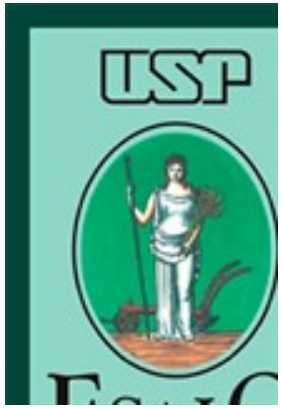


Escola Superior de Agricultura
"Luiz de Queiroz"
Universidade de São Paulo

LCE2112 – Estatística Aplicada às Ciências Sociais e Ambientais

Taciana Villela Savian
Sala 304, pav. Engenharia, ramal 237
tvsavian@usp.br
tacianavillela@gmail.com

Já vimos

- **Tipos de variáveis**
  - Qualitativas nominais (sexo, cor dos olhos, espécies arbóreas, etc...);
  - Qualitativas ordinais (grau de instrução, estado civil, grau de severidade de doenças, etc...);
  - Quantitativas discretas (número de filhos, número de insetos/folha, número de empresas sustentáveis, etc..);
  - Quantitativas contínuas (peso, altura, volume, diâmetro, produção, produtividade, etc....).
- **Construção de Tabelas de Distribuição de Frequências para os três primeiros tipos de variáveis**.

Análise Exploratória – Aula 1

Tabela 2. Distribuição de frequência da capacidade de uso do solo em diferentes localidades em Piracicaba em 2016 .

Uso do solo	Frequência
SR	4
RCA	7
P	9
RCP	4
F	6
Total	30

O que se pode dizer sobre a capacidade de uso do solo com base nesses dados?

Análise Exploratória – Aula 1

Qualitativas (nominal ou **ordinal**): as diferentes respostas são as próprias linhas da tabela.

Levantamento do grau de escolaridade de 20 funcionários de uma empresa (EM=ensino médio; ESI=ensino superior incompleto; ESC=ensino superior completo)

Tabela 3. Distribuição de frequência.

Grau de escolaridade	Frequência
EM	8
ESI	6
ESC	6
Total	20

Análise Exploratória – Aula 1

Quantitativas discretas (contagens): quando assume um número pequeno de valores (até 5 valores diferentes), cada valor se constitui em uma linha da tabela;

Tabela 4. Distribuição de frequência.

Número de brotos	Frequência
0	7
1	11
2	14
3	8
Total	40

Análise bidimensional

- Nem sempre estamos interessados em avaliar uma variável por vez;
- Analisar a distribuição de duas ou mais variáveis conjuntamente;
- Quando consideramos duas variáveis, podemos ter basicamente 3 situações:
  - As duas variáveis qualitativas (Tabelas de Contingência e Coeficiente de Contingência Pearson Corrigido);
  - As duas variáveis quantitativas (gráficos de dispersão e Coeficiente de Correlação de Pearson);
  - Uma das variáveis qualitativa e a outra quantitativa (Tabelas de Contingência e Coeficiente de Contingência Pearson Corrigido);

Análise bidimensional

- As duas variáveis qualitativas

  Espécies da fauna terrestre brasileira ameaçadas de extinção: Grupos taxonômicos (qualitativa nominal) e Biomas (qualitativa nominal).

  Tabela de Contingência = Tabela de dupla entrada considerando os níveis de uma das variáveis nas colunas da tabela e os níveis da outra variável nas linhas da tabela;

Análise bidimensional

- Tabela de Contingência

Tabela 1. Tabela de Distribuição de frequências conjunta do número de espécies da fauna terrestre brasileira ameaçada de extinção, por grupos taxonômicos e biomas. Brasil/2008.

Bioma	Grupos Taxonômicos			Total
	Mamíferos	Aves	Répteis	
Amazônia	85	20	6	111
Cerrado	16	48	15	79
Caatinga	10	25	1	36
Mata Atlântica	38	112	3	153
Pantanal	14	23	15	52
Pampas	5	20	17	42
Áreas Costeiras	8	16	6	30
Total	176	264	63	503

Análise bidimensional

- Existe ou não relação entre as duas variáveis qualitativas?
  - Grupo taxonômico x Biomas
- Qual a magnitude dessa relação?
  - Fraca;
  - Moderada;
  - Forte;
- Coeficiente de Contingência de Pearson Corrigido (C*)

Medida de Associação

- Coeficiente de Contingência de Pearson (C)

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

n é o tamanho da amostra; χ^2 é a estatística de quiquadrado.

χ^2 baseia-se na diferença entre as frequências observadas (fo_{ij}) e frequências esperadas (fe_{ij}) caso as variáveis não sejam relacionadas.

Análise bidimensional

- Tabela de Contingência

Tabela 1. Tabela de Distribuição de frequências conjunta do número de espécies da fauna terrestre brasileira ameaçada de extinção, por grupos taxonômicos e biomas. Brasil/2008.

Bioma	Grupos Taxonômicos			Total
	Mamíferos	Aves	Répteis	
Amazônia	85	20	6	111
Cerrado	16	48	15	79
Caatinga	10	25	1	36
Mata Atlântica	38	112	3	153
Pantanal	14	23	15	52
Pampas	5	20	17	42
Áreas Costeiras	8	16	6	30
Total	176	264	63	503

Medida de Associação

- fo_{ij} indica a frequência observada na linha i e coluna j na tabela de contingência; EX: $fo_{13}=6$
  - $fo_{i.}$ Indica o total da linha i; EX: $fo_{1.}=111$
  - $fo_{.j}$ Indica o total da coluna j; EX: $fo_{.1}=176$
  - $fo_{..}$ Indica o total geral da tabela; EX: $fo_{..}=503$
- fe_{ij} indica a frequência esperada na linha i e coluna j, na tabela de contingência, caso as variáveis não sejam relacionadas.

Medida de Associação

Quem são fe_{ij} ?

$$fe_{ij} = \frac{fo_{i.} \times fo_{.j}}{fo_{..}}$$

Por exemplo, quem é fe_{23}?

$$fe_{23} = \frac{fo_{2.} \times fo_{.3}}{fo_{..}}$$

Análise bidimensional

- Tabela de Contingência

Tabela 1. Tabela de Distribuição de frequências conjunta do número de espécies da fauna terrestre brasileira ameaçada de extinção, por grupos taxonômicos e biomas. Brasil/2008.

Bioma	Grupos Taxonômicos			Total
	Mamíferos	Aves	Répteis	
Amazônia	85	20	6	111
Cerrado	16	48	fo_{23}=15 (fe_{23}=9,9)	79
Caatinga	10	25	1	36
Mata Atlântica	38	112	3	153
Pantanal	14	23	15	52
Pampas	5	20	17	42
Áreas Costeiras	8	16	6	30
Total	176	264	63	503

Análise bidimensional

- Tabela de Contingência

Tabela 1. Tabela de Distribuição de frequências conjunta do número de espécies da fauna terrestre brasileira ameaçada de extinção, por grupos taxonômicos e biomas. Brasil/2008.

Bioma	Grupos Taxonômicos			Total
	Mamíferos	Aves	Répteis	
Amazônia	85 (38,8)	20 (58,2)	6 (13,9)	111
Cerrado	16 (27,6)	48 (41,5)	15 (9,9)	79
Caatinga	10 (12,6)	25 (18,9)	1 (4,5)	36
Mata Atlântica	38 (53,5)	112 (80,3)	3 (19,1)	153
Pantanal	14 (18,2)	23 (27,3)	15 (6,5)	52
Pampas	5 (14,7)	20 (22,0)	17 (5,3)	42
Áreas Costeiras	8 (10,5)	16 (15,7)	6 (3,7)	30
Total	176	264	63	503

Medida de Associação

$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{\left(fo_{ij} - fe_{ij} \right)^2}{fe_{ij}}$$

$$\chi^2 = \frac{\left(fo_{11} - fe_{11} \right)^2}{fe_{11}} + \ldots + \frac{\left(fo_{73} - fe_{73} \right)^2}{fe_{73}}$$

Medida de Associação

$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{\left(fo_{ij} - fe_{ij} \right)^2}{fe_{ij}}$$

$$\chi^2 = \frac{\left(85 - 38,8\right)^2}{38,8} + \ldots + \frac{\left(6 - 3,7\right)^2}{3,7} = 176,36$$

Coeficiente de Contingência de Pearson (C)

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{176,36}{176,36 + 503}} = 0,5095$$

Medidas de Associação

$$\chi^2 = \frac{(85-38,8)^2}{38,8} + \ldots + \frac{(6-3,7)^2}{3,7} = 176,36$$

$\chi^2 = 0$ somente quando $fo_{ij} = fe_{ij}$, indicando a não associação entre as variáveis;

$\chi^2 > 0$ significa que $fo_{ij} \neq fe_{ij}$ e temos indicação de associação entre as variáveis;

Problema: χ^2 é um valor que pode variar de zero a infinito.

Medidas de Associação

- Coeficiente de Contingência de Pearson Corrigido (C*)

$$C^* = \frac{C}{\sqrt{\frac{t-1}{t}}}$$

em que: *t* é o mínimo entre o número de linhas (s) e o número de colunas (r) da tabela de distribuição de frequência conjunta (tabela de contingência), ou seja,

t = min(s,r).

O Coeficiente C* assume valores entre 0 e 1 e quanto mais próximo de um mais forte é a associação entre as variáveis qualitativas estudadas.

Medidas de Associação

- Para o exemplo:

$$C^* = \frac{C}{\sqrt{\frac{t-1}{t}}} = \frac{0,5095}{\sqrt{\frac{3-1}{3}}} = 0,6240$$

Conclusão: O Coeficiente de Contingência de Pearson Corrigido (C*) indica uma associação **moderada à forte** entre as variáveis: Bioma e Grupos Taxonômicos em relação ao número de espécies da fauna terrestre brasileira ameaçadas de extinção.

Medidas de Associação

Verifique se existe uma relação (e qual a magnitude) entre as variáveis Grau de Instrução e Região de Procedência dos 30 funcionário da empresa.

Tabela 2. Tabela de Distribuição de frequências conjunta do Grau de Instrução e Região de Procedência de 30 funcionários de uma empresa de Piracicaba, em 2015.

Procedência	Grau de Instrução			Total
	Ensino Médio	Superior Completo	Superior Incompleto	
Interior	2	10	4	16
Capital	1	7	6	14
Total	3	17	10	30

Fórmulas

$$fe_{ij} = \frac{fo_{i.} \text{ x } fo_{.j}}{fo_{..}} \quad \textit{(Frequência esperada)}$$

$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{\left(fo_{ij} - fe_{ij}\right)^2}{fe_{ij}} \quad \textit{(Estatística de Qui-quadrado)}$$

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \quad \textit{(Coeficiente de Contingência de Pearson)}$$

$t = \min(s,r)$ *(mínimo entre número de linhas e colunas da tabela)*

$$C^* = \frac{C}{\sqrt{(t-1)/t}} \quad \textit{(Coeficiente de Contingência de Pearson Corrigido)}$$

Resultado Exercício

1º Passo) Cálculo das frequências esperadas (fe_{ij}) se não houvesse uma associação entre as variáveis.

Tabela 2. Tabela de Distribuição de frequências conjunta do Grau de Instrução e Região de Procedência de 30 funcionários de uma empresa de Piracicaba, em 2015.

Procedência	Grau de Instrução			Total
	Ensino Médio	Superior Completo	Superior Incompleto	
Interior	2 (1,6)	10 (9,1)	4 (5,3)	16
Capital	1 (1,4)	7 (7,9)	6 (4,7)	14
Total	3	17	10	30

Resultado Exercício

2º Passo) Cálculo da Estatística de Quiquadrado (χ^2)

Tabela 2. Tabela de Distribuição de frequências conjunta do Grau de Instrução e Região de Procedência de 30 funcionários de uma empresa de Piracicaba, em 2015.

Procedência	Grau de Instrução			Total
	Ensino Médio	Superior Completo	Superior Incompleto	
Interior	2 (1,6)	10 (9,1)	4 (5,3)	16
Capital	1 (1,4)	7 (7,9)	6 (4,7)	14
Total	3	17	10	30

$$\chi^2 = \frac{(2-1{,}6)^2}{1{,}6} + \frac{(10-9{,}1)^2}{9{,}1} \ldots + \frac{(6-4{,}7)^2}{4{,}7} = 1{,}08$$

Resultado Exercício

3º Passo) Cálculo do Coeficiente de Contingência de Pearson (C)

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

$$C = \sqrt{\frac{1{,}08}{1{,}08 + 30}} = 0{,}1864$$

Resultado Exercício

4º Passo) Correção do Coeficiente de Contingência de Pearson (C*) e concluir

$$C^* = \frac{C}{\sqrt{\frac{t-1}{t}}}$$

$$C^* = \frac{0{,}1864}{\sqrt{\frac{2-1}{2}}} = 0{,}2636$$

Conclusão: O Coeficiente de Contingência de Pearson Corrigido (C*) indica uma associação **fraca** entre as variáveis: Grau de Instrução e Região de Procedência dos funcionários da empresa de Piracicaba.

Análise bidimensional

- As duas variáveis quantitativas (gráficos de dispersão e Coeficiente de Correlação de Pearson);

- Gráfico de Dispersão = Gráfico de pontos onde plotamos os pares de dados (x,y) em que x é uma das variáveis quantitativas e y é a outra variável quantitativa;

Análise bidimensional

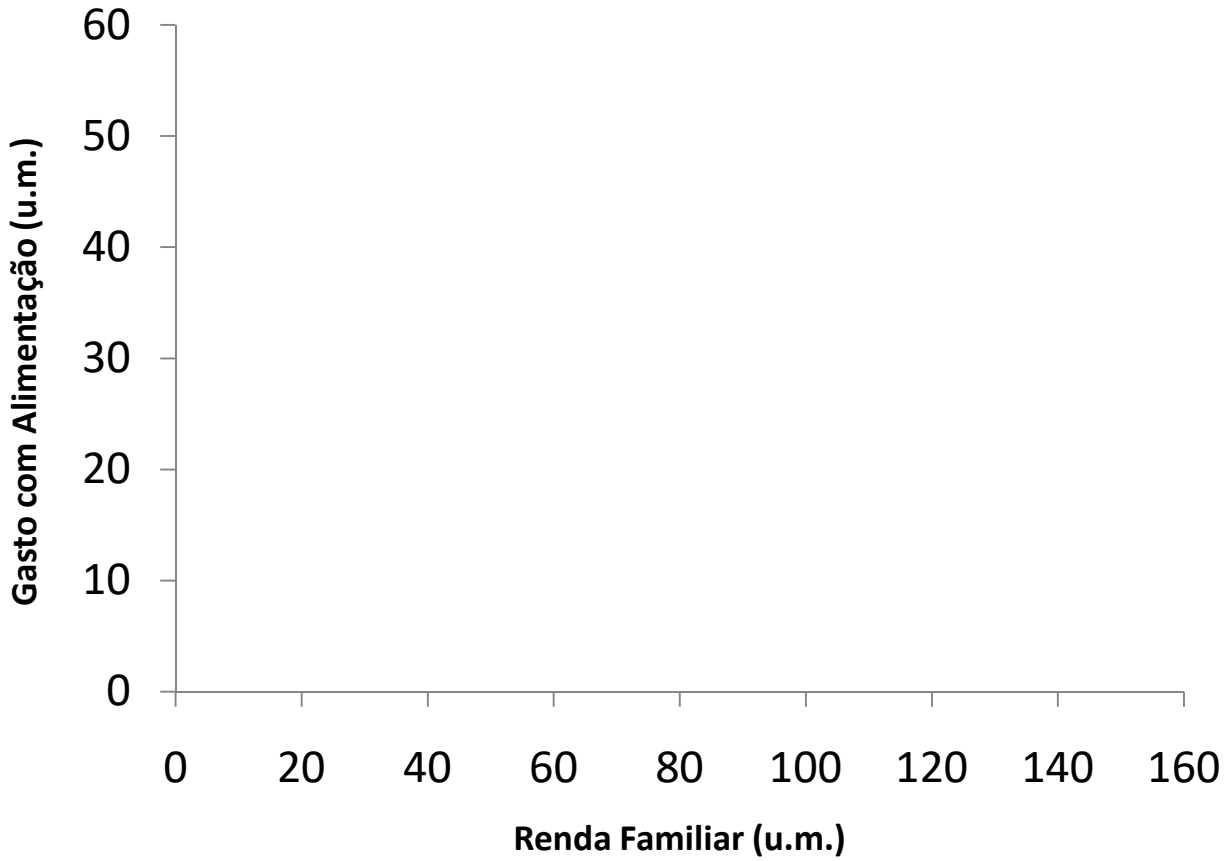
- Exemplo: Os dados a seguir correspondem à variável renda familiar e gasto com alimentação (em unidades monetárias) para uma amostra de 10 famílias.

Renda	3	5	10	20	20	40	60	70	120	150
Gasto	1,5	2	6	12	15	10	20	25	40	50

Gráfico de Dispersão = Gráfico de pontos onde plotamos os pares de dados (x,y) em que x é a renda familiar y é o gasto com alimentação

Análise bidimensional

- Gráfico de Dispersão



Análise bidimensional

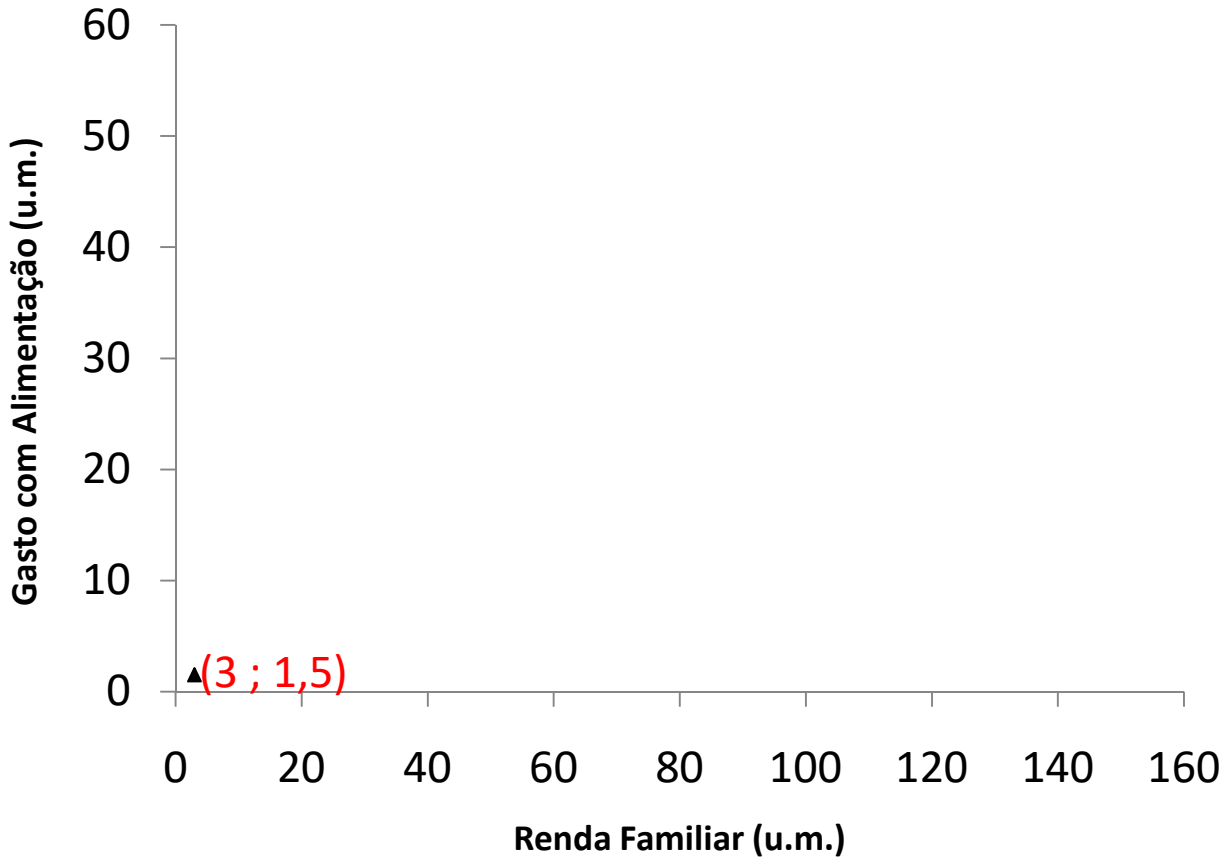
- Exemplo: Os dados a seguir correspondem à variável renda familiar e gasto com alimentação (em unidades monetárias) para uma amostra de 10 famílias.

Renda	3	5	10	20	20	40	60	70	120	150
Gasto	1,5	2	6	12	15	10	20	25	40	50

Gráfico de Dispersão = Gráfico de pontos onde plotamos os pares de dados (x,y) em que x é a renda familiar y é o gasto com alimentação

Análise bidimensional

- Gráfico de Dispersão



Análise bidimensional

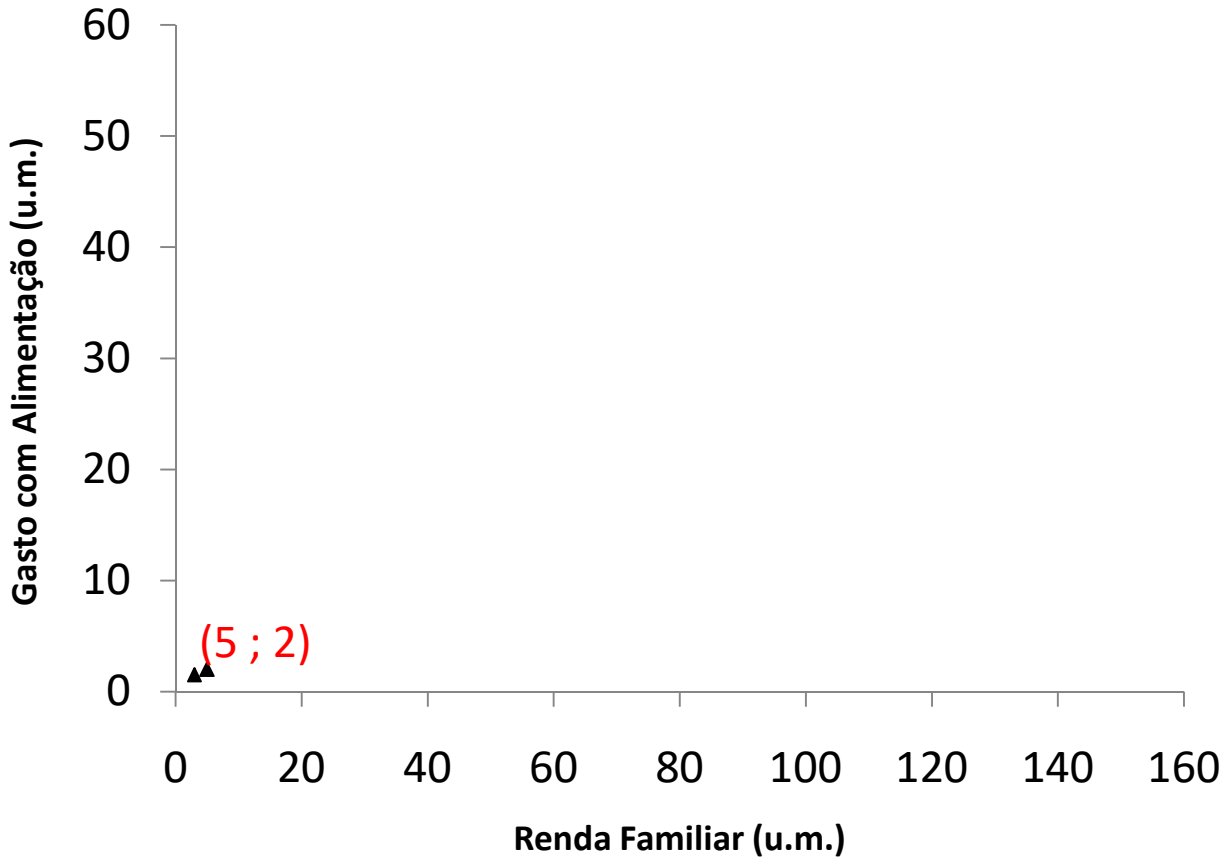
- Exemplo: Os dados a seguir correspondem à variável renda familiar e gasto com alimentação (em unidades monetárias) para uma amostra de 10 famílias.

Renda	3	5	10	20	20	40	60	70	120	150
Gasto	1,5	2	6	12	15	10	20	25	40	50

Gráfico de Dispersão = Gráfico de pontos onde plotamos os pares de dados (x,y) em que x é a renda familiar y é o gasto com alimentação

Análise bidimensional

- Gráfico de Dispersão



Análise bidimensional

- Exemplo: Os dados a seguir correspondem à variável renda familiar e gasto com alimentação (em unidades monetárias) para uma amostra de 10 famílias.

Renda	3	5	10	20	20	40	60	70	120	150
Gasto	1,5	2	6	12	15	10	20	25	40	50

Gráfico de Dispersão = Gráfico de pontos onde plotamos os pares de dados (x,y) em que x é a renda familiar y é o gasto com alimentação

Análise bidimensional

- Gráfico de Dispersão

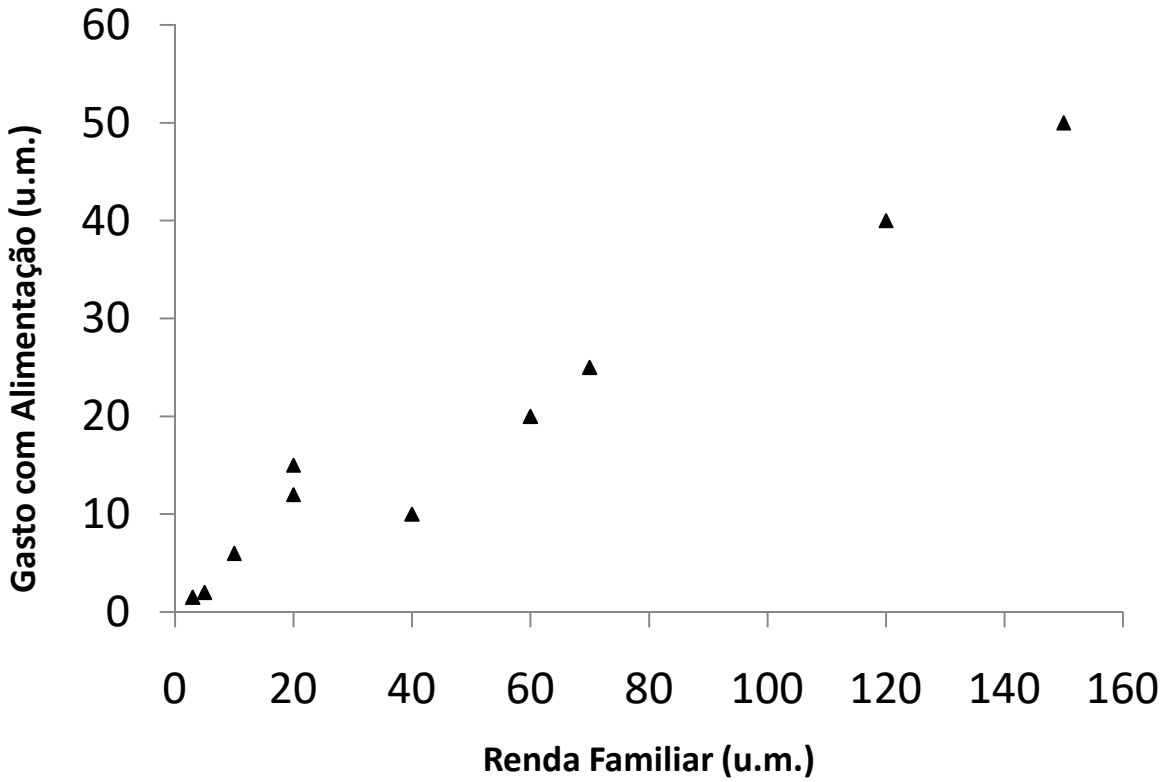


Figura 1. Gráfico de dispersão do gasto com alimentação e renda familiar (em unidades monetárias) de uma amostra de 10 famílias.

Análise bidimensional

- Exemplo: Os dados a seguir correspondem à variável renda familiar e gasto com alimentação (em unidades monetárias) para uma amostra de 10 famílias.

Renda	3	5	10	20	20	40	60	70	120	150
Gasto	1,5	2	6	12	15	10	20	25	40	50

Quantificar a associação entre essas duas variáveis?

Coeficiente de Correlação de Pearson (r)

Análise bidimensional

- **Coeficiente de Correlação de Pearson (r)**

$$r = \frac{SPXY}{\sqrt{(SQX)(SQY)}}$$

Em que: SPXY – Soma de Produtos de X e Y;

SQX – Soma de Quadrados de X;

SQY – Soma de Quadrados de Y;

Como calcular cada um dos termos?

Análise bidimensional

- **Coeficiente de Correlação de Pearson (r)**

$$SPXY = \sum_{i=1}^{n} X_i Y_i - \frac{(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n}$$

X – Renda	3	5	10	20	20	40	60	70	120	150
Y – Gasto	1,5	2	6	12	15	10	20	25	40	50
XY	4,5	10	60	240	300	400	1200	1750	4800	7500

$$\sum_{i=1}^{n} X_i = 3 + 5 + \cdots + 150 = 498$$

$$\sum_{i=1}^{n} Y_i = 1,5 + 2 + \cdots + 50 = 181,5$$

$$\sum_{i=1}^{n} X_i Y_i = 4,5 + 10 + \cdots + 7500 = 16.264,5$$

Análise bidimensional

- **Coeficiente de Correlação de Pearson (r)**

$$SPXY = 16.264,5 - \frac{498 \times 181,5}{10} = 7.245,8$$

X – Renda	3	5	10	20	20	40	60	70	120	150
Y – Gasto	1,5	2	6	12	15	10	20	25	40	50
XY	4,5	10	60	240	300	400	1200	1750	4800	7500

$$\sum_{i=1}^{n} X_i = 3 + 5 + \cdots + 150 = 498$$

$$\sum_{i=1}^{n} Y_i = 1,5 + 2 + \cdots + 50 = 181,5$$

$$\sum_{i=1}^{n} X_i Y_i = 4,5 + 10 + \cdots + 7500 = 16.264,5$$

Análise bidimensional

- **Coeficiente de Correlação de Pearson (r)**

$$SQX = \sum_{i=1}^{n} X_i^{\,2} - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}$$

X – Renda	3	5	10	20	20	40	60	70	120	150
Y – Gasto	1,5	2	6	12	15	10	20	25	40	50
XY	4,5	10	60	240	300	400	1200	1750	4800	7500

$$\sum_{i=1}^{n} X_i = 3 + 5 + \cdots + 150 = 498$$

$$\sum_{i=1}^{n} X_i^{\,2} = 3^2 + 5^2 + \cdots + 150^2 = 47.934$$

$$SQX = 47.934 - \frac{(498)^2}{10} = 23.133,6$$

Análise bidimensional

- **Coeficiente de Correlação de Pearson (r)**

$$SQY = \sum_{i=1}^{n} Y_i^{\,2} - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}$$

X – Renda	3	5	10	20	20	40	60	70	120	150
Y – Gasto	1,5	2	6	12	15	10	20	25	40	50
XY	4,5	10	60	240	300	400	1200	1750	4800	7500

$$\sum_{i=1}^{n} Y_i = 1,5 + 2 + \cdots + 50 = 181,5$$

$$\sum_{i=1}^{n} Y_i^{\,2} = 1,5^2 + 2^2 + \cdots + 50^2 = 5.636,2$$

$$SQY = 5.636,2 - \frac{(181,5)^2}{10} = 2.342,0$$

Análise bidimensional

- **Coeficiente de Correlação de Pearson (r)**

$$r = \frac{SPXY}{\sqrt{(SQX)(SQY)}}$$

Em que: SPXY – Soma de Produtos de X e Y;

SQX – Soma de Quadrados de X;

SQY – Soma de Quadrados de Y;

$$r = \frac{SPXY}{\sqrt{(SQX)(SQY)}} = \frac{7.245{,}8}{\sqrt{(23.133{,}6)(2.342{,}0)}} = 0{,}98$$

Análise bidimensional

- **Coeficiente de Correlação de Pearson (r)**

$$r = \frac{SPXY}{\sqrt{(SQX)(SQY)}}$$

Propriedades:

- **r é sempre um número entre -1 e 1;**
- **r = 0 não indica independência entre as variáveis, indica apenas que não existe uma relação LINEAR entre as variáveis;**
- **|r| próximo a 1, indica alta associação entre as variáveis (positiva/negativa)**
- **|r| próximo a 0, indica não relação linear entre as variáveis;**
- **|r| próximo a 0,5, indica associação moderada.**

Análise bidimensional

- **Coeficiente de Correlação de Pearson (r)**

$$r = \frac{SPXY}{\sqrt{(SQX)(SQY)}}$$

$$r = \frac{SPXY}{\sqrt{(SQX)(SQY)}} = \frac{7.245{,}8}{\sqrt{(23.133{,}6)(2.342{,}0)}} = 0{,}98$$

Segundo o resultado da correlação obtida (r=0,98), pode-se notar que há uma forte correlação linear entre as variáveis renda familiar e gasto com alimentação. Nota-se que à medida que a renda familiar aumenta o gasto com alimentação (em unidades monetárias) também aumenta, o que é coerente com o gráfico de dispersão apresentada anteriormente.

Análise bidimensional

Os quatro conjuntos de dados a seguir foram preparados pelo estatístico F. J. Anscombe e são usados com frequência em aulas sobre correlação.

Para cada conjunto:

a) Fazer o diagrama de dispersão;

b) Obter o coeficiente de correlação de Pearson;

c) Interpretar o resultado.

Conjunto 1		Conjunto 2		Conjunto 3		Conjunto 4	
X	Y	X	Y	X	Y	X	Y
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,10	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,10	4	5,39	19	12,50
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

Análise bidimensional

Os quatro conjuntos de dados a seguir foram preparados pelo estatístico F. J. Anscombe e são usados com frequência em aulas sobre correlação.

Para cada conjunto:

a) Fazer o diagrama de dispersão;

b) Obter o coeficiente de correlação de Pearson;

c) Interpretar o resultado.

Conjunto 1		Conjunto 2		Conjunto 3		Conjunto 4	
X	Y	X	Y	X	Y	X	Y
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,10	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,10	4	5,39	19	12,50
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89
r=0,82		r=0,82		r=0,82		r=0,82	