

# Analise Bidimensional

Analisar o comportamento conjunto de duas ou mais variáveis aleatórias.

**Tabela 4.1** Tabela de dados.

| Indivíduo | Variável |          |     |          |     |          |
|-----------|----------|----------|-----|----------|-----|----------|
|           | $X_1$    | $X_2$    | ... | $X_j$    | ... | $X_p$    |
| 1         | $x_{11}$ | $x_{12}$ | ... | $x_{1j}$ | ... | $x_{1p}$ |
| 2         | $x_{21}$ | $x_{22}$ | ... | $x_{2j}$ | ... | $x_{2p}$ |
| $\vdots$  | $\vdots$ | $\vdots$ |     | $\vdots$ |     | $\vdots$ |
| $i$       | $x_{i1}$ | $x_{i2}$ | ... | $x_{ij}$ | ... | $x_{ip}$ |
| $\vdots$  | $\vdots$ | $\vdots$ |     | $\vdots$ |     | $\vdots$ |
| $n$       | $x_{n1}$ | $x_{n2}$ | ... | $x_{nj}$ | ... | $x_{np}$ |

o objetivo é encontrar as possíveis **relações ou associações** entre as duas variáveis.

Como no caso de apenas uma variável que estudamos, **a distribuição conjunta das frequências** será um instrumento poderoso para a compreensão do comportamento dos dados.

Quando consideramos duas variáveis (ou dois conjuntos de dados), podemos ter **três situações**:

- (a) as duas variáveis são qualitativas;
- (b) as duas variáveis são quantitativas; e
- (c) uma variável é qualitativa e outra é quantitativa.

As técnicas de análise de dados nas três situações são diferentes.

## Variáveis Qualitativas

**Exemplo 4.1** Suponha que queiramos analisar o comportamento conjunto das variáveis **Y: grau de instrução** e **V: região de procedência**, cujas observações estão contidas na Tabela 2.1. A distribuição de frequências é representada por uma tabela de dupla entrada e está na Tabela 4.2.

**Tabela 4.2** Distribuição conjunta das frequências das variáveis grau de instrução (Y) e região de procedência (V).

| <b>V \ Y</b> | Ensino Fundamental | Ensino Médio | Superior | Total |
|--------------|--------------------|--------------|----------|-------|
| Capital      | 4                  | 5            | 2        | 11    |
| Interior     | 3                  | 7            | 2        | 12    |
| Outra        | 5                  | 6            | 2        | 13    |
| Total        | 12                 | 18           | 6        | 36    |

Fonte: Tabela 2.1.

Em vez de trabalharmos com as frequências absolutas, podemos construir tabelas com **as frequências relativas (proporções)**, como foi feito no caso unidimensional. Existem três possibilidades de expressarmos a proporção de cada casela:

- (a) em relação ao total geral;
- (b) em relação ao total de cada linha;
- (c) ou em relação ao total de cada coluna.

De acordo com o objetivo do problema em estudo, uma delas será a mais conveniente.

**Tabela 4.3** Distribuição conjunta das proporções (em porcentagem) em relação ao total geral das variáveis  $Y$  e  $V$  definidas no texto.

| $V \backslash Y$ | Fundamental | Médio | Superior | Total |
|------------------|-------------|-------|----------|-------|
| Capital          | 11%         | 14%   | 6%       | 31%   |
| Interior         | 8%          | 19%   | 6%       | 33%   |
| Outra            | 14%         | 17%   | 5%       | 36%   |
| Total            | 33%         | 50%   | 17%      | 100%  |

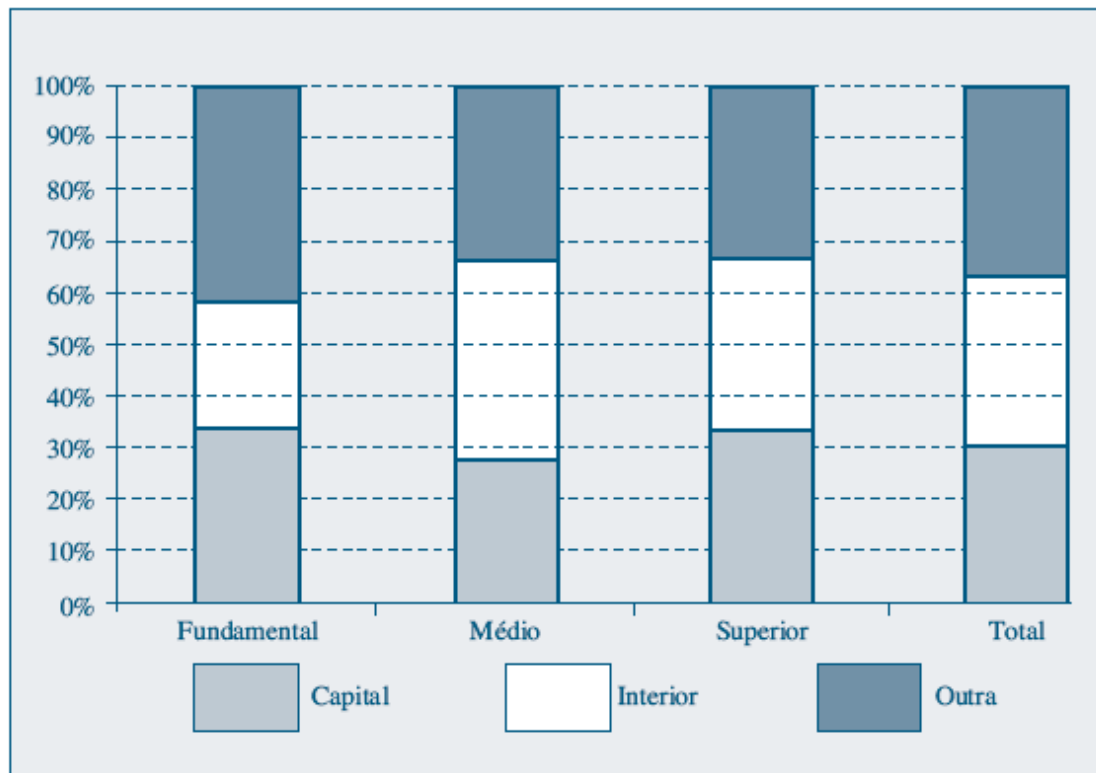
Fonte: Tabela 4.2.

**Tabela 4.4** Distribuição conjunta das proporções (em porcentagem) em relação aos totais de cada coluna das variáveis  $Y$  e  $V$  definidas no texto.

| $V \backslash Y$ | Fundamental | Médio | Superior | Total |
|------------------|-------------|-------|----------|-------|
| Capital          | 33%         | 28%   | 33%      | 31%   |
| Interior         | 25%         | 39%   | 33%      | 33%   |
| Outra            | 42%         | 33%   | 34%      | 36%   |
| Total            | 100%        | 100%  | 100%     | 100%  |

Fonte: Tabela 4.2.

**Figura 4.1** Distribuição da região de procedência por grau de instrução.



## Associação entre Variáveis Qualitativas

Um dos principais objetivos de se construir uma distribuição conjunta de duas variáveis qualitativas é descrever **a associação entre elas**, isto é, queremos conhecer **o grau de dependência** entre elas, de modo que possamos prever melhor o resultado de uma delas quando conhecermos a realização da outra.

**Exemplo 4.2** Queremos verificar se existe ou não associação entre o sexo e a carreira escolhida por 200 alunos de Economia e Administração. Esses dados estão na Tabela 4.5.

**Tabela 4.5** Distribuição conjunta de alunos segundo o sexo (X) e o curso escolhido (Y).

| $Y \backslash X$ | Masculino | Feminino | Total |
|------------------|-----------|----------|-------|
| Economia         | 85        | 35       | 120   |
| Administração    | 55        | 25       | 80    |
| Total            | 140       | 60       | 200   |

Fonte: Dados hipotéticos.

**Tabela 4.6** Distribuição conjunta das proporções (em porcentagem) de alunos segundo o sexo (X) e o curso escolhido (Y).

| $Y \backslash X$ | Masculino | Feminino | Total |
|------------------|-----------|----------|-------|
| Economia         | 61%       | 58%      | 60%   |
| Administração    | 39%       | 42%      | 40%   |
| Total            | 100%      | 100%     | 100%  |

Fonte: Tabela 4.5.

A partir dessa tabela podemos observar que, **independentemente** do sexo, 60% das pessoas preferem Economia e 40% preferem Administração (observe na coluna de total).

**Não havendo dependência** entre as variáveis, esperaríamos essas mesmas proporções para cada sexo.

Observando a tabela, vemos que as proporções do sexo masculino (61% e 39%) e do sexo feminino (58% e 42%) são próximas das marginais (60% e 40%). Esses resultados parecem indicar não haver dependência entre as duas variáveis, para o conjunto de alunos considerado. Concluimos então que, neste caso, as variáveis sexo e escolha do curso parecem ser não associadas.

**Tabela 4.7** Distribuição conjunta das frequências e proporções (em porcentagem), segundo o sexo ( $X$ ) e o curso escolhido ( $Y$ ).

| $Y \backslash X$ | Masculino  | Feminino  | Total      |
|------------------|------------|-----------|------------|
| Física           | 100 (71%)  | 20 (33%)  | 120 (60%)  |
| Ciências Sociais | 40 (29%)   | 40 (67%)  | 80 (40%)   |
| Total            | 140 (100%) | 60 (100%) | 200 (100%) |

Fonte: Dados hipotéticos.

Nesse caso, as variáveis sexo e curso escolhido parecem ser associadas.

# Medidas de Associação entre Variáveis Qualitativas

De modo geral, a quantificação do grau de associação entre duas variáveis é feita pelos chamados **coeficientes de associação ou correlação**. Essas são medidas que descrevem, por meio de um único número, a associação (ou dependência) entre duas variáveis.

Para facilitar a compreensão, esses coeficientes usualmente variam entre 0 e 1, ou entre -1 e +1, e a proximidade de zero indica falta de associação.

## Duas medidas:

coeficiente de contingência, devido a K. Pearson e uma modificação desse.

**Exemplo 4.3** Queremos verificar se a criação de determinado tipo de cooperativa está associada com algum fator regional. Coletados os dados relevantes, obtemos a Tabela 4.8.

**Tabela 4.8** Cooperativas autorizadas a funcionar por tipo e estado, junho de 1974.

| Estado        | Tipo de Cooperativa |           |           |           | Total        |
|---------------|---------------------|-----------|-----------|-----------|--------------|
|               | Consumidor          | Produtor  | Escola    | Outras    |              |
| São Paulo     | 214 (33%)           | 237 (37%) | 78 (12%)  | 119 (18%) | 648 (100%)   |
| Paraná        | 51 (17%)            | 102 (34%) | 126 (42%) | 22 (7%)   | 301 (100%)   |
| Rio G. do Sul | 111 (18%)           | 304 (51%) | 139 (23%) | 48 (8%)   | 602 (100%)   |
| Total         | 376 (24%)           | 643 (42%) | 343 (22%) | 189 (12%) | 1.551 (100%) |

Fonte: Sinopse Estatística do Brasil — IBGE, 1977.

**Tabela 4.9** Valores esperados na Tabela 4.8 assumindo a independência entre as duas variáveis.

| Estado        | Tipo de Cooperativa |           |           |           | Total        |
|---------------|---------------------|-----------|-----------|-----------|--------------|
|               | Consumidor          | Produtor  | Escola    | Outras    |              |
| São Paulo     | 157 (24%)           | 269 (42%) | 143 (22%) | 79 (12%)  | 648 (100%)   |
| Paraná        | 73 (24%)            | 124 (42%) | 67 (22%)  | 37 (12%)  | 301 (100%)   |
| Rio G. do Sul | 146 (24%)           | 250 (42%) | 133 (22%) | 73 (12%)  | 602 (100%)   |
| Total         | 376 (24%)           | 643 (42%) | 343 (22%) | 189 (12%) | 1.551 (100%) |

Fonte: Tabela 4.8.

**Tabela 4.10** Desvios entre observados e esperados.

| Estado        | Tipo de Cooperativa |            |             |            |
|---------------|---------------------|------------|-------------|------------|
|               | Consumidor          | Produtor   | Escola      | Outras     |
| São Paulo     | 57 (20,69)          | -32 (3,81) | -65 (29,55) | 40 (20,25) |
| Paraná        | -22 (6,63)          | -22 (3,90) | 59 (51,96)  | -15 (6,08) |
| Rio G. do Sul | -35 (8,39)          | 54 (11,66) | 6 (0,27)    | -25 (8,56) |

Fonte: Tabelas 4.8 e 4.9.

construir, para cada casela, a medida

$$\frac{(o_i - e_i)^2}{e_i} \quad (4.1)$$

no qual  $o_i$  é o valor observado e  $e_i$  é o valor esperado.

Uma medida do afastamento global pode ser dada pela soma de todas as medidas (4.1). Essa medida é denominada  $\chi^2$  (qui-quadrado) de Pearson, e no nosso exemplo teríamos

$$\chi^2 = 20,69 + 6,63 + \dots + 8,56 = 171,76.$$

Um valor grande de  $\chi^2$  indica associação entre as variáveis, o que parece ser o caso.

Em geral,

**Tabela 4.11** Notação para tabelas de contingência.

| $X \backslash Y$ | $B_1$    | $B_2$    | ...      | $B_j$    | ...      | $B_s$    | Total    |
|------------------|----------|----------|----------|----------|----------|----------|----------|
| $A_1$            | $n_{11}$ | $n_{12}$ | ...      | $n_{1j}$ | ...      | $n_{1s}$ | $n_{1.}$ |
| $A_2$            | $n_{21}$ | $n_{22}$ | ...      | $n_{2j}$ | ...      | $n_{2s}$ | $n_{2.}$ |
| $\vdots$         | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $A_i$            | $n_{i1}$ | $n_{i2}$ | ...      | $n_{ij}$ | ...      | $n_{is}$ | $n_{i.}$ |
| $\vdots$         | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $A_r$            | $n_{r1}$ | $n_{r2}$ | ...      | $n_{rj}$ | ...      | $n_{rs}$ | $n_{r.}$ |
| Total            | $n_{.1}$ | $n_{.2}$ | ...      | $n_{.j}$ | ...      | $n_{.s}$ | $n_{..}$ |

Suponha que temos duas variáveis qualitativas  $X$  e  $Y$ , classificadas em  $r$  categorias  $A_1, A_2, \dots, A_r$  para  $X$  e  $s$  categorias  $B_1, B_2, \dots, B_s$ , para  $Y$ .

Na tabela, temos:

$n_{ij}$  = número de elementos pertencentes à  $i$ -ésima categoria de  $X$  e  $j$ -ésima categoria de  $Y$ ;

$n_{i.} = \sum_{j=1}^s n_{ij}$  = número de elementos da  $i$ -ésima categoria de  $X$ ;

$n_{.j} = \sum_{i=1}^r n_{ij}$  = número de elementos da  $j$ -ésima categoria de  $Y$ ;

$n_{..} = n = \sum_{i=1}^r \sum_{j=1}^s n_{ij}$  = número total de elementos.

o qui-quadrado de Pearson pode ser escrito

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}, \quad (4.4)$$

$$n_{ij}^* = \frac{n_{i.} n_{.j}}{n}, \quad i = 1, \dots, r, j = 1, \dots, s. \quad (4.3)$$

Se a hipótese de não associação for verdadeira, o valor calculado de (4.4) deve estar próximo de zero. Se as variáveis forem associadas, o valor de  $\chi^2$  deve ser grande.

Pearson definiu uma medida de associação, baseada em (4.4), chamada *coeficiente de contingência*, dado por

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}. \quad (4.5)$$

Contudo, o coeficiente acima não varia entre 0 e 1. O valor máximo de  $C$  depende de  $r$  e  $s$ . Para evitar esse inconveniente, costuma-se definir um outro coeficiente, dado por

$$T = \sqrt{\frac{\chi^2/n}{(r-1)(s-1)}}, \quad (4.6)$$

que atinge o máximo igual a 1 se  $r = s$ .

**No exemplo:**

$$C = 0,32 \text{ e } T = 0,14.$$