceptual para perfil do investidor e tipo de aplicação financeira.

Com base no mapa perceptual da Figura 11.9, podemos verificar que o perfil *Conservador* apresenta mais forte associação com o tipo de aplicação financeira *Poupança*. Além disso, enquanto o perfil *Moderado* associa-se, com maior frequência, à aplicação do tipo *CDB*, o perfil *Agressivo* associa-se mais fortemente com o tipo de investimento *Ações*.

A Figura 11.70, no apêndice deste capítulo, apresenta as configurações mais comuns que um mapa perceptual de uma análise de correspondência simples pode assumir, em função das características da tabela de contingência.

Voltando à análise do mapa perceptual da Figura 11.9, os achados estão, obviamente, de acordo com o discutido quando da análise dos resíduos padronizados ajustados, reproduzidos novamente a seguir, na Tabela 11.18.

**Tabela 11.18** Resíduos padronizados ajustados.

| Perfil \ Aplicação | Poupança | CDB | Ações |
|---|---|---|---|
| Conservador | 4,1 | −1,5 | −1,4 |
| Moderado | 0,8 | 2,8 | −3,4 |
| Agressivo | −3,8 | −1,3 | 4,0 |

Seguindo a lógica apresentada por Batista, Escuder e Pereira (2004), para auxiliar a interpretação do mapa perceptual, vamos desenhar uma linha de projeção para a caracterização do tipo de aplicação financeira *Poupança* (da Origem do mapa perceptual em direção à *Poupança*), nela se projetando as categorias do perfil do investidor *Conservador*, *Moderado* e *Agressivo*, conforme mostra a Figura 11.10. As projeções das categorias do perfil do investidor sobre a linha Origem-Poupança correspondem aos resíduos padronizados ajustados, ou seja, 4,1 (*Conservador*), 0,8 (*Moderado*) e −3,8 (*Agressivo*). As diferenças de escala entre essas projeções sobre a linha Origem-Poupança e os valores dos resíduos padronizados são devidas à distorção da projeção de um espaço tridimensional original para o espaço bidimensional utilizado para que fosse construído o mapa perceptual.

Pode-se repetir o mesmo exercício imaginando linhas de projeção para quaisquer categorias do perfil do investidor ou do tipo de aplicação financeira. No mapa perceptual da Figura 11.11, são projetadas, por sua vez, as categorias do tipo de aplicação financeira sobre a linha Origem-Agressivo, em que as projeções correspondem aos resíduos padronizados ajustados −3,8 (*Poupança*), −1,3 (*CDB*) e 4,0 (*Ações*). Da mesma forma, as diferenças de escala entre essas projeções sobre a linha Origem-Agressivo e os valores dos resíduos padronizados devem-se à distorção da projeção do espaço tridimensional original para o espaço bidimensional.

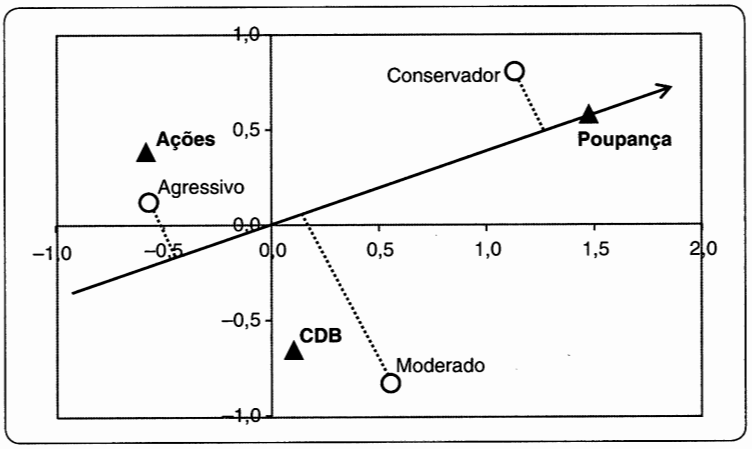



**Figura 11.10** Mapa perceptual para perfil do investidor e tipo de aplicação financeira, com foco na categoria *Poupança*.

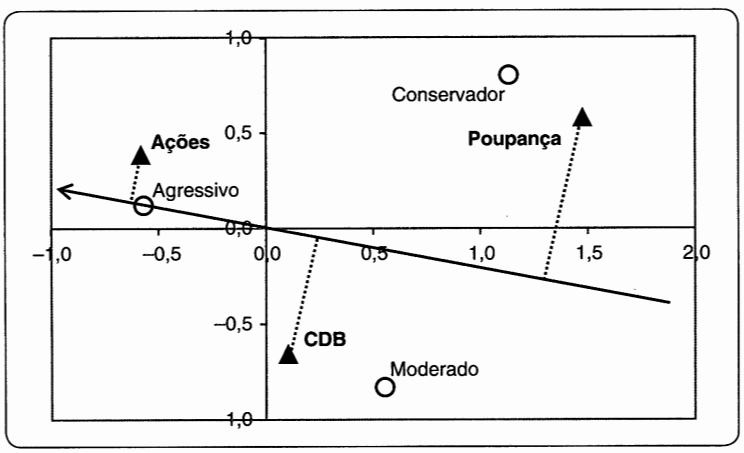



**Figura 11.11** Mapa perceptual para perfil do investidor e tipo de aplicação financeira, com foco na categoria *Agressivo*.

Podemos, portanto, concluir que há diferenças entre as formas de aplicação financeira de pessoas com diferentes perfis de investimento e que essas diferenças podem, de fato, ser identificadas e caracterizadas.

Enquanto na seção 11.4.1 serão apresentados os procedimentos para elaboração da análise de correspondência simples no SPSS, assim como seus resultados, na seção 11.5.1 serão apresentados os comandos para elaboração da técnica no Stata, com respectivos *outputs*.

Elaborado o teste $\chi^2$, avaliadas as associações entre as categorias das duas variáveis e construído o mapa perceptual, vamos partir para o estudo das relações entre categorias de mais de duas variáveis, por meio da análise de correspondência múltipla.

## 11.3. ANÁLISE DE CORRESPONDÊNCIA MÚLTIPLA

A análise de correspondência múltipla, também conhecida como **ACM**, é uma técnica de análise multivariada que representa uma extensão natural da análise de correspondência simples (Anacor), uma vez que permite que sejam estudadas as associações entre mais de duas variáveis categóricas e entre suas categorias, bem como a intensidade dessas associações.

Ao contrário da Anacor, técnica de análise bivariada, não é possível verificar a existência de associações entre mais de duas variáveis simultaneamente para a elaboração da análise de correspondência múltipla, visto que a estatística do teste $\chi^2$ é calculada apenas com base em uma tabela de contingência bidimensional. Isso não impede, por outro lado, que, em função das massas das categorias de cada uma das variáveis a serem inseridas na análise de correspondência múltipla, sejam calculados autovalores utilizados para que se definam as coordenadas daquelas categorias em um mapa perceptual. Portanto, a lógica da análise de correspondência múltipla é semelhante à estudada para a análise de correspondência simples. Ressalta-se que só devem ser inseridas na análise de

correspondência múltipla, entretanto, as variáveis que apresentarem associação, verificada por meio do teste $\chi^2$, com pelo menos uma das demais variáveis. Nesse sentido, **é recomendável que seja elaborado um teste $\chi^2$ para cada par de variáveis antes da elaboração de uma análise de correspondência múltipla**. Se uma delas não apresentar associação estatisticamente significante a nenhuma das demais variáveis, a determinado nível de significância, recomenda-se que seja excluída da análise de correspondência múltipla.

Enquanto na seção 11.3.1 serão apresentados os principais conceitos pertinentes à técnica, na seção 11.3.2 será elaborado um exemplo prático resolvido por meio de solução algébrica.

## 11.3.1. Notação

Para que seja elaborada a análise de correspondência múltipla, é necessário apresentar o conceito de **matriz binária**. Imaginemos um banco de dados com $N$ observações e $Q$ variáveis ($Q > 2$), e que cada variável $q$ ($q = 1, ..., Q$) possua $J_q$ categorias. Logo, o número total de categorias envolvidas em uma análise de correspondência múltipla é:

$$J = \sum_{q=1}^{Q} J_q \tag{11.38}$$

A Tabela 11.19 apresenta, de forma esquemática, um banco de dados com $N$ observações e $Q$ ($Q > 2$) variáveis categóricas.

**Tabela 11.19** Banco de dados com $N$ observações e $Q$ ($Q > 2$) variáveis categóricas.

| Observação | Variável $q$ | | | |
|---|---|---|---|---|
| | 1 | 2 | ... | $Q$ |
| 1 | categoria 1 | categoria 4 | | categoria 2 |
| 2 | categoria 2 | categoria 1 | | categoria 1 |
| 3 | categoria 1 | categoria 3 | ... | categoria 1 |
| 4 | categoria 3 | categoria 2 | | categoria 2 |
| ⋮ | ⋮ | ⋮ | | ⋮ |
| $N$ | categoria 2 | categoria 4 | | categoria 2 |
| **Número de categorias $J_q$** | **3** | **4** | **...** | **2** |

Note, com base no banco de dados apresentado na Tabela 11.19, que, por exemplo, $J_1 = 3, J_2 = 4$ e $J_Q = 2$. Por meio desse banco de dados, é possível construir um novo banco de dados apenas com variáveis binárias, criadas com base na codificação das categorias das variáveis para cada observação. Assim, por exemplo, para a observação 1, com respostas para as categorias das variáveis 1, 2, ..., $Q$ sendo, respectivamente, 1, 4, ..., 2, teremos a codificação binária representada, respectivamente, por (1 0 0), (0 0 0 1), ..., (0 1). A Tabela 11.20 apresenta a codificação binária para as observações apresentadas na Tabela 11.19.

**Tabela 11.20** Codificação binária das categorias das variáveis originais.

| Observação | Variável 1 | | | Variável 2 | | | | ... | Variável Q | |
|---|---|---|---|---|---|---|---|---|---|---|
| | cat. 1 | cat. 2 | cat. 3 | cat. 1 | cat. 2 | cat. 3 | cat. 4 | | cat. 1 | cat. 2 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | | 0 | 1 |
| 2 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | | 1 | 0 |
| 3 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | | 1 | 0 |
| 4 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | ... | 0 | 1 |
| ⋮ | ⋮ | | | ⋮ | | | | | ⋮ | |
| $N$ | 0 | 1 | 0 | 0 | 0 | 0 | 1 | | 0 | 1 |

A Tabela 11.20 com a codificação binária das categorias das variáveis originais é chamada de **matriz binária Z**, por meio da qual pode ser definida a inércia principal total da análise de correspondência múltipla, cujo cálculo é bastante simples e depende apenas da quantidade total de variáveis inseridas na análise e do número de categorias de cada uma delas, não dependendo das frequências absolutas das categorias. Conforme discute Greenacre (2008), a matriz binária **Z** é composta por matrizes $\mathbf{Z}_q$ agrupadas lateralmente, uma para cada variável $q$. Como cada matriz $\mathbf{Z}_q$ apresenta somente um valor 1 em cada linha, todos os perfis linha se situam nos vértices de um sistema de coordenadas, e, portanto, estamos diante de um exemplo de matriz em que ocorrem as maiores associações possíveis entre linhas e colunas, conforme discutimos na seção 11.2.5. Como consequência, para cada matriz $\mathbf{Z}_q$, a inércia principal parcial da dimensão principal será sempre igual a 1, e a inércia principal total, igual a $J_q - 1$. Dessa forma, a inércia principal total de **Z** corresponde à média das inércias principais totais das matrizes $\mathbf{Z}_q$ que a compõem, ou seja, pode ser obtida por meio da seguinte expressão:

$$I_T = \frac{\sum_{q=1}^{Q} (J_q - 1)}{Q} = \frac{J - Q}{Q} \tag{11.39}$$

Por meio do método da codificação binária, **pode-se supor que a matriz Z seja uma tabela de contingência de uma análise de correspondência simples**, a partir da qual podem ser definidos os valores das inércias principais parciais de cada uma das $J - Q$ dimensões. Consequentemente, conforme estudamos na seção 11.2, por meio dos autovalores e autovetores calculados a partir da matriz binária **Z** (considerada uma tabela de contingência de uma Anacor), podem ser definidas as coordenadas de cada uma das categorias das variáveis inseridas na análise de correspondência múltipla, o que permite que seja construído o mapa perceptual. **As coordenadas geradas por meio do método da matriz binária são conhecidas como coordenadas-padrão**.

Ainda segundo Greenacre (2008), a análise de correspondência múltipla pode também ser elaborada por meio de método alternativo, combinadas, em uma única matriz, as tabelas de contingência com os cruzamentos de todos os pares de variáveis. Essa matriz resultante, quadrada e simétrica, é conhecida por **matriz de Burt**.

Considerando a matriz binária $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_Q]$, a matriz de Burt pode ser definida, portanto, de acordo como segue:

$$\mathbf{B} = \mathbf{Z}' \cdot \mathbf{Z} \tag{11.40}$$

ou seja:

$$\mathbf{B} = \begin{pmatrix} \mathbf{Z}'_1 \cdot \mathbf{Z}_1 & \mathbf{Z}'_1 \cdot \mathbf{Z}_2 & \cdots & \mathbf{Z}'_1 \cdot \mathbf{Z}_Q \\ \mathbf{Z}'_2 \cdot \mathbf{Z}_1 & \mathbf{Z}'_2 \cdot \mathbf{Z}_2 & \cdots & \mathbf{Z}'_2 \cdot \mathbf{Z}_Q \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Z}'_Q \cdot \mathbf{Z}_1 & \mathbf{Z}'_Q \cdot \mathbf{Z}_2 & \cdots & \mathbf{Z}'_Q \cdot \mathbf{Z}_Q \end{pmatrix}_{J \times J} \tag{11.41}$$

Segundo Naito (2007), enquanto cada submatriz $\mathbf{Z}'_q \cdot \mathbf{Z}_q$ é uma matriz diagonal, cujos elementos são, respectivamente, iguais à soma das colunas da matriz $\mathbf{Z}_q$, cada submatriz $\mathbf{Z}'_q \cdot \mathbf{Z}_{q'}$ $(q \neq q')$ corresponde a uma tabela de contingência com os cruzamentos de cada variável $q$ com cada variável $q'$. Essa estrutura permite comparar os comportamentos das frequências absolutas observadas para todos os pares de variáveis, ao contrário do que ocorre com a matriz binária **Z**.

**Considerando a matriz de Burt (B) uma tabela de contingência**, podemos também elaborar uma análise de correspondência simples, da qual se pode verificar que as coordenadas das categorias das variáveis corresponderão às coordenadas-padrão geradas por meio do método da matriz binária **Z**, porém com valores em escala reduzida. Esse fato, segundo discute Greenacre (2008), faz os mapas perceptuais construídos a partir das coordenadas geradas pelo método da matriz de Burt serem mais reduzidos e com pontos mais concentrados em torno da Origem, o que, em alguns casos, pode prejudicar a análise visual das associações entre as categorias, embora isso não afete o estudo da relação entre as variáveis.

**As coordenadas geradas por meio do método da matriz de Burt são conhecidas por coordenadas principais**, e a relação entre essas coordenadas principais e as coordenadas-padrão obtidas pelo método da matriz binária é dada pela seguinte expressão:

$$(\textbf{coord. principal}_{\textbf{dim.}k})_{\mathbf{B}} = \lambda_k \cdot (\textbf{coord. padrão}_{\textbf{dim.}k})_{\mathbf{Z}} \tag{11.42}$$

ou seja, as coordenadas principais de determinada dimensão são as coordenadas-padrão multiplicadas pela raiz quadrada da inércia principal parcial daquela dimensão. Como as inércias principais parciais são menores que 1, explica-se a redução de escala do mapa perceptual construído a partir do método da matriz de Burt.

Enquanto elaboraremos a análise de correspondência múltipla fazendo uso das coordenadas principais no SPSS, a mesma técnica será elaborada com base nas coordenadas-padrão obtidas pelo método da matriz binária no Stata, conforme poderemos analisar nas seções 11.4.2 e 11.5.2, respectivamente.

Introduzidos esses conceitos, vamos apresentar um exemplo com o mesmo banco de dados utilizado quando da elaboração da análise de correspondência simples, porém com a inclusão de uma terceira variável categórica.

## 11.3.2. Exemplo prático da análise de correspondência múltipla (ACM)

Imagine agora que nosso professor tenha o interesse em estudar as associações eventualmente existentes entre o perfil de investidor de seus alunos, o tipo de aplicação financeira em que alocam seus recursos e uma terceira variável categórica, correspondente ao estado civil de cada um deles. Portanto, o banco de dados, parcialmente apresentado na Tabela 11.21, traz, além das variáveis estudadas quando da elaboração da análise de correspondência simples (*perfil* e *aplicação*), uma nova variável correspondente ao estado civil de cada estudante, com apenas duas categorias (solteiro ou casado).

**Tabela 11.21** Exemplo: Perfil do investidor, tipo de aplicação financeira e estado civil.

| Estudante | Perfil do Investidor | Tipo de Aplicação Financeira | Estado Civil |
|---|---|---|---|
| Gabriela | Conservador | Poupança | Casado |
| Luiz Felipe | Conservador | Poupança | Casado |
| ⋮ | | | |
| Renata | Conservador | CDB | Casado |
| Guilherme | Conservador | Ações | Solteiro |
| ⋮ | | | |
| Kamal | Moderado | Poupança | Solteiro |
| Rodolfo | Moderado | CDB | Solteiro |
| ⋮ | | | |
| Raquel | Moderado | CDB | Casado |
| Anna Luiza | Moderado | Ações | Solteiro |
| ⋮ | | | |
| Nuno | Agressivo | Poupança | Solteiro |
| Bráulio | Agressivo | CDB | Solteiro |
| ⋮ | | | |
| Estela | Agressivo | Ações | Solteiro |

O banco de dados completo pode ser acessado no arquivo **Perfil_Investidor × Aplicação × Estado_Civil.xls**. Nesse exemplo, temos $N = 100$ observações e $Q = 3$ variáveis, sendo que cada variável possui, respectivamente, $J_1 = 3$ categorias, $J_2 = 3$ categorias e $J_3 = 2$ categorias. Portanto, o número total de categorias envolvidas nessa análise de correspondência múltipla é $J = 8$.

Antes de elaborarmos a análise de correspondência múltipla propriamente dita, apresentamos, nas Tabelas 11.22, 11.23 e 11.24, as tabelas de contingência entre cada par de variáveis, com destaque para os resultados dos respectivos testes $\chi^2$.

**Tabela 11.22** Tabela de contingência para perfil do investidor e tipo de aplicação financeira.

| Perfil \ Aplicação | Poupança | CDB | Ações | **Total** |
|---|---|---|---|---|
| Conservador | 8 | 4 | 5 | **17** |
| Moderado | 5 | 16 | 4 | **25** |
| Agressivo | 2 | 20 | 36 | **58** |
| **Total** | **15** | **40** | **45** | **100** |
| $\chi^2 = 31{,}764$ (*valor-P* $\chi^2_{cal} = 0{,}000$) | | | | |

**Tabela 11.23** Tabela de contingência para perfil do investidor e estado civil.

| Perfil \ Estado Civil | Solteiro | Casado | **Total** |
|---|---|---|---|
| Conservador | 5 | 12 | **17** |
| Moderado | 11 | 14 | **25** |
| Agressivo | 41 | 17 | **58** |
| **Total** | **57** | **43** | **100** |
| $\chi^2 = 11{,}438$ ($valor\text{-}P\ \chi^2_{cal} = 0{,}003$) | | | |

**Tabela 11.24** Tabela de contingência para tipo de aplicação financeira e estado civil.

| Aplicação \ Estado Civil | Solteiro | Casado | **Total** |
|---|---|---|---|
| Poupança | 5 | 10 | **15** |
| CDB | 16 | 24 | **40** |
| Ações | 36 | 9 | **45** |
| **Total** | **57** | **43** | **100** |
| $\chi^2 = 17{,}857$ ($valor\text{-}P\ \chi^2_{cal} = 0{,}000$) | | | |

Com base nos resultados dos testes $\chi^2$, podemos afirmar que existem associações estatisticamente significantes, ao nível de significância de 5%, entre cada par de variáveis e, portanto, as três variáveis serão incluídas na análise de correspondência múltipla. Caso uma delas não se associasse a nenhuma outra a determinado nível de significância, seria recomendável sua exclusão da análise de correspondência múltipla.

Conforme discutimos na seção 11.3.1, por meio desse banco de dados é possível construir uma matriz **Z**, que possui apenas variáveis binárias criadas com base na codificação das categorias das variáveis originais para cada estudante. Assim, por exemplo, para a observação 1 (**Gabriela**), que apresenta perfil de investidor *Conservador*, aplica seus recursos em *Poupança* e encontra-se no estado civil *Casado*, temos a codificação binária representada, respectivamente, por (1 0 0), (1 0 0), ..., (0 1). A Tabela 11.25 apresenta a codificação binária para as observações apresentadas na Tabela 11.21.

**Tabela 11.25** Codificação binária das categorias das variáveis originais – Matriz binária **Z**.

| | **Perfil do Investidor ($Z_1$)** | | | **Tipo de Aplicação Financeira ($Z_2$)** | | | **Estado Civil ($Z_3$)** | |
|---|---|---|---|---|---|---|---|---|
| **Observação** | **Conservador** | **Moderado** | **Agressivo** | **Poupança** | **CDB** | **Ações** | **Solteiro** | **Casado** |
| Gabriela | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| Luiz Felipe | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| ⋮ | | | | | | | | |
| Renata | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| Guilherme | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| ⋮ | | | | | | | | |
| Kamal | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| Rodolfo | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| ⋮ | | | | | | | | |
| Raquel | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| Anna Luiza | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| ⋮ | | | | | | | | |
| Nuno | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| Bráulio | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| ⋮ | | | | | | | | |
| Estela | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |

A matriz binária **Z** completa também pode ser acessada no arquivo **Perfil_Investidor × Aplicação × Estado_Civil.xls**. Inicialmente, com base na expressão (11.39), podemos calcular a inércia principal total de **Z**. Assim, temos que:

$$I_T = \frac{8-3}{3} = 1{,}666$$

Supondo que a matriz binária **Z** seja uma tabela de contingência de uma análise de correspondência simples, podem ser definidos os valores das inércias principais parciais de cada uma das $J - Q = 8 - 3 = 5$ dimensões. Assim, fazendo uso dos conceitos estudados na seção 11.2, chegamos aos seguintes valores das inércias principais parciais, que são autovalores obtidos a partir da matriz binária **Z**:

$$\begin{cases} \lambda_1^2 = 0{,}602 \\ \lambda_2^2 = 0{,}436 \\ \lambda_3^2 = 0{,}276 \\ \lambda_4^2 = 0{,}180 \\ \lambda_5^2 = 0{,}172 \end{cases}$$

de onde podemos comprovar que $I_T = \lambda_1^2 + \lambda_2^2 + \lambda_3^2 + \lambda_4^2 + \lambda_5^2 = 1{,}666$.

Conforme discute Greenacre (2008), **somente é interessante que sejam plotadas no mapa perceptual as coordenadas das dimensões que apresentarem valores de inércia principal parcial superiores à média da inércia principal total por dimensão** que, em nosso exemplo, é igual a (1,666/5) = 0,333. Portanto, para a análise de correspondência múltipla de nosso exemplo, será construído um mapa perceptual com duas dimensões, visto que $\lambda_3^2 < 0{,}333$. A Tabela 11.26 apresenta as coordenadas-padrão das categorias de cada uma das variáveis para as duas dimensões, calculadas da mesma forma que no exemplo apresentado na seção 11.2.5, com base nos conceitos e expressões estudados ao longo da seção 11.2.

**Tabela 11.26** Coordenadas-padrão das categorias das variáveis – Método da matriz binária **Z**.

| Variável | Categoria | Coordenadas da 1ª Dimensão (Abcissas) | Coordenadas da 2ª Dimensão (Ordenadas) |
|---|---|---|---|
| Perfil do Investidor | Conservador | $\mathbf{x}_{11} = 1{,}456$ | $\mathbf{y}_{11} = 2{,}247$ |
| | Moderado | $\mathbf{x}_{12} = 0{,}962$ | $\mathbf{y}_{12} = -1{,}476$ |
| | Agressivo | $\mathbf{x}_{13} = -0{,}841$ | $\mathbf{y}_{13} = -0{,}022$ |
| Tipo de Aplicação Financeira | Poupança | $\mathbf{x}_{21} = 1{,}780$ | $\mathbf{y}_{21} = 2{,}016$ |
| | CDB | $\mathbf{x}_{22} = 0{,}538$ | $\mathbf{y}_{22} = -1{,}416$ |
| | Ações | $\mathbf{x}_{23} = -1{,}071$ | $\mathbf{y}_{23} = 0{,}587$ |
| Estado Civil | Solteiro | $\mathbf{x}_{31} = -0{,}820$ | $\mathbf{y}_{31} = 0{,}150$ |
| | Casado | $\mathbf{x}_{32} = 1{,}086$ | $\mathbf{y}_{32} = -0{,}199$ |

Conforme discutimos na seção 11.3.1, a análise de correspondência múltipla também pode ser realizada por meio da elaboração de uma matriz quadrada e simétrica que agrupa as frequências absolutas observadas provenientes dos cruzamentos de todos os pares de variáveis, conhecida por matriz de Burt. A matriz de Burt do nosso exemplo, que pode ser construída tanto por meio da expressão (11.40), fazendo-se uso da matriz binária **Z**, quanto por meio das tabelas de contingência apresentadas nas Tabelas 11.22, 11.23 e 11.24, encontra-se na Tabela 11.27.

Note, na Tabela 11.27, que as submatrizes $\mathbf{Z}_1' \cdot \mathbf{Z}_1$, $\mathbf{Z}_2' \cdot \mathbf{Z}_2$ e $\mathbf{Z}_3' \cdot \mathbf{Z}_3$, em destaque, são matrizes diagonais cujos elementos correspondem, respectivamente, à soma das colunas das matrizes $\mathbf{Z}_1$, $\mathbf{Z}_2$ e $\mathbf{Z}_3$ (perfil do investidor, tipo de aplicação financeira e estado civil, respectivamente). Já as matrizes $\mathbf{Z}_1' \cdot \mathbf{Z}_2$, $\mathbf{Z}_1' \cdot \mathbf{Z}_3$ e $\mathbf{Z}_2' \cdot \mathbf{Z}_3$, e correspondem, respectivamente, às tabelas de contingência apresentadas nas Tabelas 11.22, 11.23 e 11.24.

**Tabela 11.27** Matriz de Burt (**B**).

| | | Perfil do Investidor | | | Tipo de Aplicação Financeira | | | Estado Civil | |
|---|---|---|---|---|---|---|---|---|---|
| | | Conservador | Moderado | Agressivo | Poupança | CDB | Ações | Solteiro | Casado |
| Perfil do Investidor | Conservador | 17 | 0 | 0 | 8 | 4 | 5 | 5 | 12 |
| | Moderado | 0 | 25 | 0 | 5 | 16 | 4 | 11 | 14 |
| | Agressivo | 0 | 0 | 58 | 2 | 20 | 36 | 41 | 17 |
| Tipo de Aplicação Financeira | Poupança | 8 | 5 | 2 | 15 | 0 | 0 | 5 | 10 |
| | CDB | 4 | 16 | 20 | 0 | 40 | 0 | 16 | 24 |
| | Ações | 5 | 4 | 36 | 0 | 0 | 45 | 36 | 9 |
| Estado Civil | Solteiro | 5 | 11 | 41 | 5 | 16 | 36 | 57 | 0 |
| | Casado | 12 | 14 | 17 | 10 | 24 | 9 | 0 | 43 |
| **Massas** | | **0,057** | **0,083** | **0,193** | **0,050** | **0,133** | **0,150** | **0,190** | **0,143** |

Considerando a matriz de Burt (**B**) uma tabela de contingência, podemos também elaborar uma análise de correspondência simples, que gera as coordenadas principais das categorias das variáveis, conforme apresentado na Tabela 11.28.

**Tabela 11.28** Coordenadas principais das categorias das variáveis – Método da matriz de Burt **B**.

| Variável | Categoria | Coordenadas da 1ª Dimensão (Abcissas) | Coordenadas da 2ª Dimensão (Ordenadas) |
|---|---|---|---|
| Perfil do Investidor | Conservador | $\mathbf{x}_{11} = 1{,}130$ | $\mathbf{y}_{11} = 1{,}484$ |
| | Moderado | $\mathbf{x}_{12} = 0{,}747$ | $\mathbf{y}_{12} = -0{,}975$ |
| | Agressivo | $\mathbf{x}_{13} = -0{,}653$ | $\mathbf{y}_{13} = -0{,}015$ |
| Tipo de Aplicação Financeira | Poupança | $\mathbf{x}_{21} = 1{,}381$ | $\mathbf{y}_{21} = 1{,}331$ |
| | CDB | $\mathbf{x}_{22} = 0{,}417$ | $\mathbf{y}_{22} = -0{,}935$ |
| | Ações | $\mathbf{x}_{23} = -0{,}831$ | $\mathbf{y}_{23} = 0{,}388$ |
| Estado Civil | Solteiro | $\mathbf{x}_{31} = -0{,}636$ | $\mathbf{y}_{31} = 0{,}099$ |
| | Casado | $\mathbf{x}_{32} = 0{,}843$ | $\mathbf{y}_{32} = -0{,}131$ |

Com base nas coordenadas apresentadas nas Tabelas 11.26 (método da matriz binária **Z**) e 11.28 (método da matriz de Burt **B**), podemos facilmente verificar a relação existente entre elas, apresentada na expressão (11.42). Assim, para a primeira dimensão da categoria *Conservador* temos, por exemplo, que:

$$\mathbf{(coord.\ principal_1)_B} = \sqrt{\lambda_1^2} \cdot \mathbf{(coord.\ padrão_1)_Z} = \sqrt{0{,}602} \cdot (1{,}456) = 1{,}130$$

e, para a segunda dimensão da mesma categoria, temos que:

$$\mathbf{(coord.\ principal_2)_B} = \sqrt{\lambda_2^2} \cdot \mathbf{(coord.\ padrão_2)_Z} = \sqrt{0{,}436} \cdot (2{,}247) = 1{,}484$$

Isso mostra que as coordenadas obtidas pelo método da matriz de Burt realmente apresentam escala reduzida, em especial para a segunda dimensão, pelo fato de a inércia principal parcial ser ainda menor.

Enquanto na seção 11.4.2 serão apresentados os resultados dos procedimentos para elaboração da análise de correspondência múltipla no SPSS, em que são geradas as coordenadas principais das categorias, na seção 11.5.2 serão apresentados os resultados dos procedimentos para elaboração da técnica no Stata, por meio dos quais será possível analisar as coordenadas-padrão obtidas pelo método da matriz binária **Z**.

Como o método da matriz de Burt gera coordenadas com escala reduzida, optamos por apresentar, na Figura 11.12, o mapa perceptual construído com base nas coordenadas-padrão obtidas pelo método da matriz binária e apresentadas na Tabela 11.26.

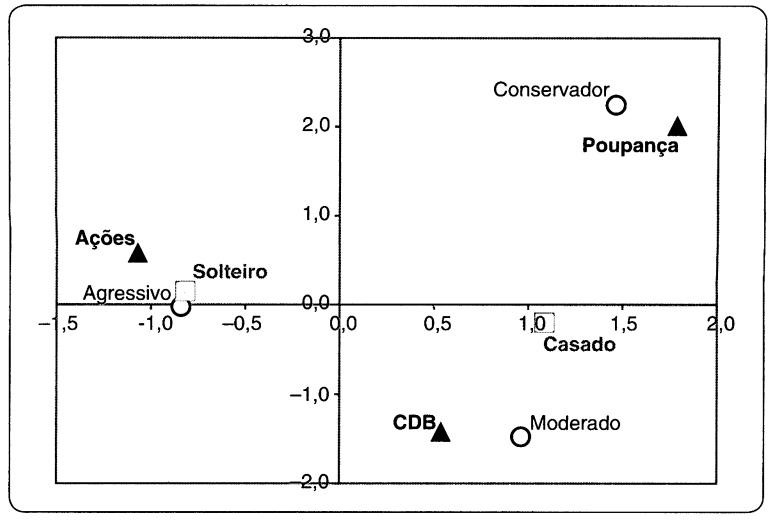



**Figura 11.12** Mapa perceptual da análise de correspondência múltipla – Coordenadas-padrão.

Com base no mapa perceptual da Figura 11.12, podemos verificar que a categoria *Solteiro* apresenta forte associação com as categorias *Agressivo* e *Ações*. Por outro lado, a categoria *Casado* encontra-se entre as categorias *Conservador* e *Moderado* e entre *Poupança* e *CDB*, porém com maior proximidade de *Moderado* e *CDB*. Esse fato é provavelmente caracterizado pela maior aversão ao risco que passam a ter aqueles que se tornam responsáveis por uma família, como os casados.

Interessante também seria se incluíssemos na análise uma variável que permitisse identificar se cada estudante possui ou não filhos, independentemente da quantidade. Será que o fato de ter filhos aumenta ainda mais a aversão ao risco? Há associação entre o fato de ter um ou mais filhos, o perfil do investidor e o tipo de aplicação financeira? Deixaremos essas perguntas para um exercício ao final do capítulo.

## 11.4. ANÁLISE DE CORRESPONDÊNCIA SIMPLES E MÚLTIPLA NO SOFTWARE SPSS

Nesta seção, apresentaremos o passo a passo para a elaboração de nossos exemplos no IBM SPSS Statistics Software®. Seguindo a lógica proposta no livro, o principal objetivo é propiciar ao pesquisador uma oportunidade de elaborar análises de correspondências simples e múltiplas neste software, dada sua facilidade de manuseio e a didática das operações. A cada apresentação de um *output*, faremos menção ao respectivo resultado obtido quando da solução algébrica das técnicas nas seções anteriores, a fim de que o pesquisador possa compará-los e formar seu conhecimento e erudição sobre o tema. A reprodução das imagens nessa seção tem autorização da International Business Machines Corporation©.

### 11.4.1. Elaboração da análise de correspondência simples no software SPSS

Voltando ao exemplo apresentado na seção 11.2.5, lembremos que nosso professor tem o interesse em estudar se o perfil de investidor de seus alunos relaciona-se com o tipo de aplicação financeira realizada, ou seja, se existe associação estatisticamente significante, a determinado nível de significância, entre os perfis dos investidores e a forma como são alocados seus recursos financeiros. Os dados encontram-se no arquivo **Perfil_Investidor × Aplicação.sav** e são exatamente iguais aos apresentados parcialmente na Tabela 11.7 da seção 11.2.5. Note que os rótulos das categorias das variáveis *perfil* e *aplicação* já estão definidos no banco de dados.

A fim de que sejam geradas as tabelas de frequências absolutas observadas (*cross-tabulations*) e esperadas e, consequentemente, a tabela de resíduos e o valor da estatística $\chi^2$, vamos inicialmente clicar em **Analyze** → **Descriptive Statistics** → **Crosstabs...**, para elaborarmos o primeiro diagnóstico sobre a interdependência entre as duas variáveis categóricas. A caixa de diálogo da Figura 11.13 será aberta.

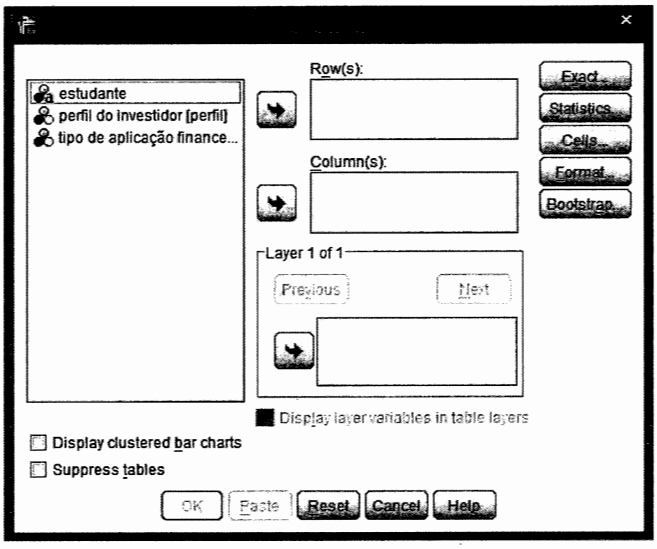



**Figura 11.13** Caixa de diálogo para elaboração das tabelas de frequências absolutas observadas e esperadas, dos resíduos e do teste $\chi^2$.

Conforme mostra a Figura 11.14, devemos inserir a variável *perfil* em **Row(s)**, e a variável *aplicação* em **Column(s)**. No botão **Statistics...**, devemos selecionar a opção **Chi-square**, conforme mostra a Figura 11.15.

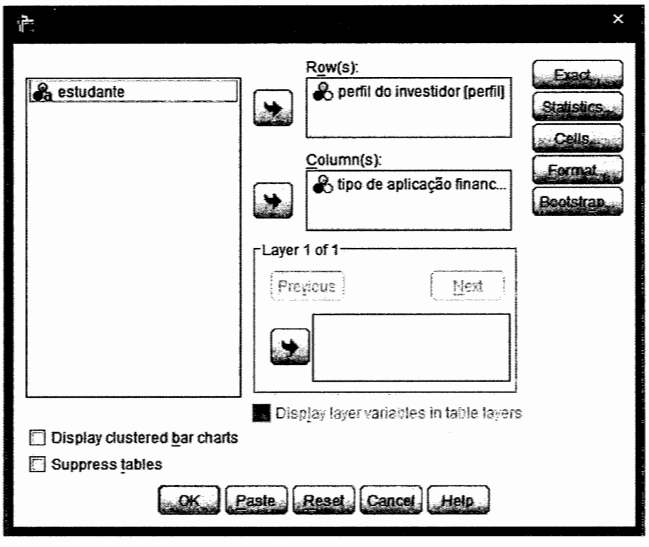



**Figura 11.14** Seleção das variáveis em **Row(s)** e em **Column(s).**

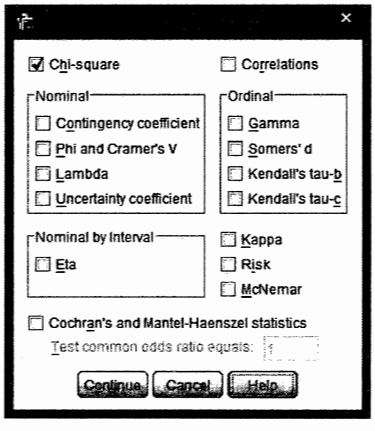



**Figura 11.15** Seleção da estatística $\chi^2$.

Ao clicarmos em **Continue**, voltaremos à caixa de diálogo anterior. No botão **Cells...**, marcaremos as opções **Observed** e **Expected**, em **Counts**, e **Unstandardized**, **Standardized** e **Adjusted standardized**, em **Residuals**, conforme mostra a Figura 11.16.

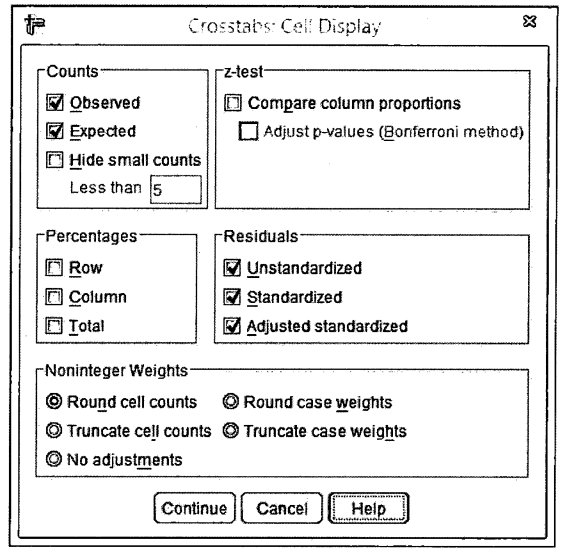



**Figura 11.16** Seleção das opções para elaboração das tabelas de frequências e dos resíduos.

Na sequência, podemos clicar em **Continue** e em **OK**. Os primeiros *outputs* encontram-se nas Figuras 11.17 e 11.18.

Conforme estudamos nas seções anteriores, a fim de verificarmos inicialmente a existência de associação estatisticamente significante entre as variáveis *perfil* e *aplicação*, devemos fazer uso do teste $\chi^2$. A Figura 11.17 apresenta a estatística correspondente, cujo cálculo é feito com base na somatória, para todas as células, da razão entre o resíduo ao quadrado e a respectiva frequência esperada, de acordo com a expressão (11.6).

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 31,764[a] | 4 | ,000 |
| Likelihood Ratio | 30,777 | 4 | ,000 |
| Linear-by-Linear Association | 20,352 | 1 | ,000 |
| N of Valid Cases | 100 | | |

a. 2 cells (22,2%) have expected count less than 5. The minimum expected count is 2,55.

**Figura 11.17** Resultado do teste $\chi^2$ para verificação de associação entre *perfil* e *aplicação*.

Logo, temos que:

$$\chi^2_{4\,g.l.} = \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{(resíduos_{ij})^2}{(frequências\ esperadas_{ij})} = 31{,}764$$

que é exatamente igual ao valor calculado algebricamente na seção 11.2.5. Assim, de acordo com a Figura 11.17, o *valor-P* (*Asymp. Sig.*) da estatística $\chi^2_{cal}$ é consideravelmente menor que 0,05 (*valor-P* $\chi^2_{cal}$ = 0,000). Logo, para $(I-1) \times (J-1) = (3-1) \times (3-1) = 4$ graus de liberdade, podemos rejeitar a hipótese nula de que as duas variáveis categóricas se associam de forma aleatória, ou seja, existe associação estatisticamente significante, ao nível de significância de 5%, entre o perfil do investidor e o tipo de aplicação financeira.

Conforme discutimos na seção 11.2.5, tão importante quanto avaliar a existência de associação estatisticamente significante entre essas duas variáveis é estudar a relação de dependência entre cada par de categorias. A Figura 11.18 permite que essa análise seja elaborada.

**perfil do investidor * tipo de aplicação financeira Crosstabulation**

| | | | tipo de aplicação financeira | | | |
|---|---|---|---|---|---|---|
| | | | Poupança | CDB | Ações | Total |
| perfil do investidor | Conservador | Count | 8 | 4 | 5 | 17 |
| | | Expected Count | 2,6 | 6,8 | 7,7 | 17,0 |
| | | Residual | 5,5 | -2,8 | -2,7 | |
| | | Std. Residual | 3,4 | -1,1 | -1,0 | |
| | | Adjusted Residual | 4,1 | -1,5 | -1,4 | |
| | Moderado | Count | 5 | 16 | 4 | 25 |
| | | Expected Count | 3,8 | 10,0 | 11,3 | 25,0 |
| | | Residual | 1,3 | 6,0 | -7,3 | |
| | | Std. Residual | ,6 | 1,9 | -2,2 | |
| | | Adjusted Residual | ,8 | 2,8 | -3,4 | |
| | Agressivo | Count | 2 | 20 | 36 | 58 |
| | | Expected Count | 8,7 | 23,2 | 26,1 | 58,0 |
| | | Residual | -6,7 | -3,2 | 9,9 | |
| | | Std. Residual | -2,3 | -,7 | 1,9 | |
| | | Adjusted Residual | -3,8 | -1,3 | 4,0 | |
| Total | | Count | 15 | 40 | 45 | 100 |
| | | Expected Count | 15,0 | 40,0 | 45,0 | 100,0 |

**Figura 11.18** Tabela de frequências e de resíduos para *perfil* e *aplicação*.

A Figura 11.18 mostra, para cada uma das células, as frequências absolutas observadas (*Count*), as frequências absolutas esperadas (*Expected Count*), os resíduos (*Residual*), os resíduos padronizados (*Std. Residual*) e os resíduos padronizados ajustados (*Adjusted Residual*), bem como os valores totais em linha e em coluna de *Count* e de *Expected Count* que, obviamente, são iguais. Note que, enquanto os valores de *Count* correspondem aos apresentados na Tabela 11.8, os valores de *Expected Count* e de *Residual* são os calculados e apresentados nas Tabelas 11.9 e 11.10, respectivamente. Além disso, os valores de *Std. Residual* e de *Adjusted Residual* correspondem, respectivamente, aos apresentados nas Tabelas 11.12 e 11.13.

Podemos verificar que, enquanto há uma maior proporção de estudantes que se consideram agressivos em termos de perfil de investidor, há também uma quantidade maior de estudantes que aplicam seus recursos financeiros em ações. No perfil *Conservador*, os resíduos são maiores para a categoria *Poupança*, o que indica que as diferenças entre as frequências absolutas observadas e esperadas nessa célula são maiores que para as demais células do perfil *Conservador* e, como o valor do resíduo padronizado ajustado nessa célula é igual a 4,1 (positivo e maior que 1,96), podemos concluir que há dependência entre as categorias *Conservador* e *Poupança*. O mesmo também pode ser dito em relação às categorias *Moderado* e *CDB* (resíduo padronizado ajustado igual a 2,8) e entre as categorias *Agressivo* e *Ações* (resíduo padronizado ajustado igual a 4,0).

Em muitos casos, o pesquisador pode restringir a análise apenas com base nos resultados do teste $\chi^2$ e nos resíduos padronizados ajustados, já que esses já oferecem muitos subsídios para a elaboração de uma interessante análise dos dados com foco para a tomada de decisão. Entretanto, para que seja construído o mapa perceptual no SPSS, é necessário elaborar mais alguns passos. Para tanto, vamos clicar em **Analyze → Dimension Reduction → Correspondence Analysis...**. Uma caixa de diálogo como a apresentada na Figura 11.19 será aberta.

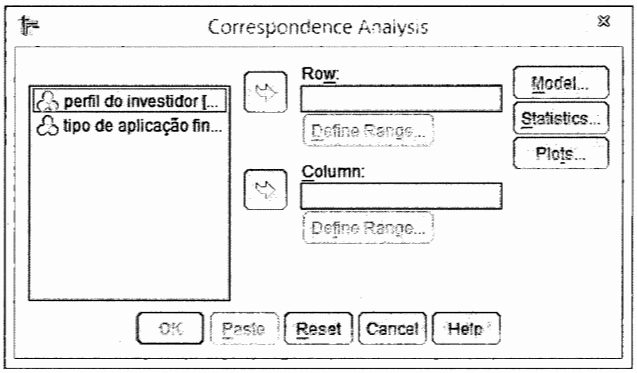



**Figura 11.19** Caixa de diálogo para elaboração da análise de correspondência simples no SPSS.

Devemos inicialmente selecionar a variável *perfil* e inseri-la em **Row**, conforme mostra a Figura 11.20.

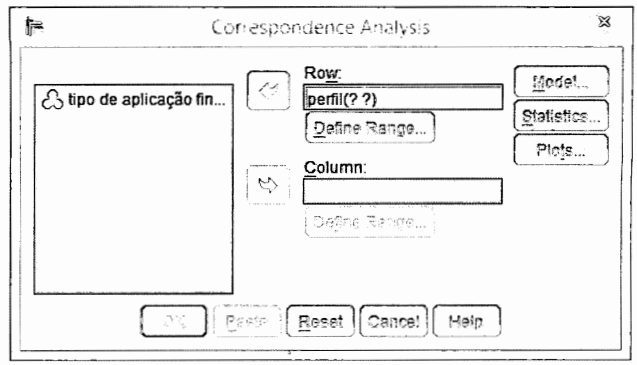



**Figura 11.20** Inclusão da variável *perfil* em **Row**.

Ao clicarmos em **Define Range...**, abrirá uma caixa de diálogo. Como a variável *perfil* apresenta três categorias (*Conservador*, *Moderado* e *Agressivo*), e nossa intenção é incluí-las, sem exceção, na análise de correspondência, devemos digitar 1 em **Minimum value**, 3 em **Maximum value** e clicar em **Update**, conforme mostra a Figura 11.21. É importante lembrar que os valores 1, 2 e 3 foram inseridos inicialmente no banco de dados, e, a eles, foram atribuídas, respectivamente, as categorias *Conservador*, *Moderado* e *Agressivo* como rótulos (*labels*). O pesquisador poderá, como bem entender, alterar os valores iniciais de preenchimento no banco de dados; porém, nesse momento, precisará digitar os valores correspondentes às categorias a serem incluídas na análise. Para retornarmos à caixa de diálogo principal, devemos clicar em **Continue**.

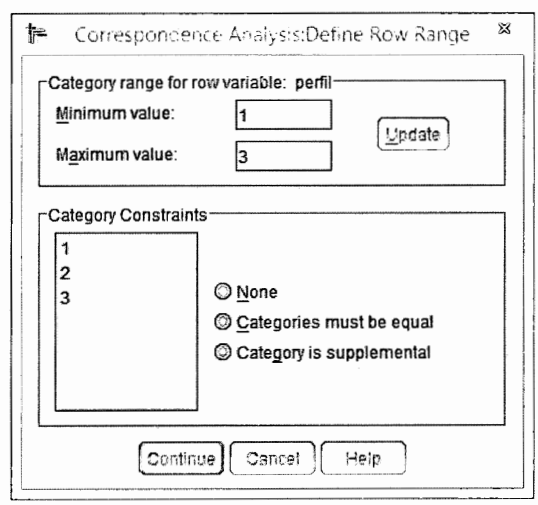



**Figura 11.21** Seleção das categorias da variável *perfil*.

Na sequência, vamos elaborar o mesmo procedimento para a variável *aplicação*. Conforme mostra a Figura 11.22, devemos inseri-la em **Column**.

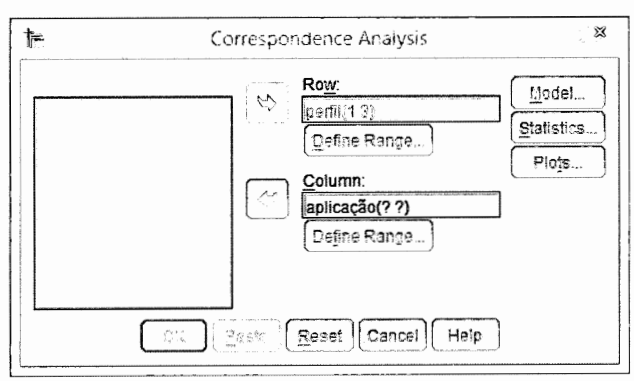



**Figura 11.22** Inclusão da variável *aplicação* em **Column**.

Analogamente, em **Define Range...**, devemos digitar 1 em **Minimum value**, 3 em **Maximum value** e clicar em **Update**, como mostra a Figura 11.23, visto que a variável *aplicação* também apresenta três categorias (*Poupança*, *CDB* e *Ações*). Na sequência, vamos clicar em **Continue** para voltarmos à caixa de diálogo inicial.

Na caixa de diálogo inicial, vamos agora clicar em **Model...**. Abrirá uma caixa em que deverão ser selecionadas as opções **Chi square** (em **Distance Measure**), **Row and column means are removed** (em **Standardization Method**) e **Symmetrical** (em **Normalization Method**), de acordo com a Figura 11.24. Por meio dessa mesma figura, é possível verificar que há o valor 2 em **Dimensions in solution**, correspondente ao número de dimensões do mapa perceptual. Nesse caso, o número de dimensões é, de fato, 2, uma vez que, conforme estudamos, o número de dimensões é igual a mín$(I-1, J-1)$. Caso tivéssemos mais categorias em cada uma das variáveis, ainda assim poderíamos elaborar um mapa perceptual bidimensional, plotando apenas as duas dimensões com as maiores inércias principais parciais.

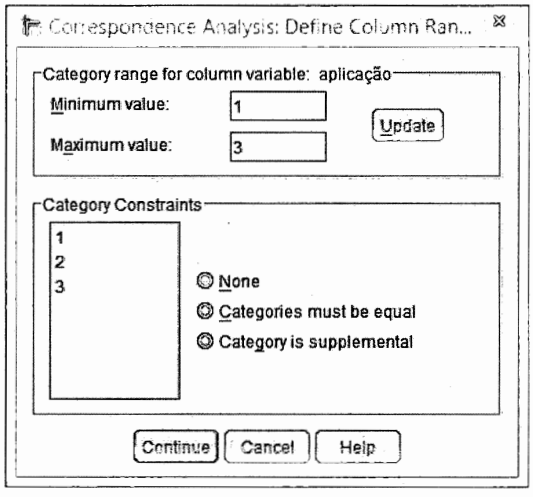



**Figura 11.23** Seleção das categorias da variável *aplicação*.

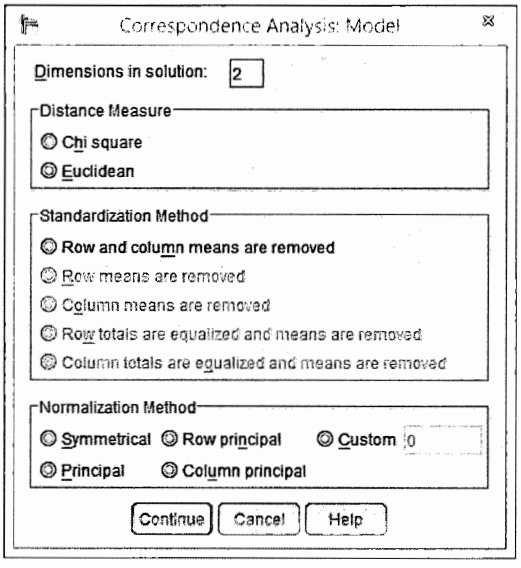



**Figura 11.24** Definição das características da análise de correspondência.

Conforme discutimos na seção 11.2.4, é possível que o pesquisador deseje privilegiar exclusivamente a visualização das massas em linha ou em coluna de determinada tabela de contingência para a construção do mapa perceptual. Nesse sentido, poderá abrir mão da normalização simétrica (**Symmetrical**) e optar pelas normalizações principal linha ou principal coluna, clicando, respetivamente, nas opções **Row principal** ou **Column principal** em **Normalization Method** (Figura 11.24). Nesses casos, as coordenadas das categorias serão calculadas com base nas expressões apresentadas no Quadro 11.1. Não apresentaremos, todavia, esses mapas específicos.

Para dar sequência à análise, devemos clicar em **Continue**. Na caixa de diálogo inicial, vamos clicar em **Statistics...** e, na caixa que será aberta, vamos marcar as opções **Correspondence table**, **Row profiles** e **Column profiles**, a fim de que sejam geradas, nos *outputs*, a tabela de contingência (tabela de frequências absolutas observadas) e as tabelas de massas *row profiles* e *column profiles*. Além disso, vamos também selecionar as opções **Overview of row points** e **Overview of column points**, por meio das quais serão apresentados os quadros com as coordenadas das categorias das variáveis. A Figura 11.25 apresenta essas opções selecionadas. Na sequência, devemos clicar em **Continue**.

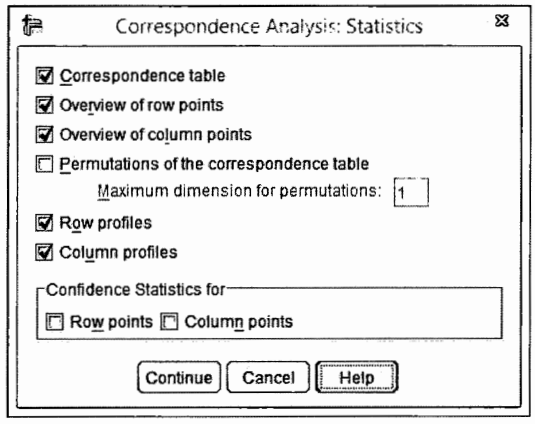



**Figura 11.25** Definição dos *outputs* a serem gerados.

Por fim, em **Plots...** (caixa de diálogo inicial), devemos apenas clicar em **Biplot**, conforme mostra a Figura 11.26. Caso o pesquisador deseje elaborar gráficos com as categorias de apenas uma das variáveis, poderá também selecionar as opções **Row points** ou **Column points**. Na sequência, podemos clicar em **Continue** e em **OK**.

Os primeiros *outputs* gerados encontram-se nas Figuras 11.27, 11.28 e 11.29 e referem-se, respectivamente, à tabela de contingência e às tabelas de massas *column profile* e *row profile*. Os valores nessas figuras correspondem, respectivamente, aos apresentados nas Tabelas 11.8, 11.14 e 11.15.

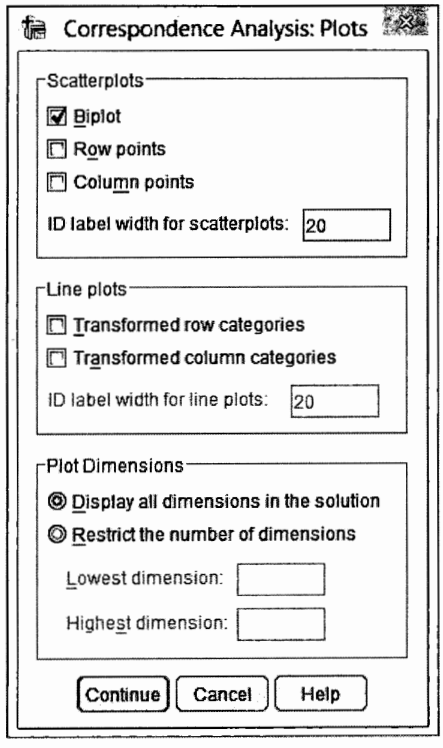



**Figura 11.26** Definição do mapa perceptual.

**Correspondence Table**

| perfil do investidor | tipo de aplicação financeira | | | |
|---|---|---|---|---|
| | Poupança | CDB | Ações | Active Margin |
| Conservador | 8 | 4 | 5 | 17 |
| Moderado | 5 | 16 | 4 | 25 |
| Agressivo | 2 | 20 | 36 | 58 |
| Active Margin | 15 | 40 | 45 | 100 |

**Figura 11.27** Tabela de contingência com frequências absolutas observadas para *perfil* e *aplicação*.

**Column Profiles**

| perfil do investidor | tipo de aplicação financeira | | | |
|---|---|---|---|---|
| | Poupança | CDB | Ações | Mass |
| Conservador | ,533 | ,100 | ,111 | ,170 |
| Moderado | ,333 | ,400 | ,089 | ,250 |
| Agressivo | ,133 | ,500 | ,800 | ,580 |
| Active Margin | 1,000 | 1,000 | 1,000 | |

**Figura 11.28** Massas – *Column profiles*.

**Row Profiles**

| perfil do investidor | tipo de aplicação financeira | | | |
|---|---|---|---|---|
| | Poupança | CDB | Ações | Active Margin |
| Conservador | ,471 | ,235 | ,294 | 1,000 |
| Moderado | ,200 | ,640 | ,160 | 1,000 |
| Agressivo | ,034 | ,345 | ,621 | 1,000 |
| Mass | ,150 | ,400 | ,450 | |

**Figura 11.29** Massas – *Row profiles*.

Logo, conforme também discutimos, a tabela de massas *column profiles* apresenta o cálculo das razões entre as frequências absolutas observadas de cada célula da tabela de contingência e a soma total de cada coluna (chamada, pelo SPSS, de *Active Margin*). Logo, a massa da categoria *Conservador* da variável *perfil* é dada pela relação 17/100 = 0,170.

Analogamente, a tabela de massas *row profiles* apresenta o cálculo das razões entre as frequências absolutas observadas de cada célula da tabela de contingência e a soma total de cada linha (também chamada, pelo SPSS, de *Active Margin*). Logo, a massa da categoria *CDB* da variável *aplicação* é dada pela relação 40/100 = 0,400.

Na sequência, são apresentados os *outputs* referentes à decomposição inercial (Figura 11.30), com destaque para os valores singulares e as inércias principais parciais de cada dimensão. Além disso, também são apresentados os valores da inércia principal total e da estatística $\chi^2$.

**Summary**

| Dimension | Singular Value | Inertia | Chi Square | Sig. | Proportion of Inertia | | Confidence Singular Value | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Accounted for | Cumulative | Standard Deviation | Correlation 2 |
| 1 | ,483 | ,233 | | | ,734 | ,734 | ,088 | ,179 |
| 2 | ,291 | ,084 | | | ,266 | 1,000 | ,100 | |
| Total | | ,318 | 31,764 | ,000[a] | 1,000 | 1,000 | | |

a. 4 degrees of freedom

**Figura 11.30** Decomposição inercial para as duas dimensões e estatística $\chi^2$.

Assim como mostra o *output* da Figura 11.17, podemos inicialmente verificar, com base nos *outputs* da Figura 11.30, que o perfil do investidor e o tipo de aplicação financeira não se combinam aleatoriamente, visto que o

*valor-P* da estatística $\chi^2_{cal}$ é menor que 0,05 (*Sig.* $\chi^2_{cal}$ = 0,000). Além disso, temos, para cada dimensão, os seguintes valores das inércias principais parciais:

$$\begin{cases} \lambda_1^2 = 0{,}233 \\ \lambda_2^2 = 0{,}084 \end{cases}$$

e, portanto, a inércia principal total é $I_T = \lambda_1^2 + \lambda_2^2 = 0{,}318$. Conforme estudamos na seção 11.2.5, podemos também verificar, por meio da expressão (11.7), que:

$$I_T = \frac{\chi^2}{N} = \frac{31{,}764}{100} = 0{,}318$$

Os valores singulares de cada dimensão são iguais a:

$$\begin{cases} \lambda_1 = 0{,}483 \\ \lambda_2 = 0{,}291 \end{cases}$$

Ainda com base nos *outputs* apresentados na Figura 11.30, podemos afirmar que as dimensões 1 e 2 explicam, respectivamente, 73,4% (0,233 / 0,318) e 26,6% (0,084 / 0,318) da inércia principal total. Esses valores já haviam sido calculados e apresentados na Tabela 11.16.

As Figuras 11.31 e 11.32 apresentam as coordenadas (abcissas e ordenadas) das categorias das duas variáveis. Enquanto as abcissas são denominadas *Score in Dimension* 1, as ordenadas são denominadas *Score in Dimension* 2.

**Overview Row Points[a]**

| | | Score in Dimension | | | Contribution | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Of Point to Inertia of Dimension | | Of Dimension to Inertia of Point | | |
| perfil do investidor | Mass | 1 | 2 | Inertia | 1 | 2 | 1 | 2 | Total |
| Conservador | ,170 | -1,132 | ,805 | ,137 | ,451 | ,379 | ,767 | ,233 | 1,000 |
| Moderado | ,250 | -,553 | -,829 | ,087 | ,158 | ,592 | ,425 | ,575 | 1,000 |
| Agressivo | ,580 | ,570 | ,122 | ,094 | ,391 | ,029 | ,973 | ,027 | 1,000 |
| Active Total | 1,000 | | | ,318 | 1,000 | 1,000 | | | |

a. Symmetrical normalization

**Figura 11.31** Coordenadas (*scores*) das categorias da variável *perfil*.

**Overview Column Points[a]**

| | | Score in Dimension | | | Contribution | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Of Point to Inertia of Dimension | | Of Dimension to Inertia of Point | | |
| tipo de aplicação financeira | Mass | 1 | 2 | Inertia | 1 | 2 | 1 | 2 | Total |
| Poupança | ,150 | -1,475 | ,582 | ,172 | ,675 | ,175 | ,914 | ,086 | 1,000 |
| CDB | ,400 | -,102 | -,655 | ,052 | ,009 | ,591 | ,039 | ,961 | 1,000 |
| Ações | ,450 | ,582 | ,389 | ,093 | ,316 | ,234 | ,789 | ,211 | 1,000 |
| Active Total | 1,000 | | | ,318 | 1,000 | 1,000 | | | |

a. Symmetrical normalization

**Figura 11.32** Coordenadas (*scores*) das categorias da variável *aplicação*.

Note, a partir dos *outputs* apresentados nas Figuras 11.31 e 11.32, que o SPSS apresenta as coordenadas das abcissas de cada categoria (*Score in Dimension* 1) com sinais invertidos em relação aos calculados algebricamente no final da seção 11.2.5. Isso faz o mapa perceptual ser construído de forma verticalmente espelhada se comparado ao mapa apresentado na Figura 11.9, porém não altera absolutamente a interpretação dos resultados da análise de correspondência. Ressalta-se que isso acontece apenas para algumas versões do SPSS.

Conforme discutimos, as coordenadas das categorias da variável em linha podem ser calculadas a partir das coordenadas das categorias da variável em coluna para determinada dimensão e vice-versa. Para tanto, devemos multiplicar a matriz de massas pelo vetor de coordenadas de uma variável e dividir pelo correspondente valor singular da dimensão em análise, para que sejam obtidas as coordenadas das categorias da outra variável, de acordo com as expressões (11.33) e (11.34). Assim, a abcissa da categoria *Ações* pode ser calculada da seguinte forma:

$$\mathbf{x}_{Ações} = \frac{[0{,}111 \times (-1{,}132)] + [0{,}089 \times (-0{,}553)] + [0{,}800 \times 0{,}570]}{0{,}483} = 0{,}582$$

e, analogamente, a ordenada da categoria *Moderado* pode ser calculada por meio da seguinte expressão:

$$\mathbf{y}_{Moderado} = \frac{[0{,}200 \times 0{,}582] + [0{,}640 \times (-0{,}655)] + [0{,}160 \times 0{,}389]}{0{,}291} = -0{,}829$$

Além disso, também mostramos, com base nas expressões (11.35) e (11.36), que os valores singulares de cada dimensão podem ser obtidos pela soma, em linha ou em coluna, da multiplicação da coordenada ao quadrado de cada categoria pela respectiva massa. Assim, para a primeira dimensão, e fazendo uso das coordenadas da variável *perfil*, podemos obter o valor singular da seguinte maneira:

$$\lambda_1 = [(-1{,}132)^2 \times 0{,}170] + [(-0{,}553)^2 \times 0{,}250] + [(0{,}570)^2 \times 0{,}580] = 0{,}483$$

e o mesmo resultado pode ser encontrado se forem utilizadas as coordenadas da variável *aplicação* e respectivas massas.

Analogamente, para a segunda dimensão, e fazendo uso das coordenadas da variável *aplicação*, podemos obter o valor singular da seguinte maneira:

$$\lambda_2 = [(0{,}582)^2 \times 0{,}150] + [(-0{,}655)^2 \times 0{,}400] + [(0{,}389)^2 \times 0{,}450] = 0{,}291$$

sendo o mesmo resultado obtido se utilizadas as coordenadas da variável *perfil* e respectivas massas.

As Figuras 11.31 e 11.32 apresentam também um importante *output*, chamado de ***Contribution of Point to Inertia of Dimension***, que oferece uma possibilidade de que sejam analisadas as categorias mais representativas de cada variável para a composição inercial de cada dimensão. Segundo Olariaga e Hernández (2000), se determinada categoria de uma variável apresentar, por exemplo, um valor de abcissa bastante alto em módulo, ou seja, mais distante horizontalmente da Origem, e possuir massa elevada, mais representativa essa categoria será para a composição inercial da primeira dimensão. Analogamente, se outra categoria apresentar, por exemplo, um valor de ordenada bastante alto em módulo, ou seja, mais distante verticalmente da Origem, e também possuir massa elevada, mais representativa essa outra categoria será para a composição inercial da segunda dimensão.

Por exemplo, a contribuição da categoria *Conservador* para a inércia da primeira dimensão pode ser calculada da seguinte forma:

$$\frac{[(-1{,}132)^2 \times 0{,}170]}{0{,}483} = 0{,}451$$

que torna a categoria *Conservador* a mais representativa da variável *perfil* para a composição inercial da primeira dimensão (45,1%). Para essa mesma variável, a categoria *Moderado* é a mais representativa para a composição inercial da segunda dimensão, com uma contribuição de 59,2% da inércia principal total. Já para a variável *aplicação*, enquanto a categoria *Poupança* é a mais representativa para a composição inercial da primeira dimensão (67,5%), a categoria *CDB* é a mais representativa para a composição inercial da segunda dimensão (59,1%).

Com base nas abcissas e ordenadas apresentadas nas Figuras 11.31 e 11.32, pode ser construído o mapa perceptual apresentado na Figura 11.33.

Conforme discutido, como as abcissas das categorias calculadas pelo SPSS apresentam sinais opostos aos das abcissas calculadas algebricamente na seção 11.2.5, o mapa perceptual da Figura 11.33 é horizontalmente espelhado em relação ao mapa apresentado na Figura 11.9 (esse fato ocorre apenas para algumas versões do SPSS). Entretanto, em nada altera a análise e não impede que se comprove a existência de associação entre as variáveis *perfil* e *aplicação* e, mais que isso, a associação entre as categorias *Conservador* e *Poupança*, entre *Moderado* e *CDB*, e entre *Agressivo* e *Ações*.

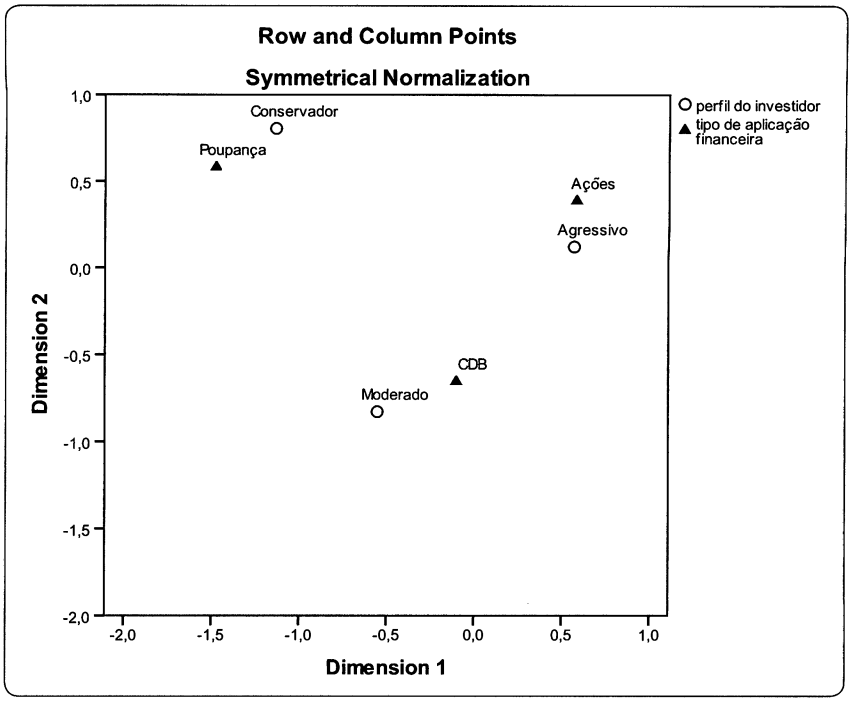



**Figura 11.33** Mapa perceptual para perfil do investidor e tipo de aplicação financeira.

Como são calculadas duas inércias principais parciais e, na sequência, é construído um mapa perceptual com duas dimensões (*biplot*), é importante enfatizar que 100% da inércia principal total estão representados no mapa bidimensional. Esse fato não ocorre para os casos em que há uma quantidade maior de categorias em ambas as variáveis e, na sequência, o pesquisador constrói um mapa perceptual bidimensional. Nessa situação, apenas as dimensões com as duas maiores inércias principais parciais serão plotadas no mapa.

## 11.4.2. Elaboração da análise de correspondência múltipla no software SPSS

Seguindo a lógica apresentada na seção 11.3.2, vamos elaborar a análise de correspondência múltipla no SPSS. Os dados encontram-se no arquivo **Perfil_Investidor × Aplicação × Estado_Civil.sav** e são exatamente iguais aos apresentados parcialmente na Tabela 11.21. Note que os rótulos das categorias das variáveis *perfil*, *aplicação* e *estado_civil* já estão definidos no banco de dados.

Inicialmente, é recomendável que sejam geradas as tabelas de frequências absolutas observadas (*cross-tabulations*) e os valores da estatística $\chi^2$ para cada par de variáveis, a fim de que seja elaborado um primeiro diagnóstico sobre a existência de associação entre elas e, consequentemente, sobre a eventual necessidade de que alguma precise ser eliminada da análise. Conforme procedimento adotado na seção 11.4.1, para essa análise preliminar, devemos clicar em **Analyze → Descriptive Statistics → Crosstabs...**. Como sabemos que existe associação entre as variáveis *perfil* e *aplicação*, vamos apresentar os resultados gerados para o par *perfil – estado_civil* e para o par *aplicação – estado_civil*. Esses *outputs* encontram-se nas Figuras 11.34 a 11.37.

**perfil do investidor * estado civil Crosstabulation**

Count

| | | estado civil | | Total |
|---|---|---|---|---|
| | | Solteiro | Casado | |
| perfil do investidor | Conservador | 5 | 12 | 17 |
| | Moderado | 11 | 14 | 25 |
| | Agressivo | 41 | 17 | 58 |
| Total | | 57 | 43 | 100 |

**Figura 11.34** Tabela de contingência com frequências absolutas observadas para *perfil* e *estado_civil*.

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 11,438[a] | 2 | ,003 |
| Likelihood Ratio | 11,600 | 2 | ,003 |
| Linear-by-Linear Association | 11,073 | 1 | ,001 |
| N of Valid Cases | 100 | | |

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 7,31.

**Figura 11.35** Resultado do teste $\chi^2$ para verificação de associação entre *perfil* e *estado_civil*.

**tipo de aplicação financeira * estado civil Crosstabulation**

Count

| | | estado civil | | Total |
|---|---|---|---|---|
| | | Solteiro | Casado | |
| tipo de aplicação financeira | Poupança | 5 | 10 | 15 |
| | CDB | 16 | 24 | 40 |
| | Ações | 36 | 9 | 45 |
| Total | | 57 | 43 | 100 |

**Figura 11.36** Tabela de contingência com frequências absolutas observadas para *aplicação* e *estado_civil*.

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 17,857[a] | 2 | ,000 |
| Likelihood Ratio | 18,690 | 2 | ,000 |
| Linear-by-Linear Association | 15,302 | 1 | ,000 |
| N of Valid Cases | 100 | | |

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 6,45.

**Figura 11.37** Resultado do teste $\chi^2$ para verificação de associação entre *aplicação* e *estado_civil*.

Com base nos *outputs* das Figuras 11.35 e 11.37, podemos afirmar que a variável *estado_civil* apresenta associação estatisticamente significante, ao nível de significância de 5%, com as variáveis *perfil* e *aplicação*, o que dá suporte à sua inclusão na análise de correspondência. Conforme discutimos no início da seção 11.3, se a variável *estado_civil* não apresentasse associação às demais, não faria sentido sua inclusão na análise, que voltaria a ser, nesse caso, bivariada.

Vamos, portanto, partir para a elaboração da análise de correspondência múltipla propriamente dita. Para tanto, devemos clicar em **Analyze → Dimension Reduction → Optimal Scaling...**. Uma caixa de diálogo como a apresentada na Figura 11.38 será aberta e devemos manter as opções selecionadas inicialmente, ou seja, **All variables are multiple nominal** em **Optimal Scaling Level** e **One set** em **Number of Sets of Variables**. Note que a análise escolhida é a **Multiple Corrrespondence Analysis**.

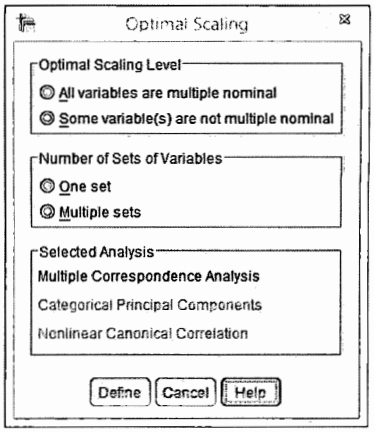



**Figura 11.38** Caixa de diálogo para seleção da análise de correspondência múltipla no SPSS.

Ao clicarmos em **Define**, será aberta uma caixa de diálogo como a apresentada na Figura 11.39.

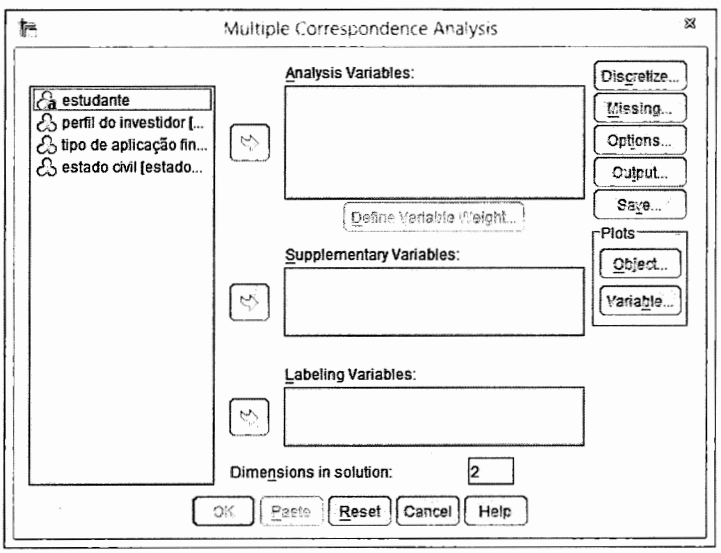



**Figura 11.39** Caixa de diálogo para elaboração da análise de correspondência múltipla no SPSS.

Primeiramente, devemos selecionar as três variáveis e inseri-las em **Analysis Variables**, conforme mostra a Figura 11.40.

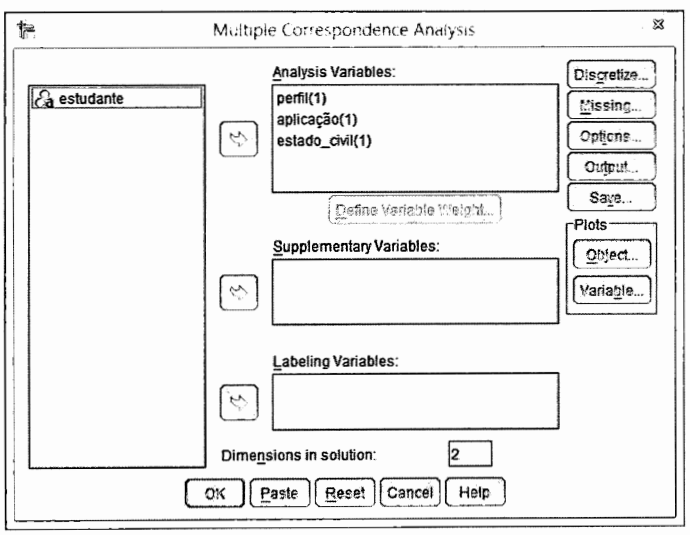



**Figura 11.40** Seleção das variáveis a serem incluídas na análise de correspondência múltipla.

Na sequência, ao clicarmos em **Output...**, será aberta uma caixa de diálogo como a da Figura 11.41. Nessa caixa, a fim de que sejam apresentadas as coordenadas de cada uma das categorias, devemos selecionar as três variáveis e inseri-las em **Category Quantifications and Contributions**. Em seguida, podemos clicar em **Continue**, a fim de retornarmos à caixa de diálogo principal.

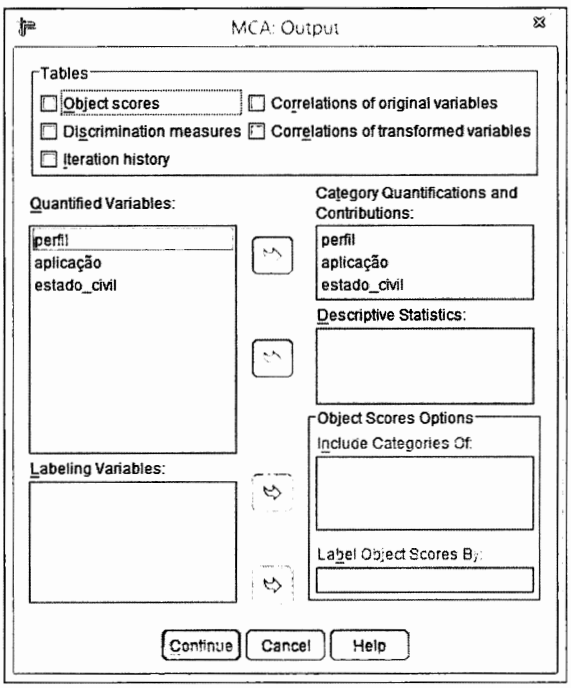



**Figura 11.41** Caixa de diálogo para geração das coordenadas das categorias nos *outputs*.

No botão **Save...**, devemos apenas selecionar a opção **Save object scores to the active dataset** em **Object Scores**, conforme mostra a Figura 11.42. Esse procedimento gerará as coordenadas para cada uma das observações da amostra no próprio banco de dados, conforme discutiremos adiante. Na sequência, podemos clicar em **Continue**.

Na caixa de diálogo principal, podemos agora clicar em **Object...**. Na caixa que será aberta, devemos selecionar as opções **Object points** e **Objects and centroids (biplot)** em **Plots**. Além disso, também devemos selecionar a opção **Variable** em **Label Objects** e incluir todas as variáveis em **Selected**, conforme mostra a Figura 11.43. Na sequência, podemos clicar em **Continue**.

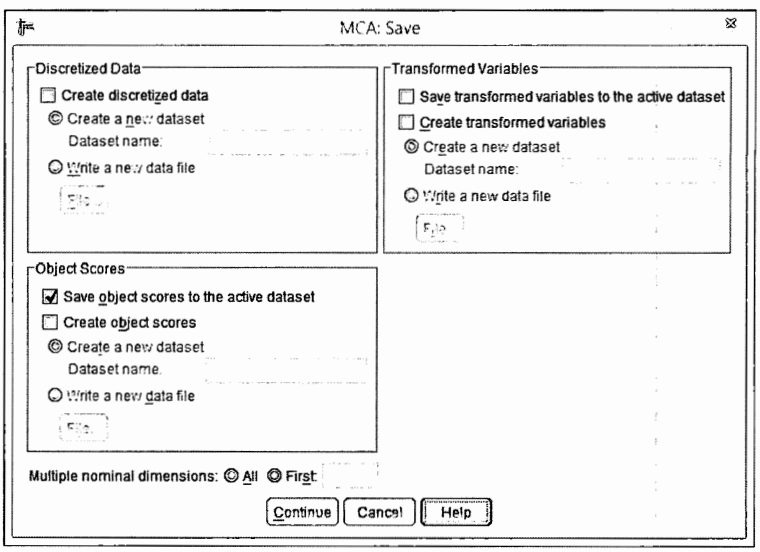



**Figura 11.42** Caixa de diálogo para geração das coordenadas das observações no banco de dados.

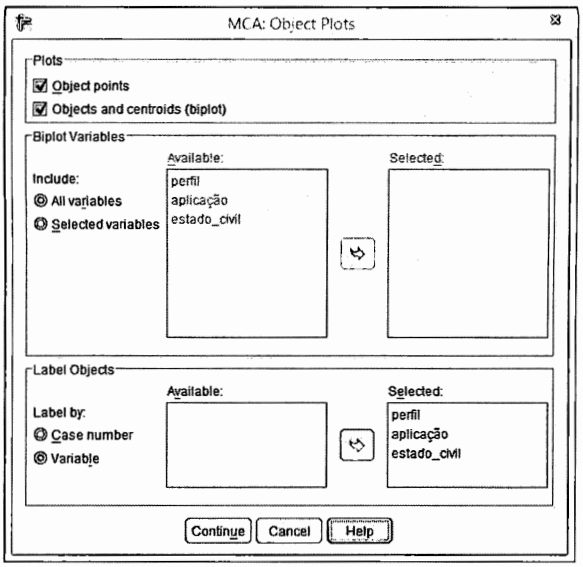



**Figura 11.43** Seleção das opções para elaboração dos gráficos.

Por fim, em **Variable...**, devemos selecionar as três variáveis e inseri-las em **Joint Category Plots**, conforme mostra a Figura 11.44. Esse procedimento gera nos *outputs* o mapa perceptual completo com as coordenadas de todas as categorias envolvidas na análise.

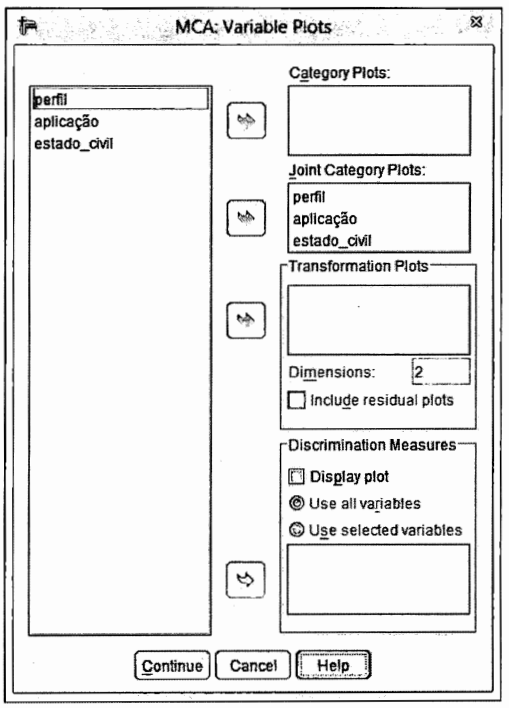



**Figura 11.44** Caixa de diálogo para elaboração do mapa perceptual com as coordenadas das categorias.

Na sequência, podemos clicar em **Continue** e em **OK**.

O primeiro *output* relevante encontra-se na Figura 11.45, em que são apresentados os valores das inércias principais parciais das duas primeiras dimensões, cujos valores são iguais aos apresentados na seção 11.3.2, ou seja:

$$\begin{cases} \lambda_1^2 = 0{,}602 \\ \lambda_2^2 = 0{,}436 \end{cases}$$

**Model Summary**

| Dimension | Cronbach's Alpha | Variance Accounted For | | |
|---|---|---|---|---|
| | | Total (Eigenvalue) | Inertia | % of Variance |
| 1 | ,670 | 1,807 | ,602 | 60,230 |
| 2 | ,353 | 1,308 | ,436 | 43,598 |
| Total | | 3,115 | 1,038 | |
| Mean | ,537[a] | 1,557 | ,519 | 51,914 |

a. Mean Cronbach's Alpha is based on the mean Eigenvalue.

**Figura 11.45** Inércias principais parciais.

É importante frisarmos que os procedimentos adotados para a elaboração da análise de correspondência no SPSS geram coordenadas principais das categorias das variáveis. As Figuras 11.46, 11.47 e 11.48 apresentam as coordenadas de cada categoria, por variável.

**perfil do investidor**

Points:Coordinates

| Category | Frequency | Centroid Coordinates | |
|---|---|---|---|
| | | Dimension | |
| | | 1 | 2 |
| Conservador | 17 | 1,130 | -1,481 |
| Moderado | 25 | ,747 | ,970 |
| Agressivo | 58 | -,653 | ,016 |

Variable Principal Normalization.

**Figura 11.46** Coordenadas principais – Variável *perfil*.

**tipo de aplicação financeira**

Points:Coordinates

| Category | Frequency | Centroid Coordinates | |
|---|---|---|---|
| | | Dimension | |
| | | 1 | 2 |
| Poupança | 15 | 1,382 | -1,335 |
| CDB | 40 | ,417 | ,937 |
| Ações | 45 | -,831 | -,388 |

Variable Principal Normalization.

**Figura 11.47** Coordenadas principais – Variável *aplicação*.

**estado civil**

Points:Coordinates

| Category | Frequency | Centroid Coordinates | |
|---|---|---|---|
| | | Dimension | |
| | | 1 | 2 |
| Solteiro | 57 | -,636 | -,101 |
| Casado | 43 | ,843 | ,134 |

Variable Principal Normalization.

**Figura 11.48** Coordenadas principais – Variável *estado_civil*.

Conforme discutimos na seção 11.3, as coordenadas principais geradas na análise de correspondência múltipla apresentam escala reduzida se comparadas às coordenadas-padrão, o que colabora para a construção de um mapa perceptual com pontos mais concentrados em torno da Origem. Além disso, podemos também perceber, a partir dos *outputs* apresentados nas Figuras 11.46, 11.47 e 11.48, que o SPSS apresenta as coordenadas das ordenadas de cada categoria (*Centroid Coordinates Dimension* 2) com sinais invertidos em relação aos calculados algebricamente no final da seção 11.3.2 e apresentados na Tabela 11.28 (esse fato ocorre apenas para algumas versões do SPSS). Isso, entretanto, não altera absolutamente a interpretação dos resultados da análise de correspondência. Com base nessas coordenadas principais, pode ser construído o mapa perceptual, apresentado na Figura 11.49.

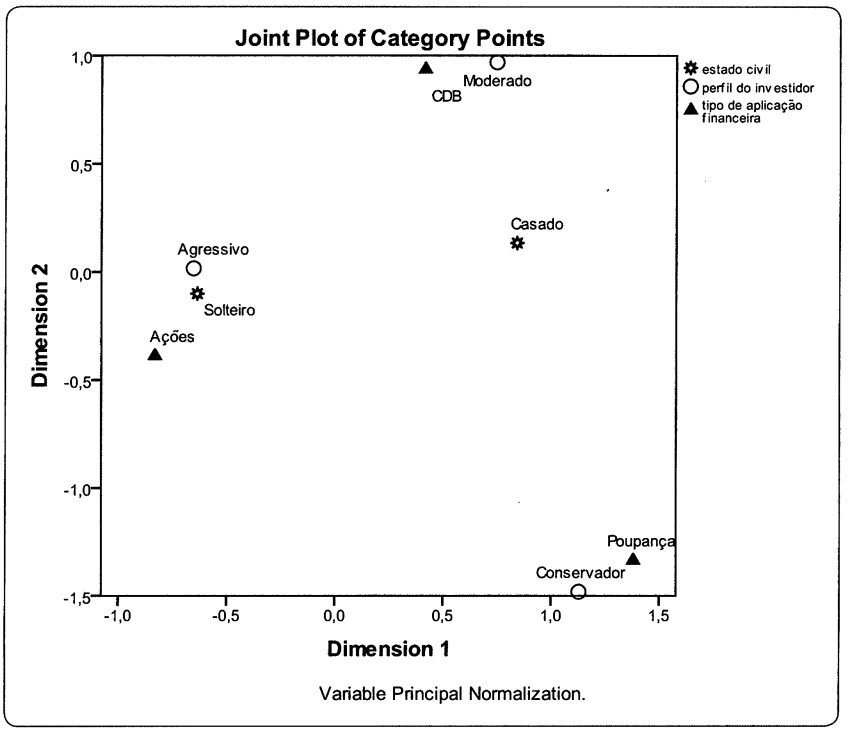



**Figura 11.49** Mapa perceptual para perfil do investidor, tipo de aplicação financeira e estado civil.

Com base no mapa perceptual da Figura 11.49, podemos verificar que a categoria *Solteiro* apresenta forte associação com as categorias *Agressivo* e *Ações*. Por outro lado, a categoria *Casado* encontra-se entre *Conservador* e *Moderado* e entre *Poupança* e *CDB*, porém com maior proximidade de *Moderado* e *CDB*.

Para fins didáticos, caso o pesquisador queira reproduzir os achados do exemplo desta seção por meio da elaboração de uma análise de correspondência simples (inércias, coordenadas principais e mapa perceptual), poderá fazer uso do arquivo **Burt.sav**, que mostra os dados oriundos da matriz de Burt, apresentada na Tabela 11.27 da seção 11.3.2. Nesse caso, o pesquisador irá perceber que os valores singulares de cada dimensão serão iguais aos valores das inércias principais parciais geradas por meio da análise de correspondência múltipla para as respectivas dimensões.

Por fim, podemos verificar, ao elaborarmos o procedimento descrito, que são geradas duas novas variáveis no banco de dados, chamadas pelo SPSS de *OBSCO1_1* e *OBSCO2_1*, conforme mostra a Figura 11.50 para as 20 primeiras observações. Essas variáveis referem-se às coordenadas da primeira e da segunda dimensões para cada uma das observações do banco de dados (*object scores*).

A partir das coordenadas de cada observação, é possível elaborar um gráfico, que se encontra na Figura 11.51, com as posições relativas dos estudantes e por meio do qual podemos estudar as similaridades entre eles com base no comportamento das variáveis *perfil*, *aplicação* e *estado_civil*. Ao contrário do que poderia ser feito a partir de um procedimento errado de ponderação arbitrária das categorias das variáveis originais, essas similaridades podem, de fato, ser avaliadas fazendo-se uso das coordenadas (*object scores*) de cada observação, visto que são variáveis métricas e, portanto, quantitativas. Note, inclusive, que essas novas variáveis (*OBSCO1_1* e *OBSCO2_1*) são ortogonais, isto

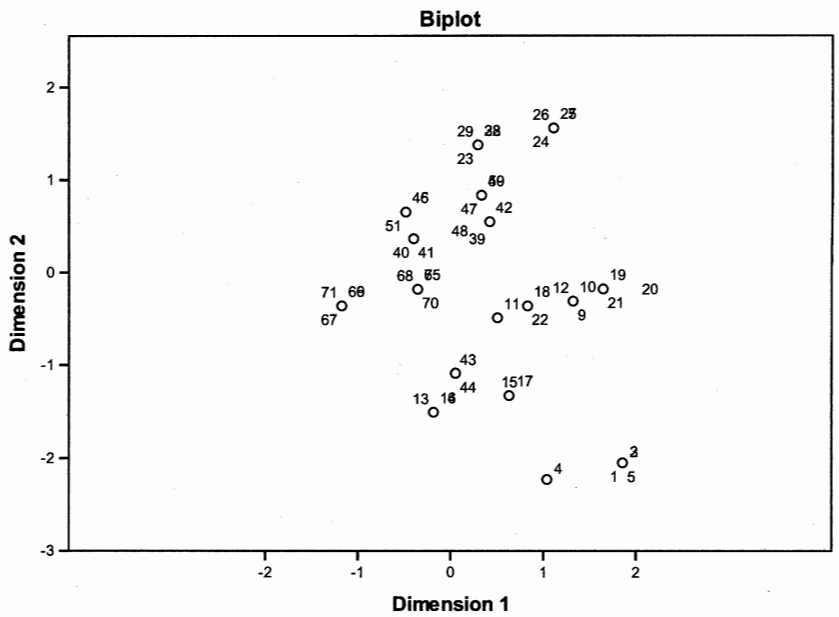
é, apresentam correlação de Pearson igual a 0, em conformidade com a ortogonalidade dos eixos do gráfico. Neste momento, é suscitada uma analogia com os fatores gerados a partir da elaboração de uma análise fatorial por componentes principais, estudada no capítulo anterior, que também podem ser ortogonais para determinados métodos de rotação.


49 :

| | estudante | perfil | aplicação | estado_civil | OBSCO1_1 | OBSCO2_1 |
|---|---|---|---|---|---|---|
| 1 | Gabriela | Conservador | Poupança | Casado | 1,86 | -2,05 |
| 2 | Luiz Felipe | Conservador | Poupança | Casado | 1,86 | -2,05 |
| 3 | Patrícia | Conservador | Poupança | Casado | 1,86 | -2,05 |
| 4 | Gustavo | Conservador | Poupança | Solteiro | 1,04 | -2,23 |
| 5 | Letícia | Conservador | Poupança | Casado | 1,86 | -2,05 |
| 6 | Ovídio | Conservador | Poupança | Casado | 1,86 | -2,05 |
| 7 | Leonor | Conservador | Poupança | Casado | 1,86 | -2,05 |
| 8 | Dalila | Conservador | Poupança | Casado | 1,86 | -2,05 |
| 9 | Antônio | Conservador | CDB | Casado | 1,32 | -,31 |
| 10 | Júlia | Conservador | CDB | Casado | 1,32 | -,31 |
| 11 | Roberto | Conservador | CDB | Solteiro | ,50 | -,49 |
| 12 | Renata | Conservador | CDB | Casado | 1,32 | -,31 |
| 13 | Guilherme | Conservador | Ações | Solteiro | -,19 | -1,51 |
| 14 | Rodrigo | Conservador | Ações | Solteiro | -,19 | -1,51 |
| 15 | Giulia | Conservador | Ações | Casado | ,63 | -1,33 |
| 16 | Felipe | Conservador | Ações | Solteiro | -,19 | -1,51 |
| 17 | Karina | Conservador | Ações | Casado | ,63 | -1,33 |
| 18 | Pietro | Moderado | Poupança | Solteiro | ,83 | -,36 |
| 19 | Cecília | Moderado | Poupança | Casado | 1,65 | -,18 |
| 20 | Gisele | Moderado | Poupança | Casado | 1,65 | -,18 |

**Figura 11.50** Banco de dados com as coordenadas das observações (*object scores*).

Essa é uma das principais contribuições da análise de correspondência múltipla, uma vez que, a partir dessas coordenadas, pode-se, por exemplo, elaborar uma análise de agrupamentos. A própria inclusão das coordenadas como variáveis explicativas em técnicas confirmatórias, como análise de regressão, pode fazer algum sentido para efeitos de diagnóstico sobre o comportamento de determinado fenômeno em estudo, dependendo dos interesses e dos objetivos do pesquisador.




Variable Principal Normalization.

**Figura 11.51** Posições relativas das observações da amostra.

Como as variáveis *perfil* e *aplicação* possuem três categorias, e a variável *estado_civil*, duas categorias, existem 18 possibilidades de combinação para cada uma das observações da amostra (3 × 3 × 2 = 18), sendo que, dessas, 17 combinações ocorrem em nosso exemplo, uma vez que não há qualquer estudante que apresente perfil agressivo, aplique seus recursos financeiros em poupança e seja casado. Note, no gráfico da Figura 11.51, que realmente 17 pontos são plotados, e a maioria deles representa o comportamento de mais de um estudante.

Além disso, o pesquisador também pode desejar estudar as posições relativas dos estudantes com base na explicitação, no gráfico, das categorias de cada uma das variáveis, em vez da identificação de cada observação. Os gráficos das Figuras 11.52, 11.53 e 11.54 explicitam, para cada um dos 17 pontos, as categorias das variáveis *perfil*, *aplicação* e *estado_civil*, respectivamente.

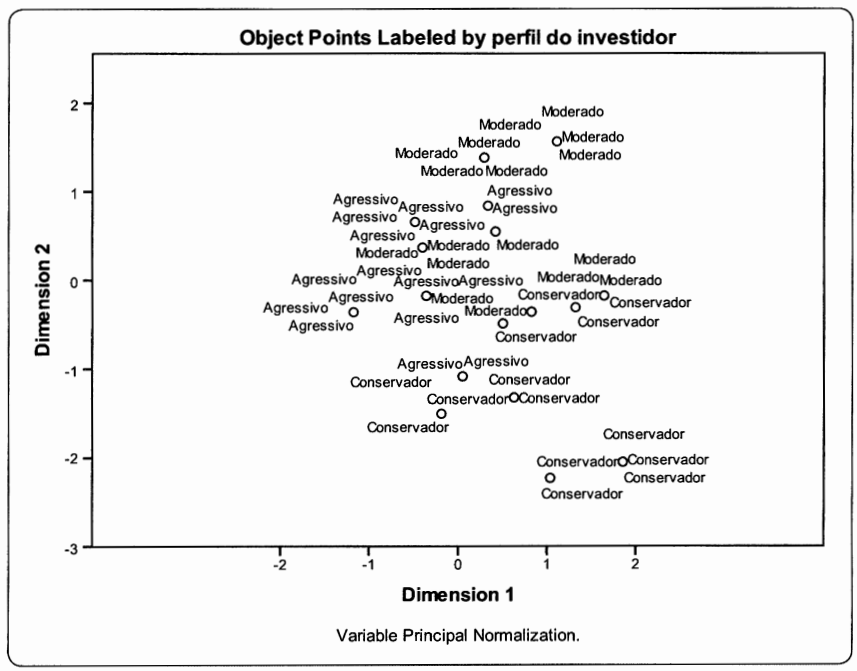



**Figura 11.52** Posições relativas das observações da amostra – Categorias da variável *perfil*.

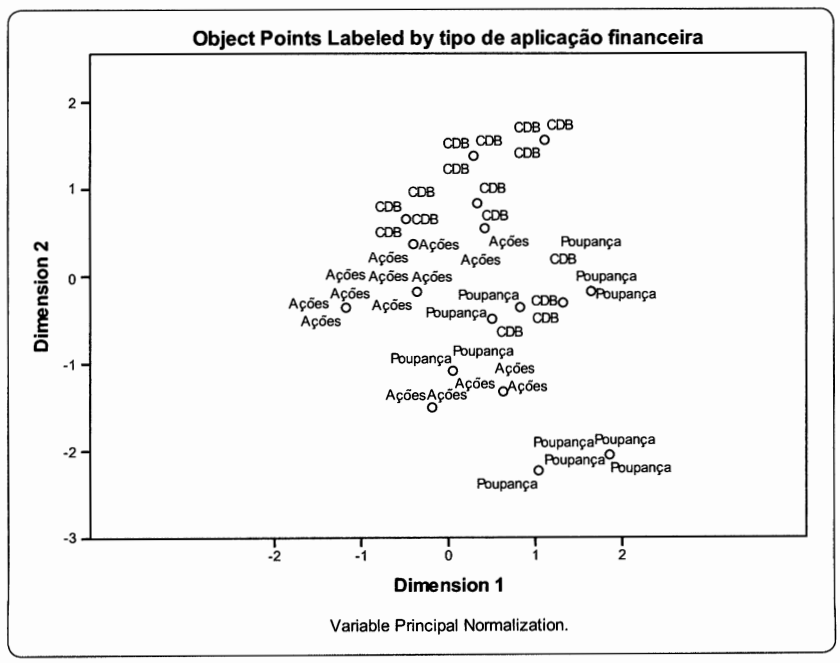



**Figura 11.53** Posições relativas das observações da amostra – Categorias da variável *aplicação*.

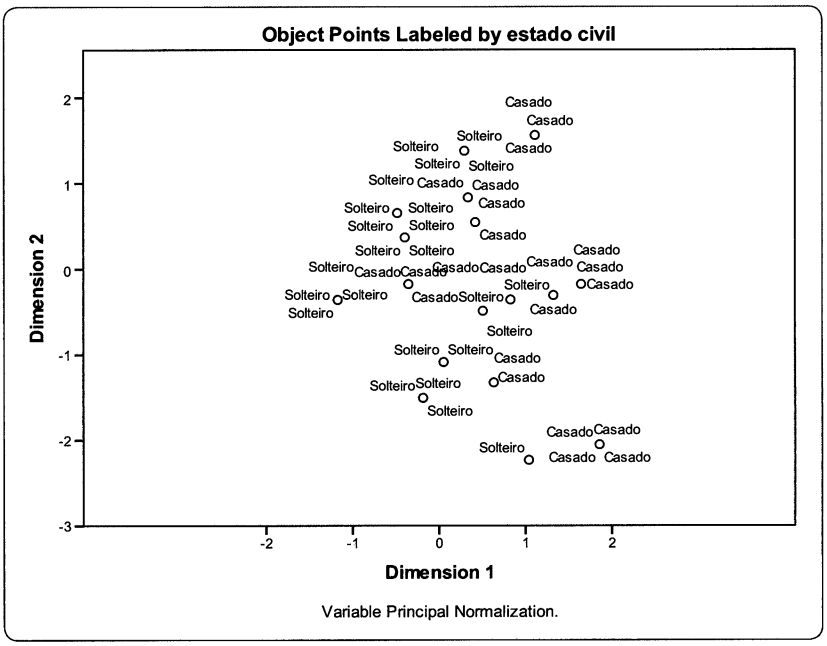



**Figura 11.54** Posições relativas das observações da amostra – Categorias da variável *estado_civil*.

Note que há certa separação entre as categorias das variáveis nos gráficos das Figuras 11.52, 11.53 e 11.54, principalmente para coordenadas mais afastadas da Origem, o que reforça ainda mais a existência de associação entre o perfil do investidor, o tipo de aplicação financeira e seu estado civil.

Apresentados os procedimentos para aplicação da análise de correspondência simples e da análise de correspondência múltipla no SPSS, partiremos para a elaboração das técnicas no Stata.

## 11.5. ANÁLISE DE CORRESPONDÊNCIA SIMPLES E MÚLTIPLA NO SOFTWARE STATA

Apresentaremos agora o passo a passo para a elaboração dos nossos exemplos no Stata Statistical Software®. Nosso objetivo, nesta seção, não é discutir novamente os conceitos pertinentes à análise de correspondência, porém propiciar ao pesquisador uma oportunidade de elaborar as técnicas por meio dos comandos desse software. A cada apresentação de um *output*, faremos menção ao respectivo resultado obtido quando da elaboração da técnica de forma algébrica e também por meio do SPSS. A reprodução das imagens apresentadas nesta seção tem autorização da StataCorp LP©.

### 11.5.1. Elaboração da análise de correspondência simples no software Stata

Seguindo, portanto, a mesma lógica proposta quando da elaboração da técnica de análise de correspondência simples no software SPSS, já partiremos para o banco de dados construído pelo professor a partir dos questionamentos feitos a cada um de seus 100 estudantes. O banco de dados encontra-se no arquivo **Perfil_Investidor × Aplicação.dta** e é exatamente igual ao apresentado parcialmente na Tabela 11.7 da seção 11.2.5. Note que os rótulos das categorias das variáveis *perfil* e *aplicação* já estão definidos no banco de dados.

Inicialmente, podemos digitar o comando **desc**, que possibilita analisarmos as características do banco de dados, como a quantidade de observações, a quantidade de variáveis e a descrição de cada uma delas. A Figura 11.55 apresenta esse primeiro *output* do Stata.

```
. desc

  obs:           100
 vars:             3
 size:         1,700 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
estudante       str11  %11s
perfil          byte   %11.0g      perfil     perfil do investidor
aplicação       byte   %14.0g      aplicação  tipo de aplicação financeira
-------------------------------------------------------------------------------
Sorted by:
```

**Figura 11.55** Descrição do banco de dados **Perfil_Investidor × Aplicação.dta**.

O comando **tab2** permite gerar a tabela de contingência correspondente ao cruzamento das categorias de duas variáveis. Ao digitarmos o seguinte comando, poderemos analisar a distribuição das frequências absolutas observadas por categoria, bem como avaliar a significância estatística da associação entre as duas variáveis (termo **chi2**).

```
tab2 perfil aplicação, chi2
```

A Figura 11.56 apresenta o *output* gerado.

```
. tab perfil aplicação, chi2

  perfil do |   tipo de aplicação financeira
 investidor |  Poupança        CDB       Ações |     Total
------------+---------------------------------+----------
Conservador |         8          4           5 |        17
   Moderado |         5         16           4 |        25
  Agressivo |         2         20          36 |        58
------------+---------------------------------+----------
      Total |        15         40          45 |       100

          Pearson chi2(4) =  31.7642   Pr = 0.000
```

**Figura 11.56** Tabela de contingência com frequências absolutas observadas e teste $\chi^2$.

A partir do resultado do teste $\chi^2$, podemos afirmar, para o nível de significância de 5% e para 4 graus de liberdade, que existe associação estatisticamente significante entre as variáveis *perfil* e *aplicação*, visto que $\chi^2_{cal} = 31{,}76$ ($\chi^2$ calculado para 4 graus de liberdade) e *Prob.* $\chi^2_{cal} < 0{,}05$. Dado que a associação entre as duas variáveis não se dá de forma aleatória, podemos, por meio da análise dos resíduos padronizados ajustados, estudar a relação de dependência entre cada par de categorias. No Stata, o comando **tab2** não permite gerar esses resíduos nos *outputs*, porém o comando **tabchi**, desenvolvido a partir de um módulo de tabulação criado por Nicholas J. Cox, faz os resíduos padronizados ajustados serem calculados. Para que esse comando seja utilizado, devemos inicialmente digitar:

```
findit tabchi
```

e instalá-lo no link **tab chi from http://fmwww.bc.edu/RePEc/bocode/t**. Feito isso, podemos digitar o seguinte comando:

```
tabchi perfil aplicação, a
```

Os *outputs* encontram-se na Figura 11.57, que mostra, além do apresentado na Figura 11.56, as frequências absolutas esperadas e os resíduos padronizados ajustados por célula, em conformidade com o apresentado nas Tabelas 11.9 e 11.13 da seção 11.2.5, e também na Figura 11.18 quando da elaboração da técnica no SPSS (seção 11.4.1).

```
. tabchi perfil aplicação, a

          observed frequency
          expected frequency
          adjusted residual

-----------------------------------------
perfil do   | tipo de aplicação financeira
investidor  | Poupança       CDB      Ações
------------+----------------------------
Conservador |        8         4          5
            |    2.550     6.800      7.650
            |    4.063    -1.522     -1.418
            |
   Moderado |        5        16          4
            |    3.750    10.000     11.250
            |    0.808     2.828     -3.366
            |
  Agressivo |        2        20         36
            |    8.700    23.200     26.100
            |   -3.802    -1.323      4.032
-----------------------------------------

2 cells with expected frequency < 5

          Pearson chi2(4) =  31.7642   Pr = 0.000
 likelihood-ratio chi2(4) =  30.7767   Pr = 0.000
```

**Figura 11.57** Tabela de frequências e de resíduos padronizados ajustados para *perfil* e *aplicação*.

Assim como discutido anteriormente, podemos verificar que há dependência entre as categorias *Conservador* e *Poupança*, entre *Moderado* e *CDB* e entre *Agressivo* e *Ações*, uma vez que os resíduos padronizados das células correspondentes são, respectivamente, iguais a 4,063, 2,828 e 4,032 (positivos e maiores que 1,96).

Verificada a existência de associação estatisticamente significante entre as variáveis *perfil* e *aplicação* e identificadas as relações de dependência entre suas categorias, podemos digitar o comando da análise de correspondência simples, que faz com que sejam calculadas as coordenadas de cada categoria a partir das quais pode ser construído o mapa perceptual no Stata. O comando é:

```
ca perfil aplicação
```

Os *outputs* gerados encontram-se na Figura 11.58.

```
. ca perfil aplicação

Correspondence analysis                         Number of obs      =       100
                                                Pearson chi2(4)    =     31.76
                                                Prob > chi2        =    0.0000
                                                Total inertia      =    0.3176
    3 active rows                               Number of dim.     =         2
    3 active columns                            Expl. inertia (%)  =    100.00

                |   singular    principal                              cumul
  Dimension     |    value       inertia          chi2      percent    percent
----------------+-----------------------------------------------------------
      dim 1     |   .4829233     .2332149         23.32       73.42      73.42
      dim 2     |   .2905629     .0844268          8.44       26.58     100.00
----------------+-----------------------------------------------------------
      total     |                .3176416         31.76         100

Statistics for row and column categories in symmetric normalization

                |          overall                |        dimension_1              |        dimension_2
   Categories   |    mass  quality   %inert       |   coord   sqcorr   contrib      |   coord   sqcorr   contrib
----------------+---------------------------------+---------------------------------+---------------------------
perfil          |                                 |                                 |
  Conservador   |   0.170    1.000     0.432      |   1.132    0.767     0.451      |   0.805    0.233     0.379
     Moderado   |   0.250    1.000     0.274      |   0.553    0.425     0.158      |  -0.829    0.575     0.592
    Agressivo   |   0.580    1.000     0.295      |  -0.570    0.973     0.391      |   0.122    0.027     0.029
----------------+---------------------------------+---------------------------------+---------------------------
aplicação       |                                 |                                 |
     Poupança   |   0.150    1.000     0.542      |   1.475    0.914     0.675      |   0.582    0.086     0.175
          CDB   |   0.400    1.000     0.164      |   0.102    0.039     0.009      |  -0.655    0.961     0.591
        Ações   |   0.450    1.000     0.294      |  -0.582    0.789     0.316      |   0.389    0.211     0.234
----------------+---------------------------------+---------------------------------+---------------------------
```

**Figura 11.58** *Outputs* da análise de correspondência simples no Stata.

Note, com base na análise dos *outputs* da Figura 11.58, que as inércias principais parciais correspondem às calculadas algebricamente na seção 11.2.5 e também apresentadas na Figura 11.30 da seção 11.4.1 e, por meio delas, é possível afirmar que as dimensões 1 e 2 explicam, respectivamente, 73,42% (0,2332 / 0,3176) e 26,58% (0,0844 / 0,3176) da inércia principal total. Além disso, as coordenadas (**dimension_1 coord** e **dimension_2 coord**) também correspondem às calculadas algebricamente, bem como às apresentadas pelo SPSS, conforme discutido na seção 11.4.1.

Ainda com base nos *outputs* da Figura 11.58, é possível afirmar, para a variável *perfil*, que, enquanto a categoria *Conservador* é a mais representativa para a composição inercial da primeira dimensão (**dimension_1 contrib** = 45,1%), a categoria *Moderado* é a mais representativa para a composição inercial da segunda dimensão (**dimension_2 contrib** = 59,2%). Já para a variável *aplicação*, enquanto a categoria *Poupança* é a mais representativa para a composição inercial da primeira dimensão (**dimension_1 contrib** = 67,5%), a categoria *CDB* é a mais representativa para a composição inercial da segunda dimensão (**dimension_2 contrib** = 59,1%).

Um primeiro gráfico pode ser construído a partir das coordenadas apresentadas na Figura 11.58 e é conhecido por **gráfico de projeção das coordenadas nas dimensões**, pois permite analisar isoladamente o comportamento de cada categoria em cada dimensão. Para elaborarmos esse gráfico, que se encontra na Figura 11.59, precisamos digitar o seguinte comando:

```
caprojection
```

que somente pode ser aplicado após a elaboração da Figura 11.58 (comando **ca**).

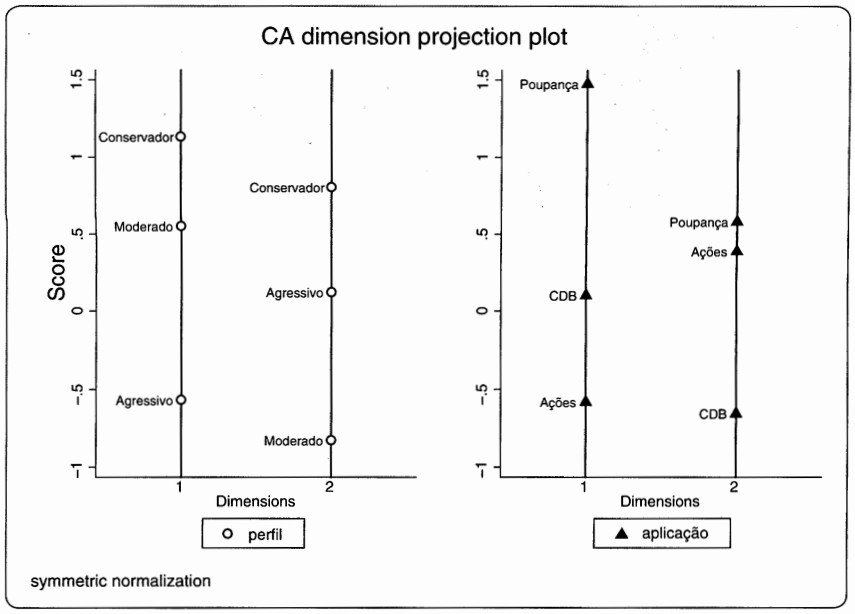



**Figura 11.59** Gráfico de projeção das coordenadas nas dimensões.

O gráfico de projeção das coordenadas nas dimensões pode ser bastante útil para estudar a lógica do sequenciamento das categorias, principalmente em variáveis qualitativas ordinais. Para os dados de nosso exemplo, podemos verificar que existe lógica na ordenação dos pontos referentes às categorias das variáveis para a primeira dimensão, com destaque para a variável *perfil*, de fato, ordinal. Além disso, também podemos observar que os pontos se encontram em lados opostos e relativamente afastados da Origem para o eixo da primeira dimensão, o que é adequado para a elaboração da análise de correspondência simples, pois permite melhor visualização do mapa perceptual.

O mapa perceptual propriamente dito pode ser construído a partir da digitação do seguinte comando:

```
cabiplot, origin
```

que também só pode ser aplicado após a elaboração da Figura 11.58 (comando **ca**). O mapa perceptual que mostra a relação entre as categorias de *perfil* e *aplicação* encontra-se na Figura 11.60.

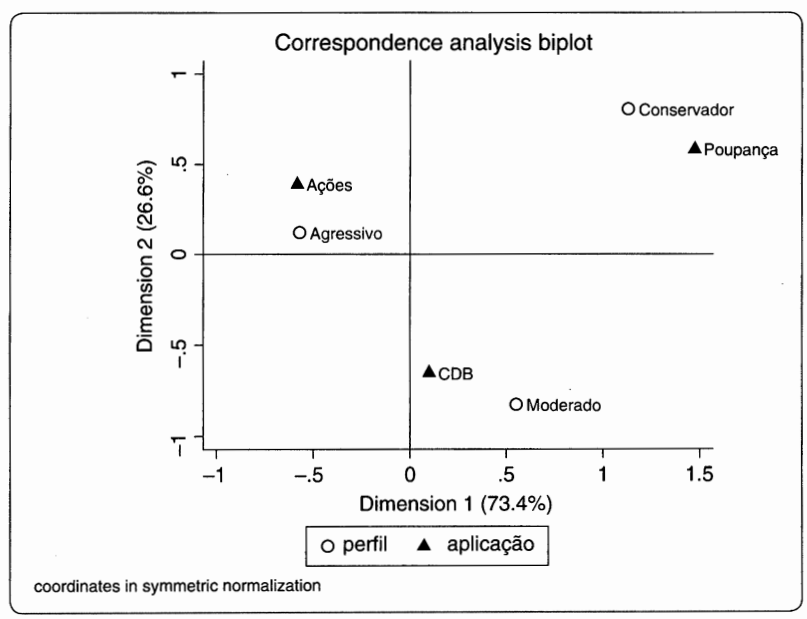



**Figura 11.60** Mapa perceptual para perfil do investidor e tipo de aplicação financeira.

Apresentados os comandos para a realização da análise de correspondência simples no Stata, partiremos para a elaboração da análise de correspondência múltipla no mesmo software.

### 11.5.2. Elaboração da análise de correspondência múltipla no software Stata

Seguindo a mesma lógica proposta quando da elaboração da técnica de análise de correspondência múltipla no SPSS, já partiremos para o banco de dados construído pelo professor a partir dos questionamentos feitos a cada um de seus 100 estudantes. O banco de dados encontra-se no arquivo **Perfil_Investidor × Aplicação × Estado_Civil.dta** e é exatamente igual ao apresentado parcialmente na Tabela 11.21 da seção 11.3.2. Note que os rótulos das categorias das variáveis *perfil*, *aplicação* e *estado_civil* já estão definidos no banco de dados.

O primeiro *output*, que se encontra na Figura 11.61, gerado a partir do comando **desc**, apresenta as características do banco de dados, como a quantidade de observações e a descrição de cada variável.

```
. desc

  obs:           100
 vars:             4
 size:         2,100 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
estudante       str11  %11s
perfil          byte   %11.0g      perfil     perfil do investidor
aplicação       byte   %14.0g      aplicação  tipo de aplicação financeira
estado_civil    float  %9.0g       est_civil  estado civil
-------------------------------------------------------------------------------
Sorted by:
```

**Figura 11.61** Descrição do banco de dados **Perfil_Investidor × Aplicação × Estado_Civil.dta**.

Conforme discutimos, a fim de que seja elaborado o diagnóstico sobre a existência de associação entre as variáveis e, consequentemente, sobre a eventual necessidade de que alguma delas precise ser eliminada da análise, devemos gerar as tabelas de frequências absolutas observadas para cada par de variáveis com os respectivos testes $\chi^2$. Para tanto, devemos digitar o seguinte comando:

```
tab2 perfil aplicação estado_civil, chi2
```

Os *outputs* encontram-se na Figura 11.62, por meio dos quais podemos verificar que todos os pares de variáveis apresentam associação estatisticamente significante, ao nível de significância de 5%. Para que determinada variável seja incluída em uma análise de correspondência múltipla, é preciso que se associe de maneira estatisticamente significante a pelo menos uma das demais variáveis.

```
. tab2 perfil aplicação estado_civil, chi2

-> tabulation of perfil by aplicação

  perfil do |   tipo de aplicação financeira
 investidor |  Poupança        CDB      Ações |     Total
------------+---------------------------------+----------
Conservador |         8          4          5 |        17
   Moderado |         5         16          4 |        25
  Agressivo |         2         20         36 |        58
------------+---------------------------------+----------
      Total |        15         40         45 |       100

          Pearson chi2(4) =  31.7642   Pr = 0.000

-> tabulation of perfil by estado_civil

  perfil do |     estado civil
 investidor |  Solteiro     Casado |     Total
------------+----------------------+----------
Conservador |         5         12 |        17
   Moderado |        11         14 |        25
  Agressivo |        41         17 |        58
------------+----------------------+----------
      Total |        57         43 |       100

          Pearson chi2(2) =  11.4376   Pr = 0.003

-> tabulation of aplicação by estado_civil

   tipo de |
 aplicação |     estado civil
financeira |  Solteiro     Casado |     Total
-----------+----------------------+----------
  Poupança |         5         10 |        15
       CDB |        16         24 |        40
     Ações |        36          9 |        45
-----------+----------------------+----------
     Total |        57         43 |       100

          Pearson chi2(2) =  17.8567   Pr = 0.000
```

**Figura 11.62** Tabelas de contingência com frequências absolutas observadas e testes $\chi^2$.

Visto que todas as variáveis devem ser incluídas na análise de correspondência múltipla, podemos partir para a elaboração propriamente dita da técnica, digitando o seguinte comando:

```
mca perfil aplicação estado_civil, method(indicator)
```

em que o termo **method(indicator)** corresponde ao método da matriz binária **Z**, discutido na seção 11.3, que gera coordenadas-padrão para cada uma das categorias das variáveis. Os *outputs* encontram-se na Figura 11.63.

```
. mca perfil aplicação estado_civil, method(indicator)

Multiple/Joint correspondence analysis          Number of obs      =       100
                                                Total inertia      =  1.666667
    Method: Indicator matrix                    Number of axes     =         2

             |   principal             cumul
   Dimension |    inertia     percent  percent
-------------+------------------------------
       dim 1 |     .6023045     36.14     36.14
       dim 2 |     .4359878     26.16     62.30
       dim 3 |     .2764728     16.59     78.89
       dim 4 |     .1798371     10.79     89.68
       dim 5 |     .1720645     10.32    100.00
-------------+------------------------------
       Total |    1.666667    100.00

Statistics for column categories in standard normalization

             |          overall                |        dimension_1              |        dimension_2
  Categories |    mass  quality   %inert       |    coord   sqcorr   contrib     |    coord   sqcorr   contrib
-------------+---------------------------------+---------------------------------+---------------------------
perfil       |                                 |                                 |
 Conservador |   0.057    0.712    0.166       |    1.456    0.262    0.093     |    2.247    0.451    0.189
    Moderado |   0.083    0.503    0.150       |    0.962    0.186    0.060     |   -1.476    0.317    0.120
   Agressivo |   0.193    0.589    0.084       |   -0.841    0.589    0.106     |   -0.022    0.000    0.000
-------------+---------------------------------+---------------------------------+---------------------------
aplicação    |                                 |                                 |
    Poupança |   0.050    0.649    0.170       |    1.780    0.337    0.123     |    2.016    0.313    0.134
         CDB |   0.133    0.699    0.120       |    0.538    0.116    0.030     |   -1.416    0.583    0.177
       Ações |   0.150    0.688    0.110       |   -1.071    0.565    0.134     |    0.587    0.123    0.034
-------------+---------------------------------+---------------------------------+---------------------------
estado_civil |                                 |                                 |
     Solteiro |   0.190    0.549    0.086       |   -0.820    0.536    0.099     |    0.150    0.013    0.003
      Casado |   0.143    0.549    0.114       |    1.086    0.536    0.131     |   -0.199    0.013    0.004
-------------------------------------------------------------------------------------------------------------
```

**Figura 11.63** *Outputs* da análise de correspondência múltipla no Stata – Coordenadas-padrão.

Note, com base nos *outputs* da Figura 11.63, que as coordenadas das categorias das variáveis *perfil*, *aplicação* e *estado_civil* para as duas dimensões (**dimension_1 coord** e **dimension_2 coord**) são exatamente iguais às calculadas algebricamente na seção 11.3.2 e apresentadas na Tabela 11.26 (coordenadas-padrão). Além disso, a inércia principal total da matriz binária **Z** é igual a:

$$I_T = \frac{J-Q}{Q} = \frac{8-3}{3} = 1{,}6667$$

em que $J$ representa o número de categorias de todas as variáveis envolvidas na análise ($J = 8$), e $Q$, o número de variáveis ($Q = 3$). Portanto, podem ser calculadas as inércias principais parciais das $J - Q = 8 - 3 = 5$ dimensões, cujos valores são:

$$\begin{cases} \lambda_1^2 = 0{,}6023 \\ \lambda_2^2 = 0{,}4360 \\ \lambda_3^2 = 0{,}2765 \\ \lambda_4^2 = 0{,}1798 \\ \lambda_5^2 = 0{,}1721 \end{cases}$$

de onde podemos comprovar que $I_T = \lambda_1^2 + \lambda_2^2 + \lambda_3^2 + \lambda_4^2 + \lambda_5^2 = 1{,}6667$, conforme também calculado algebricamente na seção 11.3.2.

Analogamente ao realizado na seção 11.5.1, podemos inicialmente construir, a partir das coordenadas-padrão apresentadas na Figura 11.63, o gráfico de projeção das coordenadas nas dimensões, que se encontra na Figura 11.64. Para tanto, devemos digitar o seguinte comando:

```
mcaprojection, normalize(standard)
```

Para os dados do nosso exemplo, podemos verificar, a partir do gráfico de projeção das coordenadas nas dimensões, que existe lógica na ordenação dos pontos referentes às categorias das variáveis para a primeira dimensão, com destaque para a variável *perfil*, de fato, ordinal. Além disso, também podemos observar que os pontos se encontram em lados opostos e relativamente afastados da Origem para o eixo da primeira dimensão, o que pode ser bastante adequado para melhor visualização do mapa perceptual da análise de correspondência múltipla.

Dando sequência à análise, caso o pesquisador queira obter a matriz binária **Z**, deve simplesmente digitar o comando a seguir:

```
xi i.perfil i.aplicação i.estado_civil, noomit
```

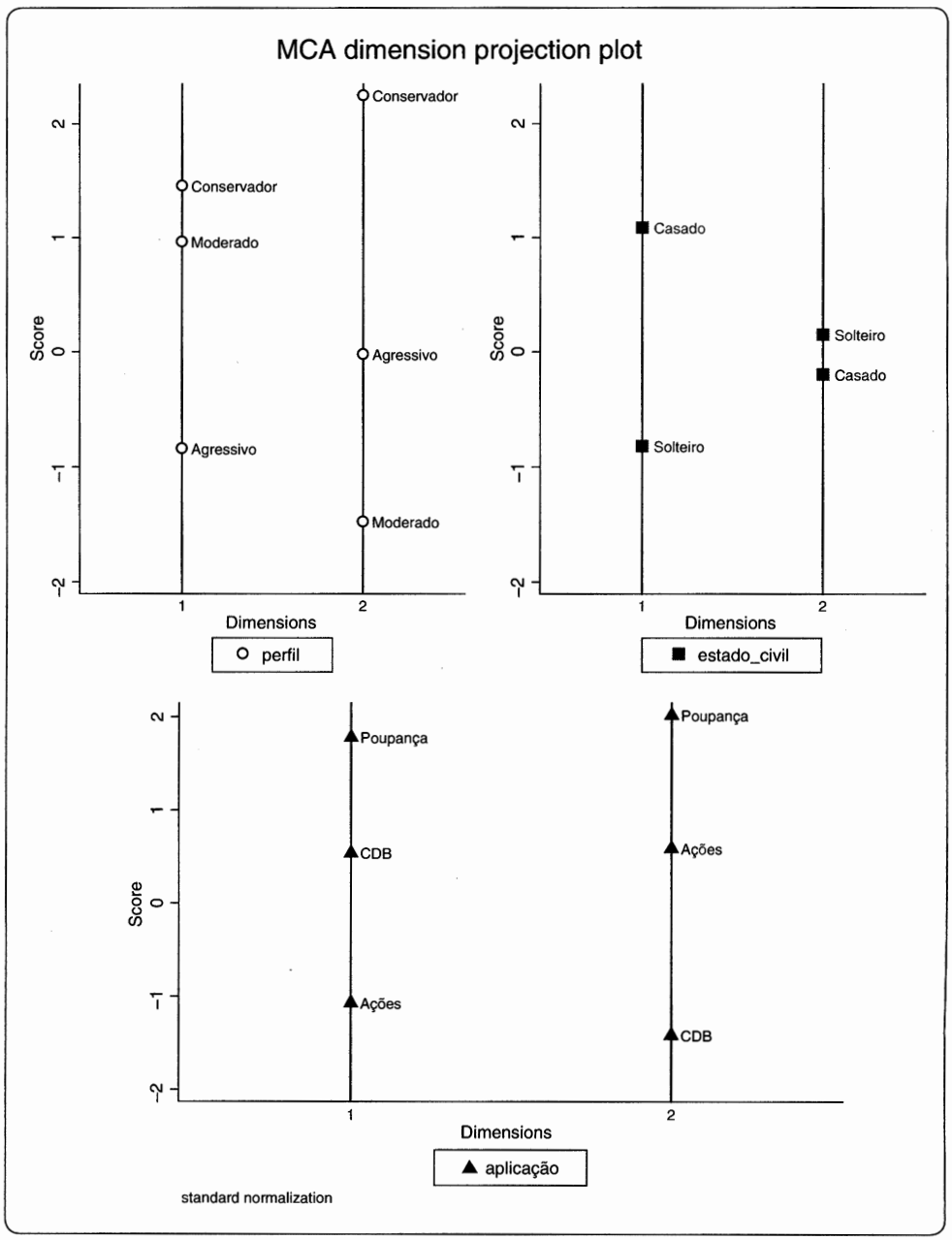



**Figura 11.64** Gráfico de projeção das coordenadas nas dimensões.

A Figura 11.65 mostra a matriz binária **Z** gerada no próprio banco de dados, para as 20 primeiras observações. É importante salientar que essa matriz pode ser utilizada para o cálculo das inércias principais parciais das cinco dimensões do nosso exemplo, desde que considerada uma tabela de contingência. Em outras palavras, para aplicar uma análise de correspondência simples e calcular as inércias principais parciais apresentadas na Figura 11.63, a matriz binária

**Z** deve ser transformada em um banco de dados bivariado, que deverá possuir 300 linhas. O arquivo **Matriz Binária Z.dta** contém o banco de dados correspondente à matriz binária **Z** do nosso exemplo, e, caso o pesquisador deseje aplicar a análise de correspondência simples às suas duas variáveis, para efeitos didáticos, irá verificar que serão geradas exatamente as mesmas cinco inércias principais parciais obtidas quando da elaboração da análise de correspondência múltipla no banco de dados original.

| | estudante | perfil | aplicação | estado_civil | _Iperfil_1 | _Iperfil_2 | _Iperfil_3 | _Iaplicaçã~1 | _Iaplicaçã~2 | _Iaplicaçã~3 | _Iestado_c~1 | _Iestado_c~2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Gabriela | Conservador | Poupança | Casado | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 2 | Luiz Felipe | Conservador | Poupança | Casado | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 3 | Patrícia | Conservador | Poupança | Casado | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 4 | Gustavo | Conservador | Poupança | Solteiro | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 5 | Letícia | Conservador | Poupança | Casado | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 6 | Ovídio | Conservador | Poupança | Casado | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 7 | Leonor | Conservador | Poupança | Casado | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 8 | Dalila | Conservador | Poupança | Casado | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 9 | Antônio | Conservador | CDB | Casado | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 10 | Júlia | Conservador | CDB | Casado | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 11 | Roberto | Conservador | CDB | Solteiro | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 12 | Renata | Conservador | CDB | Casado | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 13 | Guilherme | Conservador | Ações | Solteiro | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 14 | Rodrigo | Conservador | Ações | Solteiro | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 15 | Giulia | Conservador | Ações | Casado | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 16 | Felipe | Conservador | Ações | Solteiro | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 17 | Karina | Conservador | Ações | Casado | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 18 | Pietro | Moderado | Poupança | Solteiro | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| 19 | Cecília | Moderado | Poupança | Casado | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| 20 | Gisele | Moderado | Poupança | Casado | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |

matriz binária Z

**Figura 11.65** Banco de dados com a matriz binária **Z.**

A partir das coordenadas-padrão apresentadas na Figura 11.63, podemos construir o mapa perceptual propriamente dito, que se encontra na Figura 11.66, por meio da digitação do seguinte comando:

```
mcaplot, overlay origin dim(2 1)
```

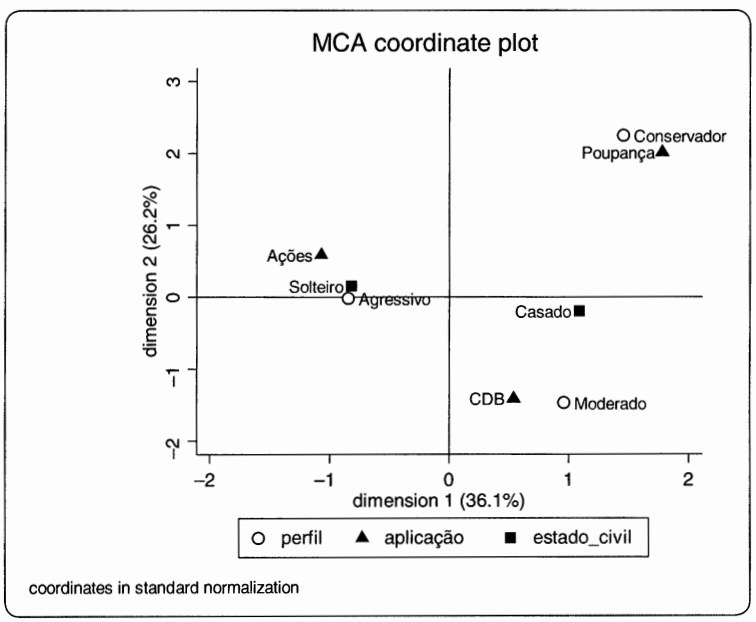



**Figura 11.66** Mapa perceptual para perfil do investidor, tipo de aplicação financeira e estado civil.

O mapa perceptual construído pelo Stata é o mesmo apresentado na Figura 11.12 da seção 11.3.2, porém possui uma escala menos reduzida se comparado àquele construído pelo SPSS, visto que, para o procedimento adotado na seção 11.4.2, o SPSS gera coordenadas principais para as categorias das variáveis. Conforme também discutido na seção 11.3.2, são somente plotadas no mapa perceptual as coordenadas-padrão das dimensões que apresentam inércias principais parciais superiores a 0,3333, valor da média da inércia principal total por dimensão (1,6667 / 5 = 0,3333). Portanto, como as inércias principais parciais das duas primeiras dimensões são iguais a

0,6023 e 0,4360, essas dimensões explicam, respectivamente, 36,1% e 26,2% da inércia principal total, conforme mostra o mapa perceptual da Figura 11.66.

Caso o pesquisador deseje elaborar o mapa perceptual destacando as massas das categorias no próprio mapa, poderá recorrer ao comando **svmat2**, desenvolvido por Nicholas J. Cox. Para usá-lo, devemos inicialmente digitar:

```
findit svmat2
```

e instalá-lo no link **dm79 from http://www.stata.com/stb/stb56&#8243**. Feito isso, podemos digitar a seguinte sequência de comandos:

```
mca perfil aplicação estado_civil, method(indicator)
mat mcamat=e(cGS)
mat colnames mcamat = mass qual inert co1 rel1 abs1 co2 rel2 abs2
svmat2 mcamat, rname(varname) name(co1)
```

Esses comandos criam novas variáveis no banco de dados que trazem informações sobre as matrizes geradas após a elaboração da análise de correspondência múltipla, entre as quais as massas e as coordenadas de cada categoria. O novo mapa perceptual pode, portanto, ser construído, com os pontos referentes a cada categoria apresentando diâmetros proporcionais às respectivas massas. Para tanto, devemos digitar o seguinte comando:

```
graph twoway scatter co2 co1 [aweight=mass], xline(0) yline(0) ||
scatter co2 co1, mlabel(varname) legend(off)
```

O novo mapa perceptual encontra-se na Figura 11.67.

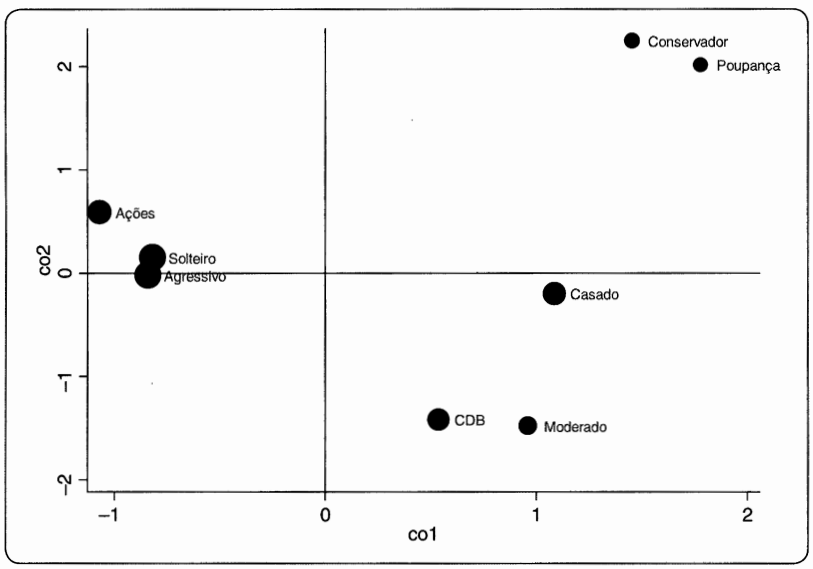



**Figura 11.67** Mapa perceptual para perfil do investidor, tipo de aplicação financeira e estado civil, com ponderações pelas massas de cada categoria.

Assim como realizado na seção 11.4.1 quando da elaboração da técnica no SPSS, podemos criar duas novas variáveis no banco de dados, correspondentes às coordenadas de cada uma das observações da amostra, digitando o seguinte comando:

```
predict a1 a2
```

Note que as coordenadas de cada observação são exatamente as mesmas geradas pelo SPSS, embora as coordenadas das categorias tenham sido calculadas por meio de procedimentos distintos (coordenadas-padrão para o Stata e coordenadas principais para o SPSS). Portanto, a partir das coordenadas de cada observação, é possível elaborar um gráfico, que se encontra na Figura 11.68, com as posições relativas dos estudantes. As variáveis que contêm essas coordenadas são ortogonais e análogas aos fatores criados por meio de uma análise fatorial por componentes principais, e, a partir delas, podem ser elaboradas técnicas como análise de agrupamentos, a fim de que sejam, por exemplo, agrupados estudantes com características similares entre si. Para que esse gráfico seja construído, precisamos digitar o seguinte comando:

```
graph twoway scatter a2 a1, xline(0) yline(0) mlabel(estudante)
```

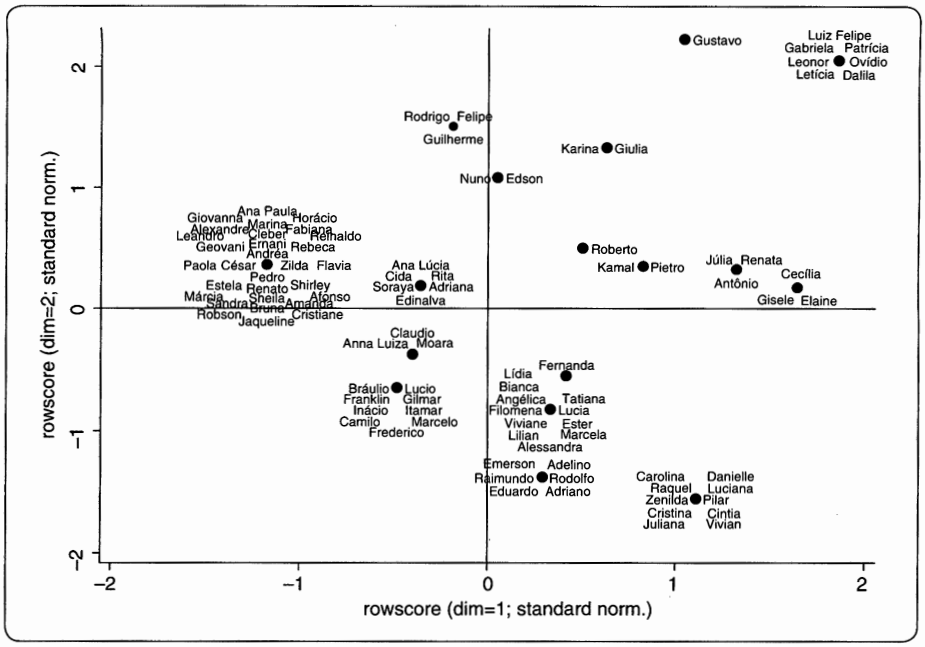



**Figura 11.68** Posições relativas das observações da amostra.

Por fim, caso o pesquisador queira obter as coordenadas principais calculadas a partir do método da matriz de Burt, poderá digitar o seguinte comando, que gerará os *outputs* da Figura 11.69. Note que as coordenadas apresentadas nessa figura correspondem às apresentadas nas Figuras 11.46, 11.47 e 11.48, obtidas quando da aplicação da técnica no SPSS, com exceção dos sinais invertidos para as ordenadas e de pequenos erros de arredondamento.

```
mca perfil aplicação estado_civil, method(indicator) normalize
(principal)
```

em que o termo **normalize(principal)** faz com que sejam geradas as coordenadas principais, em vez das coordenadas-padrão apresentadas na Figura 11.63.

```
. mca perfil aplicação estado_civil, method(indicator) normalize(principal)

Multiple/Joint correspondence analysis          Number of obs      =       100
                                                Total inertia      =  1.666667
    Method: Indicator matrix                    Number of axes     =         2

                |  principal              cumul
      Dimension |    inertia     percent  percent
    ------------+--------------------------------
          dim 1 |     .6023045     36.14    36.14
          dim 2 |     .4359878     26.16    62.30
          dim 3 |     .2764728     16.59    78.89
          dim 4 |     .1798371     10.79    89.68
          dim 5 |     .1720645     10.32   100.00
    ------------+--------------------------------
          Total |     1.666667    100.00

Statistics for column categories in principal normalization

                |          overall                |        dimension_1              |        dimension_2
     Categories |    mass  quality   %inert |    coord   sqcorr   contrib |    coord   sqcorr   contrib
    ------------+---------------------------------+---------------------------------+---------------------------------
    perfil      |                                 |                                 |
     Conservador |   0.057    0.712    0.166 |    1.130    0.262    0.093 |    1.484    0.451    0.189
        Moderado |   0.083    0.503    0.150 |    0.747    0.186    0.060 |   -0.975    0.317    0.120
       Agressivo |   0.193    0.589    0.084 |   -0.653    0.589    0.106 |   -0.015    0.000    0.000
    ------------+---------------------------------+---------------------------------+---------------------------------
    aplicação   |                                 |                                 |
         Poupança |   0.050    0.649    0.170 |    1.381    0.337    0.123 |    1.331    0.313    0.134
              CDB |   0.133    0.699    0.120 |    0.417    0.116    0.030 |   -0.935    0.583    0.177
            Ações |   0.150    0.688    0.110 |   -0.831    0.565    0.134 |    0.388    0.123    0.034
    ------------+---------------------------------+---------------------------------+---------------------------------
    estado_civil |                                 |                                 |
         Solteiro |   0.190    0.549    0.086 |   -0.636    0.536    0.099 |    0.099    0.013    0.003
           Casado |   0.143    0.549    0.114 |    0.843    0.536    0.131 |   -0.131    0.013    0.004
    ------------------------------------------------------------------------------------------------------------------
```

**Figura 11.69** *Outputs* da análise de correspondência múltipla no Stata – Coordenadas principais.

Conforme discutimos na seção 11.3, as coordenadas principais de determinada dimensão são calculadas multiplicando-se as coordenadas-padrão pela raiz quadrada da inércia principal parcial daquela dimensão. Isso pode ser facilmente verificado a partir dos resultados apresentados nas Figuras 11.63 e 11.69.

Além disso, caso o pesquisador também queira obter as coordenadas principais das categorias das variáveis aplicando uma análise de correspondência simples aos dados gerados a partir da matriz de Burt do nosso exemplo, poderá utilizar o arquivo **Burt.dta**. Nesse caso, é importante apenas atentar para o fato de que os valores singulares de cada dimensão corresponderão aos valores das inércias principais parciais geradas por meio da análise de correspondência múltipla para as respectivas dimensões.

## 11.6. CONSIDERAÇÕES FINAIS

As tabelas de contingência se apresentam com bastante frequência em diversos campos do conhecimento, pela forte presença de variáveis categóricas, como sexo, faixas de idade ou de renda e características comportamentais, setoriais ou de localidade. O estudo aprofundado dessas tabelas, no entanto, ainda é pouco explorado no sentido de se construírem mapas perceptuais que permitem ao pesquisador avaliar, visualmente, as associações entre variáveis e entre suas categorias.

Nesse sentido, as técnicas de análise de correspondência simples e de análise de correspondência múltipla têm, por principal objetivo, avaliar a significância da associação entre variáveis categóricas e entre suas categorias, gerar coordenadas das categorias e construir, a partir dessas coordenadas, mapas perceptuais. Enquanto a primeira é uma técnica que permite avaliar a associação entre apenas duas variáveis categóricas e entre suas categorias, a segunda é uma técnica multivariada em que são estudadas as associações entre mais de duas variáveis categóricas e entre cada par de categorias. Essas técnicas permitem, portanto, aprimorar os processos decisórios com base no comportamento e na relação de interdependência entre variáveis que apresentam alguma forma de categorização.

Enfatiza-se que a aplicação de técnicas exploratórias, como a análise de correspondência, deve ser feita por meio do correto e consciente uso do software escolhido para a modelagem, com base na teoria subjacente e na experiência e intuição do pesquisador.

## 11.7. EXERCÍCIOS

**1.** Com o intuito de estudar a associação entre a percepção dos clientes sobre a qualidade do atendimento prestado e a percepção sobre o nível de preços praticados em relação à concorrência, um estabelecimento supermercadista realizou uma pesquisa com 3.000 consumidores dentro da loja, coletando dados de variáveis com as seguintes características:

| Variável | Descrição |
|---|---|
| *id* | Variável *string* (de 0001 a 3000) que identifica o consumidor e que não será utilizada na modelagem. |
| *atendimento* | Variável qualitativa ordinal com cinco categorias, correspondente à percepção sobre a qualidade do atendimento prestado pelo estabelecimento (péssimo = 1; ruim = 2; regular = 3; bom = 4; ótimo = 5). |
| *preço* | Variável qualitativa ordinal com cinco categorias, correspondente à percepção sobre o nível de preços praticados em relação à concorrência (péssimo = 1; ruim = 2; regular = 3; bom = 4; ótimo = 5). |

Por meio da análise do banco de dados presente nos arquivos **Atendimento × Preço.sav** e **Atendimento × Preço.dta**, pede-se:

**a.** Elabore uma tabela de contingência com os valores das frequências absolutas observadas em cada célula a partir do cruzamento das categorias das variáveis *atendimento* e *preço*.

**b.** Apresente a tabela de frequências absolutas esperadas a partir do mesmo cruzamento.

**c.** Com base na estatística $\chi^2$, é possível afirmar que existe associação estatisticamente significante, ao nível de significância de 5%, entre as variáveis *atendimento* e *preço*?

**d.** Apresente a tabela de resíduos padronizados ajustados. Com base nela, discuta a relação de dependência entre cada par de categorias.

**e.** A partir da elaboração da análise de correspondência simples entre *atendimento* e *preço*, pergunta-se: Quais os valores das inércias principais parciais de cada dimensão? Quais os percentuais da inércia principal total explicados por dimensão?

**f.** Com base nas coordenadas das categorias das variáveis *atendimento* e *preço*, obtidas a partir da elaboração da análise de correspondência simples, elabore o mapa perceptual bidimensional e faça uma breve discussão sobre o comportamento dos pontos correspondentes às categorias de cada variável.

**g.** Elabore o gráfico de projeção das coordenadas nas dimensões (Stata) e discuta, para a primeira dimensão, a lógica da ordenação das categorias das duas variáveis qualitativas ordinais (*atendimento* e *preço*).

**2.** O Ministério da Saúde de determinado país deseja implementar uma campanha para alertar a população sobre a importância de se praticar exercícios físicos para a redução do índice de colesterol LDL (mg/dL). Para tanto, realizou uma pesquisa com 2.304 indivíduos, em que foram levantadas as seguintes variáveis:

| Variável | Descrição |
|---|---|
| *colestclass* | Classificação do índice de colesterol LDL (mg/dL), a saber:<br>– Muito elevado: superior a 189 mg/dL;<br>– Elevado: de 160 a 189 mg/dL;<br>– Limítrofe: de 130 a 159 mg/dL;<br>– Subótimo: de 100 a 129 mg/dL;<br>– Ótimo: inferior a 100 mg/dL. |
| *esporte* | Número de vezes em que pratica atividades físicas semanalmente. |

Ao divulgar os resultados da pesquisa, o Ministério da Saúde apresentou a seguinte tabela de contingência, com as frequências absolutas observadas para cada cruzamento de categorias das duas variáveis.

| Classificação do índice de colesterol LDL (mg/dL) | Atividades físicas semanais (número de vezes) | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 |
| **Muito elevado** | 32 | 158 | 264 | 140 | 40 | 0 |
| **Elevado** | 22 | 108 | 178 | 108 | 58 | 0 |
| **Limítrofe** | 0 | 26 | 98 | 190 | 86 | 36 |
| **Subótimo** | 0 | 16 | 114 | 166 | 104 | 54 |
| **Ótimo** | 0 | 0 | 82 | 118 | 76 | 30 |

Note que, enquanto a variável *colestclass* é qualitativa ordinal, a variável *esporte* é quantitativa, porém discreta e com poucas possibilidades de resposta e, portanto, pode ser considerada categórica para efeitos de análise de correspondência.

Nesse sentido, pede-se:

**a.** Apresente a tabela com frequências absolutas esperadas.

**b.** Elabore a tabela de resíduos.

**c.** Apresente a tabela de valores de $\chi^2$ por célula e calcule o valor total da estatística $\chi^2$.

**d.** Com base no valor calculado da estatística $\chi^2$ e nos graus de liberdade da tabela de contingência, é possível afirmar que o índice de colesterol LDL e a quantidade semanal de atividades esportivas não se associam de forma aleatória, ao nível de significância de 5%?

**e.** Construa o banco de dados a partir da tabela de contingência apresentada, e, por meio dele, elabore uma análise de correspondência simples entre *colestclass* e *esporte*. Quais os valores das inércias principais parciais de cada dimensão? Quais os percentuais da inércia principal total explicados por dimensão?

**f.** Com base nas coordenadas das categorias das variáveis *colestclass* e *esporte* obtidas a partir da elaboração da análise de correspondência simples, elabore o mapa perceptual bidimensional e faça uma breve discussão sobre o comportamento dos pontos correspondentes às categorias de cada variável.

**g.** Elabore o gráfico de projeção das coordenadas nas dimensões (Stata) e discuta, para a primeira dimensão, a lógica da ordenação das categorias das duas variáveis.

**3.** O prefeito de determinado município, com a intenção de avaliar a evolução anual de sua popularidade, encomendou a um instituto, em cada um dos três últimos anos (20X1, 20X2, 20X3), a realização de uma pesquisa aplicada a 3.000 cidadãos escolhidos aleatoriamente. Nas três pesquisas realizadas, foi coletada apenas uma variável, no formato Likert, a partir da seguinte afirmativa:

***Estou satisfeito com a gestão do atual prefeito!***

A variável coletada apresenta as seguintes categorias de resposta:

| Variável | Descrição |
|---|---|
| *avaliação* | – Discordo totalmente;<br>– Discordo parcialmente;<br>– Nem concordo, nem discordo;<br>– Concordo parcialmente;<br>– Concordo totalmente. |

A partir dos resultados das pesquisas, foi elaborada a seguinte tabela de contingência, porém os dados também podem ser acessados nos arquivos **Gestão do Prefeito.sav** e **Gestão do Prefeito.dta**.

| Estou satisfeito com a gestão do atual prefeito! | Ano | | |
|---|---|---|---|
| | 20X1 | 20X2 | 20X3 |
| **Discordo totalmente** | 0 | 1 | 997 |
| **Discordo parcialmente** | 1 | 998 | 1.005 |
| **Nem concordo, nem discordo** | 967 | 1.005 | 998 |
| **Concordo parcialmente** | 1.066 | 996 | 0 |
| **Concordo totalmente** | 966 | 0 | 0 |
| **TOTAL** | **3.000** | **3.000** | **3.000** |

Pede-se:

**a.** É possível afirmar que a evolução anual da popularidade do prefeito não se dá de forma aleatória, ao nível de significância de 5%?

**b.** Apresente a tabela de resíduos padronizados ajustados. Com base nela, discuta a relação de dependência entre as categorias da variável Likert e cada um dos anos em que foi aplicada a pesquisa?

**c.** Com base nas coordenadas das categorias das variáveis *avaliação* e *ano*, obtidas a partir da elaboração da análise de correspondência simples, elabore o mapa perceptual bidimensional. É possível afirmar que a popularidade do prefeito vem piorando com o decorrer dos anos?

**4.** Conforme propusemos ao final da resolução do exercício elaborado na seção 11.3.2, seria interessante também se avaliássemos a existência de associação entre o fato de se ter um ou mais filhos, o perfil do investidor e o tipo de aplicação financeira. Nesse sentido, foi elaborado o banco de dados presente nos arquivos **Perfil_Investidor × Aplicação × Filhos.sav** e **Perfil_Investidor × Aplicação × Filhos.dta**. Pede-se:

**a.** Apresente as tabelas de contingência e os resultados dos testes $\chi^2$ para cada par de variáveis. Há associação entre o fato de se ter um ou mais filhos, o perfil do investidor e o tipo de aplicação financeira, ao nível de significância de 5%, ou alguma das variáveis deve ser excluída da análise?

**b.** Caso nenhuma variável seja excluída da análise, elabore a análise de correspondência múltipla com as três variáveis (*perfil*, *aplicação* e *filhos*). Quais as coordenadas principais e padrão das categorias de cada uma delas?

**c.** Elabore o mapa perceptual bidimensional (com coordenadas-padrão) e faça uma breve discussão sobre o comportamento dos pontos correspondentes às categorias de cada variável. É possível afirmar que o fato de ter filhos aumenta a aversão ao risco?

**5.** Uma pesquisa com 500 executivos de empresas multinacionais foi realizada com o intuito de avaliar a percepção sobre a qualidade geral do serviço prestado e sobre o respeito aos prazos de projeto de três grandes empresas de consultoria (*Gabicks*, *Lipehigh* e *Montvero*). Cada executivo respondeu sobre sua percepção em relação a cada uma das três empresas, e as variáveis coletadas encontram-se a seguir:

| Variável | Descrição |
|---|---|
| *qualidade* | Percepção sobre a qualidade geral do serviço prestado, a saber:<br>– Péssima;<br>– Ruim;<br>– Regular;<br>– Boa;<br>– Ótima. |
| *pontualidade* | Respeito aos prazos de projeto:<br>– Não;<br>– Sim. |

Por meio da análise do banco de dados presente nos arquivos **Consultoria.sav** e **Consultoria.dta**, pede-se:

**a.** Apresente as tabelas de contingência e os resultados dos testes $\chi^2$ para as variáveis *qualidade* e *empresa* e para *pontualidade* e *empresa*. Há associação entre a variável *empresa* e as outras variáveis, ao nível de significância de 5%?

**b.** Se a resposta do item anterior for positiva, elabore uma análise de correspondência múltipla com as três variáveis. Quais as coordenadas principais e padrão das categorias de cada uma delas?

**c.** Elabore o gráfico de projeção das coordenadas-padrão nas dimensões (Stata) e discuta, para a primeira dimensão, a lógica da ordenação das categorias da variável *qualidade*.

**d.** Elabore o mapa perceptual bidimensional (com coordenadas-padrão) e discorra sobre a leitura que os executivos fazem sobre as três empresas de consultoria.

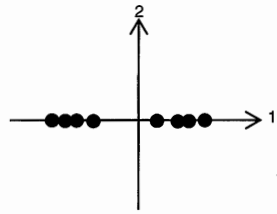
**e.** A partir das coordenadas de cada uma das respostas dadas (1.500 observações), geradas após a aplicação da análise de correspondência múltipla, elabore dois gráficos (SPSS) com as posições relativas dessas observações, tendo em vista a explicitação das categorias das variáveis *qualidade* e *empresa*, respectivamente. Há lógica nas respostas dadas pelos executivos em relação às categorias dessas variáveis?

APÊNDICE

# Configurações do mapa perceptual de uma análise de correspondência simples

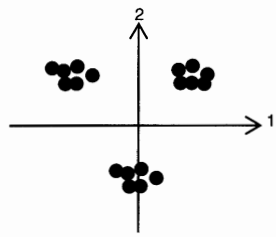
Muitas são as configurações que podem assumir os mapas perceptuais, em função das características das tabelas de contingência. A Figura 11.70 apresenta as configurações mais comuns. Enquanto as células em destaque e com setas ↑ representam valores elevados de frequências absolutas observadas, as células com setas ↓ representam valores baixos, ou até mesmo nulos, dessas frequências.

a) Nuvem de Pontos Dividida em Grupos sobre a Primeira Dimensão
(pelo menos uma variável com duas categorias)



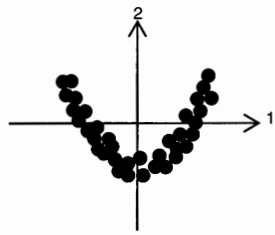

| | 1 | 2 |
|---|---|---|
| 1 | ↑ | ↓ |
| 2 | ↓ | ↑ |

b) Nuvem de Pontos Divida em Grupos nas Duas Dimensões
(variáveis com pelo menos três categorias)
(corresponde aos dados do exemplo da seção 3.2.5)



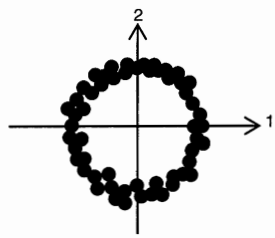

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | ↑ | ↓ | ↓ |
| 2 | ↓ | ↑ | ↓ |
| 3 | ↓ | ↓ | ↑ |

c) Forma Parabólica da Nuvem de Pontos
(estrutura diagonal da tabela de contingência para mais de três categorias em cada variável)




| | 1 | 2 | ... | *J* |
|---|---|---|---|---|
| 1 | ↑ | ↓ | ... | ↓ |
| 2 | ↓ | ↑ | | ↓ |
| ⋮ | | ⋮ | ↑ | ⋮ |
| *I* | ↓ | ↓ | ... | ↑ |

d) Forma Circular da Nuvem de Pontos
(mais de uma estrutura diagonal na tabela de contingência)




| | 1 | 2 | 3 | 4 | ... | *J* |
|---|---|---|---|---|---|---|
| 1 | ↓ | ↑ | ↓ | ↓ | | ↓ |
| 2 | ↑ | ↓ | ↑ | ↓ | ... | ↓ |
| 3 | ↓ | ↑ | ↓ | ↑ | | ↓ |
| 4 | ↓ | ↓ | ↑ | ↓ | | ↓ |
| ⋮ | | | ⋮ | | ↓ | ↑ |
| *I* | ↓ | ↓ | ↓ | ↓ | ↑ | ↓ |

**Figura 11.70** Configurações do mapa perceptual de uma análise de correspondência simples em função das características da tabela de contingência.

*Fonte:* Pereira e Sousa (2015).

PARTE III

# TÉCNICAS MULTIVARIADAS CONFIRMATÓRIAS: MODELOS DE REGRESSÃO

Talvez a mais famosa equação já desenvolvida na história da humanidade seja aquela atribuída a Albert Einstein, $E = m.c^2$. Embora Einstein não a tenha formulado exatamente dessa forma em seu seminal artigo "A inércia de um corpo depende da sua quantidade de energia?", publicado no *annus mirabilis* de 1905 na *Annalen der Physik*, tal equação tornou-se mundialmente famosa por sua simplicidade ao tentar relacionar massa e energia de corpos físicos e, com esse propósito, pode ser classificada como um **modelo de regressão**.

O conjunto de **técnicas de regressão** é muito provavelmente o mais utilizado em análises de dados que procuram entender a relação entre o comportamento de determinado fenômeno e o comportamento de uma ou mais variáveis potencialmente preditoras, sem que haja, entretanto, uma relação obrigatória de causa e efeito. Por exemplo, a relação entre a quantidade de horas de estudo na preparação e as notas no vestibular para Medicina é, obviamente, de natureza causal, ou seja, quanto maior a dedicação aos estudos, maiores serão as notas no vestibular, mesmo que também existam outros fatores que possam influenciar as notas no exame, como ansiedade e poder de concentração do candidato.

Por outro lado, existem situações em que o fenômeno em estudo apresenta relação com determinada variável inserida no modelo, sem que essa relação seja, de fato, de natureza causal. Nesses casos, é comum que uma terceira variável não observada esteja influenciando o comportamento tanto do fenômeno em estudo quanto da variável preditora. Gustav Fischer, em 1936, apresentou um estudo bastante interessante sobre esse fato ao investigar ao longo de 7 anos a relação entre a quantidade de cegonhas e o número de recém-nascidos em pequenas cidades da Dinamarca. Curiosamente, essa relação mostrava-se forte e positiva. Entretanto, essas duas variáveis eram causadas pelo tamanho das cidades, variável não considerada no modelo, visto que em cidades maiores, onde nasciam mais crianças, também havia uma quantidade maior de chaminés, onde as cegonhas faziam seus ninhos. **Nesse sentido, é de fundamental importância que o pesquisador seja bastante cuidadoso e criterioso ao interpretar os resultados de uma modelagem de regressão. A existência de um modelo de regressão não significa que ocorra, obrigatoriamente, relação de causa e efeito entre as variáveis consideradas!**

O termo **regressão** é uma homenagem aos trabalhos realizados por Francis Galton e Karl Pearson na tentativa de se estimar uma função linear que procurava investigar a relação entre a altura dos filhos e a altura dos pais, de modo a se estabelecer uma eventual **lei universal de regressão**.

Segundo Stanton (2001), embora Pearson tivesse desenvolvido um tratamento matemático rigoroso acerca do que se convencionou chamar de **correlação**, foi a imaginação de Galton que originalmente concebeu as noções de correlação e de regressão. Sir Francis Galton, primo de Charles Darwin, foi bastante criticado no final do século XIX por defender a eugenia, e a própria fama de seu primo acabou por ofuscar suas profundas contribuições científicas nos campos da biologia, psicologia e estatística aplicada. Seu fascínio por genética e hereditariedade forneceu a inspiração necessária que levou à regressão.

Em 1875, Galton teve a ideia de distribuir pacotes de sementes de ervilha doce a sete amigos e, embora cada pacote contivesse sementes com peso uniforme, havia variação substancial entre os diferentes pacotes. Após algum tempo, sementes da nova geração foram colhidas das plantas que brotaram a partir das sementes originais, para que pudessem ser elaborados gráficos que relacionavam os pesos das sementes da nova geração e os pesos das sementes originais. Galton percebeu que os pesos médios das novas sementes geradas a partir de sementes originais com um peso específico descreviam, aproximadamente, uma reta com inclinação positiva e inferior a 1.

Duas décadas mais tarde, em 1896, Pearson publicou seu primeiro rigoroso tratado sobre correlação e regressão no *Philosophical Transactions of the Royal Society of London*. Nesse trabalho, Pearson creditou Bravais (1846) por ser o primeiro a estudar as formulações matemáticas iniciais da correlação, enfatizando que Bravais, embora tivesse se deparado com um método adequado para o cálculo do **coeficiente de correlação**, acabou não conseguindo provar que isso proporcionaria o melhor ajuste aos dados. Por meio do mesmo método, porém fazendo uso de avançada prova estatística com base na **expansão de Taylor**, Pearson acabou por chegar aos valores ótimos da inclinação e do coeficiente de correlação de um modelo de regressão.

Em 1911, com a morte de Galton, Karl Pearson tornou-se seu biógrafo e, descreve, de forma primorosa, como se deu o desenvolvimento do conceito da inclinação em um modelo de regressão.

Com o transcorrer do tempo, os modelos de regressão passaram a ser mais estudados e aplicados em diversos campos do conhecimento humano e, com o desenvolvimento tecnológico e o aprimoramento computacional, verificou-se, principalmente a partir da segunda metade do século XX, o surgimento de novos e cada vez mais complexos tipos de modelagens de regressão. As técnicas de regressão inserem-se dentro do que é conhecido por **técnicas de dependência**, em que há a intenção de que sejam estimados modelos (equações) que permitam ao pesquisador estudar o comportamento dos dados e a relação entre as variáveis e elaborar previsões do fenômeno em estudo, com intervalos de confiança. São, portanto, consideradas **técnicas confirmatórias**.

Optamos, com base em razões didáticas e conceituais por abordar na Parte III as principais técnicas pertinentes aos modelos de regressão, ficando os capítulos estruturados em três subpartes distintas, a saber:

**PARTE III.1: MODELOS LINEARES GENERALIZADOS**

**Capítulo 12:** Modelos de Regressão Simples e Múltipla

**Capítulo 13:** Modelos de Regressão Logística Binária e Multinomial

**Capítulo 14:** Modelos de Regressão para Dados de Contagem: Poisson e Binomial Negativo

**PARTE III.2: MODELOS DE REGRESSÃO PARA DADOS EM PAINEL**

**Capítulo 15:** Modelos Longitudinais de Regressão para Dados em Painel

**Capítulo 16:** Modelos Multinível de Regressão para Dados em Painel

**PARTE III.3: OUTROS MODELOS DE REGRESSÃO**

**Capítulo 17:** Modelos de Regressão para Dados de Sobrevivência: Riscos Proporcionais de Cox

**Capítulo 18:** Modelos de Regressão com Múltiplas Variáveis Dependentes: Correlação Canônica

Cada capítulo da Parte III está estruturado dentro de uma mesma lógica de apresentação. Inicialmente, são introduzidos os conceitos pertinentes a cada modelo, bem como os critérios para estimação de seus parâmetros, sempre com o uso de bases de dados que possibilitam, em um primeiro momento, a resolução de exercícios práticos, na maioria dos casos, em Excel. Na sequência, os mesmos exercícios são resolvidos nos pacotes estatísticos Stata Statistical Software® e IBM SPSS Statistics Software®. Acreditamos que essa lógica facilita o estudo e o entendimento sobre a utilização correta de cada um dos modelos de regressão, a estimação dos respectivos parâmetros e a análise dos resultados. Além disso, a aplicação prática das modelagens em Stata e SPSS também traz benefícios ao pesquisador, à medida que os resultados podem, a todo instante, ser comparados com aqueles já estimados ou calculados algebricamente nas seções iniciais de cada capítulo, além de propiciar uma oportunidade de manuseio desses importantes softwares. Ao término dos capítulos, são propostos exercícios complementares, com respostas apresentadas por meio de *outputs* gerados em Stata, disponibilizadas no final do livro.

PARTE III.1

# MODELOS LINEARES GENERALIZADOS

O estudo das distribuições estatísticas não é recente, e desde o início do século XIX, até aproximadamente o início do século XX, os modelos lineares que envolvem a distribuição normal praticamente dominou o cenário da modelagem de dados.

Entretanto, a partir do período entre guerras, começam a surgir modelos para fazer frente a situações em que as modelagens lineares normais não se adequavam satisfatoriamente. McCullagh e Nelder (1989), Turkman e Silva (2000) e Cordeiro e Demétrio (2007) citam, neste contexto, os trabalhos de Berkson (1944), Dyke e Patterson (1952) e Rasch (1960) sobre os modelos logísticos envolvendo as distribuições de Bernoulli e binomial, de Birch (1963) sobre os modelos para dados de contagem envolvendo a distribuição Poisson, de Feigl e Zelen (1965), Zippin e Armitage (1966) e Glasser (1967) sobre os modelos exponenciais, e de Nelder (1966) sobre modelos polinomiais envolvendo a distribuição Gama.

Todos estes modelos acabaram por ser consolidados, do ponto de vista teórico e conceitual, por meio do seminal trabalho de Nelder e Wedderburn (1972), em que foram definidos os **Modelos Lineares Generalizados** (***Generalized Linear Models***), que representam um grupo de modelos de regressão lineares e exponenciais não lineares, em que a variável dependente possui, por exemplo, distribuição normal, Bernoulli, binomial, Poisson ou Poisson-Gama. São casos particulares dos Modelos Lineares Generalizados os seguintes modelos:

- Modelos de Regressão Lineares e Modelos com Transformação de Box-Cox;
- Modelos de Regressão Logística Binária e Multinomial;
- Modelos de Regressão Poisson e Binomial Negativo para Dados de Contagem;

e a estimação de cada um deles deve ser elaborada respeitando-se as características dos dados e a distribuição da variável que representa o fenômeno que se deseja estudar, chamada de variável dependente.

Um Modelo Linear Generalizado é definido da seguinte forma:

$$\eta_i = \alpha + \beta_1 . \ X_{1i} + \beta_2 . X_{2i} + ... + \beta_k . X_{ki} \qquad \text{(III.1.1)}$$

em que $\eta$ é conhecido por função de ligação canônica, $\alpha$ representa a constante, $\beta_j$ ($j = 1, 2, ..., k$) são os coeficientes de cada variável explicativa e correspondem aos parâmetros a serem estimados, $X_j$ são as variáveis explicativas (métricas ou *dummies*) e os subscritos $i$ representam cada uma das observações da amostra em análise ($i = 1, 2, ..., n$, em que $n$ é o tamanho da amostra).

O Quadro III.1.1 relaciona cada caso particular dos modelos lineares generalizados com a característica da variável dependente, a sua distribuição e a respectiva função de ligação canônica.

Logo, para uma dada variável dependente $Y$ que representa o fenômeno em estudo (variável dependente), podemos especificar cada um dos modelos apresentados no Quadro III.1.1 da seguinte maneira:

**Modelo de Regressão Linear:**

$$\hat{Y}_i = \alpha + \beta_1 . X_{1i} + \beta_2 . X_{2i} + ... + \beta_k . X_{ki} \qquad \text{(III.1.2)}$$

em que $\hat{Y}$ é o valor esperado da variável dependente $Y$.

**Quadro III.1.1** Modelos lineares generalizados, características da variável dependente e funções de ligação canônica.

| Modelo de Regressão | Característica da Variável Dependente | Distribuição | Função de Ligação Canônica ($\eta$) |
|---|---|---|---|
| **Linear** | Quantitativa | Normal | $\hat{Y}$ |
| **Com Transformação de Box-Cox** | Quantitativa | Normal Após a Transformação | $\frac{\hat{Y}^{\lambda}-1}{\lambda}$ |
| **Logística Binária** | Qualitativa com 2 Categorias (*Dummy*) | Bernoulli | $\ln\left(\frac{p}{1-p}\right)$ |
| **Logística Multinomial** | Qualitativa $M$ ($M > 2$) Categorias | Binomial | $\ln\left(\frac{p_m}{1-p_m}\right)$ |
| **Poisson** | Quantitativa com Valores Inteiros e Não Negativos (Dados de Contagem) | Poisson | $\ln(\lambda)$ |
| **Binomial Negativo** | Quantitativa com Valores Inteiros e Não Negativos (Dados de Contagem) | Poisson-Gama | $\ln(u)$ |

**Modelo de Regressão com Transformação de Box-Cox:**

$$\frac{\hat{Y}_i^{\lambda}-1}{\lambda} = \alpha + \beta_1.X_{1i} + \beta_2.X_{2i} + ... + \beta_k.X_{ki} \quad \text{(III.1.3)}$$

em que $\hat{Y}$ é o valor esperado da variável dependente $Y$ e $\lambda$ é o parâmetro da transformação de Bo x-Cox que maximiza a aderência à normalidade da distribuição da nova variável gerada a partir da variável $Y$ original.

**Modelo de Regressão Logística Binária:**

$$\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1.X_{1i} + \beta_2.X_{2i} + ... + \beta_k.X_{ki} \quad \text{(III.1.4)}$$

em que $p$ é a probabilidade de ocorrência do evento de interesse definido por $Y = 1$, dado que a variável dependente $Y$ é *dummy*.

**Modelo de Regressão Logística Multinomial:**

$$\ln\left(\frac{p_{i_m}}{1-p_{i_m}}\right) = \alpha_m + \beta_{1m}.X_{1i} + \beta_{2m}.X_{2i} + ... + \beta_{km}.X_{ki} \quad \text{(III.1.5)}$$

em que $p_m$ ($m = 0, 1, ..., M-1$) é a probabilidade de ocorrência de cada uma das $M$ categorias da variável dependente $Y$.

**Modelo de Regressão Poisson para Dados de Contagem:**

$$\ln(\lambda_i) = \alpha + \beta_1 . X_{1i} + \beta_2 . X_{2i} + ... + \beta_k . X_{ki} \quad \text{(III.1.6)}$$

em que $\lambda$ é o valor esperado da quantidade de ocorrências do fenômeno representado pela variável dependente $Y$, que apresenta dados de contagem com distribuição Poisson.

**Modelo de Regressão Binomial Negativo para Dados de Contagem:**

$$\ln(u_i) = \alpha + \beta_1 . X_{1i} + \beta_2 . X_{2i} + ... + \beta_k . X_{ki} \quad \text{(III.1.7)}$$

em que $u$ é o valor esperado da quantidade de ocorrências do fenômeno representado pela variável dependente $Y$, que apresenta dados de contagem com distribuição Poisson-Gama.

Portanto, a Parte III.1 trata dos Modelos Lineares Generalizados. Enquanto o Capítulo 12 aborda os modelos de regressão linear e os modelos com transformação de Box-Cox, os dois capítulos seguintes abordam, respectivamente, os modelos de regressão logística binária e multinomial e os modelos de regressão para dados de contagem do tipo Poisson e binomial negativo, que são modelos exponenciais não lineares, também chamados de modelos log-lineares ou semilogarítmicos à esquerda. A Figura III.1.1 apresenta esta lógica.

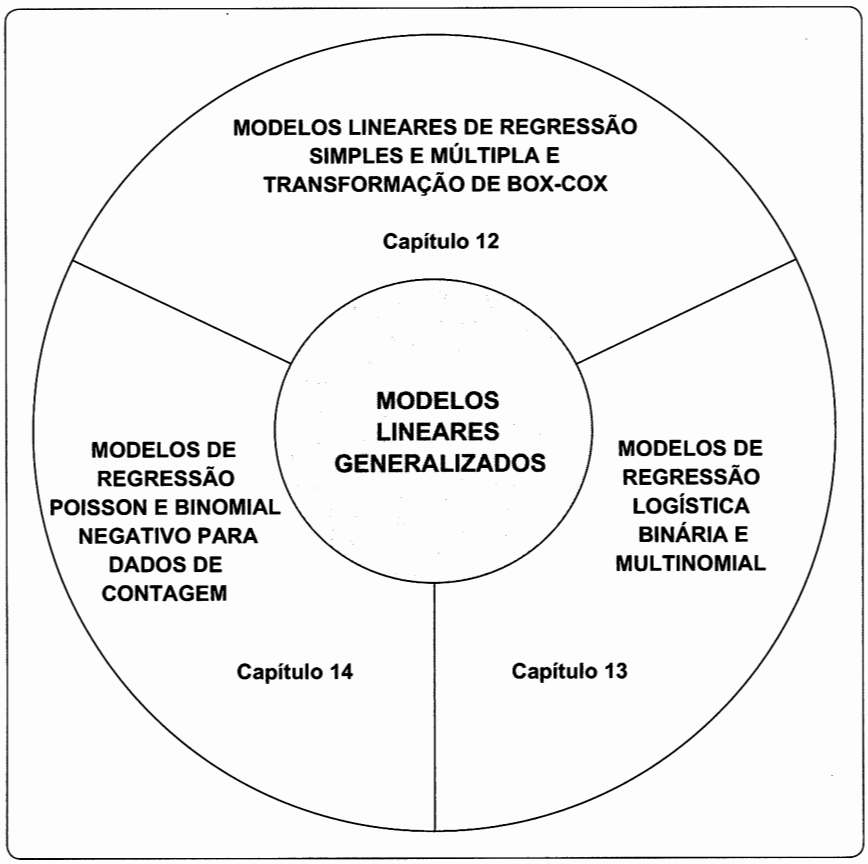



**Figura III.1.1** Modelos lineares generalizados e estruturação dos capítulos da Parte III.1.

Os capítulos da Parte III.1 estão estruturados dentro de uma mesma lógica de apresentação, em que são, inicialmente, introduzidos os conceitos pertinentes a cada modelo e apresentados os critérios para estimação de seus parâmetros, sempre com o uso de bases de dados que possibilitam a resolução de exercícios práticos em Excel. Na sequência, os mesmos exercícios são resolvidos, passo a passo, nos softwares Stata e SPSS. Ao final de cada capítulo, são propostos exercícios complementares, cujas respostas estão disponibilizadas ao final do livro.