# Modulo II - Introdução à Estatística Básica

Umberto Mignozzetti

6/1/2020

# Modulo II

# Análise de dados

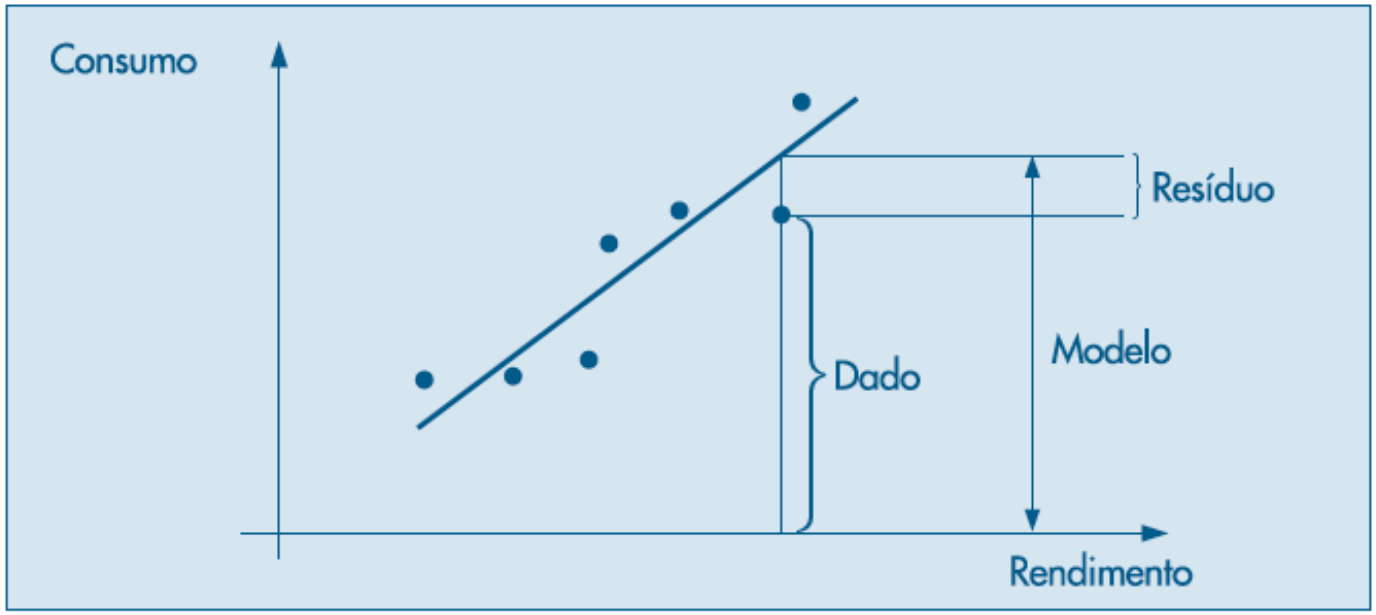
# Análise de dados

- Objetivo Stat é analisar dados.
- Três etapas:
  1. Entender os dados: análise descritiva
  2. Modelar os dados: probabilidade
  3. Formular hipóteses: inferencia estatística

# Análise de dados

- Modelagem: propor uma representação que explique a maior parte da variabilidade dos dados

**Figura 1.1:** Relação entre consumo e rendimento.




Podemos, então, escrever de modo esquemático:

$$\text{Dados} = \text{Modelo} + \text{Resíduos}$$

Figure 1: f1

# Análise de dados

- Gráficos: visualizar os dados que temos.
- Objetivos:
  - Buscar padrões
  - Checar expectativas
  - Descobrir fenômenos
  - Confirmar suposições
  - Apresentar resultados
- Altamente recomendável!

# Análise de dados

- Softwares estatísticos:
  - R / S+
  - SPSS / PSPP
  - Excel / Calc
  - SAS
  - Stata
- Qual usar? Qual vc preferir. (Esse tipo de pergunta importa?!)
- Eu uso R. Motivo: de graça e bom!

# Medidas Resumo

# Tipos de variáveis

- Qualitativas: descrevem atributos dos casos:
  - Pessoa casada
  - Votou no Bolsonaro
  - Cidade com mais Corona no Brasil
  - Superior completo...
- Quantitativas: realizações de uma contagem / mensuração
  - Idade
  - Renda
  - Numero de ligações

# Tipos de variáveis

- Qualitativas:
  - Nominais: sexo
  - Ordinais: escolaridade
- Quantitativas:
  - Discretas: numero de filhos
  - Contínuas: salário

# Tipos de variáveis

Classifique o banco:

```
dat <- read.csv('https://raw.githubusercontent.com/umberto
head(dat)
```

```
##   N Estado.Civil  Grau.de.Instrução N.de.Filhos Salario
## 1 1     solteiro ensino fundamental          NA
## 2 2       casado ensino fundamental           1
## 3 3       casado ensino fundamental           2
## 4 4     solteiro       ensino médio          NA
## 5 5     solteiro ensino fundamental          NA
## 6 6       casado ensino fundamental           0
##   Região.de.Procedência
## 1              interior
## 2               capital
## 3               capital
## 4                 outra
## 5                 outra
## 6              interior
```

# Tabela de frequência

- Contagem de valores para cada um dos níveis pré-definidos
- E.g., Grau de Instrução:

**Tabela 2.2:** Freqüências e porcentagens dos 36 empregados da seção de orçamentos da Companhia MB segundo o grau de instrução.

| Grau de instrução | Freqüência $n_i$ | Proporção $f_i$ | Porcentagem $100\ f_i$ |
|---|---|---|---|
| Fundamental | 12 | 0,3333 | 33,33 |
| Médio | 18 | 0,5000 | 50,00 |
| Superior | 6 | 0,1667 | 16,67 |
| Total | 36 | 1,0000 | 100,00 |

Figure 2: f2

# Tabela de frequência

**Tabela 2.2:** Freqüências e porcentagens dos 36 empregados da seção de orçamentos da Companhia MB segundo o grau de instrução.

| Grau de instrução | Freqüência $n_i$ | Proporção $f_i$ | Porcentagem $100\ f_i$ |
|---|---|---|---|
| Fundamental | 12 | 0,3333 | 33,33 |
| Médio | 18 | 0,5000 | 50,00 |
| Superior | 6 | 0,1667 | 16,67 |
| Total | 36 | 1,0000 | 100,00 |

Figure 3: f2

- Stats:
  - Contagem
  - Frequencia (relativa): $f_i = \frac{n_i}{n}$
  - Porcentagem: $prop_i = 100 \times \frac{n_i}{n}$

# Tabela de frequência

- Para uma variável quanti, temos o seguinte:
  1. Criamos intervalos
  2. Contamos valores nos intervalos

**Tabela 2.4:** Freqüências e porcentagens dos 36 empregados da seção de orçamentos da Companhia MB por faixa de salário.

| Classe de salários | Freqüência $n_i$ | Porcentagem $100\ f_i$ |
|---|---|---|
| 4,00 ⊢ 8,00 | 10 | 27,78 |
| 8,00 ⊢ 12,00 | 12 | 33,33 |
| 12,00 ⊢ 16,00 | 8 | 22,22 |
| 16,00 ⊢ 20,00 | 5 | 13,89 |
| 20,00 ⊢ 24,00 | 1 | 2,78 |
| Total | 36 | 100,00 |

Fonte: Tabela 2.1.

Figure 4: f3

# Gráficos

- Basta colocar as tabelas que montamos em figuras!?

**Figura 2.2:** Gráfico em barras para a variável $Y$: grau de instrução.

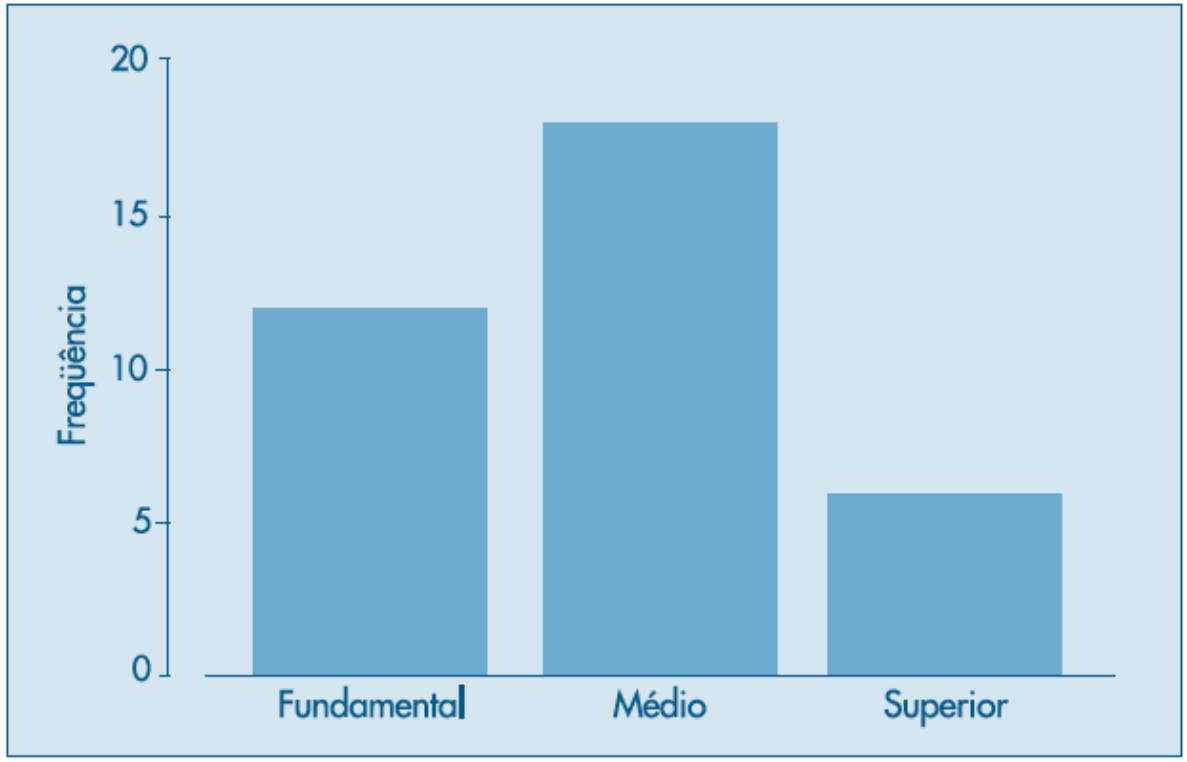
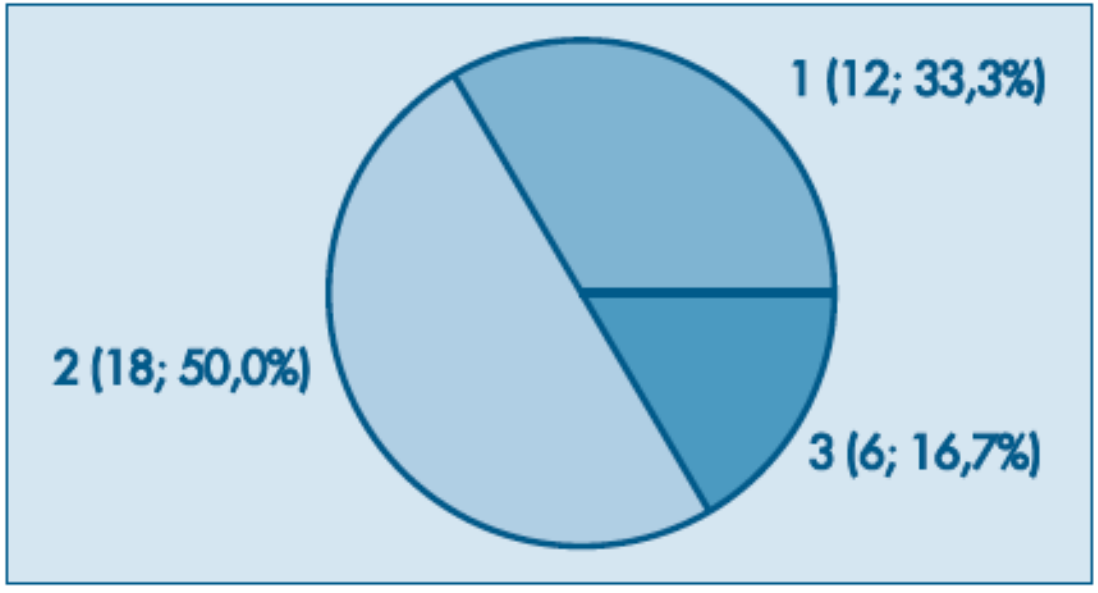



Figure 5: f4

# Gráficos

- Basta colocar as tabelas que montamos em figuras!?

**Figura 2.3:** Gráfico em setores para a variável *Y*: grau de instrução.




1 = Fundamental, 2 = Médio e 3 = Superior

Figure 6: f5

# Gráficos

- Basta colocar as tabelas que montamos em figuras!?

**Figura 2.7:** Histograma da variável *S*: salários.

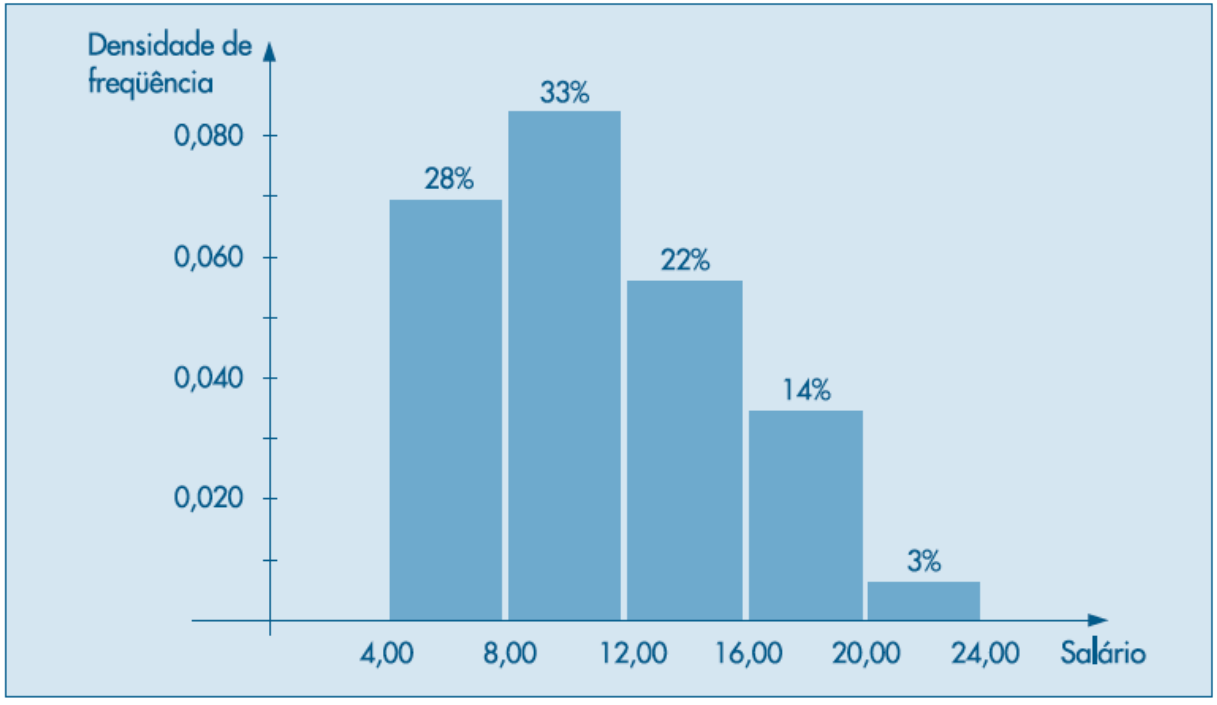



Figure 7: f6

# Exercício

6. As taxas médias geométricas de incremento anual (por 100 habitantes) dos 30 maiores municípios do Brasil estão dadas abaixo.

| | | | | |
|---|---|---|---|---|
| 3,67 | 1,82 | 3,73 | 4,10 | 4,30 |
| 1,28 | 8,14 | 2,43 | 4,17 | 5,36 |
| 3,96 | 6,54 | 5,84 | 7,35 | 3,63 |
| 2,93 | 2,82 | 8,45 | 5,28 | 5,41 |
| 7,77 | 4,65 | 1,88 | 2,12 | 4,26 |
| 2,78 | 5,54 | 0,90 | 5,09 | 4,07 |

(a) Construa um histograma.

Figure 8: f7

# Medidas-Resumo

# Medidas Resumo

- Dois tipos mais importantes:
  - Posição
  - Dispersão
- Além dessas, temos algumas outras que são boas para analisar os dados.

# Medidas de posição

- Média:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Média (com frequências relativas):

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}$$

- Exercício: calcule a média dos dados: 1,5,2,3,2,4,10

# Medidas de posição

- Posição e medidas de ordem: em que lugar está o dado se ordenarmos?

Consideremos, agora, as observações ordenadas em ordem crescente. Vamos denotar a menor observação por $X_{(1)}$, a segunda por $X_{(2)}$, e assim por diante, obtendo-se

$$X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n-1)} \leqslant X_{(n)}. \qquad (3.4)$$

Por exemplo, se $X_1 = 3$, $X_2 = -2$, $X_3 = 6$, $X_4 = 1$, $X_5 = 3$, então $-2 \leqslant 1 \leqslant 3 \leqslant 3 \leqslant 6$, de modo que $X_{(1)} = -2$, $X_{(2)} = 1$, $X_{(3)} = 3$, $X_{(4)} = 3$ e $X_{(5)} = 6$.

Figure 9: f8

- Ex.:

```
x <- c(3,-2,6,1,3)
x
```

```
## [1]  3 -2  6  1  3
```

```
sort(x)
```

# Medidas de posição

- Mediana:

```
x <- c(3,-2,6,1,3)
x
```

```
## [1]  3 -2  6  1  3
```

```
sort(x)
```

```
## [1] -2  1  3  3  6
```

```
median(x)
```

```
## [1] 3
```

$$\mathrm{md}(X) = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ ímpar;} \\ \dfrac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se } n \text{ par.} \end{cases}$$

# Medidas de dispersão

- Suponha as notas dos alunos em cinco grupos:

grupo A (variável $X$): 3, 4, 5, 6, 7
grupo B (variável $Y$): 1, 3, 5, 7, 9
grupo C (variável $Z$): 5, 5, 5, 5, 5
grupo D (variável W): 3, 5, 5, 7
grupo E (variável V): 3, 5, 5, 6, 6

Figure 11: f10

- Exercício: quais são as médias? Elas ajudam a diferenciar esses dados?

# Medidas de dispersão

- Não ajudam nesses casos: os dados acima eram claramente diferentes!
- Duas medidas mais usadas: desvio-médio absoluto e variância.

$$\mathrm{dm}(X) = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n},$$

$$\mathrm{var}(X) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n},$$

Figure 12: f11

- Exercício: vamos fazer no R? Considere os dados do exercício acima.

# Medidas de dispersão: exercício

1. Quer se estudar o número de erros de impressão de um livro. Para isso escolheu-se uma amostra de 50 páginas, encontrando-se o número de erros por página da tabela abaixo.
   (a) Qual o número médio de erros por página?
   (b) E o número mediano?
   (c) Qual é o desvio padrão?
   (d) Faça uma representação gráfica para a distribuição.
   (e) Se o livro tem 500 páginas, qual o número total de erros esperado no livro?

| Erros | Freqüência |
|---|---|
| 0 | 25 |
| 1 | 20 |
| 2 | 3 |
| 3 | 1 |
| 4 | 1 |

Figure 13: f12

# Quantís empíricos

- Apenas com média e desvio-padrão não temos ideia do que está acontecendo nos dados:
  - Valores extremos?
  - Assimetria?
- Quantís: boas medidas de resumo dos dados
- Posição e medidas de ordem: em que lugar está o dado se ordenarmos?

Consideremos, agora, as observações ordenadas em ordem crescente. Vamos denotar a menor observação por $X_{(1)}$, a segunda por $X_{(2)}$, e assim por diante, obtendo-se

$$X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n-1)} \leqslant X_{(n)}. \tag{3.4}$$

Por exemplo, se $X_1 = 3$, $X_2 = -2$, $X_3 = 6$, $X_4 = 1$, $X_5 = 3$, então $-2 \leqslant 1 \leqslant 3 \leqslant 3 \leqslant 6$, de modo que $X_{(1)} = -2$, $X_{(2)} = 1$, $X_{(3)} = 3$, $X_{(4)} = 3$ e $X_{(5)} = 6$.

Figure 14: f8

- Ex.:

# Quantís empíricos

- Quantís: medidas de posição, para uma dada ordem nos dados.
- E.g.: mediana: $q(0.5)$: valor que divide os dados pela metade.
- E.g.: percentil 0.95: $q(0.95)$: valor que divide os dados com 95% dos casos abaixo e 5% acima desse valor.

# Quantís empíricos

```
x <- c(15, 5, 3, 8, 10, 2, 7, 11, 12)
sort(x)
```

```
## [1]  2  3  5  7  8 10 11 12 15
```

```
quantile(x)
```

```
##   0%  25%  50%  75% 100%
##    2    5    8   11   15
```

```
quantile(x, probs = 0.95)
```

```
##  95%
## 13.8
```

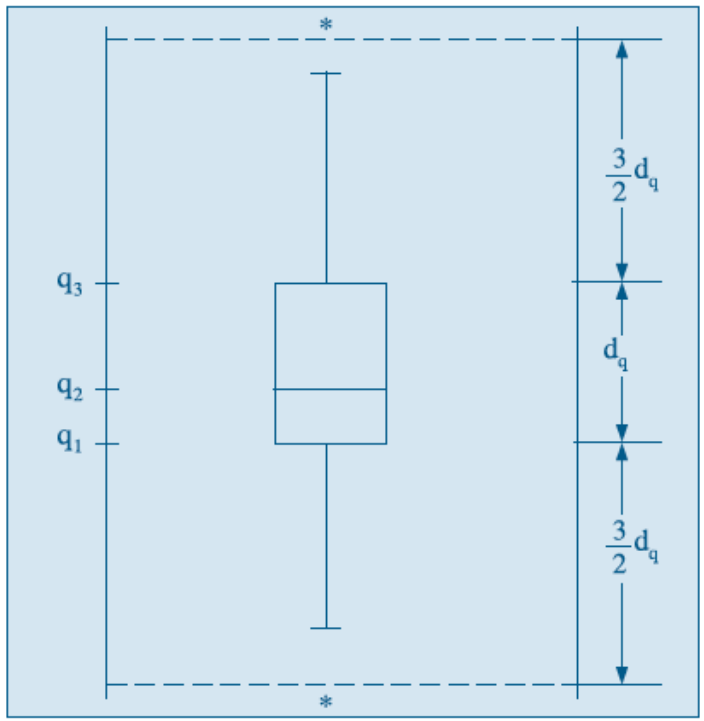
```
summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   5.000   8.000   8.111  11.000  15.000
```

# Quantís empíricos

- Box-plot: jeito de apresentar os quantís que dá uma noção da distribuição e disperção dos dados.
- $LS = MD + 1.5 \times IIQ$
- $LI = MD - 1.5 \times IIQ$
- $IIQ = q(0.75) - q(0.25)$

**Figura 3.4:** *Box Plot.*



# Quantís empíricos

- Motivo estatístico

**Figura 3.8:** Área sob a curva normal entre LI e LS.

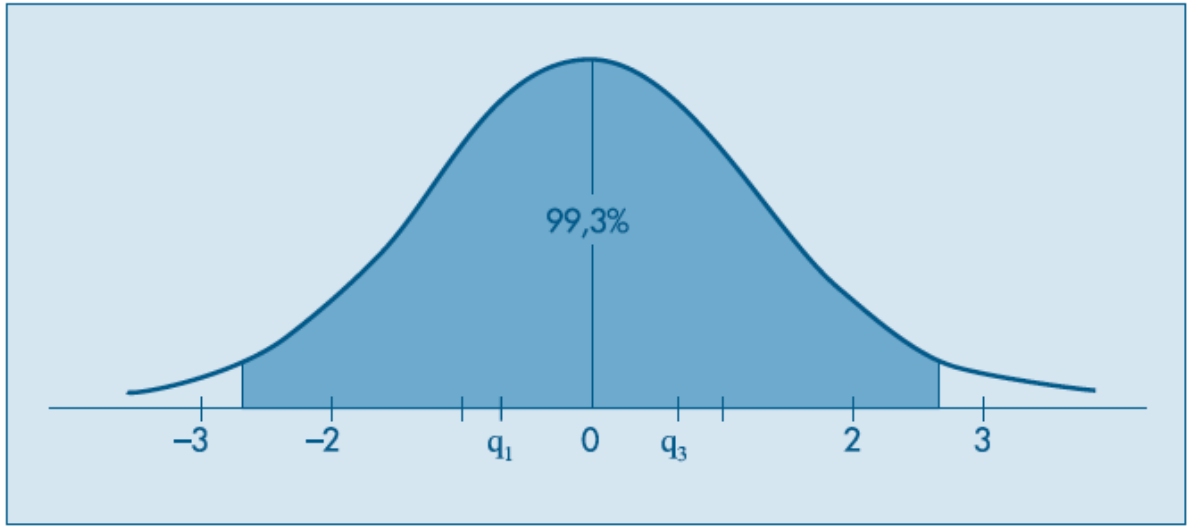



Figure 16: f14

# Exercício

- Faça uma análise dos dados da empresa MB.

# Análise bidimensional

# Análise Bidimensional

- Três tipos:
  - Quali x Quali
  - Quali x Quant
  - Quant x Quant

# Associação Quali-Quali

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

| $V$ \ $Y$ | 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.

Figure 17: f15

# Associação Quali-Quali

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

| $V$ \ $Y$ | 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.

Figure 18: f16

# Associação Quali-Quali

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

| $V$ \ $Y$ | 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.

Figure 19: f17

# Associação Quali-Quali

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

| $Y$ \ $X$ | 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.

Figure 20: f18

# Associação Quali-Quali

**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 da Brasil — IBGE, 1977.

Figure 21: f19

# Associação Quali-Quali

**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.

Figure 22: f20

# Associação Quali-Quali

**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.

Figure 23: f21

# Associação Quali-Quali

6. Uma companhia de seguros analisou a freqüência com que 2.000 segurados (1.000 homens e 1.000 mulheres) usaram o hospital. Os resultados foram:

| | Homens | Mulheres |
|---|---|---|
| Usaram o hospital | 100 | 150 |
| Não usaram o hospital | 900 | 850 |

(a) Calcule a proporção de homens entre os indivíduos que usaram o hospital.
(b) Calcule a proporção de homens entre os indivíduos que não usaram o hospital.
(c) O uso do hospital independe do sexo do segurado?

Figure 24: f22b

# Quanti x Quanti

# Quanti x Quanti: Correlação

**Figura 4.2:** Gráfico de dispersão para as variáveis $X$: anos de serviço e $Y$: número de clientes.

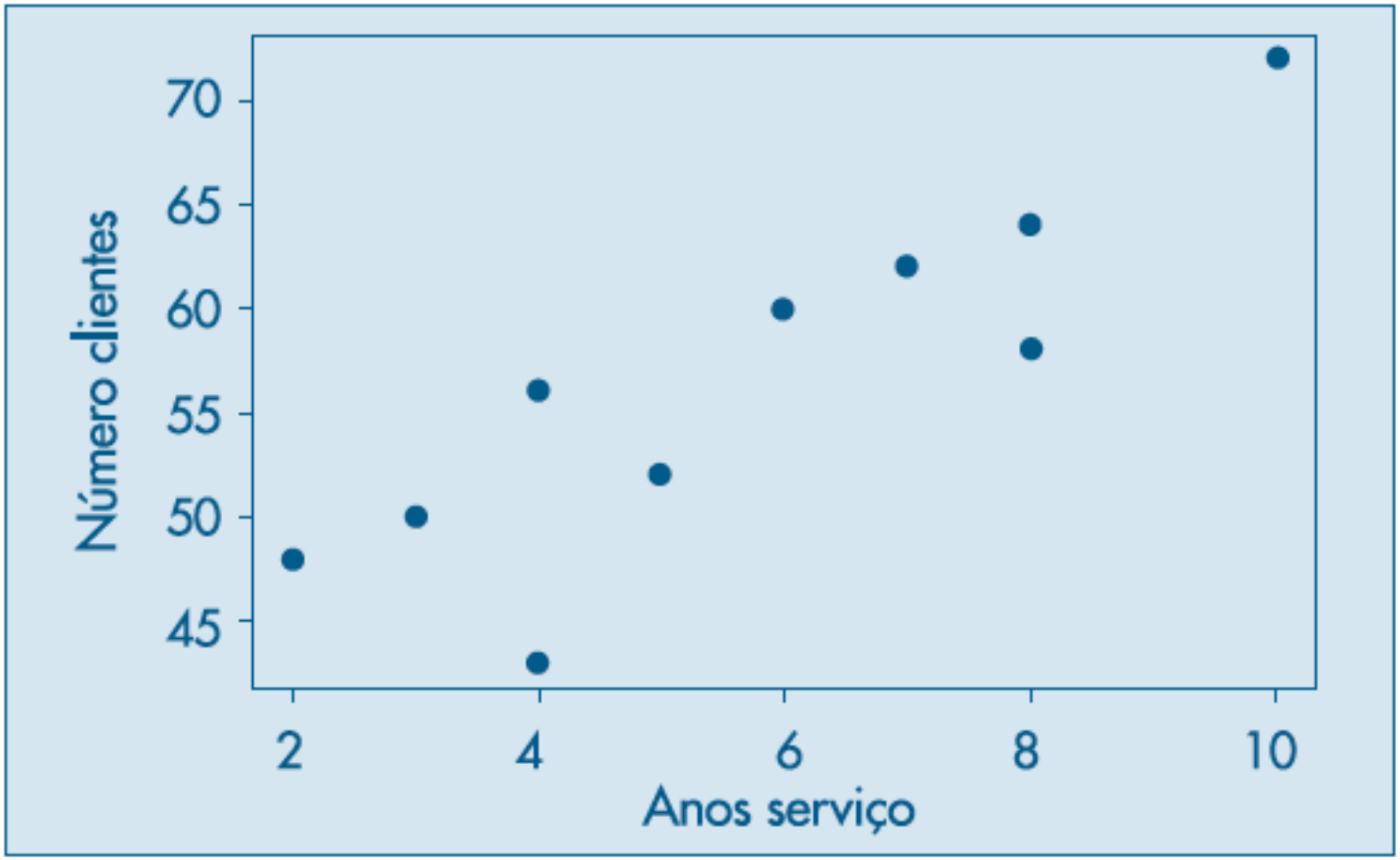



Figure 25: f22

# Quanti x Quanti: Correlação

**Figura 4.6:** Tipos de associações entre duas variáveis.

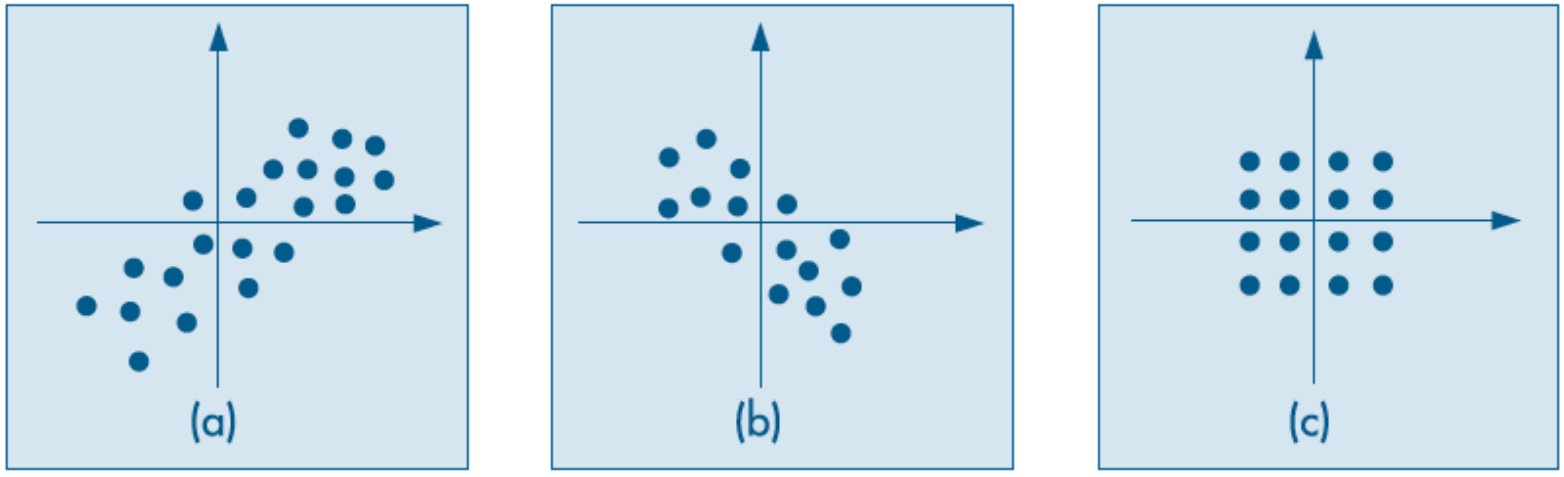



Figure 26: f23

# Quanti x Quanti: Correlação

**Figura 4.7:** Mudança de escalas para o cálculo do coeficiente de correlação.

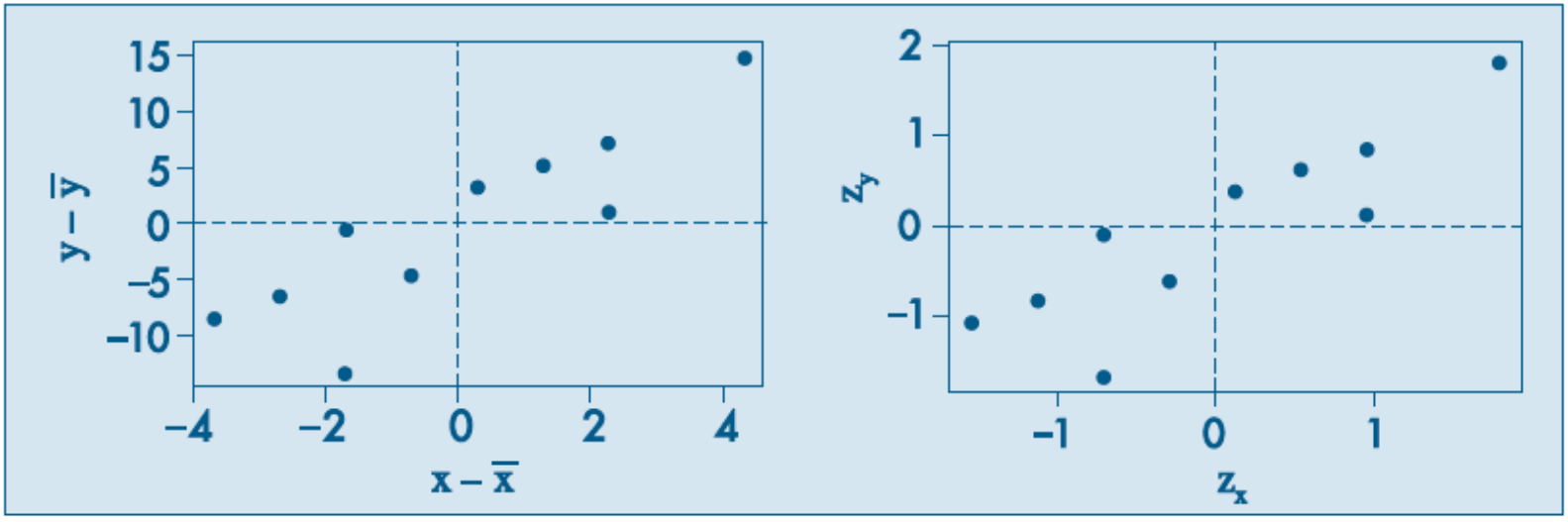



Figure 27: f25

# Quanti x Quanti: Correlação

**Tabela 4.15:** Cálculo do coeficiente de correlação.

| Agente | Anos $x$ | Clientes $y$ | $x - \bar{x}$ | $y - \bar{y}$ | $\frac{x - \bar{x}}{dp(x)} = z_x$ | $\frac{y - \bar{y}}{dp(y)} = z_y$ | $z_x \cdot z_y$ |
|---|---|---|---|---|---|---|---|
| A | 2 | 48 | –3,7 | –8,5 | –1,54 | –1,05 | 1,617 |
| B | 3 | 50 | –2,7 | –6,5 | –1,12 | –0,80 | 0,846 |
| C | 4 | 56 | –1,7 | –0,5 | –0,71 | –0,06 | 0,043 |
| D | 5 | 52 | –0,7 | –4,5 | –0,29 | –0,55 | 0,160 |
| E | 4 | 43 | –1,7 | –13,5 | –0,71 | –1,66 | 1,179 |
| F | 6 | 60 | 0,3 | 3,5 | 0,12 | 0,43 | 0,052 |
| G | 7 | 62 | 1,3 | 5,5 | 0,54 | 0,68 | 0,367 |
| H | 8 | 58 | 2,3 | 1,5 | 0,95 | 0,19 | 0,181 |
| I | 8 | 64 | 2,3 | 7,5 | 0,95 | 0,92 | 0,874 |
| J | 10 | 72 | 4,3 | 15,5 | 1,78 | 1,91 | 3,400 |
| Total | 57 | 565 | 0 | 0 | | | 8,769 |

$\bar{x} = 5{,}7, \quad dp(X) = 2{,}41, \quad \bar{y} = 56{,}5, \quad dp(Y) = 8{,}11$

Figure 28: f24

# Quanti x Quanti: Correlação

**Definição.** Dados $n$ pares de valores $(x_1, y_1)$, $(x_2, y_2)$, ..., $(x_n, y_n)$, chamaremos de coeficiente de correlação entre as duas variáveis $X$ e $Y$ a

$$\text{corr}\,(X, Y) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{dp(X)}\right)\left(\frac{y_i - \bar{y}}{dp(Y)}\right), \tag{4.7}$$

ou seja, a média dos produtos dos valores padronizados das variáveis.

Não é difícil provar que o coeficiente de correlação satisfaz

$$-1 \leqslant \text{corr}\,(X, Y) \leqslant 1. \tag{4.8}$$

Figure 29: f26

# Exercício

11. Abaixo estão os dados referentes à porcentagem da população economicamente ativa empregada no setor primário e o respectivo índice de analfabetismo para algumas regiões metropolitanas brasileiras.

| Regiões metropolitanas | Setor primário | Índice de analfabetismo |
|---|---|---|
| São Paulo | 2,0 | 17,5 |
| Rio de Janeiro | 2,5 | 18,5 |
| Belém | 2,9 | 19,5 |
| Belo Horizonte | 3,3 | 22,2 |
| Salvador | 4,1 | 26,5 |
| Porto Alegre | 4,3 | 16,6 |
| Recife | 7,0 | 36,6 |
| Fortaleza | 13,0 | 38,4 |

Fonte: Indicadores Sociais para Áreas Urbanas — IBGE — 1977.

(a) Faça o diagrama de dispersão.
(b) Você acha que existe uma dependência linear entre as duas variáveis?
(c) Calcule o coeficiente de correlação.
(d) Existe alguma região com comportamento diferente das demais? Se existe, elimine o valor correspondente e recalcule o coeficiente de correlação.

Figure 30: f26b