Modulo II - Introdução à Estatística Básica

Umberto Mignozzetti

6/1/2020

Modulo II

Análise de dados

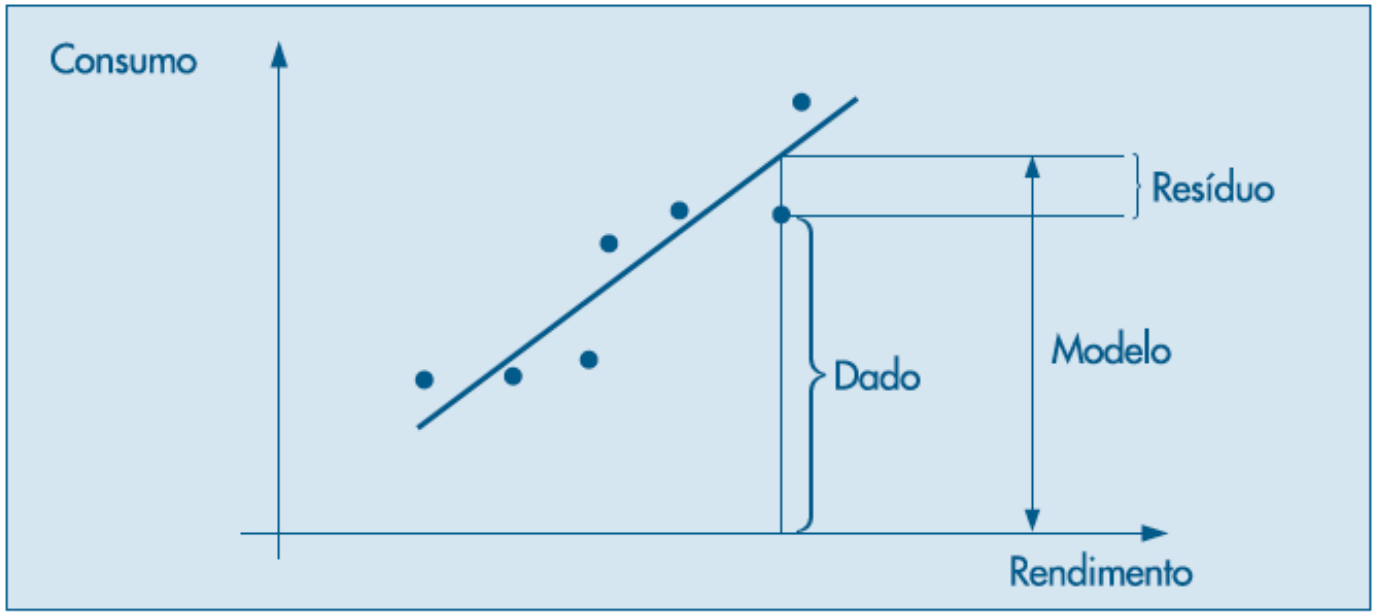
Análise de dados

- 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.

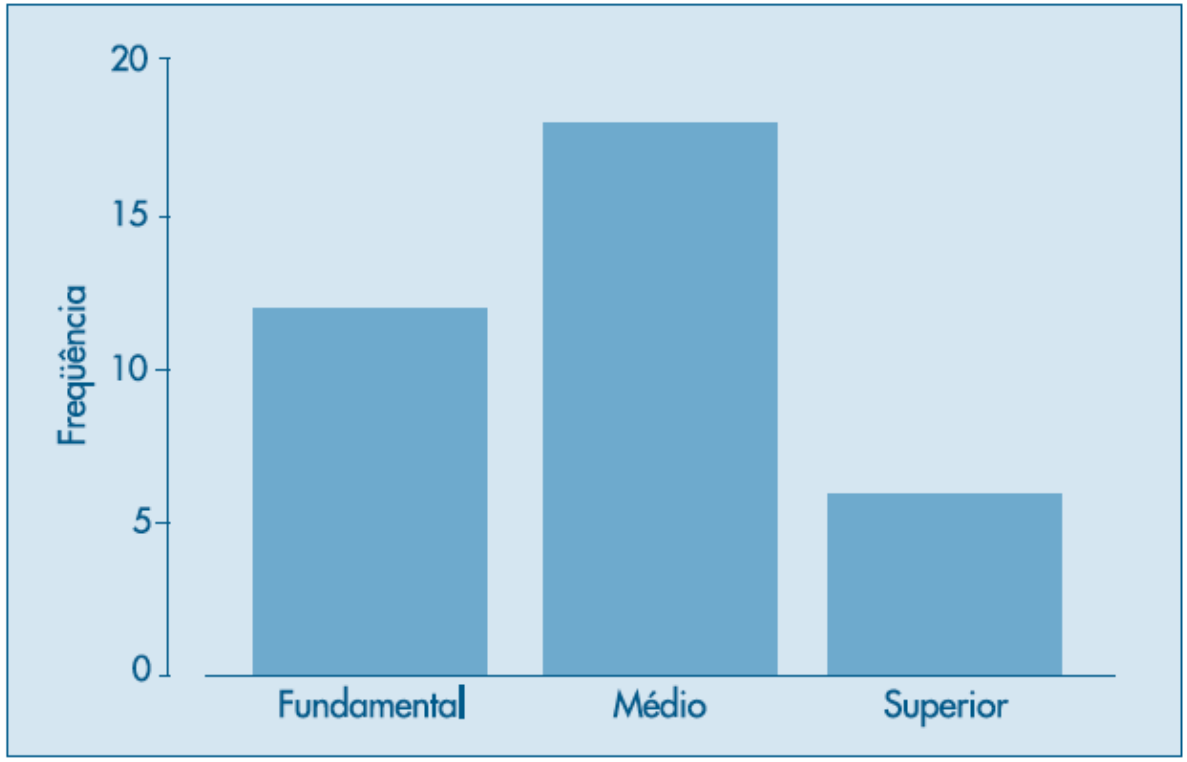
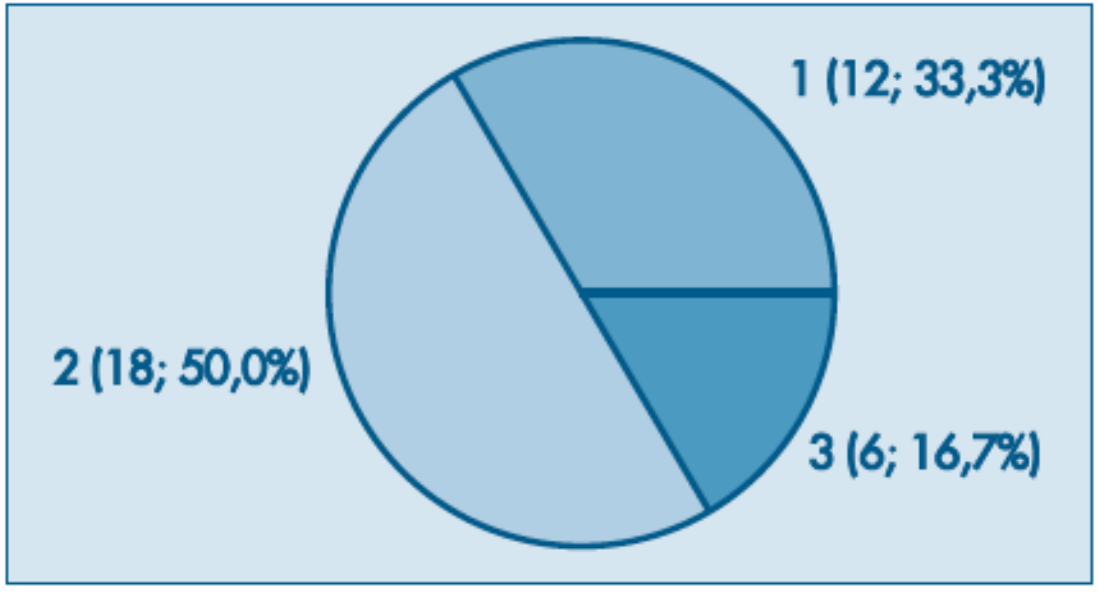


Figure 5: f4

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.

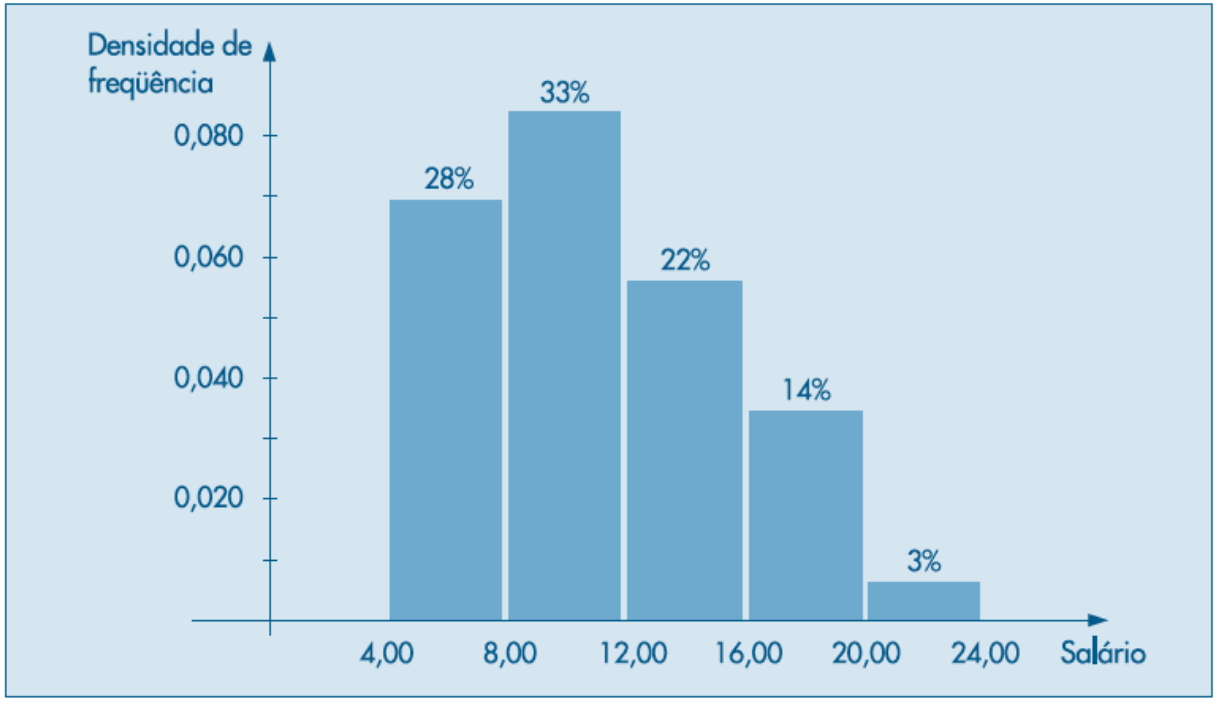


Figure 7: f6

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
```

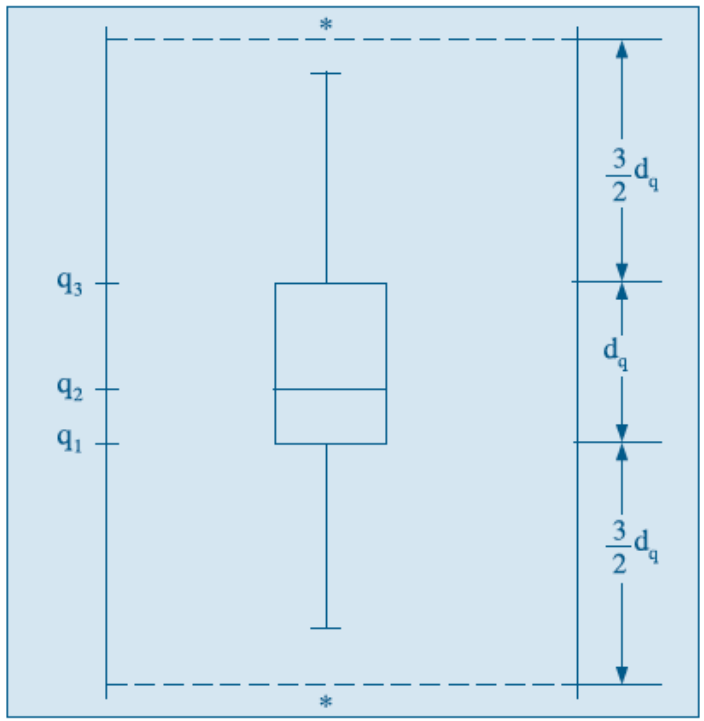
```
summary(x)
```

```
##    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.

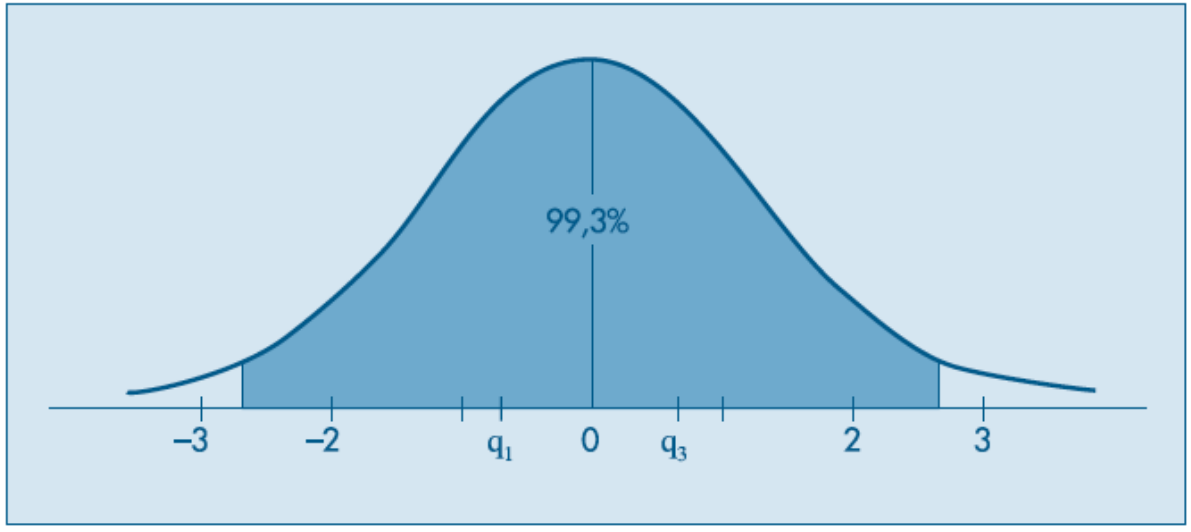


Figure 16: f14

Exercício

- Faça uma análise dos dados da empresa MB.