



UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO  
INSTITUTO TRÊS RIOS  
DEPARTAMENTO DE CIÊNCIAS ECONÔMICAS E EXATAS

---

# ESTATÍSTICA I

## NOTAS DE AULA

Prof. Dr. Diógenes Ferreira Filho

---

## Sumário

### I - Análise Exploratória de Dados. Organização, Resumo e Apresentação de Dados Estatísticos 1

1- Resumo de dados .....	1
<b>1.1 Tipos de variáveis .....</b>	<b>1</b>
<b>1.2 Notação de somatório .....</b>	<b>3</b>
<b>1.3 Tabelas: Construção e normas; Distribuição de frequências .....</b>	<b>6</b>
<b>1.4 Gráficos: barras/colunas, setores, linhas, histograma, polígono de frequências .....</b>	<b>16</b>
<b>1.5 Ramo-e-folhas .....</b>	<b>29</b>

### 2- Medidas de posição para dados brutos e agrupados: média aritmética, moda, mediana, quantis .....

31

<b>2.1 Média aritmética (dados brutos) .....</b>	<b>31</b>
<b>2.2 Média aritmética (dados agrupados) .....</b>	<b>32</b>
<b>2.3 Mediana (dados brutos) .....</b>	<b>36</b>
<b>2.4 Mediana (dados agrupados) .....</b>	<b>38</b>
<b>2.5 Moda (dados brutos) .....</b>	<b>41</b>
<b>2.6 Moda (dados agrupados) .....</b>	<b>42</b>
<b>2.7 Moda (dados qualitativos) .....</b>	<b>44</b>
<b>2.8 Quantis (dados brutos) .....</b>	<b>46</b>
<b>2.9 Quantis (dados agrupados) .....</b>	<b>48</b>
<b>2.10 Boxplot .....</b>	<b>51</b>

### 3 - Medidas de dispersão para dados brutos e agrupados: amplitude, desvio médio absoluto, variância, desvio padrão, coeficiente de variação. ....

60

<b>3.1 Amplitude .....</b>	<b>61</b>
<b>3.2 Desvio Médio Absoluto .....</b>	<b>62</b>
<b>3.3 Variância (dados brutos) .....</b>	<b>64</b>
<b>3.4 Variância (dados agrupados) .....</b>	<b>65</b>
<b>3.5 Desvio Padrão .....</b>	<b>69</b>
<b>3.6 Coeficiente de Variação .....</b>	<b>70</b>

### 4 - Simetria e Curtose .....

74

<b>4.1 Simetria .....</b>	<b>74</b>
<b>4.2 Curtose .....</b>	<b>83</b>

### 5 - Análise bidimensional .....

88

<b>5.1 Variáveis Qualitativas: Tabelas de Contingência e Coeficiente de Contingência .....</b>	<b>88</b>
<b>5.2 Medidas de dependência entre duas variáveis nominais (qui-quadrado) .....</b>	<b>93</b>

---

---

<b>5.3 Variáveis Quantitativas: Diagrama de Dispersão e Coeficiente de Correlação .....</b>	<b>97</b>
<b>II - Probabilidade .....</b>	<b>103</b>
1 - Probabilidade .....	103
<b>1.1 Espaço amostral, eventos.....</b>	<b>103</b>
<b>1.2 Probabilidade condicional, Teorema de Bayes e independência de eventos.....</b>	<b>110</b>
2. Variáveis aleatórias discretas .....	120
<b>2.1 Conceito. Valor esperado e variância de uma variável aleatória .....</b>	<b>120</b>
<b>2.2 Covariância e Correlação.....</b>	<b>128</b>
<b>2.3 Distribuição de Bernoulli.....</b>	<b>131</b>
<b>2.4 Distribuição Binomial.....</b>	<b>132</b>
<b>2.5 Distribuição de Poisson .....</b>	<b>135</b>
<b>2.6 Distribuição Geométrica.....</b>	<b>138</b>
<b>2.7 Distribuição Hipergeométrica .....</b>	<b>139</b>
3. Variáveis Aleatórias Contínuas .....	142
<b>3.1 Conceito. Noções básicas de esperança matemática e variância .....</b>	<b>142</b>
<b>3.2 Distribuição exponencial.....</b>	<b>148</b>
<b>3.3 Distribuição normal: características; distribuição normal padronizada .....</b>	<b>151</b>
<b>III - Inferência Estatística .....</b>	<b>161</b>
1 Introdução à inferência estatística. ....	161
<b>1.1 Conceitos básicos. Amostra e população. ....</b>	<b>161</b>
<b>1.2 Amostragem aleatória simples: obtenção de uma amostra aleatória.....</b>	<b>162</b>
<b>1.3 Conceito de Distribuições amostrais.....</b>	<b>164</b>
<b>1.4 Distribuição amostral da média.....</b>	<b>164</b>
2 Estimação .....	171
<b>2.1 Conceitos básicos. Estimadores não viciados .....</b>	<b>171</b>
<b>2.2 Intervalo de confiança para média de uma população Normal com variância populacional conhecida.....</b>	<b>174</b>
<b>2.3 Determinação do tamanho de uma amostra .....</b>	<b>178</b>
<b>2.4 Intervalo de confiança para a média de uma população Normal com variância populacional desconhecida.....</b>	<b>179</b>
Apêndice .....	186
<b>Apêndice 1. Tabela da distribuição Normal Padrão .....</b>	<b>186</b>
<b>Apêndice 2. Tabela (Bilateral) da distribuição <math>t</math> de Student.....</b>	<b>187</b>
Bibliografia .....	188

---

---

## **Prefácio**

Este material foi preparado com a intenção de cobrir o programa da disciplina Estatística I da Universidade Federal Rural do Rio de Janeiro. Ele é composto por notas de aula elaboradas à partir de livros e apostilas constantes na Bibliografia e não substitui a leitura dos mesmos.

O material não está livre de erros e/ou imperfeições e toda e qualquer contribuição será bem-vinda.

---

# I - Análise Exploratória de Dados. Organização, Resumo e Apresentação de Dados Estatísticos

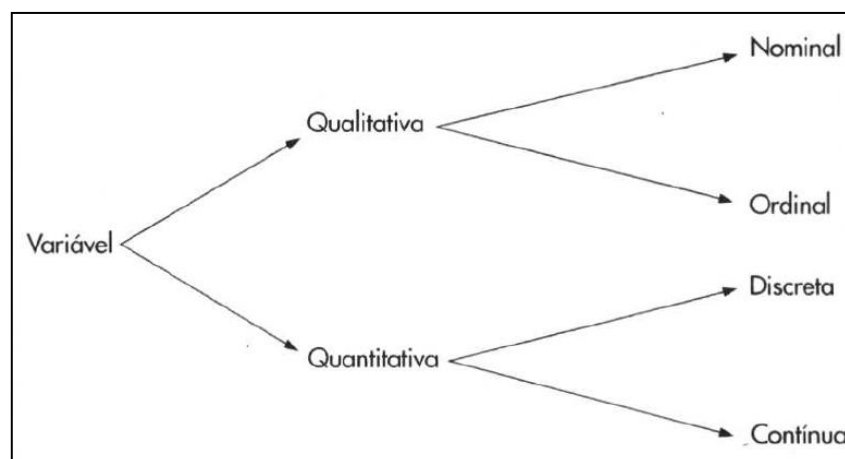
## 1- Resumo de dados

### 1.1 Tipos de variáveis

Uma característica que pode assumir diferentes valores de um indivíduo para outro é chamada de variável. Por exemplo, a característica altura é uma variável pois diferentes indivíduos podem apresentar diferentes alturas.

As variáveis podem ser classificadas em qualitativas e quantitativas. Ainda, as variáveis qualitativas podem ser classificadas em nominais e ordinais, já as variáveis quantitativas podem ser classificadas em discretas e contínuas. Pode-se observar na Figura 1.1 a classificação das variáveis.

**Figura 1.1.** Classificação das variáveis



### Variáveis Qualitativas (Categóricas)

São variáveis que apresentam como possíveis realizações uma qualidade (ou atributo) do indivíduo pesquisado.

### ***Variáveis Qualitativas Nominais***

São variáveis cujas possíveis realizações são atributos para os quais não existe nenhuma ordenação. Por exemplo, a variável *sexo*, cujas possíveis realizações são *masculino* e *feminino*, é uma variável qualitativa nominal pois suas realizações não tem nenhuma ordenação.

### ***Variáveis Qualitativas Ordinais***

São variáveis cujas possíveis realizações são atributos para os quais existe uma ordem. Por exemplo, a variável *classe social*, cujas possíveis realizações são *baixa*, *média* e *alta*, é uma variável qualitativa ordinal pois suas possíveis realizações seguem uma ordem.

### ***Variáveis Quantitativas***

São variáveis cujas possíveis realizações são números resultantes de uma contagem ou mensuração.

#### ***Variáveis Quantitativas Discretas***

São variáveis numéricas para as quais os possíveis valores formam um conjunto finito ou enumerável de números, e que resultam, frequentemente, de uma contagem. Por exemplo, a variável *número de filhos*, cujas possíveis realizações são 0, 1, 2, 3, ..., é uma variável quantitativa discreta.

#### ***Variáveis Quantitativas Contínuas***

São variáveis numéricas para as quais os possíveis valores pertencem a um intervalo de números reais e que resultam de uma mensuração. Por exemplo, a variável *altura* de um indivíduo, cujas possíveis realizações são números reais positivos (por exemplo: 1,60; 1,56; 1,75 m, ...) é uma variável quantitativa contínua.

---

## Exercícios

1) Para as variáveis apresentadas a seguir, dê exemplos de possíveis respostas e classifique-as.

- a) Peso
- b) Classe social
- c) Número de irmãos
- d) Altura
- e) Cor dos olhos
- f) Sexo
- g) Grau de instrução
- h) Idade
- i) Cidade de origem

## 1.2 Notação de somatório

### Variáveis e índices

O símbolo  $x_i$  (leia  $x$  índice  $i$ ) representa qualquer um dos  $n$  valores  $x_1, x_2, \dots, x_n$  assumidos por uma variável aleatória  $X$  no conjunto de dados. A letra  $i$ , usada como índice, indica a "posição" (de 1 a  $n$ ) do elemento  $x$  no conjunto de dados. Assim,  $x_1$  é o elemento que ocupa a 1ª posição na amostra,  $x_2$  é o elemento que ocupa a 2ª posição na amostra, ...,  $x_n$  é o elemento que ocupa a  $n$ -ésima posição na amostra.

Por exemplo, se for considerada uma amostra de tamanho  $n = 3$  pessoas e se  $X$  representa uma variável relativa ao peso em kg, então uma possibilidade de resultados é:

50,5, 64,3 e 72,6.

Logo,  $x_1 = 50,5$ ,  $x_2 = 64,3$  e  $x_3 = 72,6$ .

### Notação de somatório

Para representarmos a soma de  $n$  variáveis aleatórias podemos utilizar o símbolo  $\Sigma$ , letra grega maiúscula sigma. Assim, a soma

$$x_1 + x_2 + \dots + x_n$$

---

pode ser representada por

$$\sum_{i=1}^n x_i$$

ou seja,

$$\sum_{i=1}^n x_i = x_1 + x_2 + \cdots + x_n.$$

Por exemplo, para os dados de pesos de  $n = 3$  pessoas:

50,5    64,3    72,6

temos:

$$\sum_{i=1}^3 x_i = x_1 + x_2 + x_3 = 50,5 + 64,3 + 72,6 = 187,4.$$

A variação do índice  $i$  pode não ir de 1 a  $n$ , mas estar em qualquer subintervalo desses limites.

#### Comandos no software R para calcular o somatório:

```
#Entrando com os dados no R:
x <- c(50.5, 64.3, 72.6)

#Mostrando os dados armazenados:
x

#Somatório de x:
sum(x)
```

**Observação:** O R diferencia letras maiúsculas e minúsculas

#### Algumas propriedades

- a)  $\sum_{i=1}^n ax_i = ax_1 + ax_2 + \cdots + ax_n = a(x_1 + x_2 + \cdots + x_n) = a \sum_{i=1}^n x_i$
- b)  $\sum_{i=1}^n x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n \neq \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)$
- c)  $\sum_{i=1}^n (ax_i + by_i) = ax_1 + by_1 + ax_2 + by_2 + \cdots + ax_n + by_n = a \sum_{i=1}^n x_i + b \sum_{i=1}^n y_i$



$$d) \sum_{i=1}^n x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2 \neq \left( \sum_{i=1}^n x_i \right)^2$$

$$e) \sum_{i=1}^n k = \underbrace{k + k + \cdots + k}_{n \text{ vezes}} = nk$$

em que  $a$ ,  $b$  e  $k$  são constantes

## Exercícios

1) Sejam as amostras de tamanho  $n = 5$  dadas por:

$$X = \{2, 7, 4, 3, 2\}$$

$$Y = \{1, 2, 3, 6, 5\}$$

obter:

$$a) \sum_{i=1}^5 x_i$$

$$b) \sum_{i=2}^4 x_i$$

$$c) \sum_{i=1}^5 y_i$$

$$d) \sum_{i=1}^5 x_i y_i$$

$$e) \sum_{i=1}^5 (3x_i + 2y_i)$$

$$f) \sum_{i=1}^4 x_i y_i + \sum_{i=2}^5 x_i y_i$$

$$g) \sum_{i=1}^5 x_i^2$$

$$h) \left( \sum_{i=1}^5 x_i \right)^2$$

$$i) \sum_{i=1}^5 y_i^2 + \left( \sum_{i=1}^5 y_i \right)^2$$

2) Considerando o conjunto de dados

$$X = \{2, 4, 5, 7, 3, 9\}$$

calcule:

$$a) \bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

$$b) S^2 = \frac{1}{n-1} \left[ \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \right]$$

3) Verificar, para o conjunto de dados do exercício número 2, que

$$\sum_{i=1}^n (X_i - \bar{X}) = 0.$$

### Outros comandos no Software R:

```
#Entrando com os dados (x e y) no R:
x <- c(2,7,4,3,2)
y <- c(1,2,3,6,5)

#Multiplicando x por y (x*y):
x*y

#Soma do produto x*y:
sum(x*y)

#Entrando com os dados (x) no R:
x <- c(2, 4, 5, 7, 3, 9)

#Somatório de xi:
sum(x)

#(Somatório de xi) ao quadrado:
(sum(x))^2

#Somatório de (xi ao quadrado):
sum(x^2)

#Número de dados (n):
n <- length(x) n

#X barra:
xb <- sum(x)/n
xb

#S ao quadrado:
S2 <- 1/(n-1) * (sum(x^2) - (sum(x))^2/n)
S2

#Soma de (Xi - X barra):
sum(x-xb)
```

## 1.3 Tabelas: Construção e normas; Distribuição de frequências

### Normas para a construção de tabelas

Os dados são apresentados em tabelas colocadas perto do ponto do texto em que são mencionadas pela primeira vez. As tabelas devem conter os seguintes elementos: título, cabeçalho, indicador de linha, célula e moldura, como mostrado no exemplo a seguir.

## Exemplo

Com base na Tabela 1.1 ilustraremos os elementos de uma tabela.

**Tabela 1.1.** População residente no Brasil, segundo o sexo, de acordo com o Censo Demográfico 2010

Sexo	População residente
Homens	93.406.990
Mulheres	97.348.809
Total	190.755.799

Fonte: Censo Demográfico 2010. IBGE (2011).

- O *título* explica o tipo de dado que a tabela contém. Deve-se colocá-lo acima dos dados.

**Tabela 1.1.** População residente no Brasil, segundo o sexo, de acordo com o Censo Demográfico 2010

- O *cabeçalho* especifica o conteúdo de cada coluna.

Sexo	População residente
------	---------------------

- O indicador de linha é um conjunto de termos em que cada termo indica o conteúdo de uma linha.

Homens
Mulheres
Total

- A célula resulta do cruzamento de uma linha com uma coluna e deve conter um dado numérico.

93.406.990
97.348.809
190.755.799

Nenhuma célula da tabela deve ficar em branco. Toda célula deve apresentar um número ou, se o dado não existir, coloca-se um traço na célula (–) em que o dado deveria estar escrito.

- As tabelas devem ter *moldura*. Entende-se por moldura o conjunto de traços que delimitam a tabela.

Moldura:


As tabelas devem ser delimitadas no alto e embaixo por traços horizontais. O cabeçalho deve ser delimitado por traços horizontais. É possível fazer traços verticais no interior da tabela, separando as colunas. Não se deve fazer traços verticais "fechando" as laterais da tabela.

- A fonte indica o responsável (pessoa física ou jurídica) pelos dados. Deve ser colocada na primeira linha do rodapé da tabela e precedida pela palavra Fonte.

Fonte: Censo Demográfico 2010. IBGE (2011).

*Não se indica a fonte nos casos em que os dados foram obtidos pelo próprio pesquisador.*

## Tabela de distribuição de frequências para dados qualitativos

Quando observamos dados qualitativos, classificamos cada observação em determinada categoria. Depois, contamos o número de observações em cada categoria. A idéia é resumir as informações na forma de uma tabela que mostre essas contagens (frequências) por categoria. Temos, então uma *tabela de distribuição de frequências*.

### Exemplo

Considere os dados brutos do grau de instrução de 36 indivíduos:

Fundamental	Médio	Médio	Superior	Médio	Médio
Médio	Médio	Fundamental	Fundamental	Superior	Fundamental
Superior	Fundamental	Médio	Médio	Médio	Médio
Fundamental	Médio	Superior	Fundamental	Superior	Médio
Médio	Fundamental	Médio	Médio	Médio	Fundamental
Fundamental	Fundamental	Superior	Fundamental	Médio	Médio

Na Tabela 1.2 é apresentada a distribuição de frequências da variável grau de instrução cujos os dados foram apresentados acima.

**Tabela 1.2.** Frequências de 36 indivíduos segundo o grau de instrução

<b>Grau de instrução</b>	<b>Frequência</b>
Fundamental	12
Médio	18
Superior	6
Total	36

Observando-se os resultados da segunda coluna da Tabela 1.2, vê-se que dos 36 indivíduos, 12 têm ensino fundamental, 18 o ensino médio e 6 possuem curso superior.

#### Comandos no Software R para agrupar os dados em uma distribuição de frequências:

```
#Entrando com os dados no R:
dados <- c("Fundamental", "Médio", "Médio",
           "Superior", "Médio", "Médio",
           "Médio", "Médio", "Fundamental",
           "Fundamental", "Superior", "Fundamental",
           "Superior", "Fundamental", "Médio",
           "Médio", "Médio", "Médio",
           "Fundamental", "Médio", "Superior",
           "Fundamental", "Superior", "Médio",
           "Médio", "Fundamental", "Médio",
           "Médio", "Médio", "Fundamental",
           "Fundamental", "Fundamental", "Superior",
           "Fundamental", "Médio", "Médio")

#Mostrando os dados armazenados:
dados

#Tabela de distribuição de frequências:
tab <- table(dados)

#Mostrando a tabela de distribuição de frequências:
tab
```

As tabelas de distribuição de frequência podem conter, ainda, além das frequências, as proporções e porcentagens, como pode-se observar na Tabela 1.3. As proporções (ou frequências relativas) são calculadas dividindo-se cada frequência pelo total, enquanto as porcentagens (ou frequências percentuais) são calculadas multiplicando-se cada proporção por 100. Assim, por exemplo para o ensino fundamental, obtemos a proporção fazendo  $12/36 = 0,3333$ , e a porcentagem fazendo  $0,3333 \times 100 = 33,33 \%$ .

**Tabela 1.3.** Frequências e porcentagens de 36 indivíduos segundo o grau de instrução

Grau de instrução	Frequência	Proporção	Porcentagem
Fundamental	12	0,3333	33,33
Médio	18	0,5000	50,00
Superior	6	0,1667	16,67
Total	36	1,0000	100,00

À seguir são apresentados os comandos no *Software R* para organizar os dados em uma distribuição de frequências.

### Tabela de distribuição de frequências para dados quantitativos

Dados quantitativos também podem ser apresentados em tabelas de distribuição de frequências.

### Tabela para dados discretos

Se os dados são *discretos*, para organizar a tabela de distribuição de frequências:

- escreva os dados em ordem crescente (*rol*);
- conte quantas vezes cada valor se repete;
- organize a tabela apresentando os valores numéricos em ordem natural.

### Exemplo

Considere os dados brutos do número de faltas de trinta funcionários ao trabalho:

1	3	1	1	0	1	0	1	1	0
2	2	0	0	0	1	2	1	2	0
0	1	6	4	3	3	1	2	4	0

Construa uma tabela de distribuição de frequências do número de faltas ao trabalho.

### Solução

Organizando os dados em ordem crescente temos:

0	0	0	0	0	0	0	0	0	1
1	1	1	1	1	1	1	1	1	2
2	2	2	2	3	3	3	4	4	6

Agora basta contar as frequências e montar a Tabela 1.4:

**Tabela 1.4.** Frequências e porcentagens do número de faltas de trinta funcionários ao trabalho

Número de faltas	Frequência	Proporção	Porcentagem
0	9	0,300	30,0
1	10	0,333	33,3
2	5	0,167	16,7
3	3	0,100	10,0
4	2	0,067	6,7
5	0	0,000	0,0
6	1	0,033	3,3
Total	30	1,000	100,0

**Observação:** Quando existirem muitos valores possíveis para a variável discreta podemos construir uma tabela de distribuição de frequências do mesmo modo como faremos com uma variável contínua (agrupando os dados em classes).

#### Comandos no Software R para agrupar os dados em uma distribuição de frequências:

```
#Entrando com os dados no R:
dados <- c(1,3,1,1,0,1,0,1,1,0,
           2,2,0,0,0,1,2,1,2,0,
           0,1,6,4,3,3,1,2,4,0)

#Mostrando os dados armazenados:
dados

#Tabela de distribuição de frequências:
tab <- table(dados)

#Mostrando a tabela de distribuição de frequências:
tab
```

### Tabela para dados contínuos

Ao contrário das variáveis discretas, as variáveis contínuas assumem, em geral, muitos valores. Isto quer dizer que se usássemos as tabelas de frequências, como no caso das variáveis discretas teríamos uma tabela com muitas linhas, tornando-a pouco operacional. Para contornar este problema costuma-se descrever as variáveis

quantitativas contínuas através de **tabelas de classes de frequências** ou **tabelas de intervalos**.

### Exemplo

Considere os salários (x sal. mín.) de 36 indivíduos:

4,00	4,56	5,25	5,73	6,26	6,66
6,86	7,39	7,59	7,44	8,12	8,46
8,74	8,95	9,13	9,35	9,77	9,80
10,53	10,76	11,06	11,59	12,00	12,79
13,23	13,60	13,85	14,69	14,71	15,99
16,22	16,61	17,26	18,75	19,40	23,30

Note que se agruparmos os dados em uma distribuição de frequências da mesma forma como fizemos para dados discretos isso não resumirá as 36 observações num grupo menor, pois não existem observações iguais. Podemos observar isto na Tabela 1.5 (Observação: esta tabela está "errada", serve apenas para ilustrar como **não** deve ser feito).

**Tabela 1.5.** Construção da tabela de frequências dos salários dos 36 indivíduos (dados contínuos) como se fossem dados discretos (**TABELA ERRADA**)

Salário	Frequência
4,00	1
4,56	1
5,25	1
5,73	1
6,26	1
⋮	⋮
23,30	1
Total	36

A solução empregada é agrupar os dados em faixas de salário, como podemos observar na Tabela 1.6.

**Tabela 1.6.** Frequências e porcentagens de 36 indivíduos por faixa de salário (**TABELA CORRETA**)

Classes de salários	Frequência	Proporção	Porcentagem
4,00 – 8,00	10	0,2778	27,78
8,00 – 12,00	12	0,3333	33,33
12,00 – 16,00	8	0,2222	22,22
16,00 – 20,00	5	0,1389	13,89
20,00 – 24,00	1	0,0278	2,78
Total	36	1,000	100,0



Procedendo-se desse modo, ao resumir os dados referentes a uma variável contínua, perde-se alguma informação. Por exemplo, não sabemos quais são os oito salários da classe de 12 a 16, a não ser que investiguemos o conjunto de dados original (dados brutos). Note que estamos usando a notação  $a \vdash b$  para o intervalo de números contendo o extremo  $a$  mas não contendo o extremo  $b$ . Podemos também usar a notação  $[a, b)$  para designar o mesmo intervalo  $a \vdash b$ . Os comandos no *Software R* para a obtenção da distribuição de frequências são dados a seguir.

#### Comandos no Software R para agrupar os dados em uma distribuição de frequências:

```
#Entrando com os dados no R:
dados <- c(4.00, 4.56, 5.25, 5.73, 6.26, 6.66,
          6.86, 7.39, 7.59, 7.44, 8.12, 8.46,
          8.74, 8.95, 9.13, 9.35, 9.77, 9.80,
          10.53, 10.76, 11.06, 11.59, 12.00, 12.79,
          13.23, 13.60, 13.85, 14.69, 14.71, 15.99,
          16.22, 16.61, 17.26, 18.75, 19.40, 23.30)

#Mostrando os dados armazenados:
dados

#Organizando os dados em ordem crescente:
sort(dados)

#Frequências:
hist(dados, plot=F, breaks=c(4,8,12,16,20,24), right=F)
```

**Observações:** No resultado comando `hist()` os limites das classes são dados em `$breaks` e as frequências em `$counts`

A escolha dos intervalos pode ser feita arbitrariamente e a familiaridade do pesquisador com os dados é que lhe indicará quantas e quais classes devem ser usadas. Quando determinados intervalos de classes tiverem algum significado prático para o pesquisador, ele poderá simplesmente escolher estes intervalos.

Não existe um número "ideal" de classes para um conjunto de dados, embora existam fórmulas para estabelecer quantas classes devem ser construídas. Vejamos, a seguir um critério para a construção das classes.

## Passos para construção da Tabela de Distribuição de Frequências para dados contínuos

1. Organize os dados em ordem crescente (rol);
  2. Determine o número de classes ( $k$ ):
    - a) Se  $n \leq 100$  utilize  $k = \sqrt{n}$
    - b) Se  $n > 100$  utilize  $k = 1 + 3,22 \times \log n$
- Observação:** Arredonde o valor de  $k$  para o inteiro mais próximo;
3. Ache o valor máximo ( $X_{máx}$ ) e o valor mínimo ( $X_{mín}$ ) do conjunto de dados;
  4. Calcule a *amplitude total* ( $A$ ), que é a diferença entre o valor máximo e o valor mínimo:

$$A = X_{máx} - X_{mín}$$

5. Divida a amplitude total pelo número de classes para obter a amplitude das classes ( $c^*$ ):

$$c = \frac{A}{k}$$

**Observação:** Sempre é melhor arredondar o valor obtido para um valor mais alto;

6. Organize as classes, de maneira que a primeira classe contenha o valor observado. Deve-se iniciar a primeira classe (limite inferior da 1ª classe:  $LI_1$ ) pelo menor valor dos dados ou por um valor menor do que ele. Depois basta somar  $c$  a este valor para obter o limite superior da 1ª classe ( $LS_1 = LI_1 + c$ ). O limite inferior da 2ª classe ( $LI_2$ ) é igual ao limite superior da 1ª classe ( $LI_2 = LS_1$ ). Para obter os demais limites das classes continuamos somando  $c$ . Assim, a primeira classe de frequências é um intervalo do tipo  $L \vdash (L + c)$ , a segunda  $(L + c) \vdash (L + 2c)$  e assim sucessivamente.
7. Depois de construir todas as ( $k$ ) classes, conte quantos dados têm dentro de cada classe (frequências).

## Exemplo

Construiremos uma tabela de distribuição de frequências considerando os mesmos dados de salários (x sal. mín.) de 36 indivíduos apresentados anteriormente.

## Solução

1. Dados em ordem crescente (*rol*):

4,00	4,56	5,25	5,73	6,26	6,66
6,86	7,39	7,44	7,59	8,12	8,46
8,74	8,95	9,13	9,35	9,77	9,80
10,53	10,76	11,06	11,59	12,00	12,79
13,23	13,60	13,85	14,69	14,71	15,99
16,22	16,61	17,26	18,75	19,40	23,30

2. Número de classes:  $k = \sqrt{n} = \sqrt{36} = 6$ ;

3.  $X_{\max} = 23,30$  e  $X_{\min} = 4,00$ ;

4.  $A = X_{\max} - X_{\min} = 23,30 - 4,00 = 19,30$ ;

5.  $c = \frac{A}{k} = \frac{19,30}{6} = 3,22$  que arredondaremos para  $c = 3,5$ ;

6. Iniciaremos o limite inferior da 1ª classe ( $LI_1$ ) em 4,00 (também poderíamos ter iniciado em um valor menor do que 4) e depois iremos somando  $c = 3,5$  para obter os outros limites das demais classes:

$$LI_1 = 4,0 \quad , \quad LS_1 = 4,0 + c = 4,0 + 3,5 = 7,5$$

$$LI_2 = LS_1 = 7,5 \quad , \quad LS_2 = 7,5 + c = 7,5 + 3,5 = 11,0$$

⋮

⋮

Classes	
1ª classe	4,0 ┤ 7,5
2ª classe	7,5 ┤ 11,0
3ª classe	11,0 ┤ 14,5
4ª classe	14,5 ┤ 18,0
5ª classe	18,0 ┤ 21,5
6ª classe	21,5 ┤ 25,0

7. Agora, basta contar quantos dados têm dentro de cada classe:

4,00	4,56	5,25	5,73	6,26	6,66
6,86	7,39	7,44	7,59	8,12	8,46
8,74	8,95	9,13	9,35	9,77	9,80
10,53	10,76	11,06	11,59	12,00	12,79
13,23	13,60	13,85	14,69	14,71	15,99
16,22	16,61	17,26	18,75	19,40	23,30

e depois montar a tabela de distribuição de frequências (Tabela 1.7):

**Tabela 1.7.** Frequências de 36 indivíduos por classes de salário

Classes	Frequência
4,0 – 7,5	9
7,5 – 11,0	11
11,0 – 14,5	7
14,5 – 18,0	6
18,0 – 21,5	2
21,5 – 25,0	1
Total	36

Observe que se ao invés de arredondar o valor da amplitude da classe  $c$  para 3,5 tivéssemos arredondado para 4,0 então a sexta classe (última) seria 24 – 28 e não teria nenhum elemento; ou seja, na realidade não teríamos seis classes mas apenas cinco (pois a última classe teria frequência zero). Utilizando esta forma de montar as classes precisamos tomar cuidado ao arredondar a amplitude da classe e ao escolher o limite inferior da primeira classe.

## Exercício

Obtenha a distribuição de frequências do exemplo anterior (Tabela 1.7) utilizando o *software R*.

**Orientação:** insira os dados brutos no R e obtenha as frequências das classes usando o comando "hist()". Você deve especificar os limites das classes utilizando o comando "breaks" dentro do comando "hist()".

## 1.4 Gráficos: barras/colunas, setores, linhas, histograma, polígono de frequências

### Gráfico de barras (ou gráfico de barras horizontais)

O gráfico de barras é usado para apresentar variáveis qualitativas, sejam elas nominais ou ordinais. Um gráfico de barras é a representação gráfica de uma distribuição de frequências de dados qualitativos. Para construir um gráfico de barras:

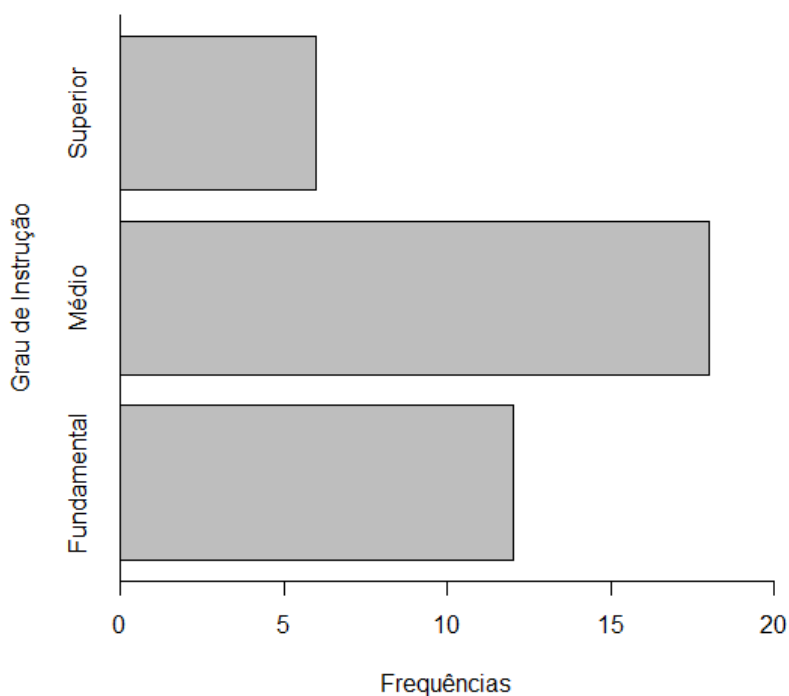
- desenhe o sistema de eixos cartesianos;
- anote as categorias da variável estudada no eixo das ordenadas (eixo vertical);

- escreva as frequências ou as frequências relativas (proporções) no eixo das abscissas (eixo horizontal), obedecendo a uma escala;
- desenhe **barras horizontais** de mesma largura, **separadas por um espaço**, para representar as categorias da variável em estudo. O comprimento de cada barra deve ser dado pela frequência da categoria;
- coloque legenda nos dois eixos (nomes dos eixos) e título na figura.

## Exemplo

Anteriormente construímos uma tabela de distribuição de frequências para a variável *grau de instrução* de 36 indivíduos (Tabela 1.2). Podemos representar graficamente esta tabela de distribuição de frequências por meio de um gráfico de barras (Figura 1.2).

**Figura 1.2.** Gráfico de barras para a variável *grau de instrução*



Recordando a Tabela 1.3

Grau de instrução	Frequência
Fundamental	12
Médio	18
Superior	6
Total	36

### Comandos no Software R para fazer o gráfico de barras:

```
#Pode-se entrar com os dados brutos no R e agrupá-los usando o
#comando table() como foi visto anteriormente ou, ainda, entrar
#diretamente com as frequências, como será feito a seguir.

#Entrando com as frequências no R:
tab <- c(12,18,6)

#Colocando os nomes das categorias:
names(tab) <- c("Fundamental","Médio","Superior")

#Mostrando a tabela:
tab

#"Plotando" o gráfico:
barplot(tab,hORIZ=T,xlim=c(0,20),xlab="Frequências",
        ylab="Grau de Instrução")
abline(v=0)
```

### Gráfico de colunas (ou gráfico de barras verticais)

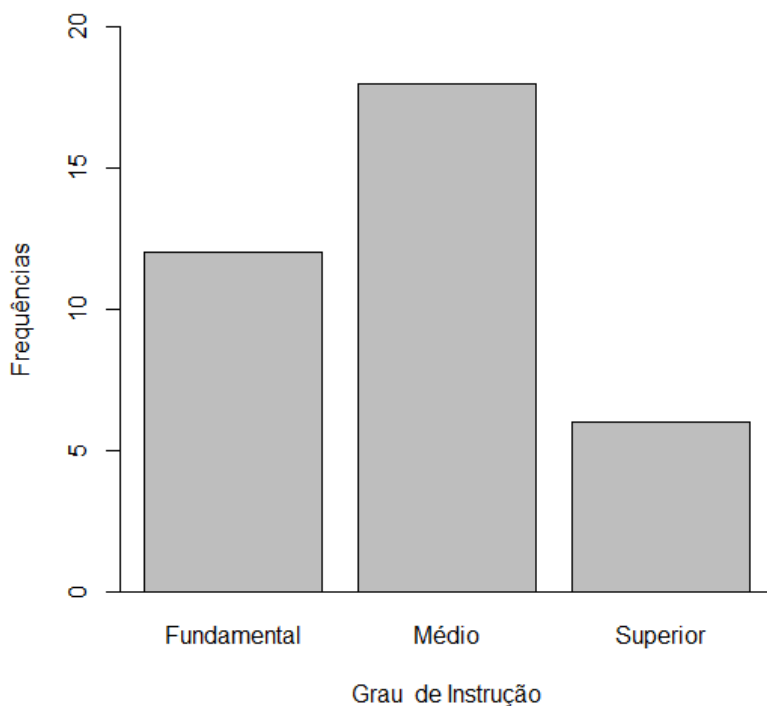
O gráfico de colunas é um tipo de gráfico de barras, porém, nesse caso as barras são verticais (colunas). Para construir um gráfico de colunas:

- desenhe o sistema de eixos cartesianos;
- anote as categorias da variável estudada no eixo das abscissas (eixo horizontal);
- escreva as frequências ou as frequências relativas (proporções) no eixo das ordenadas (eixo vertical), obedecendo a uma escala;
- desenhe **barras verticais (colunas)** de mesma largura, **separadas por um espaço**, para representar as categorias da variável em estudo. A altura de cada barra deve ser dada pela frequência da categoria;
- coloque legenda nos dois eixos (nomes dos eixos) e título na figura.

## Exemplo

Considerando a tabela de distribuição de frequências para a variável *grau de instrução* de 36 indivíduos (Tabela 1.2), podemos representar essa distribuição de frequências por meio de um gráfico de colunas (Figura 1.3).

**Figura 1.3.** Gráfico de colunas para a variável *grau de instrução*



### Comandos no Software R para fazer o gráfico de colunas:

```
#Pode-se entrar com os dados brutos no R e agrupá-los usando o
#comando table() como foi visto anteriormente ou, ainda, entrar
#diretamente com as frequências, como será feito a seguir.

#Entrando com as frequências no R:
tab <- c(12,18,6)

#Colocando os nomes das categorias:
names(tab) <- c("Fundamental","Médio","Superior")

#Mostrando a tabela:
tab

#"Plotando" o gráfico:
barplot(tab,hORIZ=F,ylim=c(0,20),xlab="Grau de Instrução",
        ylab="Frequências")
abline(h=0)
```

## Gráfico de setores

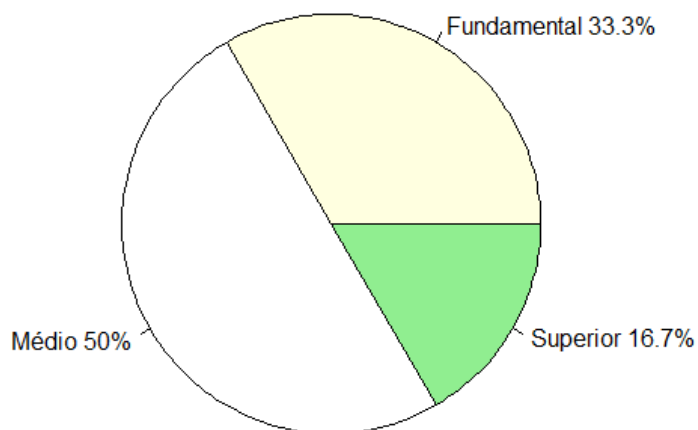
O gráfico de setores é especialmente indicado para apresentar variáveis *qualitativas*, desde que o número de categorias seja pequeno. Para construir um gráfico de setores:

- trace uma circunferência (uma circunferência tem  $360^\circ$ ). Essa circunferência representará o total, ou seja, 100%;
- divida a circunferência em tantos setores quantas sejam as categorias da variável em estudo, mas é preciso calcular o ângulo de cada setor: é igual a proporção da categoria multiplicada por  $360^\circ$ ;
- marque na circunferência os ângulos calculados; separe com o traçado dos raios;
- escreva a legenda e coloque título na figura.

## Exemplo

Considerando a tabela de distribuição de frequências para a variável *grau de instrução* de 36 indivíduos (Tabela 1.2), podemos representar essa distribuição de frequências por meio de um gráfico de setores (Figura 1.4).

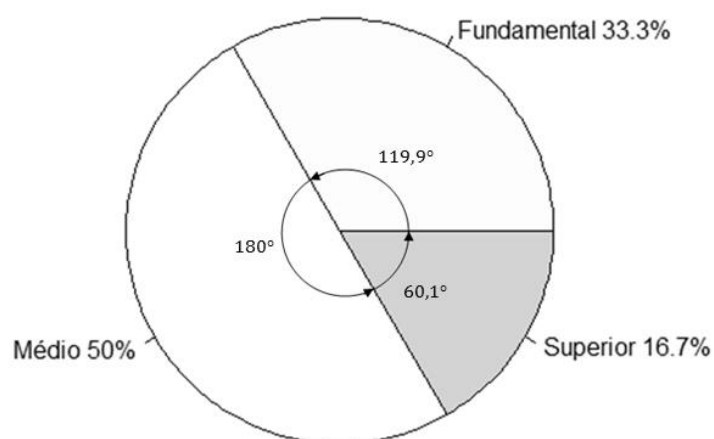
**Figura 1.4.** Gráfico de setores para a variável *grau de instrução*





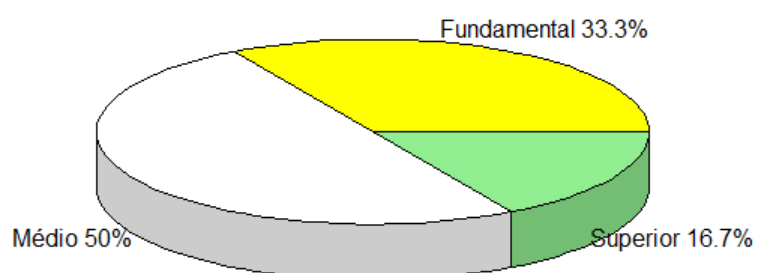
Para construir o gráfico de setores (Figura 1.4) calculamos as proporções de cada categoria e depois multiplicamos por  $360^\circ$  para encontrar o ângulo de cada setor.

Grau de instrução	Frequência	Proporção	Ângulo
Fundamental	12	$\frac{12}{36} = 0,333$	$0,333 \times 360^\circ = 119,9^\circ$
Médio	18	$\frac{18}{36} = 0,500$	$0,500 \times 360^\circ = 180^\circ$
Superior	6	$\frac{6}{36} = 0,167$	$0,167 \times 360^\circ = 60,1^\circ$



Podemos, ainda, construir um gráfico de setores 3D (Figura 1.5):

**Figura 1.5.** Gráfico de setores 3D para a variável *grau de instrução*



### Comandos no Software R para fazer o gráfico de setores:

```
#Pode-se entrar com os dados brutos no R e agrupá-los usando o
#comando table() como foi visto anteriormente ou, ainda, entrar
#diretamente com as frequências, como será feito a seguir.

#Entrando com as frequências no R:
tab <- c(12,18,6)

#Calculando as porcentagens:
prop <- paste(round(100*tab/sum(tab),1), "%", sep="")

#Mostrando as porcentagens:
prop

#Colocando os nomes das categorias com as porcentagens:
names(tab) <- paste(c("Fundamental", "Médio", "Superior"), prop)

#Mostrando a tabela de distribuição de frequências:
tab

#-----
#Gráfico simples (Figura 1.4):

#Plotando o gráfico:
pie(tab,col=c("lightyellow","white","lightgreen"))

#-----
#Gráfico 3D (Figura 1.5):

library(plotrix) #Precisa instalar o pacote plotrix

#Plotando o gráfico:
pie3D(tab,col=c("yellow","white","lightgreen"),
      labels=names(tab),labelcex=1)
```

### Diagrama de Linhas

O diagrama de linhas é utilizado para apresentar graficamente dados *quantitativos discretos* organizados em uma tabela de distribuição de frequências. Para construir um diagrama de linhas:

- desenhe o sistema de eixos cartesianos;
- escreva os valores assumidos pela variável no eixo das abscissas (eixo horizontal);
- escreva as frequências ou as porcentagens no eixo das ordenadas (eixo vertical);

- desenhe linhas verticais a partir dos pontos marcados no eixo das abscissas. Os comprimentos das barras são dados pelas frequências ou pelas porcentagens;
- coloque legendas nos dois eixos e título na figura.

## Exemplo

Considerando a tabela de distribuição de frequências para a variável *número de faltas ao trabalho* (Tabela 1.4), podemos representar essa distribuição de frequências por meio de um diagrama de linhas (Figura 1.6).

Recordando a Tabela 1.4	
Número de faltas	Frequência
0	9
1	10
2	5
3	3
4	2
5	0
6	1
Total	30

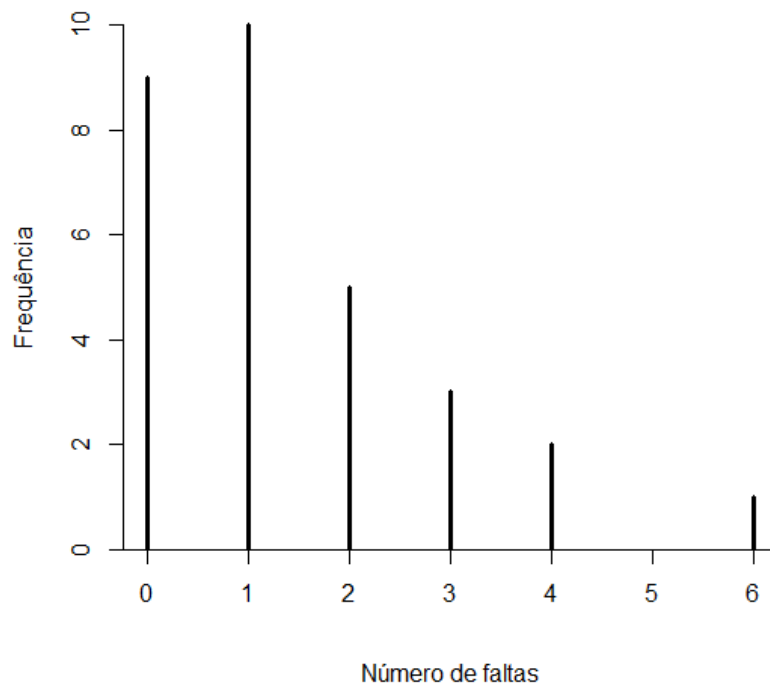
## Comandos no Software R para fazer o diagrama de linhas:

```
#Entrando com os dados no R:
dados <- c(1,3,1,1,0,1,0,1,1,0,
          2,2,0,0,0,1,2,1,2,0,
          0,1,6,4,3,3,1,2,4,0)

#Agrupando os dados em uma distribuição de frequências:
tab <- table(dados)

#Plotando o diagrama de linhas:
plot(tab,xlab="Número de faltas",ylab="Frequência",lwd=3,
      axes=F,frame.plot=F)
axis(1,pos=0); axis(2); abline(h=0)
```

Figura 1.6. Diagrama de linhas para a variável *número de faltas*



## Histograma

O histograma é utilizado para apresentar graficamente dados *quantitativos contínuos* organizados em uma tabela de distribuição de frequências. Para construir um histograma:

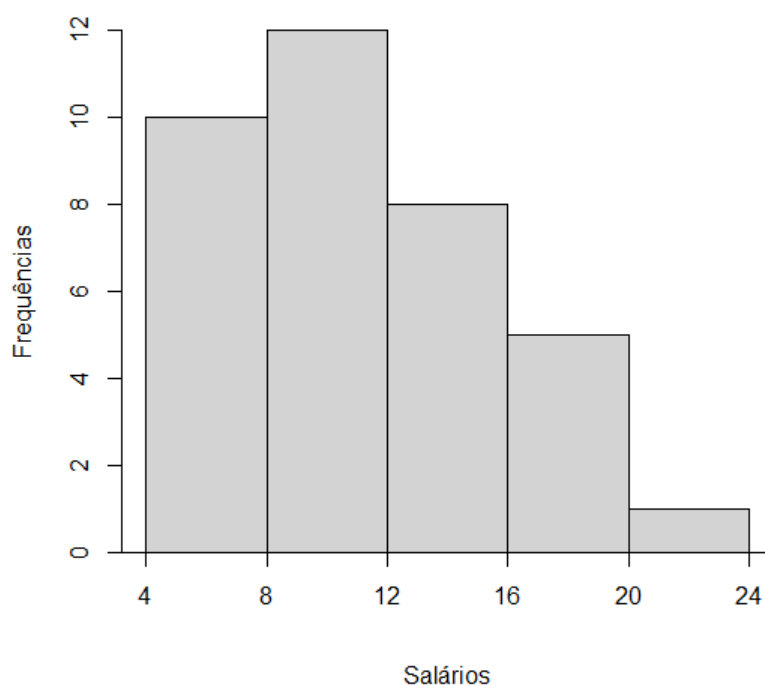
- desenhe o sistema de eixos cartesianos;
- apresente as classes no eixo das abscissas (eixo horizontal) escrevendo os limites das classes (limites inferiores e superiores). Deve-se respeitar uma escala;
- apresente as frequências no eixo das ordenadas (eixo vertical);
- desenhe barras com alturas iguais às frequências (ou às proporções ou porcentagens) das respectivas classes. As barras devem ser justapostas (**não são separadas**), a fim de evidenciar a natureza contínua da variável;
- coloque legendas nos dois eixos e título na figura.

## Exemplo

Considerando a tabela de distribuição de frequências para a variável *salário* (Tabela 1.6), podemos representar essa distribuição de frequências por meio de um histograma (Figura 1.7).

Recordando a Tabela 1.6	
Classes de salários	Frequência
4,00 ┤ 8,00	10
8,00 ┤ 12,00	12
12,00 ┤ 16,00	8
16,00 ┤ 20,00	5
20,00 ┤ 24,00	1
Total	36

**Figura 1.7.** Histograma para a variável *salário*



### Comandos no Software R para fazer o histograma:

```
#Entrando com os dados no R:
dados <- c(4.00, 4.56, 5.25, 5.73, 6.26, 6.66,
           6.86, 7.39, 7.59, 7.44, 8.12, 8.46,
           8.74, 8.95, 9.13, 9.35, 9.77, 9.80,
           10.53, 10.76, 11.06, 11.59, 12.00, 12.79,
           13.23, 13.60, 13.85, 14.69, 14.71, 15.99,
           16.22, 16.61, 17.26, 18.75, 19.40, 23.30)
```

continua...

### Comandos no Software R para fazer o histograma (continuação):

```
#Plotando o histograma:
hist(dados,breaks=c(4,8,12,16,20,24),xlab="Salários",
     ylab="Frequências",right=F,axes=F,col="lightgray",main="")
axis(1,c(4,8,12,16,20,24),pos=0); axis(2); abline(h=0)

#-----
#Pode-se, também, deixar a critério do R especificar o número de
#classes e os limites das classes:

#Plotando o histograma (sem especificar as classes):
hist(dados,xlab="Salários",ylab="Frequências",right=F,
     col="lightgray", main="")
```

**Observação:** Nas partes realçadas são especificados os limites das classes.

### Exercício

Utilizando o software R, construa o histograma referente à distribuição de frequências apresentada na Tabela 1.7.

### Polígono de frequências

Dados *quantitativos contínuos* organizados em uma tabela de distribuição de frequências também podem ser apresentados em *polígonos de frequências*. Para construir um polígono de frequências:

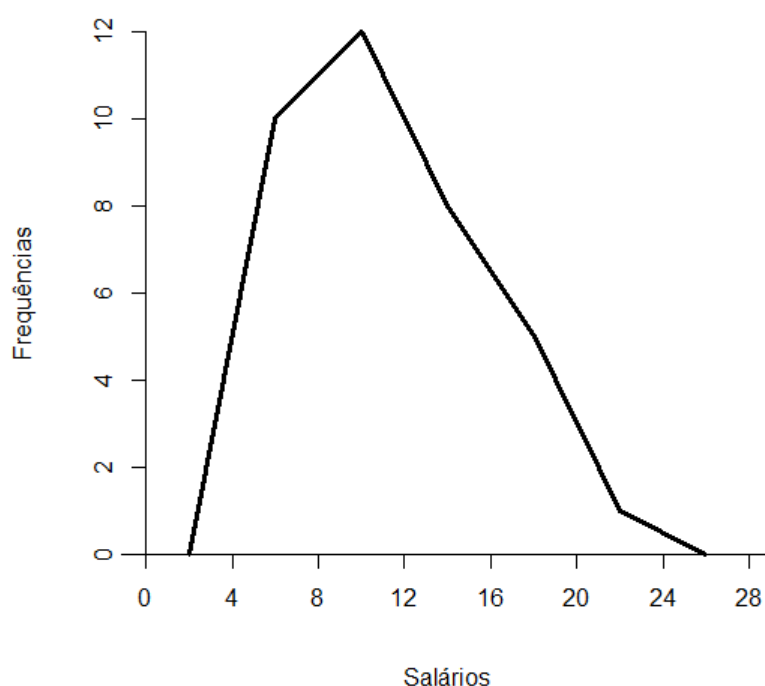
- desenhe o sistema de eixos cartesianos;
- apresente os pontos médios das classes no eixo das abscissas (eixo horizontal);
- apresente as frequências no eixo das ordenadas (eixo vertical);
- marque pontos de coordenadas  $(x, y)$  em que a coordenada  $x$  é o ponto médio da classe e a coordenada  $y$  é a frequência da classe;
- una os pontos por segmentos de reta;
- feche o polígono unindo os extremos da figura com o eixo horizontal. Utilize duas classes auxiliares, uma antes da primeira classe e outra depois da última classe, ambas com frequência zero, e utilize os pontos médios dessas classes auxiliares para fechar o polígono);
- coloque legendas nos dois eixos e título na figura.

## Exemplo

Considerando a tabela de distribuição de frequências para a variável *salário* (Tabela 1.6), podemos representar essa distribuição de frequências por meio de um polígono de frequências (Figura 1.8).

Para construir o polígono de frequências deve-se "criar" duas classes auxiliares, uma antes da primeira classe e outra depois da última classe, ambas com frequência zero, e obter os pontos médios dessas classes, como podemos observar na Tabela 1.8.

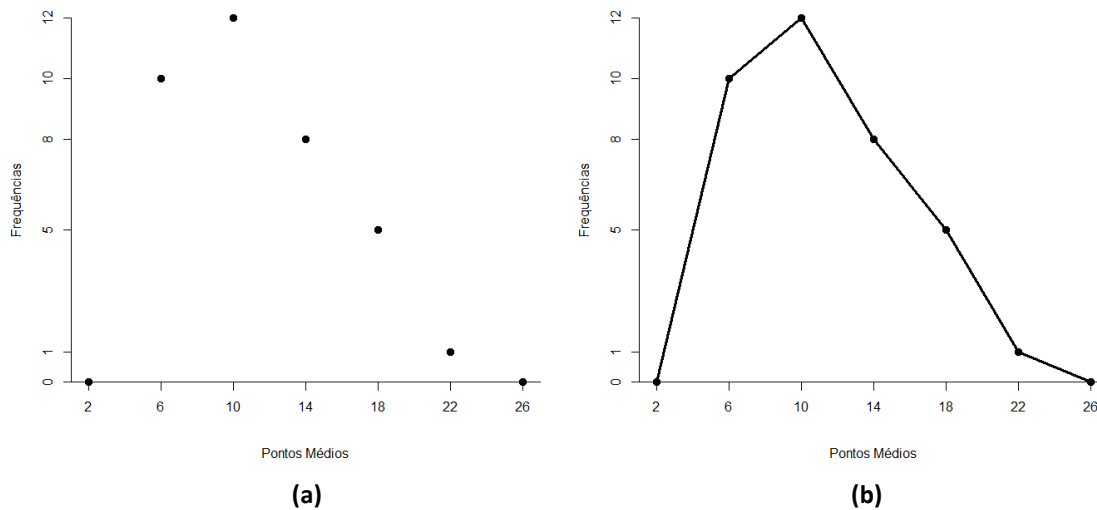
**Figura 1.8.** Polígono de frequências para a variável *salário*



**Tabela 1.8.** Tabela auxiliar para construção do polígono de frequências

Classes de salários	Ponto médio	Frequência
0,00 – 4,00	2	0
4,00 – 8,00	6	10
8,00 – 12,00	10	12
12,00 – 16,00	14	8
16,00 – 20,00	18	5
20,00 – 24,00	22	1
24,00 – 28,00	26	0

Depois devemos marcar os pontos de  $(x_i, y_i)$ , em que a coordenada  $x_i$  é o ponto médio da classe  $i$  e a coordenada  $y_i$  é a frequência da classe  $i$  (Figura 1.9a), e, por último, ligar os pontos por segmentos de retas (Figura 1.9b).

**Figura 1.9.** Construção do polígono de frequências**Comandos no Software R para fazer o polígono de frequências:**

```
#Entrando com os dados no R:
dados <- c(4.00, 4.56, 5.25, 5.73, 6.26, 6.66,
          6.86, 7.39, 7.59, 7.44, 8.12, 8.46,
          8.74, 8.95, 9.13, 9.35, 9.77, 9.80,
          10.53, 10.76, 11.06, 11.59, 12.00, 12.79,
          13.23, 13.60, 13.85, 14.69, 14.71, 15.99,
          16.22, 16.61, 17.26, 18.75, 19.40, 23.30)

#Agrupando os dados em classes:
histograma <- hist(dados, breaks=c(4,8,12,16,20,24), plot=F,
                  right=F)

#Mostrando a distribuição de frequências:
histograma

#Pontos médios das classes (Classes auxiliares pm: 2 e 26):
pontos.medios = c(2, histograma$mids, 26)
pontos.medios

#Frequências das classes (Classes auxiliares: freq. zero):
frequencias <- c(0, histograma$counts, 0)
frequencias

#Plotando o polígono de frequências:
plot(pontos.medios, frequencias, type="l", lwd=3, bty="n",
     xlab="Salários", ylab="Frequências", main="", axes=F,
     xlim=c(0,28))
axis(1,seq(0,28,4),pos=0);axis(2,pos=0);abline(h=0)
```



## 1.5 Ramo-e-folhas

Vimos que o histograma dá uma idéia da forma da distribuição da variável sob consideração. Um procedimento alternativo para resumir um conjunto de valores, com o objetivo de se obter uma idéia da forma de sua distribuição é o *ramo-e-folhas*. Uma vantagem desse diagrama sob o histograma é que não perdemos (ou perdemos pouca) informação sobre os dados em si.

Não existe uma regra fixa para construir o ramo-e-folhas, mas a idéia básica é dividir cada observação em duas partes: a primeira (o ramo) é colocada à esquerda de uma linha vertical, a segunda (as folhas) é colocada à direita. Por exemplo, para os números 4,00 e 4,56, o 4 (parte inteira) é o ramo e 00 e 56 (partes decimais) são as folhas. Para os números 91 e 97, o 9 (dezena) é o ramo e 1 e 7 (unidades) são as folhas.

### Exemplo

Considere os dados de salários de 36 indivíduos apresentados anteriormente:

4,00	4,56	5,25	5,73	6,26	6,66
6,86	7,39	7,44	7,59	8,12	8,46
8,74	8,95	9,13	9,35	9,77	9,80
10,53	10,76	11,06	11,59	12,00	12,79
13,23	13,60	13,85	14,69	14,71	15,99
16,22	16,61	17,26	18,75	19,40	23,30

Para construir o ramo-e-folhas desses dados colocaremos as partes inteiras dos dados (ramos) à esquerda de uma linha vertical e as partes decimais dos dados (folhas) à direita dessa linha (Figura 1.11).

Algumas informações que se obtêm desse ramo-e-folhas são:

- Há um grande destaque para o valor 23,30;
- Os demais valores estão razoavelmente concentrados entre 4,00 e 19,40;
- Um valor mais ou menos típico para este conjunto de dados poderia ser, por exemplo, 10,00;
- Há uma leve assimetria em direção aos valores grandes.

**Figura 1.11.** Ramo-e-folhas para a variável *salário*

4	00	56		
5	25	73		
6	26	66	86	
7	39	44	59	
8	12	46	74	95
9	13	35	77	80
10	53	76		
11	06	59		
12	00	79		
13	23	60	85	
14	69	71		
15	99			
16	22	61		
17	26			
18	75			
19	40			
20				
21				
22				
23	30			

**Comandos no Software R para fazer o ramo-e-folhas:**

```
#Entrando com os dados no R:
dados <- c(4.00, 4.56, 5.25, 5.73, 6.26, 6.66,
          6.86, 7.39, 7.44, 7.59, 8.12, 8.46,
          8.74, 8.95, 9.13, 9.35, 9.77, 9.80,
          10.53, 10.76, 11.06, 11.59, 12.00, 12.79,
          13.23, 13.60, 13.85, 14.69, 14.71, 15.99,
          16.22, 16.61, 17.26, 18.75, 19.40, 23.30)

#Mostrando os dados:
dados

#Ramo-e-folhas:
stem(dados, scale=2)
```

**Observação:** o comando "`stem()`" do R arredonda os números para uma casa decimal. Assim, por exemplo, 4,00 e 4,56 são arredondados para 4,0 e 4,6.

## 2- Medidas de posição para dados brutos e agrupados: média aritmética, moda, mediana, quantis

Vimos que o resumo de dados por meio de tabelas de frequências e ramo-e-folhas fornece muito mais informações sobre o comportamento de uma variável do que a própria tabela original de dados (dados brutos). Muitas vezes, queremos resumir ainda mais estes dados, apresentando um ou alguns valores que sejam representativos de todos os dados. Quando usamos um só valor, obtemos uma redução drástica dos dados. Usualmente, emprega-se uma das seguintes medidas de posição (ou localização) central: média, mediana ou moda.

### 2.1 Média aritmética (dados brutos)

A média aritmética, ou simplesmente média do conjunto de dados, é obtida somando-se todos os dados e dividindo-se o resultado da soma pelo número deles.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

que lê-se "x barra" é igual ao somatório de  $x$ , dividido por  $n$ .

Podemos, por simplicidade, escrever

$$\bar{x} = \frac{\sum x_i}{n}$$

em que  $\sum x_i = \sum_{i=1}^n x_i$ . Quando omitirmos o índice é porquê ele varia de 1 à  $n$ .

#### Exemplo

Se as cinco observações de uma variável forem 3, 4, 7, 8 e 8, então:

$$\bar{x} = \frac{1}{5} \sum_{i=1}^5 x_i = \frac{3 + 4 + 7 + 8 + 8}{5} = \frac{30}{5} = 6.$$

**Comandos no Software R para calcular a média (dados brutos):**

```
#Entrando com os dados no R:
dados <- c(3, 4, 7, 8, 8)

#Média:
mean(dados)
```

## 2.2 Média aritmética (dados agrupados)

### Dados discretos

Quando os dados são discretos e em grande número, pode haver repetição de valores. Nesses casos, é razoável agrupar os dados em uma tabela de distribuição de frequências. Veja a Tabela 2.1.

**Tabela 2.1.** Tabela de distribuição de frequências

Dados	Frequência
$x_1$	$f_1$
$x_2$	$f_2$
$\vdots$	$\vdots$
$x_k$	$f_k$
Total	$\sum f_i$

A média aritmética de dados agrupados em uma tabela de distribuição de frequências é dada por:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_k f_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum x_i f_i}{\sum f_i}$$

Note que  $\sum f_i = n$ , ou seja, a soma de todas as frequências é igual ao número de elementos do conjunto de dados.

### Exemplo

Considere a tabela de distribuição de frequências do número de filhos de vinte funcionários (Tabela 2.2).

**Tabela 2.2.** Distribuição de frequências para o número de filhos de vinte funcionários

Número de filhos ( $x_i$ )	Frequência ( $f_i$ )	Produto ( $x_i f_i$ )
0	6	$0 \times 6 = 0$
1	8	$1 \times 8 = 8$
2	4	$2 \times 4 = 8$
3	1	$3 \times 1 = 3$
4	0	$4 \times 0 = 0$
5	1	$5 \times 1 = 5$
Total	$\sum f_i = 20$	$\sum x_i f_i = 24$

A média do número de filhos dos 20 funcionários é calculada por:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{24}{20} = 1,2$$

**Comandos no Software R para calcular a média (dados discretos agrupados):**

```
#Entrando com os dados (xi) e as frequências (fi) no R:
xi <- c(0, 1, 2, 3, 4, 5)
fi <- c(6, 8, 4, 1, 0, 1)

#Média (dados agrupados):
weighted.mean(xi, fi)
```

Observe que se os dados não estivessem agrupados em uma tabela teríamos:

{0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1,  
1, 2, 2, 2, 2, 3, 5}

Assim:

$$\begin{aligned}
 \bar{x} &= \frac{\sum x_i}{n} = \frac{\overbrace{0+0+\dots+0}^{6 \text{ vezes}} + \overbrace{1+1+\dots+1}^{8 \text{ vezes}} + \overbrace{2+2+2+2}^{4 \text{ vezes}} + \overbrace{3}^{1 \text{ vez}} + \overbrace{5}^{1 \text{ vez}}}{20} \\
 &= \frac{0 \times 6 + 1 \times 8 + 2 \times 4 + 3 \times 1 + (4 \times 0) + 5 \times 1}{20} \\
 &= \frac{24}{20} = 1,2.
 \end{aligned}$$

## Dados contínuos

Quando os dados são contínuos e estão agrupados em uma tabela de distribuição de frequências, é preciso obter o ponto médio ( $x_i^*$ ) de cada classe. Por exemplo, a classe 4,00 – 8,00 tem dois extremos: o inferior, que é 4, e o superior, que é 8. O ponto médio dessa classe é:

$$\frac{4 + 8}{2} = \frac{12}{2} = 6.$$

Depois de obter os pontos médios de todas as classes deve-se construir uma tabela com cálculos auxiliares. Escreva as classes, os pontos médios ( $x_i^*$ ), as frequências ( $f_i$ ) de cada classe e os produtos ( $x_i^* f_i$ ).

A média é obtida por

$$\bar{x} = \frac{\sum x_i^* f_i}{\sum f_i}.$$

## Exemplo

Calcularemos a média para os dados da variável salário agrupados na Tabela 1.6. Para isso, construiremos uma tabela de cálculos auxiliares (Tabela 2.3).

**Tabela 2.3.** Tabela de cálculos auxiliares para obtenção da média

Classes de salários	Pontos médios ( $x_i^*$ )	Frequências ( $f_i$ )	Produtos ( $x_i^* f_i$ )
4,00 – 8,00	6	10	$6 \times 10 = 60$
8,00 – 12,00	10	12	$10 \times 12 = 120$
12,00 – 16,00	14	8	$14 \times 8 = 112$
16,00 – 20,00	18	5	$18 \times 5 = 90$
20,00 – 24,00	22	1	$22 \times 1 = 22$
Total	–	$\sum f_i = 36$	$\sum x_i^* f_i = 404$

Logo:

$$\bar{x} = \frac{\sum x_i^* f_i}{\sum f_i} = \frac{404}{36} = 11,22.$$

**Comandos no Software R para calcular a média (dados contínuos agrupados):**

```
#Entrando com os pontos médios (xi) e as frequências (fi) no R:  
xi <- c(6, 10, 14, 18, 22)  
fi <- c(10, 12, 8, 5, 1)  
  
#Média:  
weighted.mean(xi, fi)
```

**Observação:** Se calcularmos a média dos **dados brutos** (pág. 12) da variável salário, teremos:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{4,00 + 4,56 + 5,25 + \dots + 23,30}{36} = 11,12.$$

Note que a média calculada a partir dos dados brutos ( $\bar{x} = 11,12$ ) foi diferente da média calculada a partir dos dados agrupados ( $\bar{x} = 11,22$ ). Essa diferença ocorreu porque quando agrupamos os dados em classes perdemos informação sobre os dados. Pela Tabela 2.3 vemos que a primeira classe: 4,00 – 8,00, por exemplo, tem frequência igual a 10, ou seja, sabemos que 10 valores do conjunto de dados estão entre 4 e 8, porém, não sabemos (olhando na tabela) quais são estes valores. Assim, quando utilizamos o ponto médio da primeira classe ( $x_i = 6$ ) para representar esta classe no cálculo da média assumimos que todos os 10 valores dentro desta classe são iguais a 6 (o que não é verdade). Logo, a média calculada a partir dos dados brutos é mais precisa do que a média calculada a partir dos dados agrupados em classes.

**Exercício**

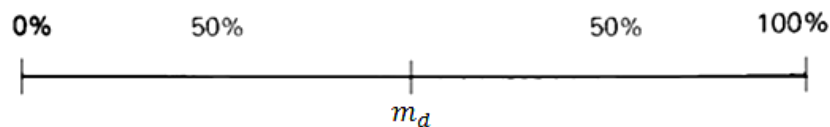
Considerando os dados agrupados da variável salário apresentados nas Tabelas 1.7 e 1.8. Pede-se:

- (a) Calcule as médias manualmente;
- (b) Calcule as médias utilizando o software R.

## 2.3 Mediana (dados brutos)

A mediana ( $m_d$ ) é o valor que ocupa a posição central do conjunto de dados ordenados. A mediana divide o conjunto de dados em duas partes:

- uma com números menores ou iguais à mediana;
- outra com números maiores ou iguais à mediana.



### Número ímpar de dados ( $n$ ímpar)

Quando o número de dados é **ímpar**, existe um único valor na posição central. Esse valor é a mediana.

#### Exemplo

O conjunto de dados

$$\{3, 5, 9\}$$

tem mediana igual a 5 ( $m_d = 5$ ), pois 5 é o valor central do conjunto de dados ordenados.

### Número par de dados ( $n$ par)

Quando o número de dados é **par**, existem dois valores na posição central. A mediana é a média desses dois valores.

#### Exemplo

O conjunto de dados

$$\{3, 5, 7, 9\}$$

tem mediana 6, pois 6 é a média de 5 e 7, que estão na posição central do conjunto de dados ordenados ( $m_d = \frac{5+7}{2} = 6$ ).



### Comandos no Software R para calcular a mediana (dados brutos):

```
#Entrando com os dados no R (n ímpar):
dados1 <- c(3,5,9)

#Mediana:
median(dados1)

#Entrando com os dados no R (n par):
dados2 <- c(3,5,7,9)

#Mediana:
median(dados2)
```

### Fórmula da Mediana

Podemos utilizar a seguinte fórmula para determinar a mediana:

$$m_d = \begin{cases} x_{(\frac{n+1}{2})} & \text{se } n \text{ for ímpar} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & \text{se } n \text{ for par} \end{cases}$$

### Exemplo

Para o conjunto de dados ordenados  $\{3, 5, 9\}$ , temos  $x_1 = 3$ ,  $x_2 = 5$  e  $x_3 = 9$ . Como  $n = 3$  é ímpar, então:

$$m_d = x_{(\frac{n+1}{2})} = x_{(\frac{3+1}{2})} = x_{(\frac{4}{2})} = x_{(2)} = 5.$$

Para o conjunto de dados ordenados  $\{3, 5, 7, 9\}$ , temos  $x_1 = 3$ ,  $x_2 = 5$ ,  $x_3 = 7$  e  $x_4 = 9$ . Como  $n = 4$  é par, então:

$$m_d = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} = \frac{x_{(\frac{4}{2})} + x_{(\frac{4}{2}+1)}}{2} = \frac{x_{(2)} + x_{(3)}}{2} = \frac{5 + 7}{2} = \frac{12}{2} = 6.$$

## 2.4 Mediana (dados agrupados)

### Dados discretos

Para dados quantitativos discretos agrupados em uma tabela de distribuição de frequências, utilizamos a frequência acumulada para verificar em que classe está o elemento da posição central  $x_{(\frac{n+1}{2})}$ , caso  $n$  for ímpar, ou os dois elementos centrais  $x_{(\frac{n}{2})}$  e  $x_{(\frac{n}{2}+1)}$  caso  $n$  for par. A frequência acumulada de uma determinada linha é a soma das frequências das linhas anteriores com a frequência desta linha (Tabela 2.4).

**Tabela 2.4.** Obtenção das frequências acumuladas

Dados ( $x_i$ )	Frequência ( $f_i$ )	Frequência Acumulada ( $F_{ac}$ )
$x_1$	$f_1$	$f_1$
$x_2$	$f_2$	$f_1 + f_2$
$x_3$	$f_3$	$f_1 + f_2 + f_3$
$\vdots$	$\vdots$	$\vdots$
$x_k$	$f_k$	$f_1 + f_2 + f_3 + \dots + f_k$
Total	$\sum f_i$	—

### Exemplo

Considerando os dados agrupados do número de filhos de vinte funcionários (Tabela 2.5), determinaremos a mediana. Usaremos as frequências acumuladas.

**Tabela 2.5.** Distribuição de frequências para o número de filhos de vinte funcionários

Número de filhos ( $x_i$ )	Frequência ( $f_i$ )	Freq. Acumulada ( $F_{ac}$ )	
0	6	6	0, 0, 0, 0, 0, 0,
1	8	14	1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
2	4	18	2, 2, 2, 2,
3	1	19	3,
4	0	19	—
5	1	20	5
Total	$n = \sum f_i = 20$	—	

Como  $n = 20$ , então a mediana é a média dos dois elementos centrais:  $x_{10}$  e  $x_{11}$ . Observando as frequências acumuladas vemos que  $x_{10}$  e  $x_{11}$  não estão na primeira linha, pois, a frequência acumulada da primeira linha é 6 (tem apenas 6 elementos até a primeira linha). Como a frequência acumulada da segunda linha é 14 (tem 14 elementos até a segunda linha) então  $x_{10}$  e  $x_{11}$  estão ambos na segunda linha (linha do número 1). Logo:

$$m_d = \frac{x_{(10)} + x_{(11)}}{2} = \frac{1 + 1}{2} = 1.$$

#### Comandos no Software R para calcular a mediana (dados discretos agrupados):

```
#Entrando com os dados (xi) e as frequências (fi) no R:
xi <- c(0,1,2,3,4,5)
fi <- c(6,8,4,1,0,1)

#Mediana:
dados <- rep(xi,fi)
median(dados)
```

### Dados contínuos

Para dados quantitativos contínuos agrupados em uma tabela de distribuição de frequências (em classes) devemos determinar a **classe mediana**, que é a classe que contém o elemento da posição  $n/2$  (independentemente de  $n$  ser par ou ímpar). Depois, determinamos o valor da mediana utilizando a fórmula:

$$m_d = LI_{md} + \frac{\frac{n}{2} - F_{ac-1}}{F_{md}} c_{md}$$

em que:

- $LI_{md}$  é o limite inferior da classe mediana;
- $c_{md}$  é a amplitude da classe mediana;
- $F_{md}$  é a frequência da classe mediana;
- $F_{ac-1}$  é a frequência acumulada da classe anterior à classe mediana. Se a classe mediana for a primeira classe então  $F_{ac-1}$  será igual a zero.

**Observação:** o número " $-1$ " que aparece em  $F_{ac-1}$  é para indicar que a frequência acumulada é da classe anterior. Não é para subtrair 1 da frequência.

## Exemplo

Considere os dados agrupados em classes na Tabela 2.6. Vamos determinar a mediana.

**Tabela 2.6.** Dados contínuos agrupados em uma distribuição de frequências

Classe	Frequência ( $f_i$ )	Freq. Acumulada ( $F_{ac}$ )
1,5 ┤ 2,0	3	3
2,0 ┤ 2,5	16	19
2,5 ┤ 3,0	32	51
3,0 ┤ 3,5	33	84
3,5 ┤ 4,0	11	95
4,0 ┤ 4,5	4	99
4,5 ┤ 5,0	1	100
Total	100	—

Como  $n = 100$ , então o elemento da posição  $n/2 = 50$  está na terceira classe, pois a frequência acumulada da terceira classe inclui o 50 ( $F_{ac} = 51$ ), ou seja, o elemento da posição 50 (ou seja,  $x_{(50)}$ ) está na terceira classe. Esta é a classe mediana (classe que contém a mediana).

Para determinar a mediana utilizaremos a fórmula:

$$m_d = LI_{md} + \frac{\frac{n}{2} - F_{ac-1}}{F_{md}} c_{md}.$$

Assim:

$$\begin{aligned}
 m_d &= 2,5 + \frac{(50 - 19)}{32} \times 0,5 \\
 &= 2,5 + 0,97 \times 0,5 \\
 &= 2,5 + 0,485 \\
 &= 2,985.
 \end{aligned}$$

Podemos ver, a seguir, como são obtidos os itens utilizados na fórmula da mediana:

Itens da fórmula da mediana:

- $LI_{md} = 2,5$ ;
- $n/2 = 100/2 = 50$ ;
- $F_{ac-1} = 19$ ;
- $F_{md} = 32$ ;
- $c_{md} = 3,0 - 2,5 = 0,5$ .

Classes	Frequência	Frequência Acumulada
1,5 – 2,0	3	3
2,0 – 2,5	16	19 $F_{ac-1}$
$LI_{md}$ 2,5 – 3,0	32 $F_{md}$	51 Contém $X_{(n/2)}$
3,0 – 3,5	33	84
3,5 – 4,0	11	95
4,0 – 4,5	4	99
4,5 – 5,0	1	100
Total	100 $n$	-

## Exercício

Calcule a mediana para os dados brutos da variável salário (pág. 12) e para os dados agrupados (Tabela 1.6).

## 2.5 Moda (dados brutos)

Moda é a realização do conjunto de dados que ocorre com maior frequência.

### Exemplo

A moda do conjunto de dados:

$$\{0, 0, 2, 5, 3, 7, 4, 7, 8, 7, 9, 6\}$$

é o número 7, porque este é o valor que ocorre o maior número de vezes.

Um conjunto de dados pode não ter moda ou ter duas ou mais modas. Assim, o conjunto de dados:

$\{0, 2, 4, 6, 8, 10\}$

não tem moda, enquanto o conjunto de dados:

$\{1, 2, 2, 3, 4, 4, 5, 6, 7\}$

tem duas modas: 2 e 4.

#### Comandos no Software R para calcular a moda (dados brutos):

```
#Entrando com os dados no R:
dados <- c(0,0,2,5,3,7,4,7,8,7,9,6)

#Moda:
tab <- table(dados)
tab
names(tab)[tab== max(tab)]
```

## 2.6 Moda (dados agrupados)

### Dados discretos

Para dados quantitativos discretos agrupados em uma tabela de distribuição de frequências a moda é o valor que tem a maior frequência (valor que ocorre maior número de vezes).

### Exemplo

Considere os dados discretos agrupados na Tabela 2.7.

**Tabela 2.7.** Dados discretos agrupados em uma distribuição de frequências

Dados ( $x_i$ )	Frequência ( $f_i$ )
0	2
1	0
2	1
3	1
4	1
5	1
6	1
<b>7</b>	<b>3</b>
8	1
9	1
Total	12

Podemos ver que a moda é o número 7, pois  $x_i = 7$  é o valor que tem a maior frequência ( $f_i = 3$ ). Portanto,  $m_o = 7$ .

## Dados contínuos

Para dados quantitativos contínuos agrupados em uma tabela de distribuição de frequências (em classes), a classe modal (classe que contém a moda) é a classe com a maior frequência. Para determinar o valor da moda utiliza-se a fórmula:

$$m_o = LI_{mo} + \frac{\Delta_1}{\Delta_1 + \Delta_2} c_{mo}$$

em que:

- $LI_{mo}$  é o limite inferior da classe modal;
- $c_{mo}$  é a amplitude da classe modal;
- $\Delta_1$  é a diferença entre a frequência da classe modal e a frequência da classe imediatamente anterior;
- $\Delta_2$  é a diferença entre a frequência da classe modal e a frequência da classe imediatamente posterior.

## Exemplo

Considere os dados agrupados em classes na Tabela 2.6. Vamos determinar a moda.

Recordando a Tabela 2.6	
Classes	Frequência
1,5 – 2,0	3
2,0 – 2,5	16
2,5 – 3,0	32
3,0 – 3,5	33
3,5 – 4,0	11
4,0 – 4,5	4
4,5 – 5,0	1
Total	100

Para determinar a moda utilizaremos a fórmula:

$$m_o = LI_{mo} + \frac{\Delta_1}{\Delta_1 + \Delta_2} c_{mo}.$$

Assim:

$$\begin{aligned} m_o &= LI_{mo} + \frac{\Delta_1}{\Delta_1 + \Delta_2} c_{mo} \\ &= 3,0 + \frac{1}{1 + 22} \times 0,5 \\ &= 3,022. \end{aligned}$$

Podemos ver, a seguir, como são obtidos os itens utilizados na fórmula da moda:

Itens da fórmula da moda:

- $LI_{mo} = 3,0$ ;
- $\Delta_1 = 33 - 32 = 1$ ;
- $\Delta_2 = 33 - 11 = 22$ ;
- $c_{mo} = 3,5 - 3,0 = 0,5$ .

Classes	Frequência
1,5 ┆ 2,0	3
2,0 ┆ 2,5	16
2,5 ┆ 3,0	32
$LI_{mo}$ 3,0 ┆ 3,5	33 <span style="color: red;">Classe modal</span>
3,5 ┆ 4,0	11 <span style="color: blue;"><math>\Delta_2 = 33 - 11 = 22</math></span>
4,0 ┆ 4,5	4
4,5 ┆ 5,0	1
Total	100

## 2.7 Moda (dados qualitativos)

A moda é a única medida de posição que também pode ser usada para descrever dados qualitativos. Nesse caso, a moda é a categoria da variável que ocorre com maior frequência.



## Exemplo

Na Tabela 2.8 é apresentada a distribuição de frequências do tipo sanguíneo de 1.167 indivíduos.

**Tabela 2.8.** Distribuição de frequências do tipo sanguíneo de 1.167 indivíduos

Grupo Sanguíneo	Frequência
O	550
A	456
B	132
AB	29
Total	1.167

A moda é o grupo sanguíneo O, pois esta foi a categoria que ocorreu com maior frequência (550).

## Utilização das medidas de tendência central

Costa (2012) propõe os seguintes critérios para escolher entre as medidas de posição:

### a) Escolha da média:

- quando a distribuição dos dados é pelo menos aproximadamente simétrica;
- quando for necessário obter posteriormente outros parâmetros que podem depender da média, como por exemplo a variância, o desvio padrão, etc.

### b) Escolha da mediana:

- quando há valores extremos;
- quando deseja-se conhecer o ponto central da distribuição;
- quando a distribuição dos dados é muito assimétrica.

### c) Escolha da moda:

- quando a medida de interesse é o ponto mais típico ou popular dos dados;
- quando precisa-se apenas de uma rápida idéia sobre a tendência central dos dados.

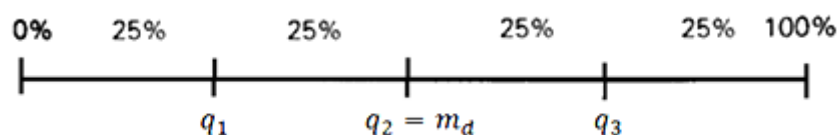
## 2.8 Quantis (dados brutos)

Vimos que a mediana é um valor que deixa metade dos dados abaixo dela e metade acima. De modo geral, podemos definir uma medida, chamada *quantil de ordem  $p$* , em que  $p$  é uma proporção qualquer,  $0 < p < 1$ , tal que  $100p\%$  das observações sejam menores do que  $q(p)$ .

Indicamos, abaixo, alguns quantis e seus nomes particulares:

$q(0,25) = q_1$ :	1° Quartil = 25° Percentil
$q(0,50) = q_2$ :	Mediana = 2° Quartil = 50° Percentil
$q(0,75) = q_3$ :	3° Quartil = 75° Percentil
$q(0,40)$ :	4° Decil = 40° Percentil
$q(0,95)$ :	95° Percentil

Note que, em particular, os quartis ( $q_1, q_2$  e  $q_3$ ) dividem o conjunto de dados em *quatro* partes iguais (ou aproximadamente iguais):



### Quartis

O primeiro quartil ( $q_1$ ) é um número tal que 25% dos dados são menores ou iguais a ele e 75% dos dados são maiores ou iguais a ele. O segundo quartil ( $q_2 = m_d$ ) é um número tal que 50% dos dados são menores ou iguais a ele e 50% dos dados são maiores ou iguais a ele. O terceiro quartil ( $q_3$ ) é um número tal que 75% dos dados são menores ou iguais a ele e 25% dos dados são maiores ou iguais a ele.

### Exemplo

**Obtendo os quartis de um conjunto com um número ímpar de dados.**

Considere o conjunto de dados:

$$\{3, 4, 1, 5, 7, 9, 2, 10, 6\}.$$

Temos que  $n = 9$  é **ímpar**. Então, a mediana é o valor central dos dados ordenados, ou seja,  $m_d = 5$ :

$$1, 2, 3, 4, \mathbf{5}, 6, 7, 9, 10.$$

Para obter o primeiro quartil,  $q_1$ , separe os dados **iguais ou menores** do que a mediana. O primeiro quartil é a mediana do novo conjunto de dados, ou seja,  $q_1 = 3$ :

$$1, 2, \mathbf{3}, 4, 5.$$

Para obter o terceiro quartil,  $q_3$ , separe os dados **iguais ou maiores** do que a mediana. O terceiro quartil é a mediana do novo conjunto de dados, ou seja,  $q_3 = 7$ :

$$5, 6, \mathbf{7}, 9, 10.$$

#### Comandos no Software R para calcular os quartis (dados brutos, n ímpar):

```
#Entrando com os dados no R:
dados <- c(3,4,1,5,7,9,2,10,6)

#Ordenando os dados:
sort(dados)

#Quartis:
quantile(dados)

#Quartis e outras medidas:
summary(dados)
```

## Exemplo

### Obtendo os quartis de um conjunto com um número par de dados.

Considere o conjunto de dados:

$$\{11, 3, 4, 1, 5, 7, 9, 2, 10, 6\}.$$

Temos que  $n = 10$  é **par**. Então, a mediana é a média dos dois valores centrais dos dados ordenados, ou seja,  $m_d = \frac{5+6}{2} = 5,5$ :

$$1, 2, 3, 4, 5, (\mathbf{5.5}), 6, 7, 9, 10, 11.$$

Para obter o primeiro quartil,  $q_1$ , separe os dados **menores** do que a mediana. O primeiro quartil é a mediana do novo conjunto de dados, ou seja,  $q_1 = 3$ :

1, 2, **3**, 4, 5.

Para obter o terceiro quartil,  $q_3$ , separe os dados **maiores** do que a mediana. O terceiro quartil é a mediana do novo conjunto de dados, ou seja,  $q_3 = 9$ :

6, 7, **9**, 10, 11.

As maneiras de se calcular os quartis apresentadas acima não são as únicas. Na realidade existem várias maneiras diferentes de se calcular quartis. Podemos ver, a seguir, que o R pode obter os quartis de formas diferentes.

#### Comandos no Software R para calcular os quartis (dados brutos, n par):

```
#Entrando com os dados no R:
dados <- c(11, 3, 4, 1, 5, 7, 9, 2, 10, 6)

#Ordenando os dados:
sort(dados)

#Quartis (método aprendido em aula, para n par):
quantile(dados, type=5)

#Quartis (método diferente do visto em aula):
quantile(dados)

#Quartis e outras medidas (método diferente do visto em aula):
summary(dados)
```

Dependendo do quantil desejado (por exemplo:  $q(0,22)$ , ou 22º Percentil) pode haver dificuldades em calculá-lo para dados brutos. Sendo assim, veremos como calcular os quantis de qualquer ordem utilizando dados agrupados.

## 2.9 Quantis (dados agrupados)

### Quartis

Vimos, anteriormente, que a mediana para dados contínuos agrupados em classes era calculada pela fórmula:

$$m_d = LI_{md} + \frac{\frac{n}{2} - F_{ac-1}}{F_{md}} c_{md}.$$

Como a mediana é o segundo quartil ( $m_d = q_2$ ), então, de maneira análoga são obtidas as fórmulas para o 1º quartil e para o 3º quartil, apresentadas a seguir.

#### **Determinação do 1º quartil ( $q_1$ ):**

**1º Passo:** Calcula-se  $n/4$ ;

**2º Passo:** Identifica-se a classe  $q_1$  pela  $F_{ac}$ ;

**3º Passo:** Aplica-se a fórmula:

$$q_1 = LI_{q_1} + \frac{\frac{n}{4} - F_{ac-1}}{F_{q_1}} c_{q_1}$$

em que:

- $LI_{q_1}$  é o limite inferior da classe do 1º quartil;
- $c_{q_1}$  é a amplitude da classe do 1º quartil;
- $F_{q_1}$  é a frequência da classe do 1º quartil;
- $F_{ac-1}$  é a frequência acumulada da classe anterior à classe do 1º quartil. Se a classe do 1º quartil for a primeira classe então  $F_{ac-1}$  será igual a zero.

**Observação:** A classe  $q_1$  é a classe que contém "o elemento da posição"  $n/4$ .

#### **Determinação do 3º quartil ( $q_3$ ):**

**1º Passo:** Calcula-se  $3n/4$ ;

**2º Passo:** Identifica-se a classe  $q_3$  pela  $F_{ac}$ ;

**3º Passo:** Aplica-se a fórmula:

$$q_3 = LI_{q_3} + \frac{\frac{3n}{4} - F_{ac-1}}{F_{q_3}} c_{q_3}$$

em que:

- $LI_{q_3}$  é o limite inferior da classe do 3º quartil;
- $c_{q_3}$  é a amplitude da classe do 3º quartil;

- $F_{q_3}$  é a frequência da classe do 3º quartil;
- $F_{ac-1}$  é a frequência acumulada da classe anterior à classe do 3º quartil. Se a classe do 3º quartil for a primeira classe então  $F_{ac-1}$  será igual a zero.

**Observação:** A classe  $q_3$  é a classe que contém "o elemento da posição"  $3n/4$ .

## Exemplo

Dada a distribuição de frequências apresentada na Tabela 2.9, iremos determinar os quartis ( $q_1$ ,  $q_2$  e  $q_3$ ).

**Tabela 2.9.** Distribuição de frequências

Classes	$f_i$	$F_{ac}$
7 – 17	6	6
17 – 27	15	21 → Classe $q_1$
27 – 37	20	41 → Classe $m_d$
37 – 47	10	51 → Classe $q_3$
47 – 57	5	56
Total	56	—

**1º e 2º passos:** Observando as frequências acumuladas encontraremos as classes que contém os quartis:

- $\frac{n}{4} = \frac{56}{4} = 14 \Rightarrow$  a 2ª classe contém "o elemento da posição"  $\frac{n}{4}$  (Classe  $q_1$ );
- $\frac{n}{2} = \frac{56}{2} = 28 \Rightarrow$  a 3ª classe contém "o elemento da posição"  $\frac{n}{2}$  (Classe  $m_d$ );
- $\frac{3n}{4} = \frac{3 \times 56}{4} = 42 \Rightarrow$  a 4ª classe contém "o elemento da posição"  $\frac{3n}{4}$  (Classe  $q_3$ ).

**3º passo:** Encontradas as classes, obteremos os itens das fórmulas:

- Para  $q_1$  temos:  $LI_{q_1} = 17$ ,  $n = 56$ ,  $F_{ac-1} = 6$ ,  $F_{q_1} = 15$ ,  $c_{q_1} = 10$ ;
- Para  $q_2 = m_d$  temos:  $LI_{m_d} = 27$ ,  $n = 56$ ,  $F_{ac-1} = 21$ ,  $F_{m_d} = 20$ ,  $c_{m_d} = 10$ ;
- Para  $q_3$  temos:  $LI_{q_3} = 37$ ,  $n = 56$ ,  $F_{ac-1} = 41$ ,  $F_{q_3} = 10$ ,  $c_{q_3} = 10$ ;

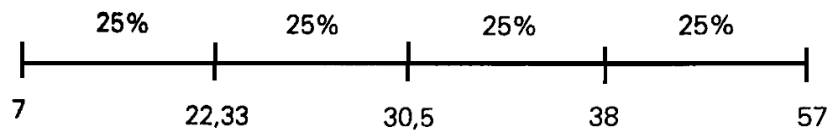
e, substituindo nas fórmulas de  $q_1$ ,  $q_2$  ( $m_d$ ) e  $q_3$ , temos:

$$\begin{aligned}
 q_1 &= LI_{q_1} + \frac{\frac{n}{4} - F_{ac-1}}{F_{q_1}} c_{q_1} \\
 &= 17 + \frac{\left(\frac{56}{4} - 6\right)}{15} \times 10 = 22,33.
 \end{aligned}$$

$$\begin{aligned}
 q_2 = m_d &= LI_{m_d} + \frac{\frac{n}{2} - F_{ac-1}}{F_{m_d}} c_{m_d} \\
 &= 27 + \frac{\left(\frac{56}{2} - 21\right)}{20} \times 10 = 30,5
 \end{aligned}$$

$$\begin{aligned}
 q_3 &= LI_{q_3} + \frac{\frac{3n}{4} - F_{ac-1}}{F_{q_3}} c_{q_3} \\
 &= 37 + \frac{\left(\frac{3 \times 56}{4} - 41\right)}{10} \times 10 = 38.
 \end{aligned}$$

Diante desses resultados, pode-se afirmar que, nesta distribuição tem-se:



Isto é:

- $q_1 = 22,33$  deixa 25% dos elementos abaixo dele;
- $q_2 = m_d = 30,5$  deixa 50% dos elementos abaixo dele;
- $q_3 = 38$  deixa 75% dos elementos abaixo dele.

## 2.10 Boxplot

As medidas que vimos anteriormente: mínimo, 1º quartil, mediana, 3º quartil e máximo - permitem traçar o diagrama de caixa (*boxplot*), que ajuda a entender a informação contida em um conjunto de dados.

Para construir o boxplot:

- desenhe um segmento de reta na posição vertical, para representar a amplitude dos dados;
- marque nesse segmento, o primeiro, o segundo e o terceiro quartis, obedecendo a escala;
- desenhe um retângulo de maneira que o lado inferior e o lado superior passem exatamente nas alturas dos pontos que marcam o primeiro e o terceiro quartis;
- faça um traço diagonal dentro do retângulo na altura do ponto que marca a mediana;
- calcule um limite inferior ( $LI$ ) e um limite superior ( $LS$ ) da seguinte maneira:

$$LI = Q_1 - 1,5d_q \quad \text{e} \quad LS = Q_3 + 1,5d_q$$

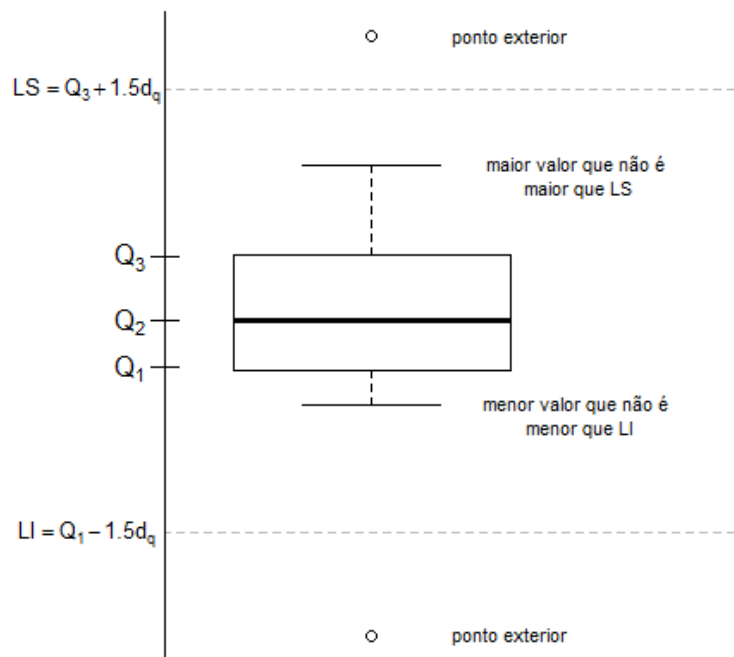
onde  $d_q = Q_3 - Q_1$  é a distância interquartílica ( $Q_3$  e  $Q_1$  são o 3º e 1º quartis);

- a partir do retângulo, para cima, segue uma linha até o maior valor dos dados que não seja maior que  $LS$ ;
- a partir do retângulo, para baixo, segue uma linha até o menor valor dos dados que não seja menor que  $LI$ ;
- as observações (dados) que forem maiores ou iguais ao limite superior ou menores ou iguais ao limite inferior são chamados *pontos exteriores* e representados por "bolinhas". Essas são observações destoante das demais e podem ou não ser o que chamamos de *outliers* ou *valores atípicos*.

O boxplot dá uma idéia da posição, dispersão, assimetria, caudas e dados discrepantes. A posição central é dada pela mediana e a dispersão por  $d_q$ . As posições relativas de  $Q_1$ ,  $Q_2$  e  $Q_3$  dão uma noção da assimetria da distribuição (voltaremos a falar sobre assimetria mais adiante).

---



Figura 2.1. Construção do *boxplot*

## Exemplo

A Figura 2.2 apresenta o boxplot para o conjunto de dados:

$$\{1, 2, 3, 4, 5, 6, 7, 10, 16\}.$$

Foram calculados:

- 1º quartil:  $Q_1 = 3$ ;
- Mediana:  $m_d = Q_2 = 5$ ;
- 3º quartil:  $Q_3 = 7$ ;

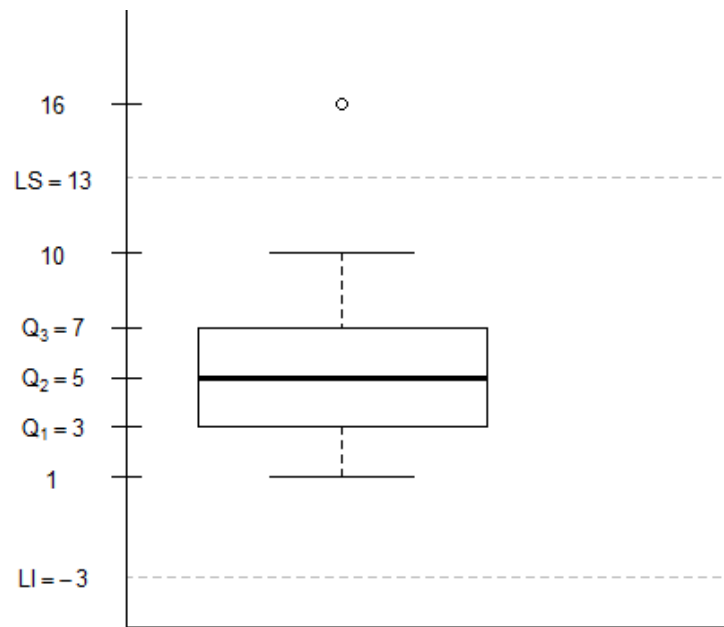
A partir desses valores podemos construir o retângulo do boxplot. Para construir os segmentos verticais acima e abaixo do retângulo, calculamos também:

- $d_q = Q_3 - Q_1 = 7 - 3 = 4$ ;
- $LI = Q_1 - 1,5d_q = 3 - 1,5 \times 4 = -3$ ;
- $LS = Q_3 + 1,5d_q = 7 + 1,5 \times 4 = 13$ ;

Porém, para construir o *boxplot* precisamos do menor valor que não é menor que  $LI$  e do maior valor que não é maior que  $LS$ :

- menor valor que não é menor que  $LI$ : 1;
- maior valor que não é maior que  $LS$ : 10;

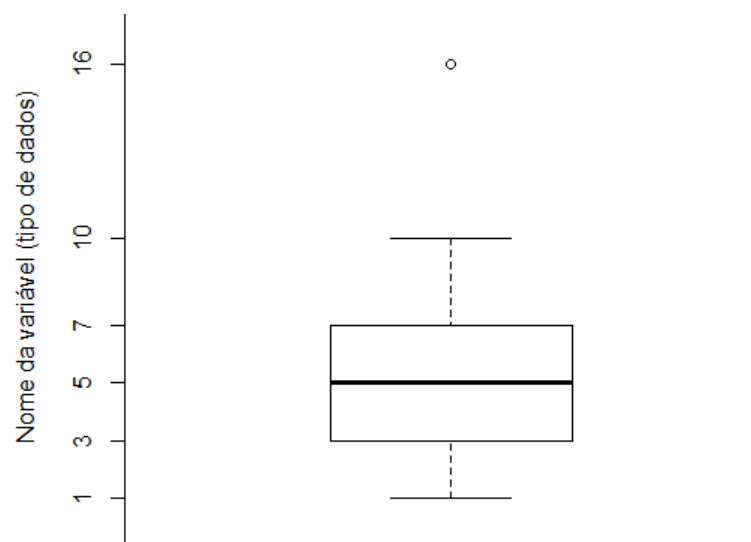
**Figura 2.2.** *Boxplot* (com limites inferior e superior apresentados)



Assim, podemos construir os segmentos de reta da parte superior do retângulo até o 10, e da parte inferior do retângulo até o 1. Como o 16 foi maior que  $LI$ , desenhamos uma "bolinha" para representar este valor. O boxplot pode ser observado na Figura 2.2.

**Observação:** Na Figura 2.2 apresentamos os limites inferior e superior, porém, não é necessário apresentar estes limites no boxplot, como podemos observar na Figura 2.3.

**Figura 2.3.** *Boxplot*

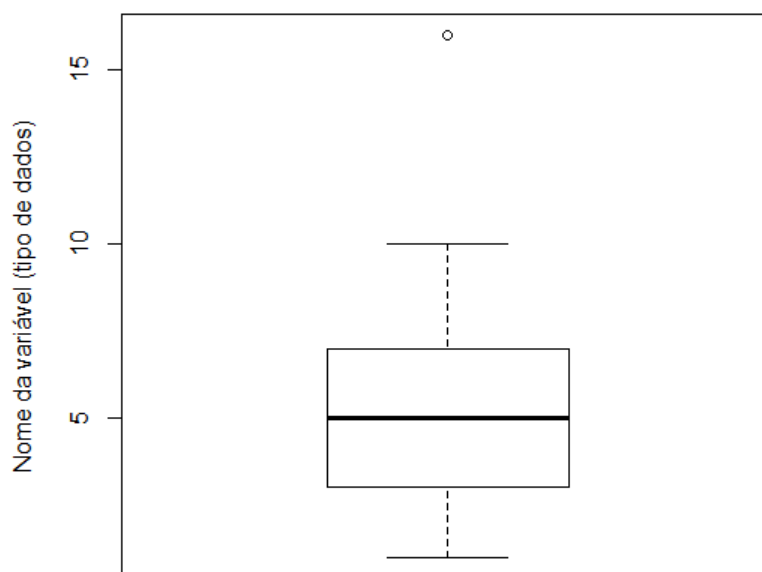


**Comandos no Software R para fazer o boxplot:**

```
#Entrando com os dados no R:  
dados <- c(1,2,3,4,5,6,7,10,16)  
  
#Plotando o boxplot:  
boxplot(dados, ylab="Nome da variável (tipo de dados)")
```

No gráfico gerado pelo R é exibida apenas a escala dos dados no eixo vertical, como pode-se observar na Figura 2.4. Podemos alterar os valores apresentados no eixo vertical utilizando o comando "axis()".

**Figura 2.4.** Boxplot gerado no R

**Comandos no Software R para fazer o boxplot alterando os valores no eixo y:**

```
#Entrando com os dados no R:  
dados <- c(1,2,3,4,5,6,7,10,16)  
  
#Plotando o boxplot:  
boxplot(dados, ylab="Nome da variável (tipo de dados)", axes=F,  
        ylim=c(0,17))  
axis(2,c(1,3,5,7,10,16))  
box()
```

## Boxplot comparativo

Outra utilidade do boxplot é na comparação de diferentes conjuntos de dados, como veremos no exemplo a seguir.

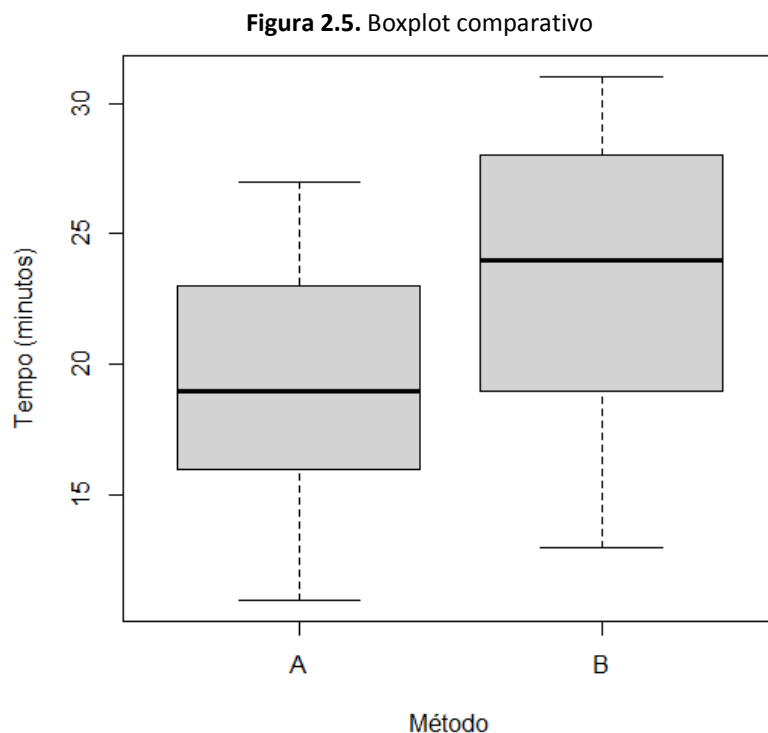
### Exemplo

Foi feito um experimento para comparar dois programas de treinamento para a execução de um serviço especializado. Vinte homens foram selecionados para esse treinamento. Dez foram escolhidos ao acaso e treinados pelo método A. Outros dez foram treinados pelo método B. Concluído o período de treinamento, todos os homens executaram o serviço e foi medido o tempo de cada um. Os dados são apresentados na Tabela 2.10.

**Tabela 2.10.** Tempo (em minutos) despendido na execução do serviço, segundo o método de treinamento

Método A	Método B
15	23
20	31
11	13
23	19
16	23
21	17
18	28
16	26
27	25
24	28

A comparação do tempo de execução de serviço pelos métodos A e B é feita utilizando dois boxplots no mesmo gráfico, que chamaremos de *boxplot comparativo* dos métodos, como é apresentado na Figura 2.5.



Pode-se observar pelo boxplot comparativo (Figura 2.5) que o tempo de execução do serviço pelo método A é menor do que o tempo pelo método B.

#### Comandos no Software R para fazer o boxplot comparativo:

```
#Entrando com os dados no R:
A <- c(15,20,11,23,16,21,18,16,27,24)
B <- c(23,31,13,19,23,17,28,26,25,28)

#Boxplot:
boxplot(A,B, names=c("A","B"), col="lightgray",
        xlab="Método", ylab="Tempo (minutos)")
```

#### Exemplo

Consideremos os dados do exemplo anterior, porém, agora dispostos como na Tabela 2.11. Podemos construir o boxplot comparativo dos dois métodos usando: Tempo em função do Método (Tempo ~ Método).

**Tabela 2.11.** Tempo (em minutos) despendido na execução do serviço, segundo o método de treinamento

Método	Tempo
A	15
A	20
A	11
A	23
A	16
A	21
A	18
A	16
A	27
A	24
B	23
B	31
B	13
B	19
B	23
B	17
B	28
B	26
B	25
B	28

Os comandos do software R são dados a seguir:

**Comandos no Software R para fazer o boxplot comparativo de Tempo ~ Método:**

```
#Entrando com os dados no R:
Tempo <- c(15, 20, 11, 23, 16, 21, 18, 16, 27, 24, 23, 31, 13, 19,
           23, 17, 28, 26, 25, 28)
Método <- c("A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "B",
            "B", "B", "B", "B", "B", "B", "B", "B", "B")

#Mostrando os dados armazenados (como na Tabela 2.11):
cbind(Método,Tempo)

#Boxplot comparativo:
boxplot(Tempo~Método, col="lightgray", xlab="Método",
        ylab="Tempo (minutos)")
```

## Exercícios

1. Considerando os dados apresentados na Tabela 2.10, calcule, para cada um dos métodos, todos os itens necessários para a construção dos boxplots (quartis, limites inferior e superior, etc.) e faça o boxplot comparativo dos dois métodos (os dois boxplots no mesmo gráfico).
2. Considerando, ainda os dois métodos de treinamento: A e B, apresentados no exemplo anterior, suponha que dois valores de A e dois valores de B tenham sido diferentes dos que foram apresentados no exemplo anterior, como podemos observar na Tabela 2.12.

**Tabela 2.12.** Tempo (em minutos) despendido na execução do serviço, segundo o método de treinamento, **com alterações nos valores**

Método	
A	B
11	23
20	46
5	13
23	19
16	23
21	17
18	28
16	36
27	25
24	28

Para esses novos dados pede-se:

- a) Calcule, para cada um dos métodos, todos os itens necessários para a construção dos boxplots (quartis, limites inferior e superior, etc.) e faça, manualmente, o boxplot comparativo dos dois métodos (no mesmo gráfico);
- b) Utilizando o R, construa o boxplot comparativo dos dois métodos (os dois boxplots no mesmo gráfico).

### 3 - Medidas de dispersão para dados brutos e agrupados: amplitude, desvio médio absoluto, variância, desvio padrão, coeficiente de variação.

As medidas de tendência central resumem a informação contida em um conjunto de dados, mas não contam toda a história. Por causa da variabilidade, a média, a mediana e a moda que estudamos no capítulo anterior não são suficientes para descrever um conjunto de dados: informam apenas a tendência central, ou seja, onde está o centro, mas nada dizem sobre a variabilidade.

Para entender esse ponto, imagine dois domicílios:

- No primeiro, moram sete pessoas, todas com 22 anos. A média de idade dos moradores desse domicílio coletivo (uma "república") é, evidentemente, 22 anos.

$$\bar{x} = \frac{22 + 22 + \dots + 22}{7} = \frac{154}{7} = 22$$

- No segundo domicílio, também moram sete pessoas: um casal - ela com 17 e ele com 23 anos, dois filhos - um com 2 e outro com 3 anos, a mãe da moça - com 38 anos, um irmão da moça - com 8 anos - e a avó da moça - com 63 anos. A média de idade nesse segundo domicílio também é de 22 anos.

$$\bar{x} = \frac{17 + 23 + 2 + 3 + 38 + 8 + 63}{7} = \frac{154}{7} = 22$$

No entanto, a "idade média de 22 anos" descreve bem a situação no primeiro domicílio, mas não no segundo.

As medidas de tendência central são tanto mais descritivas de um conjunto de dados quanto menor é a variabilidade. Então, para representar um conjunto de dados devem ser fornecidas não apenas medidas de posição, mas também uma medida de variabilidade ou dispersão.



### 3.1 Amplitude

Para medir a variabilidade, você pode fornecer o valor mínimo e o máximo do conjunto de dados. Pode, também, calcular a amplitude.

A amplitude ( $A$ ) de um conjunto de dados, definida como a diferença entre o máximo e o mínimo, é uma medida de dispersão ou variabilidade.

$$A = X_{\text{máx}} - X_{\text{mín}}$$

#### Exemplo

Cinco grupos de alunos submeteram-se a um teste, no qual obtiveram as notas apresentadas na Tabela 3.1.

**Tabela 3.1.** Notas de alunos de cinco grupos submetidos a um teste

Grupos	Alunos					Médias
	Aluno 1	Aluno 2	Aluno 3	Aluno 4	Aluno 5	
<b>Grupo A</b>	3	4	5	6	7	5
<b>Grupo B</b>	1	3	5	7	9	5
<b>Grupo C</b>	5	5	5	5	5	5
<b>Grupo D</b>	3	5	5	7	-	5
<b>Grupo E</b>	3	5	5	6	6	5

As amplitudes para cada um dos cinco grupos são dadas a seguir:

- Grupo A:  $A = 7 - 3 = 4$
- Grupo B:  $A = 9 - 1 = 8$
- Grupo C:  $A = 5 - 5 = 0$
- Grupo D:  $A = 7 - 3 = 4$
- Grupo E:  $A = 6 - 3 = 3$

A amplitude de variação é uma idéia básica em Estatística, mas um valor discrepante - por ser muito grande ou muito pequeno - aumenta muito a amplitude.

Assim, o problema em se considerar a amplitude total como medida de dispersão dos dados, é o fato dela levar em consideração em seu cálculo, apenas os valores extremos e não todos os valores. Assim, dois conjuntos de dados podem apresentar a mesma amplitude total, mesmo que tenham dispersão muito diferente. Embora fácil de calcular de interpretar, não deve ser usada normalmente como medida de dispersão.

#### Comandos no Software R para calcular a amplitude:

```
#Entrando com os dados no R:
dados <- c(3,4,5,6,7)

#Amplitude:
max(dados)-min(dados)
```

## 3.2 Desvio Médio Absoluto

Outra forma de se medir a variabilidade de uma variável é quantificando a dispersão das observações em relação a um ponto específico na distribuição, em geral, a média. À distância entre os valores observados e a média, dá-se o nome de desvio, logo

$$\text{Desvio} = x_i - \bar{x}$$

### Exemplo

Considere as notas dos alunos do Grupo A, apresentadas na Tabela 3.1.

Recordando a Tabela 3.1						
Grupos	Alunos					Média ( $\bar{x}$ )
	Aluno 1	Aluno 2	Aluno 3	Aluno 4	Aluno 5	
Grupo A	3	4	5	6	7	5

Os desvios são apresentados na Tabela 3.2.

**Tabela 3.2.** Desvios em relação à média das notas dos alunos do grupo A

<b>Nota (<math>x_i</math>)</b>	<b>Desvio (<math>x_i - \bar{x}</math>)</b>
3	$(3 - 5) = -2$
4	$(4 - 5) = -1$
5	$(5 - 5) = 0$
6	$(6 - 5) = 1$
7	$(7 - 5) = 2$
<b>Total</b>	$\sum(x_i - \bar{x}) = 0$

Observe que a soma dos desvios em relação à média é sempre zero, isto é,  $\sum(x_i - \bar{x}) = 0$ . Sendo assim, esta soma não é informativa a respeito da variabilidade dos dados, portanto, é melhor utilizar a soma dos valores absolutos (módulo) dos desvios, que será sempre positiva, isto é,  $\sum|x_i - \bar{x}|$ .

A soma dos valores absolutos será tanto maior quanto maior o número de observações ( $n$ ). Assim, o desvio absoluto médio pode ser calculado como:

$$d_m = \frac{\sum|x_i - \bar{x}|}{n}$$

Por exemplo, para as notas dos alunos do Grupo A, os desvios absolutos (módulos dos desvios) são apresentados na Tabela 3.3.

**Tabela 3.3.** Desvios e desvios absolutos das notas dos alunos do grupo A

<b>Nota (<math>x_i</math>)</b>	<b>Desvio (<math>x_i - \bar{x}</math>)</b>	<b>Desvio absoluto <math> x_i - \bar{x} </math></b>
3	$(3 - 5) = -2$	$ 3 - 5  =  -2  = 2$
4	$(4 - 5) = -1$	$ 4 - 5  =  -1  = 1$
5	$(5 - 5) = 0$	$ 5 - 5  =  0  = 0$
6	$(6 - 5) = 1$	$ 6 - 5  =  1  = 1$
7	$(7 - 5) = 2$	$ 7 - 5  =  2  = 2$
<b>Total</b>	$\sum(x_i - \bar{x}) = 0$	$\sum x_i - \bar{x}  = 6$

Para as notas dos alunos do Grupo A, temos:

$$d_m = \frac{\sum|x_i - \bar{x}|}{n} = \frac{|3 - 5| + |4 - 5| + \dots + |7 - 5|}{5} = \frac{6}{5} = 1,2.$$

### 3.3 Variância (dados brutos)

Outra forma de evitar que a soma dos desvios se anule é elevando cada desvio ao quadrado, ou seja, fazendo  $(x_i - \bar{x})^2$ . Por exemplo, para as notas dos alunos do Grupo A, os quadrados dos desvios são apresentados na Tabela 3.4.

**Tabela 3.4.** Desvios e quadrados dos desvios das notas dos alunos do grupo A

Nota ( $x_i$ )	Desvio ( $x_i - \bar{x}$ )	Quadrados dos desvios ( $(x_i - \bar{x})^2$ )
3	$(3 - 5) = -2$	$(3 - 5)^2 = (-2)^2 = 4$
4	$(4 - 5) = -1$	$(4 - 5)^2 = (-1)^2 = 1$
5	$(5 - 5) = 0$	$(5 - 5)^2 = (0)^2 = 0$
6	$(6 - 5) = 1$	$(6 - 5)^2 = (1)^2 = 1$
7	$(7 - 5) = 2$	$(7 - 5)^2 = (2)^2 = 4$
Total	$\sum(x_i - \bar{x}) = 0$	$\sum(x_i - \bar{x})^2 = 10$

A partir dos quadrados dos desvios obtemos a variância, que é a medida de variabilidade mais utilizada. A variância pode ser entendida como se fosse praticamente a "média" dos quadrados de desvios em relação à média. Numa amostra de tamanho  $n$ , este valor ( $n$ ) deveria ser usado como divisor desta soma de quadrados de desvios. No entanto, devido a motivos associados a propriedades dos estimadores, o divisor da variância amostral é dado por  $n - 1$  em lugar de  $n$  na expressão do estimador da variância. Assim, **se os dados são provenientes de uma amostra**, a **variância amostral** será denotada por  $S^2$  e será calculada da seguinte maneira:

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}.$$

Para as notas dos alunos do Grupo A, a variância será:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1} = \frac{(3 - 5)^2 + (4 - 5)^2 + \dots + (7 - 5)^2}{5 - 1} = \frac{10}{4} = 2,5.$$

**Comandos no Software R para calcular a variância:**

```
#Entrando com os dados no R:
dados <- c(3,4,5,6,7)

#Variância:
var(dados)
```

## Fórmula alternativa da variância

Outra forma de se calcular a variância é pela fórmula:

$$s^2 = \frac{1}{n-1} \left[ \sum x_i^2 - \frac{(\sum x_i)^2}{n} \right].$$

Esta fórmula é resultado de manipulações algébricas da fórmula anterior. Assim, independente de qual fórmula utilizarmos o resultado da variância será o mesmo. Considerando as notas dos alunos do Grupo A, são apresentados na Tabela 3.5 os somatórios utilizados na fórmula da variância.

**Tabela 3.5.** Tabela de cálculos auxiliares para obtenção de  $\sum x_i$  e  $\sum x_i^2$

Aluno	Nota ( $x_i$ )	$x_i^2$
1	3	$3^2 = 9$
2	4	$4^2 = 16$
3	5	$5^2 = 25$
4	6	$6^2 = 36$
5	7	$7^2 = 49$
Total	$\sum x_i = 25$	$\sum x_i^2 = 135$

Assim, obtidos  $\sum x_i = 25$  e  $\sum x_i^2 = 135$ , basta substituí-los na fórmula da variância:

$$\begin{aligned} s^2 &= \frac{1}{n-1} \left[ \sum x_i^2 - \frac{(\sum x_i)^2}{n} \right] = \frac{1}{5-1} \left[ 135 - \frac{(25)^2}{5} \right] = \frac{1}{4} \left[ 135 - \frac{625}{5} \right] \\ &= \frac{1}{4} [135 - 125] = \frac{1}{4} [10] = 2,5. \end{aligned}$$

## 3.4 Variância (dados agrupados)

### Dados discretos

Quando os dados estão dispostos em uma tabela de frequências, para se calcular a variância basta levar-se em consideração as frequências ( $f_i$ ). Temos, então:

$$S^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{n - 1}$$

ou

$$S^2 = \frac{1}{n - 1} \left[ \sum x_i^2 f_i - \frac{(\sum x_i f_i)^2}{n} \right].$$

### Exemplo

Considere a tabela de distribuição de frequências do número de filhos em idade escolar de vinte funcionários de uma empresa, apresentada na Tabela 3.6.

**Tabela 3.6.** Distribuição de frequências para o número de filhos em idade escolar de vinte funcionários

Número de filhos em idade escolar ( $x_i$ )	Frequência ( $f_i$ )
0	6
1	8
2	4
3	1
4	0
5	1

Calcularemos a variância utilizando a fórmula:

$$S^2 = \frac{1}{n - 1} \left[ \sum x_i^2 f_i - \frac{(\sum x_i f_i)^2}{n} \right].$$

Para isso, utilizaremos a Tabela 3.7 para obtenção dos cálculos auxiliares.

**Tabela 3.7.** Tabela de cálculos auxiliares para obtenção de  $\sum x_i f_i$  e  $\sum x_i^2 f_i$

$x_i$	$f_i$	$x_i f_i$	$x_i^2 f_i$
0	6	$0 \times 6 = 0$	$0^2 \times 6 = 0$
1	8	$1 \times 8 = 8$	$1^2 \times 8 = 8$
2	4	$2 \times 4 = 8$	$2^2 \times 4 = 16$
3	1	$3 \times 1 = 3$	$3^2 \times 1 = 9$
4	0	$4 \times 0 = 0$	$4^2 \times 0 = 0$
5	1	$5 \times 1 = 5$	$5^2 \times 1 = 25$
Total	$n = \sum f_i = 20$	$\sum x_i f_i = 24$	$\sum x_i^2 f_i = 58$

Assim, obtidos  $\sum x_i f_i = 24$  e  $\sum x_i^2 f_i = 58$ , basta substituí-los na fórmula da variância:

$$\begin{aligned} s^2 &= \frac{1}{n-1} \left[ \sum x_i^2 f_i - \frac{(\sum x_i f_i)^2}{n} \right] = \frac{1}{20-1} \left[ 58 - \frac{(24)^2}{20} \right] \\ &= \frac{1}{20-1} \left[ 58 - \frac{576}{20} \right] = \frac{1}{19} [58 - 28,8] = \frac{1}{19} [29,2] = 1,5 \text{ filhos}^2. \end{aligned}$$

## Dados contínuos

Para dados contínuos agrupados em classes, a fórmula da variância continua igual, porém como temos classes da forma  $[a, b)$  ao invés de números ( $x_i$ ), devemos utilizar os pontos médios das classes  $x_i^* = (a + b)/2$  para representar as classes. Assim:

$$s^2 = \frac{1}{n-1} \left[ \sum x_i^{*2} f_i - \frac{(\sum x_i^* f_i)^2}{n} \right].$$

## Exemplo

Considere a distribuição de frequências dos pesos de 86 indivíduos, apresentada na Tabela 3.8.

**Tabela 3.8.** Distribuição de frequências dos pesos (em kg) de 86 indivíduos

Pesos (Kg)	Frequências ( $f_i$ )
30 ┤ 40	8
40 ┤ 50	12
50 ┤ 60	15
60 ┤ 70	17
70 ┤ 80	14
80 ┤ 90	11
90 ┤ 100	9
Total	86

Para calcular a variância utilizaremos a fórmula:

$$s^2 = \frac{1}{n-1} \left[ \sum x_i^{*2} f_i - \frac{(\sum x_i^* f_i)^2}{n} \right]$$

em que  $x_i^*$  é o ponto médio da classe  $i$ . Utilizaremos a Tabela 3.9 para obtenção dos cálculos auxiliares.

**Tabela 3.9.** Tabela de cálculos auxiliares para obtenção de  $\sum x_i^* f_i$  e  $\sum x_i^{*2} f_i$

Classes	$x_i^*$	$f_i$	$x_i^* f_i$	$x_i^{*2}$	$x_i^{*2} f_i$
30 ┤ 40	35	8	$35 \times 8 = 280$	$35^2 = 1.225$	$1.225 \times 8 = 9.800$
40 ┤ 50	45	12	$45 \times 12 = 540$	$45^2 = 2.025$	$2.025 \times 12 = 24.300$
50 ┤ 60	55	15	$55 \times 15 = 825$	$55^2 = 3.025$	$3.025 \times 15 = 45.375$
60 ┤ 70	65	17	$65 \times 17 = 1.105$	$65^2 = 4.225$	$4.225 \times 17 = 71.825$
70 ┤ 80	75	14	$75 \times 14 = 1.050$	$75^2 = 5.625$	$5.625 \times 14 = 78.750$
80 ┤ 90	85	11	$85 \times 11 = 935$	$85^2 = 7.225$	$7.225 \times 11 = 79.475$
90 ┤ 100	95	9	$95 \times 9 = 855$	$95^2 = 9.025$	$9.025 \times 9 = 81.225$
Total	-	$\sum f_i = 86$	$\sum x_i f_i = 5.590$	-	$\sum x_i^2 f_i = 390.750$

Substituindo  $n = \sum f_i = 86$ ,  $\sum x_i^* f_i = 5.590$ , e  $\sum x_i^{*2} f_i = 390.750$  na fórmula, temos:

$$\begin{aligned}
 s^2 &= \frac{1}{n-1} \left[ \sum x_i^{*2} f_i - \frac{(\sum x_i^* f_i)^2}{n} \right] = \frac{1}{86-1} \left[ 390.750 - \frac{(5.590)^2}{86} \right] \\
 &= \frac{1}{85} \left[ 390.750 - \frac{31.248.100}{86} \right] = \frac{1}{85} [390.750 - 363.350] \\
 &= \frac{1}{85} [27.400] = 322,35 \text{ kg}^2.
 \end{aligned}$$

Observe que, no exemplo acima, a unidade de medida dos dados é dada em  $kg$  enquanto a unidade de medida da variância é dada em  $kg^2$ .

#### Comandos no Software R para calcular a variância:

```

#Entrando com os dados no R:
xi <- c(35,45,55,65,75,85,95) #Pontos médios das classes
fi <- c(8,12,15,17,14,11,9)   #Frequências

#Manipulando os dados:
dados <- rep(xi,fi)
dados
table(dados)

#Variância:
var(dados)

```



### 3.5 Desvio Padrão

No cálculo da variância, devido ao fato de se elevar os desvios ao quadrado, a unidade de medida da variância também fica elevada ao quadrado, gerando escalas sem sentido prático. Assim, se a unidade de medida dos dados seja metros ( $m$ ), a unidade de medida da variância será  $m^2$ , se a unidade de medida dos dados for  $kg$ , a unidade de medida da variância será  $kg^2$ , etc.

Uma forma de se obter uma medida de dispersão com a mesma unidade de medida dos dados observados é, simplesmente, extrair a raiz quadrada da variância, obtendo-se o desvio padrão. O desvio padrão será denotado por  $S$  e será dado por:

$$S = \sqrt{S^2}.$$

#### Exemplo

Para os dados de pesos (em quilogramas) de 86 indivíduos, apresentados na Tabela 3.8, obtivemos  $s^2 = 322,35 \text{ kg}^2$ . Assim, o desvio padrão é:

$$s = \sqrt{s^2} = \sqrt{322,35} = 17,95 \text{ kg}.$$

#### Comandos no Software R para calcular o desvio padrão:

```
#Entrando com os dados no R:
xi <- c(35,45,55,65,75,85,95) #Pontos médios das classes
fi <- c(8,12,15,17,14,11,9)   #Frequências

#Manipulando os dados:
dados <- rep(xi,fi)
dados
table(dados)

#Desvio Padrão:
sd(dados)
```

### 3.6 Coeficiente de Variação

A interpretação do desvio padrão depende da ordem de grandeza da variável em estudo. Assim, um desvio padrão igual à 10 pode ser insignificante se os valores típicos observados forem em torno de 10.000, mas pode ser muito significativo para um conjunto de dados cujos valores típicos observados sejam em torno de 100.

Logo, pode ser conveniente expressar a variabilidade dos dados de uma variável de modo independente da sua unidade de medida utilizada, tirando a influência da ordem de grandeza da variável. Tal medida é denominada coeficiente de variação. O coeficiente de variação de Pearson é a razão entre o desvio padrão e a média. Em geral, o resultado é multiplicado por 100, para que o coeficiente de variação seja dado em porcentagem. O coeficiente de variação ( $CV$ ) é dado por:

$$CV = \left( \frac{S}{\bar{X}} \cdot 100 \right) \%.$$

Sua utilidade está em fornecer uma medida para a homogeneidade de um conjunto de dados. Quanto menor o coeficiente de variação, mais homogêneo é o conjunto de dados (ou seja, mais parecidos os dados são uns com os outros). Em geral, considera-se:

- a) Baixa dispersão:  $CV < 15\%$ ;
- b) Média dispersão:  $15\% \leq CV \leq 30\%$ ;
- c) Alta dispersão:  $CV > 30\%$ .

Esta medida também pode ser bastante útil na comparação de duas variáveis ou dois grupos que, a princípio, não são comparáveis.

#### Exemplo

Na Tabela 3.10 são apresentadas a Estatura (cm), o Peso (kg) e a Idade (anos) de dez alunos aleatoriamente selecionados.

**Tabela 3.10.** Estatura (cm), Peso (kg) e Idade (anos) de dez alunos aleatoriamente selecionados

Nº do aluno	Estatura (cm)	Peso (kg)	Idade (anos)
1	177	68,0	18,0
2	162	83,0	20,1
3	188	72,0	20,5
4	157	99,9	17,7
5	166	51,0	19,2
6	153	52,0	18,9
7	158	52,0	26,9
8	176	66,5	20,1
9	168	80,0	20,7
10	163	48,0	19,3

Pede-se:

- Calcular a média ( $\bar{X}$ ), a variância ( $S^2$ ), o desvio padrão ( $S$ ) e o coeficiente de variação ( $CV$ ) para as variáveis Estatura, Peso e Idade;
- Qual variável apresenta maior variabilidade? Justifique sua resposta.
- Classifique a dispersão de cada variável como baixa, média ou alta.

### Solução (a)

Considerando os somatórios obtidos na tabela de cálculos auxiliares (Tabela 3.11) calculamos a média, a variância e o desvio padrão para cada uma das variáveis.

Para a variável "Estatura", por exemplo, temos:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1.668,00}{10} = 166,80$$

$$s^2 = \frac{1}{n-1} \left[ \sum x_i^2 - \frac{(\sum x_i)^2}{n} \right] = \frac{1}{10-1} \left[ 279.264,00 - \frac{(1.668,00)^2}{10} \right] = 115,73$$

$$s = \sqrt{s^2} = \sqrt{115,73} = 10,76$$

**Tabela 3.11.** Tabela de cálculos auxiliares

Nº do aluno	Valores da variável ( $x_i$ )			Valores ao quadrado ( $x_i^2$ )		
	Estatura ( $x_i$ )	Peso ( $x_i$ )	Idade ( $x_i$ )	Estatura ( $x_i^2$ )	Peso ( $x_i^2$ )	Idade ( $x_i^2$ )
1	177,00	68,00	18,00	31.329,00	4.624,00	324,00
2	162,00	83,00	20,10	26.244,00	6.889,00	404,01
3	188,00	72,00	20,50	35.344,00	5.184,00	420,25
4	157,00	99,90	17,70	24.649,00	9.980,01	313,29
5	166,00	51,00	19,20	27.556,00	2.601,00	368,64
6	153,00	52,00	18,90	23.409,00	2.704,00	357,21
7	158,00	52,00	26,90	24.964,00	2.704,00	723,61
8	176,00	66,50	20,10	30.976,00	4.422,25	404,01
9	168,00	80,00	20,70	28.224,00	6.400,00	428,49
10	163,00	48,00	19,30	26.569,00	2.304,00	372,49
<b>Total</b>	<b>1.668,00</b>	<b>672,40</b>	<b>201,40</b>	<b>279.264,00</b>	<b>47.812,26</b>	<b>4.116,00</b>
	$\sum x_i$	$\sum x_i$	$\sum x_i$	$\sum x_i^2$	$\sum x_i^2$	$\sum x_i^2$

Fazendo cálculos análogos para as variáveis "Peso" e "Idade" obtemos a média, a variância e o desvio padrão de todas as variáveis, os quais são apresentados na Tabela 3.12.

**Tabela 3.12.** Média, variância e desvio padrão das variáveis Estatura, Peso e Idade

Medida	Variável		
	Estatura	Peso	Idade
$\bar{X}$	166,80	67,24	20,14
$S^2$	115,73	288,90	6,64
$S$	10,76	17,00	2,58

A partir da Tabela 3.12 calcularemos o coeficiente de variação para cada uma das variáveis:

**Estatura:**

$$CV = \left( \frac{S}{\bar{X}} \cdot 100 \right) \% = \left( \frac{10,76}{166,80} \times 100 \right) \% = (0,0645 \times 100) \% = \mathbf{6,45\%}$$

**Peso:**

$$CV = \left( \frac{S}{\bar{X}} \cdot 100 \right) \% = \left( \frac{17,00}{67,24} \times 100 \right) \% = (0,2528 \times 100) \% = \mathbf{25,28\%}$$

**Idade:**

$$CV = \left( \frac{S}{\bar{X}} \cdot 100 \right) \% = \left( \frac{2,58}{20,14} \times 100 \right) \% = (0,1281 \times 100) \% = \mathbf{12,81\%}$$

**Comandos no Software R para calcular os coeficientes de variação:**

```
#Entrando com os dados no R:
Estatura <- c(177, 162, 188, 157, 166, 153, 158, 176, 168, 163)
Peso <- c(68.0, 83.0, 72.0, 99.9, 51.0, 52.0, 52.0, 66.5, 80.0,
         48.0)
Idade <- c(18.0, 20.1, 20.5, 17.7, 19.2, 18.9, 26.9, 20.1, 20.7,
          19.3)

# Coeficiente de variação para a variável Estatura:
CV1 <- 100*sd(Estatura)/mean(Estatura)
CV1

# Coeficiente de variação para a variável Peso:
CV2 <- 100*sd(Peso)/mean(Peso)
CV2

# Coeficiente de variação para a variável Idade:
CV3 <- 100*sd(Idade)/mean(Idade)
CV3
```

**Solução (b)**

A variável que apresentou maior variabilidade foi a variável Peso, pois ela foi a variável com maior coeficiente de variação ( $CV = 25,28\%$ ).

**Observação:** Note que, apesar de a variância e o desvio padrão também terem sido maiores para a variável Peso, não poderíamos ter concluído que esta variável tem maior variabilidade do que as outras a partir da variância ou do desvio padrão, pois, estas medidas (variância e desvio padrão) não podem ser utilizadas para comparar diferentes tipos de variáveis. Para essa finalidade devemos, necessariamente, utilizar o coeficiente de variação.

**Solução (c)**

**Estatura:** Dispersão baixa ( $CV < 15\%$ );

**Peso:** Dispersão média ( $15\% \leq CV \leq 30\%$ );

**Idade:** Dispersão baixa ( $CV < 15\%$ ).

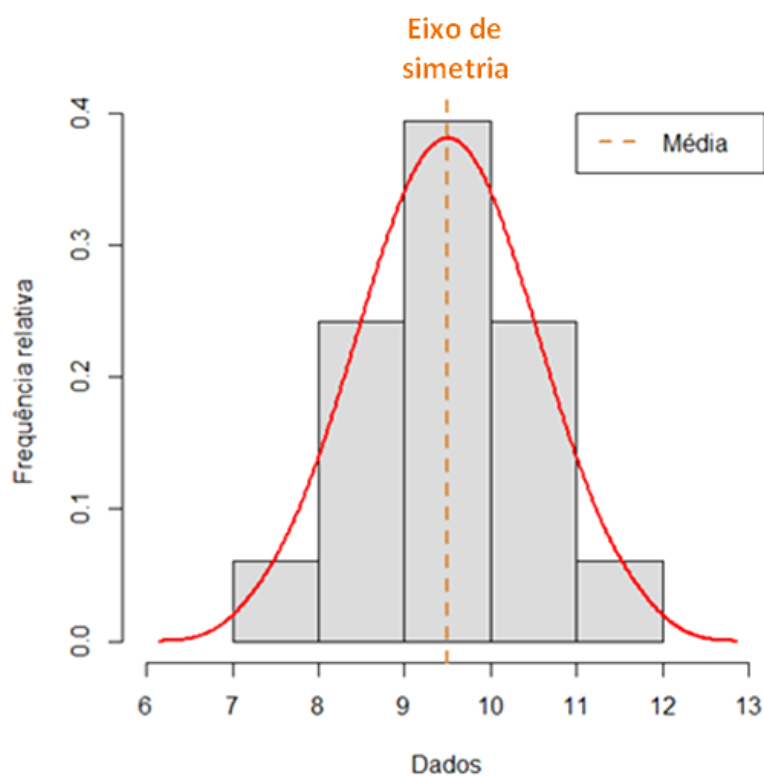
## 4 - Simetria e Curtose

### 4.1 Simetria

Numa distribuição estatística, a assimetria é o quanto sua curva de frequências se desvia ou se afasta da posição simétrica. Pode-se analisar a assimetria de uma distribuição de acordo com as relações entre suas medidas de moda, média e mediana.

Graficamente, tem-se um eixo de referência ou eixo de simetria, que é traçado sobre o valor da **média** da distribuição. Na Figura 4.1 podemos observar uma distribuição simétrica com o eixo de simetria no centro da distribuição.

**Figura 4.1.** Distribuição simétrica com eixo de simetria no centro



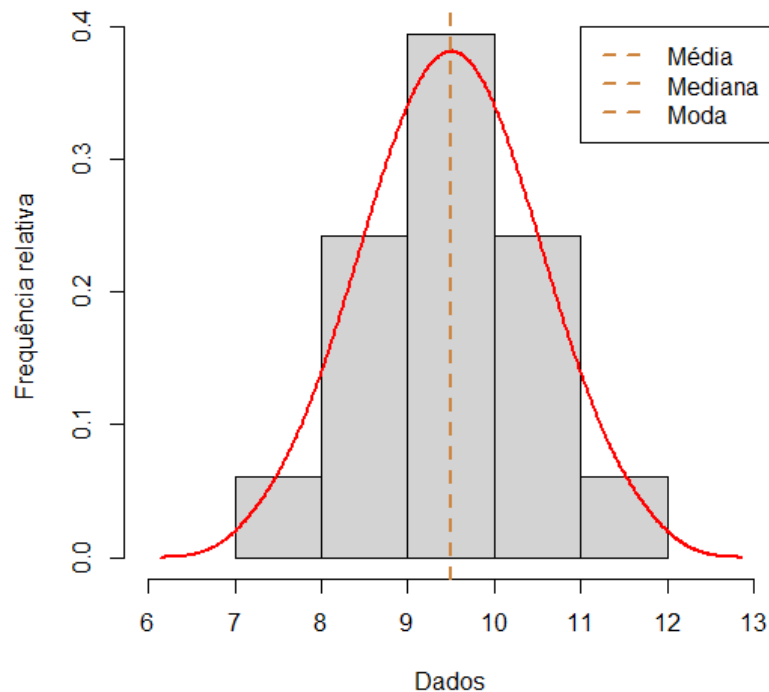
Sempre que a curva da distribuição se afastar do referido eixo, será considerada como tendo um certo grau de afastamento, que é considerado como uma assimetria da distribuição. Ou seja, assimetria é o grau de afastamento que uma distribuição apresenta do seu eixo de simetria.

Pode-se caracterizar a distribuição de frequência em:

### a) Distribuição simétrica (ou assimetria nula)

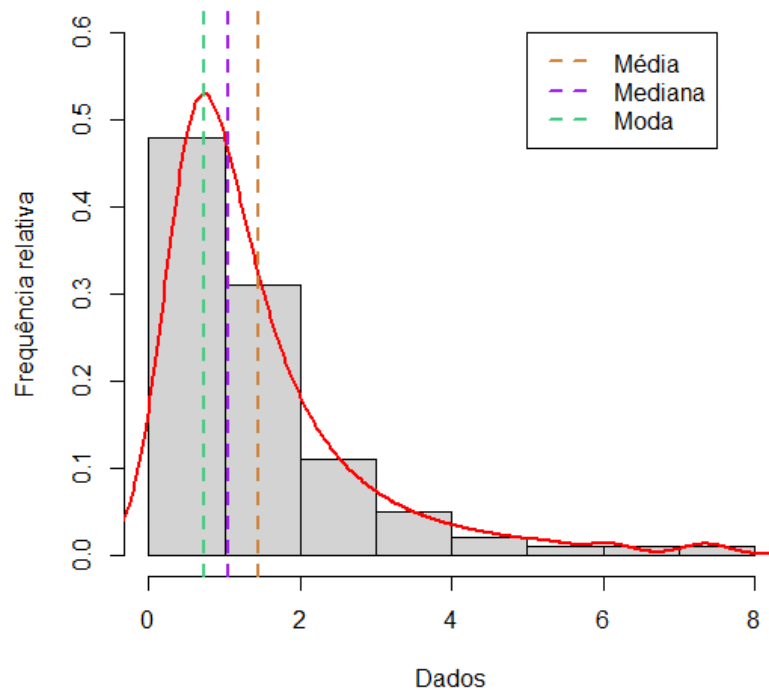
Uma distribuição é dita simétrica quando apresenta o mesmo valor para a moda, a média e a mediana, ou seja:  $\bar{x} = m_d = m_o$

Figura 4.2. Distribuição simétrica



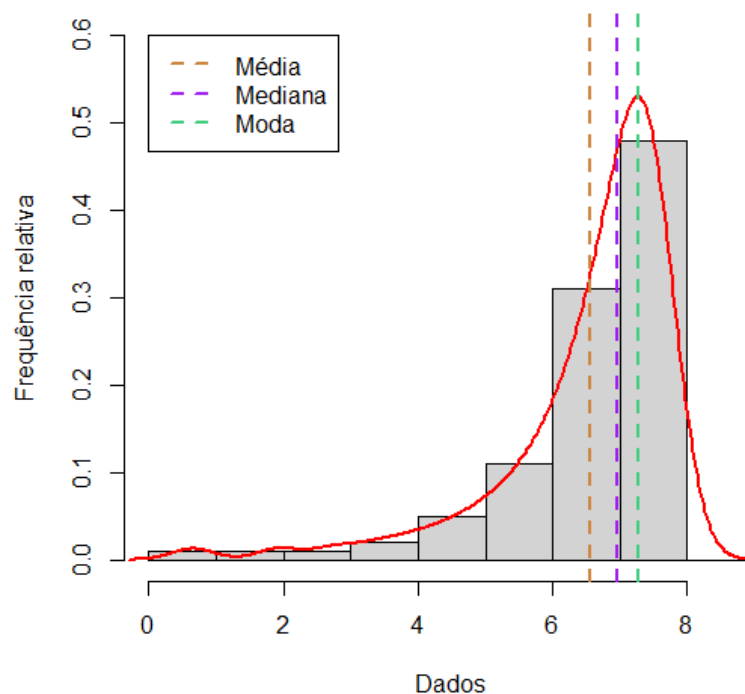
### b) Distribuição assimétrica à direita (ou positiva)

Quando a cauda da curva da distribuição declina para direita, tem-se uma distribuição com curva assimétrica positiva. Neste caso, temos:  $\bar{x} > m_d > m_o$ .

**Figura 4.3.** Distribuição assimétrica à direita

### c) Distribuição assimétrica à esquerda (ou negativa)

Analogamente, quando a cauda da curva da distribuição declina para esquerda, tem-se uma distribuição com curva assimétrica negativa. Neste caso, temos:  $\bar{x} < m_d < m_o$ .

**Figura 4.4.** Distribuição assimétrica à esquerda



Existem diversos métodos para o cálculo da medida de assimetria. Entre eles, temos:

**a) 1º coeficiente de assimetria de Pearson**

$$AS = \frac{\bar{x} - m_o}{s}.$$

**b) 2º coeficiente de assimetria de Pearson**

Quando a distribuição for quase simétrica ou moderadamente assimétrica, pode-se calcular o grau de assimetria substituindo-se a moda pela mediana, segundo a relação empírica proposta por Pearson:

$$AS = \frac{3(\bar{x} - m_d)}{s}.$$

**c) Coeficiente quartil de assimetria**

Este coeficiente, em seu cálculo, recorre apenas aos quartis. Trata-se de uma medida muito útil quando não for possível empregar o desvio-padrão como medida de dispersão. É definido por:

$$AS_q = \frac{Q_3 - 2 \cdot m_d + Q_1}{Q_3 - Q_1}.$$

**d) Coeficiente momento de assimetria**

Outra medida utilizada para avaliar a assimetria de uma distribuição de frequências é o coeficiente momento de assimetria, calculado com base nos momentos centrados de segunda e terceira ordem, definido por:

$$AS_m = \frac{m_3}{(\sqrt{m_2})^3}$$

em que:

$$m_3 = \frac{\sum (x_i - \bar{x})^3}{n} \quad e \quad m_2 = \frac{\sum (x_i - \bar{x})^2}{n},$$

ou

$$m_3 = \frac{\sum (x_i - \bar{x})^3 f_i}{n} \quad e \quad m_2 = \frac{\sum (x_i - \bar{x})^2 f_i}{n}$$

se os dados estiverem agrupados em uma distribuição de frequências.

A interpretação do coeficiente de assimetria, em qualquer dos casos é:

- $As = 0$ , então a distribuição é **simétrica**;
- $As > 0$ , a distribuição é **assimétrica positiva (à direita)**;
- $As < 0$ , a distribuição é **assimétrica negativa (à esquerda)**.

## Exemplo

Considerando a tabela de distribuição de frequências dada a seguir, determine o 1º e o 2º coeficientes de assimetria de Pearson.

**Tabela 4.1.** Pesos (em quilogramas) de 86 indivíduos

Pesos (Kg)	Frequências ( $f_i$ )
30 ┤ 40	8
40 ┤ 50	12
50 ┤ 60	15
60 ┤ 70	17
70 ┤ 80	14
80 ┤ 90	11
90 ┤ 100	9
Total	86

## Solução

- **1º coeficiente de assimetria de Pearson:**

$$AS = \frac{\bar{x} - m_o}{s}.$$

Primeiro, calcula-se  $\bar{x}$ ,  $m_o$  e  $s$ :

$$\bar{x} = \frac{\sum x_i f_i}{n} = \frac{5.590}{86} = 65$$

$$m_o = LI_{m_o} + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot c_{m_o} = 60 + \frac{2}{2 + 3} \cdot 10 = 64$$

$$s^2 = \frac{1}{n-1} \left[ \sum x_i^2 f_i - \frac{(\sum x_i f_i)^2}{n} \right] = \frac{1}{86-1} \left[ 390.750 - \frac{(5.590)^2}{86} \right] = 322,35$$

$$s = \sqrt{s^2} = \sqrt{322,35} = 17,95$$

Substituindo na fórmula:

$$AS = \frac{\bar{x} - m_o}{s} = \frac{65 - 64}{17,95} = 0,0557.$$

Portanto, a distribuição é **levemente** (AS próximo de zero) assimétrica à direita.

- **2º coeficiente de assimetria de Pearson:**

$$AS = \frac{3(\bar{x} - m_d)}{s}.$$

Primeiramente, calcula-se  $\bar{x}$ ,  $m_d$  e  $s$ . Já vimos que  $\bar{x} = 65$  e  $s = 17,95$ . Assim, resta calcular a mediana:

**Tabela 4.2.** Tabela auxiliar para o cálculo da mediana

Pesos (Kg)	Frequências ( $f_i$ )	Freq. Acumuladas ( $f_{ac}$ )
30 ┤ 40	8	8
40 ┤ 50	12	20
50 ┤ 60	15	35
60 ┤ 70	17	52
70 ┤ 80	14	66
80 ┤ 90	11	77
90 ┤ 100	9	86
Total	86	-

$$m_d = LI_{md} + \frac{\frac{n}{2} - F_{ac}}{F_{md}} \cdot c_{md} = 60 + \frac{43 - 35}{17} \cdot 10 = 64,71$$

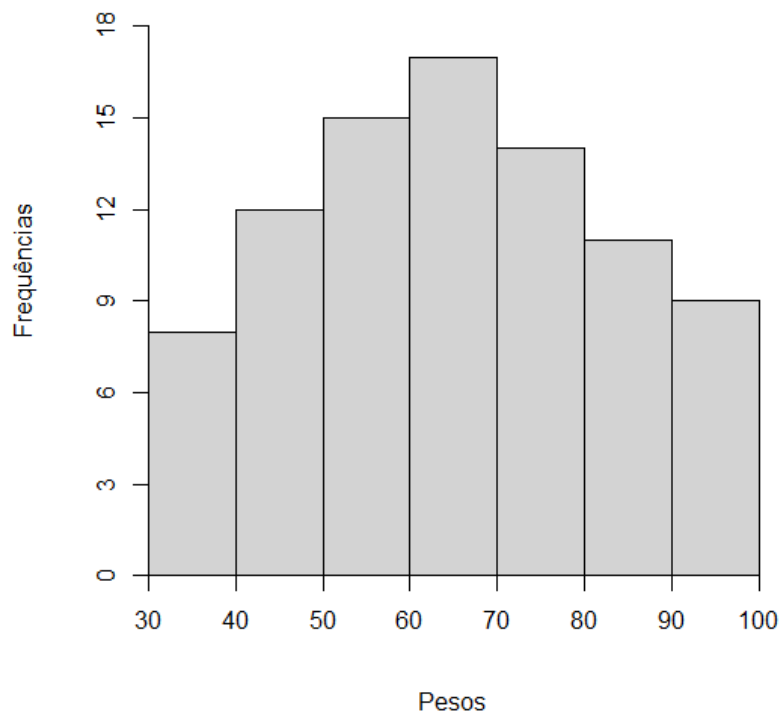
Substituindo na fórmula:

$$AS = \frac{3(\bar{x} - m_d)}{s} = \frac{3(65 - 64,71)}{17,95} = 0,04847.$$

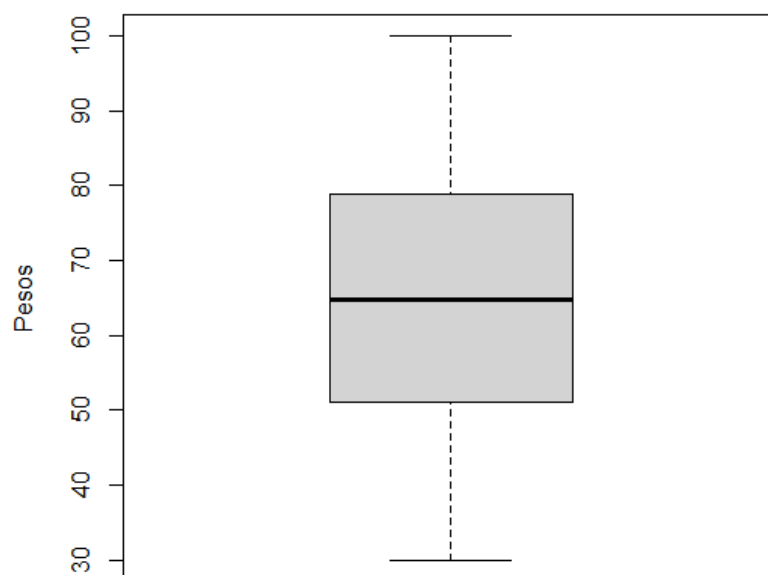
Portanto, a distribuição é **levemente** (AS próximo de zero) assimétrica à direita.

Podemos, também, observar que a distribuição é "quase simétrica" (levemente assimétrica) pelo histograma (Figura 4.5), e pelo boxplot (Figura 4.6).

**Figura 4.5.** Histograma dos pesos de 86 indivíduos



**Figura 4.6.** Boxplot dos pesos de 86 indivíduos



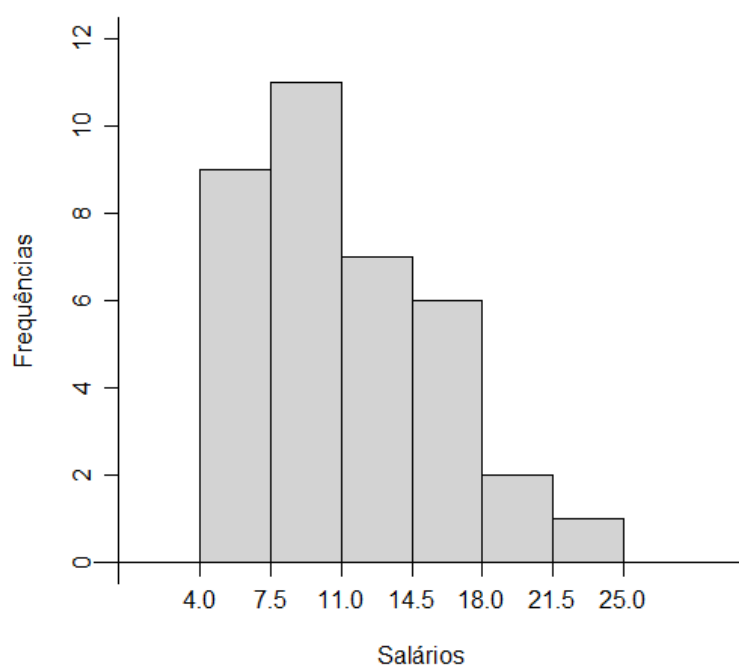
## Exemplo

Considere os salários (x sal. mín.) de 36 indivíduos:

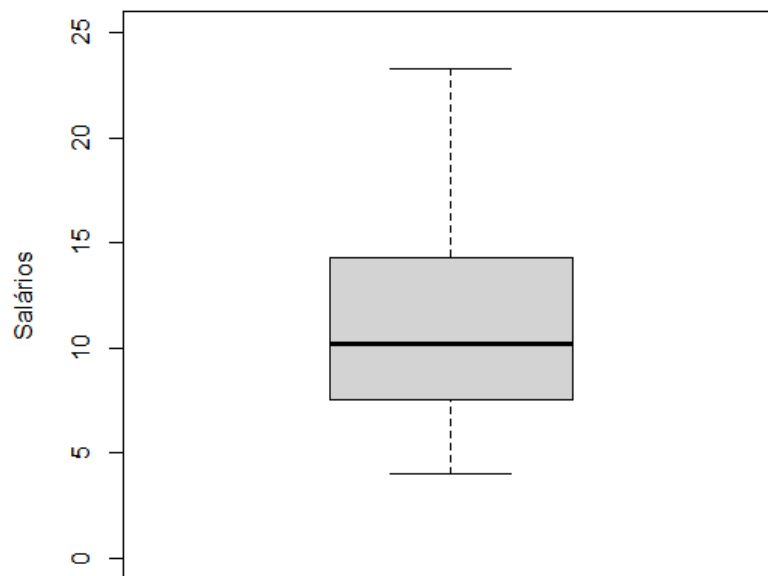
4,00	4,56	5,25	5,73	6,26	6,66
6,86	7,39	7,59	7,44	8,12	8,46
8,74	8,95	9,13	9,35	9,77	9,80
10,53	10,76	11,06	11,59	12,00	12,79
13,23	13,60	13,85	14,69	14,71	15,99
16,22	16,61	17,26	18,75	19,40	23,30

Fazendo o histograma dos dados de salários (Figura 4.7) observamos que a distribuição é assimétrica à direita.

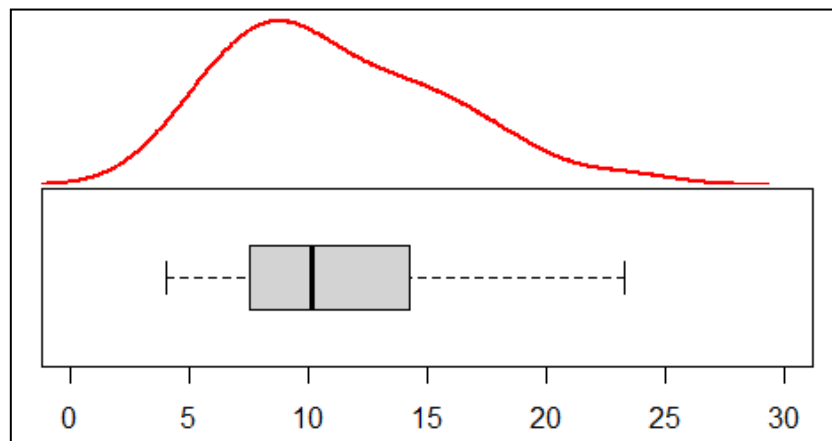
**Figura 4.7.** Histograma dos salários de 36 indivíduos



Podemos, também, avaliar a assimetria dos dados através do boxplot. Podemos ver pelo boxplot desses dados (Figura 4.8) que a mediana está mais próxima do 1º quartil do que do 3º quartil; além disso, a distância entre o 3º quartil e o máximo (segmento superior) é maior que a distância entre o 1º quartil e o mínimo (segmento inferior) do gráfico. Isso indica que a curva da distribuição tem uma cauda mais longa à direita, ou seja, indica uma assimetria à direita dos dados.

**Figura 4.8.** Boxplot dos salários de 36 indivíduos

Na Figura 4.9 podemos observar a relação entre o boxplot e a curva da distribuição dos dados. Podemos observar que o segmento superior mais longo do boxplot corresponde a uma cauda direita mais longa da curva da distribuição (assimetria à direita).

**Figura 4.9.** Boxplot e curva da distribuição dos dados

## Exercício

Considerando os dados de salários de 36 indivíduos, pede-se:

- Calcule o 2º coeficiente de assimetria de Pearson e interprete o resultado;
- Calcule o coeficiente momento de assimetria e interprete o resultado.

### Comandos no Software R para calcular o coeficiente momento de assimetria:

```
#Entrando com os dados no R:
dados <- c(4.00, 4.56, 5.25, 5.73, 6.26, 6.66,
          6.86, 7.39, 7.59, 7.44, 8.12, 8.46,
          8.74, 8.95, 9.13, 9.35, 9.77, 9.80,
          10.53, 10.76, 11.06, 11.59, 12.00, 12.79,
          13.23, 13.60, 13.85, 14.69, 14.71, 15.99,
          16.22, 16.61, 17.26, 18.75, 19.40, 23.30)

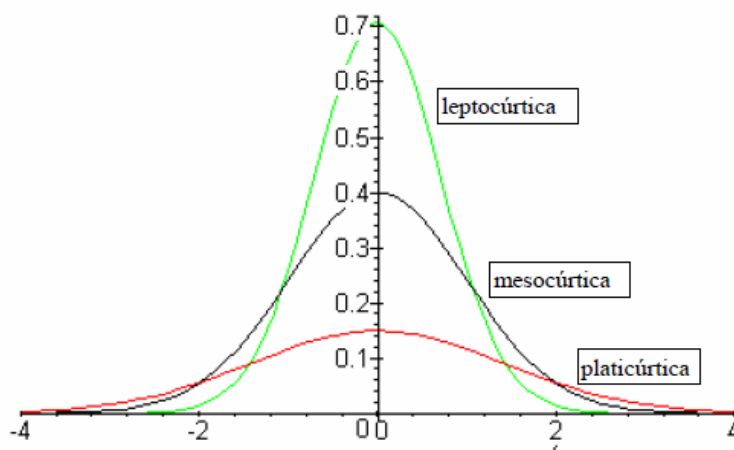
#Carregando o pacote "moments" (precisa instalar)
library(moments)

#Coeficiente momento de assimetria:
skewness(dados)
```

## 4.2 Curtose

A curtose é uma medida do grau de achatamento da distribuição quando comparada ao de uma distribuição conhecida como distribuição normal (que será vista mais adiante).

**Figura 4.10.** Distribuições com diferentes graus de curtose



Para avaliar o grau de curtose de uma curva ou distribuição de frequências, pode-se adotar dois tipos de medidas:

### a) Coeficiente percentílico de curtose

É a medida mais elementar usada para avaliar o grau de curtose de uma distribuição ou curva de frequências. É definido por:

$$k_p = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$$

em que:  $Q_1$  e  $Q_3$  são o 1º e 3º quartis e,  $P_{10}$  e  $P_{90}$  são o 10º e 90º percentis.

Neste caso, tem-se que:

- se  $k_p = 0,263$ , a curva ou distribuição é **mesocúrtica**;
- se  $k_p > 0,263$ , a curva ou distribuição é **platicúrtica**;
- se  $k_p < 0,263$ , a curva ou distribuição é **leptocúrtica**.

**Observação:** Assim como os quartis dividem a amostra em quatro partes iguais, os percentis são medidas que dividem a amostra em 100 partes iguais. O cálculo de um percentil é dado por:

$$P_i = LI_{P_i} + \frac{\left(\frac{i \times n}{100} - F_{ac-1}\right)}{F_{P_i}} \cdot c_{P_i}$$

em que:

- $LI_{P_i}$ : é o limite inferior da classe  $P_i$ ;
- $n$ : é o tamanho da amostra;
- $F_{ac}$ : é a frequência acumulada das classes anteriores à classe  $P_i$ ;
- $F_{P_i}$ : é a frequência da classe  $P_i$ ;
- $c_{P_i}$ : é a amplitude da classe  $P_i$ .

### b) coeficiente momento de curtose

Utiliza-se do quociente entre o momento centrado de quarta ordem e o quadrado do momento centrado de segunda ordem, dado por:

$$k_m = \frac{m_4}{(m_2)^2}$$



em que  $m_2$  é o segundo momento central  $m_4$  é o quarto momento central.

Temos:

$$m_2 = \frac{\sum (x_i - \bar{x})^2}{n} \quad \text{e} \quad m_4 = \frac{\sum (x_i - \bar{x})^4}{n},$$

ou

$$m_2 = \frac{\sum (x_i - \bar{x})^2 f_i}{n} \quad \text{e} \quad m_4 = \frac{\sum (x_i - \bar{x})^4 f_i}{n}$$

se os dados estiverem agrupados em uma distribuição de frequências.

A interpretação do coeficiente momento de curtose é:

- se  $k_m = 3$ , a curva ou distribuição é **mesocúrtica**;
- se  $k_m > 3$ , a curva ou distribuição é **leptocúrtica**.
- se  $k_m < 3$ , a curva ou distribuição é **platicúrtica**;

**Obs.:** A curtose calculada usando o R é baseada no coeficiente momento de curtose.

## Exemplo

Determinar o coeficiente percentílico de curtose da distribuição a seguir:

**Tabela 4.3.** Tabela de distribuição de frequências

Classes	Frequências
3 – 8	5
8 – 13	15
13 – 18	20
18 – 23	10
Total	50

## Solução

- Posição do elemento  $Q_1$ :  $\frac{n}{4} = \frac{50}{4} = 12,5^{\text{a}}$  posição;
- Posição do elemento  $Q_3$ :  $\frac{3n}{4} = \frac{3 \times 50}{4} = 37,5^{\text{a}}$  posição;
- Posição do elemento  $P_{10}$ :  $\frac{10n}{100} = \frac{50}{10} = 5^{\text{a}}$  posição;
- Posição do elemento  $P_{90}$ :  $\frac{90n}{100} = \frac{90 \times 50}{100} = 45^{\text{a}}$  posição;

**Tabela 4.4.** Tabela de cálculos auxiliares

Classes	$f_i$	$f_{ac}$	
3 – 8	5	5	← classe $P_{10}$
8 – 13	15	20	← classe $Q_1$
13 – 18	20	40	← classe $Q_3$
18 – 23	10	50	← classe $P_{90}$
Total	50	–	

Primeiramente, calcula-se  $P_{10}$ ,  $Q_1$ ,  $Q_3$  e  $P_{90}$ :

$Q_1 = P_{25}$ $= LI_{P_{25}} + \frac{\left(\frac{25 \times n}{100} - F_{ac-1}\right)}{F_{P_{25}}} \cdot c_{P_{25}}$ $= 8 + \frac{\left(\frac{25 \times 50}{100} - 5\right)}{15} \cdot 5$ $= \mathbf{10,5}$	$P_{10} = LI_{P_{10}} + \frac{\left(\frac{10 \times n}{100} - F_{ac-1}\right)}{F_{P_{10}}} \cdot c_{P_{10}}$ $= 3 + \frac{\left(\frac{10 \times 50}{100} - 0\right)}{5} \cdot 5$ $= \mathbf{8}$
$Q_3 = P_{75}$ $= LI_{P_{75}} + \frac{\left(\frac{75 \times n}{100} - F_{ac-1}\right)}{F_{P_{75}}} \cdot c_{P_{75}}$ $= 13 + \frac{\left(\frac{75 \times 50}{100} - 20\right)}{20} \cdot 5$ $= \mathbf{17,38}$	$P_{90} = LI_{P_{90}} + \frac{\left(\frac{90 \times n}{100} - F_{ac-1}\right)}{F_{P_{90}}} \cdot c_{P_{90}}$ $= 18 + \frac{\left(\frac{90 \times 50}{100} - 40\right)}{10} \cdot 5$ $= \mathbf{20,5}$

Substituindo na equação:

$$k_p = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})} = \frac{17,38 - 10,5}{2(20,5 - 8)} = \mathbf{0,27}.$$

Portanto,  $k_p > 0,263$ , logo a distribuição é **platicúrtica**.

## Exemplo

Considere os salários (x sal. mín.) de 36 indivíduos:

4,00	4,56	5,25	5,73	6,26	6,66
6,86	7,39	7,59	7,44	8,12	8,46
8,74	8,95	9,13	9,35	9,77	9,80
10,53	10,76	11,06	11,59	12,00	12,79
13,23	13,60	13,85	14,69	14,71	15,99
16,22	16,61	17,26	18,75	19,40	23,30

Determine o coeficiente momento de curtose.

## Solução

Precisamos calcular

$$m_2 = \frac{\sum (x_i - \bar{x})^2}{n} \quad \text{e} \quad m_4 = \frac{\sum (x_i - \bar{x})^4}{n}$$

para substituir na fórmula:

$$k_m = \frac{m_4}{(m_2)^2}.$$

Temos:

$$\bar{x} = 11,12;$$

$$m_2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{736,57}{36} = 20,46;$$

$$m_4 = \frac{\sum (x_i - \bar{x})^4}{n} = \frac{42.583,77}{36} = 1.182,88;$$

Assim,

$$k_m = \frac{m_4}{(m_2)^2} = \frac{1.182,88}{(20,46)^2} = 2,83.$$

Portanto,  $k_m < 3$ , logo a distribuição é **platicúrtica**.

### Comandos no Software R para calcular o coeficiente momento de curtose:

```
#Entrando com os dados no R:
dados <- c(4.00, 4.56, 5.25, 5.73, 6.26, 6.66,
          6.86, 7.39, 7.59, 7.44, 8.12, 8.46,
          8.74, 8.95, 9.13, 9.35, 9.77, 9.80,
          10.53, 10.76, 11.06, 11.59, 12.00, 12.79,
          13.23, 13.60, 13.85, 14.69, 14.71, 15.99,
          16.22, 16.61, 17.26, 18.75, 19.40, 23.30)

#Carregando o pacote "moments" (precisa instalar)
library(moments)

#Coeficiente momento de curtose:
kurtosis(dados)
```

## 5 - Análise bidimensional

Até agora vimos como organizar e resumir informações pertinentes a uma única variável, mas frequentemente estamos interessados em analisar o comportamento conjunto de duas ou mais variáveis aleatórias.

### 5.1 Variáveis Qualitativas: Tabelas de Contingência e Coeficiente de Contingência.

#### Exemplo

Suponha que queiramos analisar o comportamento conjunto das variáveis  $X$ : grau de instrução e  $Y$ : região de procedência. A distribuição de frequências é representada por uma tabela de dupla entrada e está na Tabela 5.1.

**Tabela 5.1.** Distribuição conjunta das frequências das variáveis grau de instrução ( $X$ ) e região de procedência ( $Y$ ).

$Y \backslash X$	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

A linha dos totais fornece a distribuição da variável  $X$ :

$X$	Ensino Fundamental	Ensino Médio	Superior	
<b>Total</b>	12	18	6	

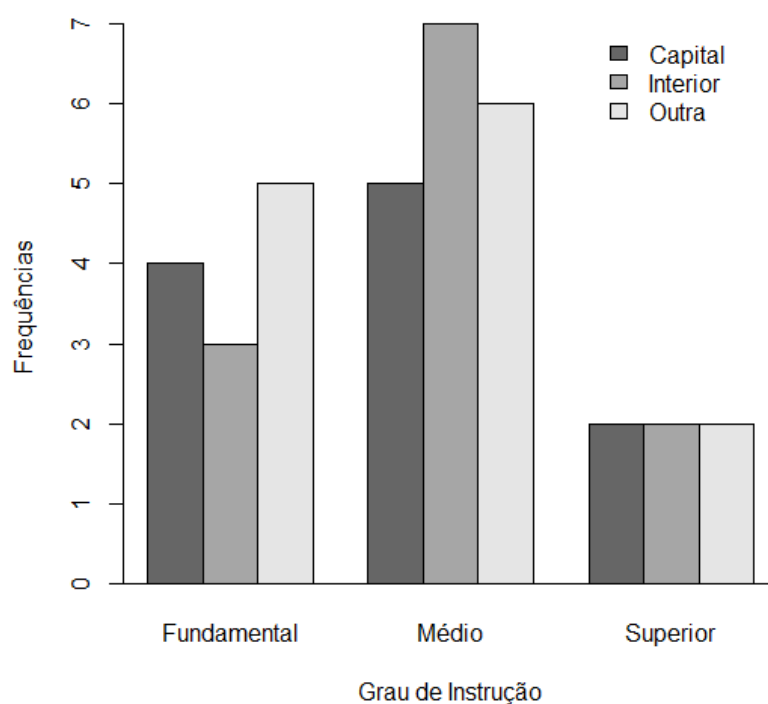
ao passo que a coluna dos totais fornece a distribuição da variável  $Y$ :

$Y$		<b>Total</b>
<b>Capital</b>		11
<b>Interior</b>		12
<b>Outra</b>		13

As distribuições assim obtidas são chamadas tecnicamente de *distribuições marginais*, enquanto que a Tabela 5.1 constitui a distribuição conjunta de  $X$  e  $Y$ .

A comparação entre duas variáveis também pode ser feita utilizando-se representações gráficas. Na Figura 5.1 temos a representação gráfica da distribuição conjunta das variáveis grau de instrução ( $X$ ) e região de procedência ( $Y$ ).

**Figura 5.1.** Distribuição conjunta das frequências das variáveis grau de instrução ( $X$ ) e região de procedência ( $Y$ ).



### Comandos no Software R para fazer o gráfico da distribuição conjunta:

```
#Entrando com a matriz da distribuição conjunta no R:
tabela <- matrix(c(4, 5, 2,
                  3, 7, 2,
                  5, 6, 2), 3, 3, byrow=T)

#Adicionando os nomes das linhas e colunas:
rownames(tabela) <- c("Capital", "Interior", "Outra")
colnames(tabela) <- c("Fundamental", "Médio", "Superior")

#Mostrando a tabela:
tabela

#Plotando o gráfico da distribuição conjunta:
barplot(tabela, beside=T, col=c("gray40", "gray65", "gray90"),
        xlab="Grau de Instrução", ylab="Frequências")
abline(h=0)

#Adicionando as legendas (nomes das colunas por cores):
legend("topright", legend=rownames(tabela),
       fill=c("gray40", "gray65", "gray90"))
```

## 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 o resultado de uma delas quando conhecermos a realização da outra.

### Exemplo

Queremos verificar se existe ou não associação entre o sexo e a carreira escolhida por 200 alunos de Economia e Administração (Tabela 5.2).

**Tabela 5.2.** 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

Inicialmente, verificamos que fica muito difícil tirar alguma conclusão, devido à diferença entre os totais marginais. Devemos, pois, construir as proporções segundo as linhas ou as colunas para podermos fazer comparações. Calcularemos as proporções segundo os totais das colunas (Tabela 5.3).

**Tabela 5.3.** Distribuição conjunta das proporções (em porcentagem) , segundo os totais das colunas, de alunos segundo o sexo ( $X$ ) e o curso escolhido ( $Y$ ).

$Y \backslash X$	Masculino	Feminino	Total
<b>Economia</b>	$85/140 = 61\%$	$35/60 = 58\%$	$120/200 = 60\%$
<b>Administração</b>	$55/140 = 39\%$	$25/60 = 42\%$	$80/200 = 40\%$
<b>Total</b>	$140 = 100\%$	$60 = 100\%$	$200 = 100\%$

Podemos observar que, independentemente do sexo, 60% das pessoas preferem economia e 40% das pessoas preferem administração (observe na coluna de totais da Tabela 5.3). Não havendo dependência entre as variáveis, esperaríamos essas mesmas proporções para cada sexo. 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 dados considerado. Concluimos então que, neste caso, as variáveis sexo e escolha do curso parecem **não estar associadas**.

## Exemplo

Vamos considerar, agora, um problema semelhante, mas envolvendo alunos de Física e Ciências Sociais (Tabela 5.4).

**Tabela 5.4.** Distribuição conjunta de alunos segundo o sexo ( $X$ ) e o curso escolhido ( $Y$ ).

$Y \backslash X$	Masculino	Feminino	Total
<b>Física</b>	100	20	120
<b>Ciências sociais</b>	40	40	80
<b>Total</b>	140	60	200

Novamente, devemos construir as proporções segundo as linhas ou as colunas para podermos fazer comparações. Calcularemos as proporções segundo os totais das colunas (Tabela 5.5).

**Tabela 5.5.** Distribuição conjunta das proporções (em porcentagem), segundo os totais das colunas, de alunos segundo o sexo ( $X$ ) e o curso escolhido ( $Y$ ).

$Y \backslash X$	Masculino	Feminino	Total
Física	$100/140 = 71\%$	$20/60 = 33\%$	$120/200 = 60\%$
Ciências sociais	$40/140 = 29\%$	$40/60 = 67\%$	$80/200 = 40\%$
Total	$140 = 100\%$	$60 = 100\%$	$200 = 100\%$

Comparando a distribuição da preferência pelos cursos independentemente do sexo (coluna de totais da Tabela 5.5) com as distribuições diferenciadas por sexo (colunas de masculino e feminino), observamos uma disparidade bem acentuada nas proporções: Masculino (71% e 29%), Feminino (33% e 67%) e Total (60% e 40%). Parece, pois, haver maior concentração de homens no curso de Física e de mulheres no de Ciências Sociais. Portanto, nesse caso, as variáveis sexo e curso escolhido parecem **estar associadas**.

#### Comandos no Software R para obter as proporções segundo as colunas:

```
#Entrando com a tabela da distribuição conjunta no R:
tabela <- matrix(c(100, 20,
                   40, 40), 2, 2, byrow=T)

#Adicionando os nomes das linhas e colunas:
rownames(tabela) <- c("Física", "Ciências Sociais")
colnames(tabela) <- c("Masculino", "Feminino")

#Mostrando a tabela da distribuição conjunta:
tabela

#Tabela das proporções segundo os totais das colunas:
tabela2 <- addmargins(prop.table(addmargins(tabela,2),2),1)

#Mostrando a tabela (com 4 casas decimais - comando "round"):
round(tabela2, 4)
```



## 5.2 Medidas de dependência entre duas variáveis nominais (qui-quadrado)

Como vimos no exemplo anterior, a análise da Tabela 5.5 mostra a existência de certa dependência entre as variáveis. Na Tabela 5.6 podemos observar as frequências observadas e as proporções em relação às colunas.

**Tabela 5.6.** Distribuição conjunta das frequências observadas e porcentagens (segundo as colunas) observadas de alunos 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%)

Se não houvesse associação, esperaríamos que 60% de homens e 60% de mulheres tivessem optado pelo curso de Física e que 40% de homens e 40% de mulheres tivessem optado pelo curso de Ciências Sociais. Ou seja, deveríamos ter as frequências apresentadas na Tabela 5.7. Estas frequências são denominadas "frequências esperadas".

**Tabela 5.7.** Distribuição conjunta das frequências esperadas e proporções esperadas (em porcentagem) de alunos segundo o sexo ( $X$ ) e o curso escolhido ( $Y$ ).

$Y \backslash X$	Masculino	Feminino	Total
Física	84 (60%)	36 (60%)	120 (60%)
Ciências sociais	56 (40%)	24 (40%)	80 (40%)
Total	140 (100%)	60 (100%)	200 (100%)

As frequências esperadas podem ser calculadas fazendo-se:

$$e_{ij} = \frac{(\text{Total da linha } i) \times (\text{Total da coluna } j)}{\text{Total de observações}}$$

Assim, temos:

- número esperado de homens a optar pelo curso de Física:

$$e_{11} = \frac{120 \times 140}{200} = 84$$

- número esperado de mulheres a optar pelo curso de Física:

$$e_{12} = \frac{120 \times 60}{200} = 36$$

- número esperado de homens a optar pelo curso de Ciências Sociais:

$$e_{21} = \frac{80 \times 140}{200} = 56$$

- número esperado de mulheres a optar pelo curso de Ciências Sociais:

$$e_{22} = \frac{80 \times 60}{200} = 24$$

Na Tabela 5.8 podemos observar as frequências observadas ( $o_{ij}$ ) em contraste com as frequências esperadas ( $e_{ij}$ ).

**Tabela 5.8.** Valores observados ( $o_{ij}$ ) e valores esperados ( $e_{ij}$ ) de alunos segundo o sexo ( $X$ ) e o curso escolhido ( $Y$ ).

$Y \backslash X$	Masculino	Feminino	Total
Física	100 (84)	20 (36)	120
Ciências sociais	40 (56)	40 (24)	80
Total	140	60	200

Consideremos, agora, a medida

$$\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

onde  $o_{ij}$  é a frequência observada e  $e_{ij}$  é a frequência esperada. Essa medida mede o grau de afastamento entre a frequência observada e a esperada. Uma medida de afastamento global pode ser dada pela soma de todas essas medidas. Essa medida global é denominada  $\chi^2$  (qui-quadrado). Assim, definimos o qui-quadrado de Pearson como

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

onde cada  $e_{ij}$  é determinado por

$$e_{ij} = \frac{(\text{Total da linha } i) \times (\text{Total da coluna } j)}{(\text{Total de observações})},$$

e  $r$  e  $s$  são o número de linhas e colunas, respectivamente.

No nosso exemplo (utilizando a Tabela 5.8) teríamos:

$$\chi^2 = \frac{(100 - 84)^2}{84} + \frac{(20 - 36)^2}{36} + \frac{(40 - 56)^2}{56} + \frac{(40 - 24)^2}{24} = 25,4.$$

### **Coefficiente de Contingência**

Pearson definiu, ainda, uma medida de associação baseada no qui-quadrado, chamada de coeficiente de contingência (de Pearson), dado por:

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

Quanto mais próximo de zero (0)  $C$  estiver, menor será a associação entre as variáveis  $X$  e  $Y$  e quanto mais próximo de um (1)  $C$  estiver, maior será a associação entre as duas variáveis. Contudo, o coeficiente acima nunca atinge o valor 1. O valor máximo de  $C$  depende de  $r$  e  $s$  (número de linhas e colunas, respectivamente). Para evitar esse inconveniente, costuma-se definir um outro coeficiente de contingência (de Tschuprow), dado por:

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

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

## Exemplo

Considerando o exemplo anterior, cujas frequências observadas e esperadas são apresentadas na Tabela 5.8, calcularemos os coeficientes de contingência.

Como já foi visto no exemplo anterior, o qui-quadrado foi  $\chi^2 = 25,4$ . Portanto:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{25,4}{25,4 + 200}} = 0,336$$

e

$$T = \sqrt{\frac{\chi^2/n}{\sqrt{(r-1)(s-1)}}} = \sqrt{\frac{25,4/200}{\sqrt{(2-1)(2-1)}}} = 0,3563$$

indicando uma associação "moderada" entre as variáveis sexo ( $X$ ) e curso escolhido ( $Y$ ).

### Comandos no Software R para obter o qui-quadrado e o coeficiente de contingência:

```
#Entrando com a matriz da distribuição conjunta no R:
matriz <- matrix(c(100, 20,
                  40, 40), 2, 2, byrow=T)

#Adicionando os nomes das linhas e colunas:
rownames(matriz) <- c("Física", "Ciências Sociais")
colnames(matriz) <- c("Masculino", "Feminino")

#Mostrando a matriz:
matriz

#Qui-quadrado usando o comando "chisq.test":
chisq.test(tabela,correct=F)

#Carregando o pacote vcd (precisa instalar):
library(vcd)

#Qui-quadrado e coeficiente de contingência:
assocstats(matriz)
```

### Saída do Software R para o qui-quadrado e o coeficiente de contingência:

```
> #Qui-quadrado usando o comando "chisq.test":
> chisq.test(tabela,correct=F)

        Pearson's Chi-squared test

data:  tabela
X-squared = 25.397, df = 1, p-value = 4.667e-07

> #Qui-quadrado e coeficiente de contingência:
> assocstats(matriz)
              X^2 df    P(> X^2)
Likelihood Ratio 25.307  1 4.8881e-07
Pearson          25.397  1 4.6669e-07

Phi-Coefficient   : 0.356
Contingency Coeff.: 0.336
Cramer's V        : 0.356
```

**Observação:** Estamos interessados nos coeficientes que foram destacados ( $\chi^2$  e  $C$ ).

## 5.3 Variáveis Quantitativas: Diagrama de Dispersão e Coeficiente de Correlação

Um dispositivo bastante útil para se verificar a associação entre duas variáveis quantitativas é o gráfico de dispersão, que vamos introduzir por meio de exemplos.

### Exemplo

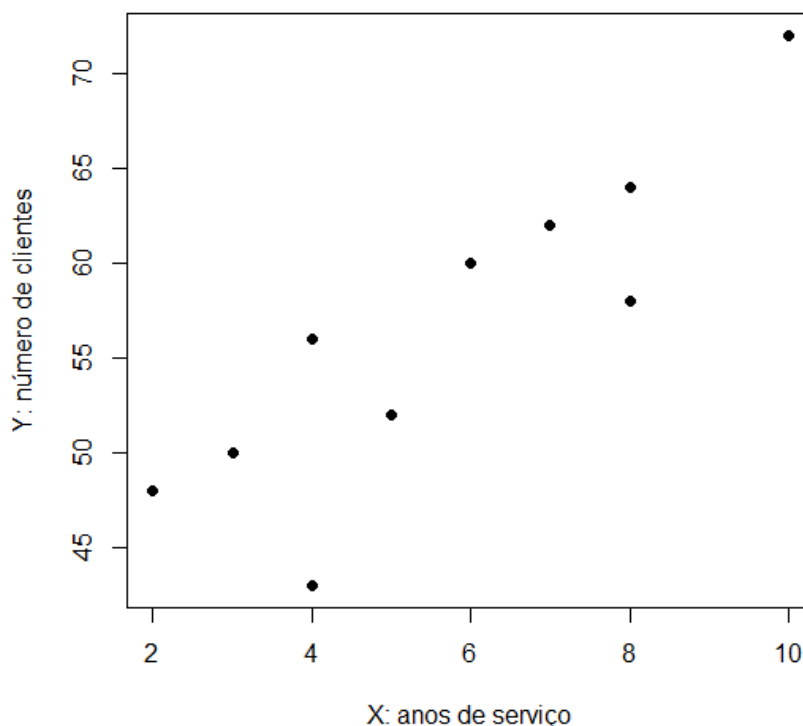
A Tabela 5.9 apresenta o número de anos de serviço ( $X$ ) e o número de clientes ( $Y$ ) de agentes de uma companhia de seguros.

**Tabela 5.9.** Número de anos de serviço ( $X$ ) e o número de clientes ( $Y$ ) de agentes de uma companhia de seguros

Agente	Anos de serviço ( $X$ )	Número de clientes ( $Y$ )
A	2	48
B	3	50
C	4	56
D	5	52
E	4	43
F	6	60
G	7	62
H	8	58
I	8	64
J	10	72

Na Figura 5.2 temos o gráfico de dispersão das variáveis  $X$  e  $Y$  da Tabela 5.9. Nesse tipo de gráfico temos os possíveis valores  $(x, y)$ , na ordem que aparecem.

**Figura 5.2.** Gráfico de dispersão das variáveis  $X$ : anos de serviço e  $Y$ : número de clientes.



Para este exemplo, vemos que parece haver uma associação entre as variáveis porque à medida que aumenta o tempo de serviço, o número de clientes também aumenta.

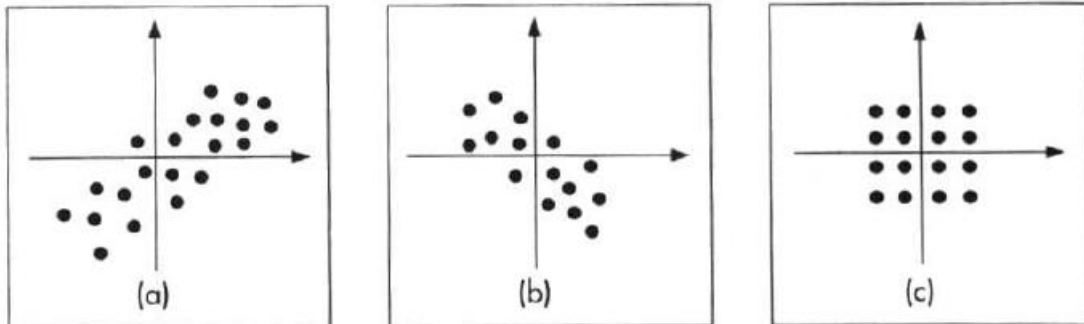
#### Comandos no Software R para fazer o gráfico de dispersão de $X$ e $Y$ :

```
#Entrando com os dados no R:  
X <- c(2, 3, 4, 5, 4, 6, 7, 8, 8, 10)  
Y <- c(48, 50, 56, 52, 43, 60, 62, 58, 64, 72)  
  
#Plotando o gráfico de dispersão:  
plot(X, Y, pch=19, xlab="X: anos de serviço",  
      ylab="Y: número de clientes", main="")
```

Na Figura 5.3 podemos observar possíveis tipos de associação (correlação) entre as variáveis. Na Figura 5.3 (a) temos uma associação linear positiva, ou seja, ao passo que uma variável aumenta a outra também aumentará. Na Figura 5.3 (b) temos uma associação linear negativa, ou seja, ao passo que uma variável aumenta a outra variável diminuirá. Finalmente, na Figura 5.3 (c) as variáveis não tem nenhuma

associação linear, ou seja, quando uma variável aumenta a outra variável não aumenta e nem diminui linearmente.

**Figura 5.3.** Tipos de associação linear entre duas variáveis



Quanto mais próximos de uma reta os pontos do diagrama de dispersão estiverem mais forte será a correlação entre as duas variáveis e quanto mais dispersos os pontos estiverem mais fraca será a correlação entre as duas variáveis. Pode-se determinar o grau de correlação entre duas variáveis utilizando-se, para isso, uma medida.

### Coeficiente de correlação

Iremos definir agora o coeficiente de correlação (linear) entre duas variáveis, que é uma medida do grau de associação entre elas.

**Definição:** Dados  $n$  pares de valores  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , chamamos de coeficiente de correlação entre as duas variáveis  $X$  e  $Y$ :

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

em que  $dp(X)$  é o desvio padrão de  $X$  e  $dp(Y)$  é o desvio padrão de  $Y$ .

O coeficiente de correlação pode variar de  $-1$  a  $1$ , ou seja:

$$-1 \leq r \leq 1.$$

Um coeficiente de correlação positivo indica uma associação linear positiva, um coeficiente de correlação negativo indica uma associação linear negativa, já um coeficiente de correlação nulo (igual a zero) indica que não existe associação linear entre as duas variáveis. Quanto mais próximo de 1 estiver  $r$ , mais forte é o grau de associação linear positiva entre  $X$  e  $Y$ , e, quanto mais próximo de -1 estiver  $r$ , mais forte é o grau de associação linear negativa entre  $X$  e  $Y$ .

A fórmula anterior pode ser operacionalizada de modo mais conveniente pela seguinte fórmula:

$$r = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sqrt{(\sum x_i^2 - n \bar{x}^2)(\sum y_i^2 - n \bar{y}^2)}}$$

## Exemplo

Considerando os dados apresentados na Tabela 5.9, iremos calcular o coeficiente de correlação. Para isso, construímos a tabela de cálculos auxiliares (Tabela 5.10):

**Tabela 5.10.** Tabela de cálculos auxiliares

Agente	$X$	$Y$	$X^2$	$Y^2$	$X \cdot Y$
A	2	48	4	2.304	96
B	3	50	9	2.500	150
C	4	56	16	3.136	224
D	5	52	25	2.704	260
E	4	43	16	1.849	172
F	6	60	36	3.600	360
G	7	62	49	3.844	434
H	8	58	64	3.364	464
I	8	64	64	4.096	512
J	10	72	100	5.184	720
Total	57	565	383	32.581	3.392

Assim, temos:

$$n = 10, \quad \bar{x} = 5,7, \quad \bar{y} = 56,5, \quad \sum x_i^2 = 383, \quad \sum y_i^2 = 32.581, \quad \sum x_i y_i = 3.392.$$



Logo:

$$\begin{aligned}
 r &= \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sqrt{(\sum x_i^2 - n \bar{x}^2)(\sum y_i^2 - n \bar{y}^2)}} \\
 &= \frac{3.392 - 10 \times 5,7 \times 56,5}{\sqrt{(383 - 10 \times (5,7)^2)(32.581 - 10 \times (56,5)^2)}} \\
 &= 0,8768.
 \end{aligned}$$

Como o coeficiente de correlação foi próximo de 1, significa que existe uma forte correlação linear positiva entre o número de anos de serviço ( $X$ ) e o número de clientes ( $Y$ ) dos agentes da companhia de seguros.

#### Comandos no Software R para calcular o coeficiente de correlação entre X e Y:

```
#Entrando com os dados no R:
X <- c(2, 3, 4, 5, 4, 6, 7, 8, 8, 10)
Y <- c(48, 50, 56, 52, 43, 60, 62, 58, 64, 72)

#Correlação entre X e Y:
cor(X,Y)
```

### Exemplo

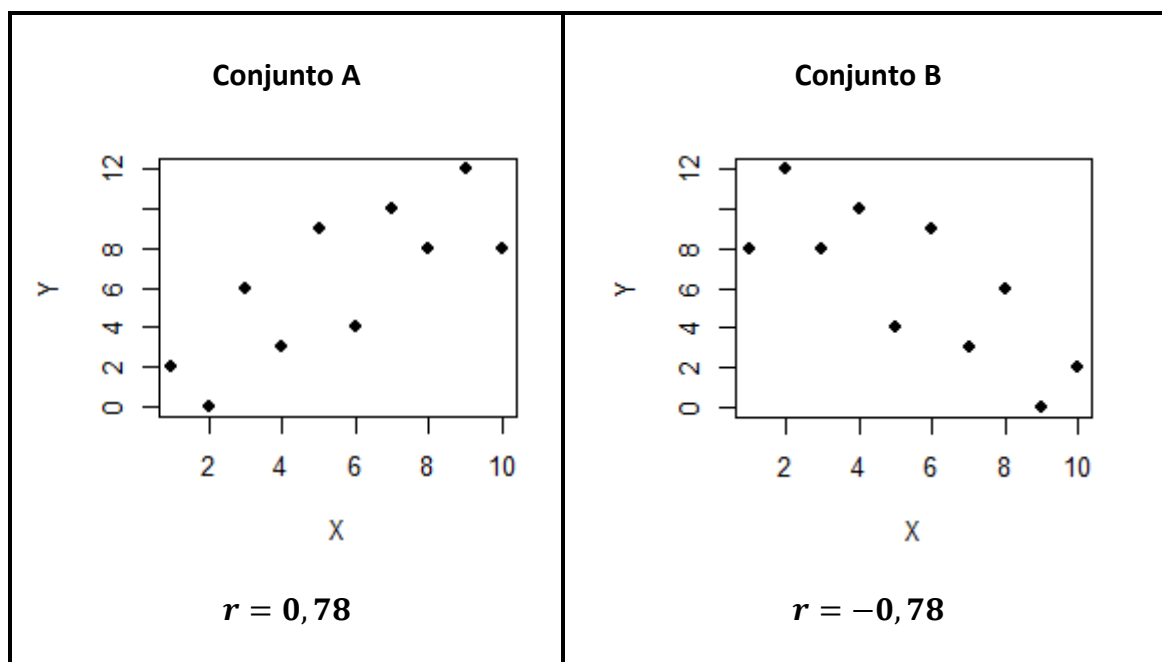
A seguir são apresentados dois conjuntos de dados (A e B), cada um com duas variáveis ( $X$  e  $Y$ ), apresentando correlações positiva e negativa, respectivamente.

**Tabela 5.11.** Conjuntos de dados (A e B) com variáveis ( $X$  e  $Y$ ) correlacionadas positivamente e negativamente

Conjunto A		Conjunto B	
$X$	$Y$	$X$	$Y$
1	2	1	8
2	0	2	12
3	6	3	8
4	3	4	10
5	9	5	4
6	4	6	9
7	10	7	3
8	8	8	6
9	12	9	0
10	8	10	2

Na Figura 5.4, a seguir, são mostrados os gráficos de dispersão de  $X$  e  $Y$  e as correlações para os conjuntos A e B.

**Figura 5.4.** Gráficos de dispersão de  $X$  e  $Y$  e correlações entre  $X$  e  $Y$  para os conjuntos A e B



## II - Probabilidade

### 1 - Probabilidade

#### 1.1 Espaço amostral, eventos

##### Experimento Aleatório

Um experimento aleatório é um experimento que:

- a) pode ser repetido indefinidamente sob as mesmas condições;
- b) não se conhece um particular valor do experimento "a priori", porém pode-se descrever todos os possíveis resultados - as possibilidades;
- c) Quando o experimento for repetido um grande número de vezes surgirá uma regularidade, isto é, haverá uma estabilidade da fração  $f = \frac{r}{n}$  (frequência relativa), em que  $n$  é o número de repetições e  $r$  o número de sucessos de um particular resultado estabelecido antes da realização.

##### Exemplos

- $E_1$ : Retirar uma carta de um baralho com 52 cartas e observar seu "naipe".
- $E_2$ : Jogar uma moeda 10 vezes e observar o número de caras obtidas.
- $E_3$ : Jogar um dado e observar o número mostrado na face de cima.

##### Espaço amostral

**Definição:** Para cada experimento aleatório  $E$ , define-se Espaço Amostral  $\Omega$  o conjunto de todos os possíveis resultados desse experimento.

---

## Exemplos

- a) Considere o experimento  $E$ : "jogar um dado e observar o nº da face de cima", então

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

- b) Seja  $E$ : "jogar duas moedas e observar o resultado", então

$$\Omega = \{(c, c), (c, k), (k, c), (k, k)\}$$

em que  $c$  = cara e  $k$  = coroa.

Observe que sendo  $\Omega$  um conjunto, poderá ser finito ou infinito. Também pode ser discreto ou contínuo (intervalos). A princípio consideraremos apenas conjuntos finitos.

## Evento

**Definição:** Evento é um conjunto de resultados do experimento, em termos de conjuntos, é um subconjunto de  $\Omega$ . Em particular,  $\Omega$  e  $\emptyset$  (conjunto vazio) são eventos,  $\Omega$  é dito o evento certo e  $\emptyset$  o evento impossível.

## Exemplos

- a) Seja o experimento  $E$ : "jogar 3 moedas e observar os resultados". Então

$$\Omega = \{(c, c, c), (c, c, k), (c, k, c), (k, c, c), (c, k, k), (k, c, k), (k, k, c), (k, k, k)\}$$

Seja o evento  $A$ : "ocorrer pelo menos 2 caras".

$$\text{Então, } A = \{(c, c, c), (c, c, k), (c, k, c), (k, c, c)\}$$

### Comandos no Software R para obter o espaço amostral (lançamento de 3 moedas):

```
#Carregando o pacote prob:
library(prob)

#Todas as combinações dos lançamentos de três moedas:
tosscoin(3)

#Outra forma:
urnsamples(1:2, x = c('C', 'K'), size=3, replace=TRUE, ordered=TRUE)
```

b) Seja o experimento  $E$ : "lançar um dado e observar o número de cima". Então

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Seja o evento  $B$ : "ocorrer múltiplo de 2". Então

$$B = \{2, 4, 6\}$$

Usando as operações com conjuntos, podem-se formar novos eventos. Assim:

- (i)  $A \cup B \rightarrow$  é o evento que ocorre se  $A$  ou  $B$  ocorrem;
- (ii)  $A \cap B \rightarrow$  é o evento que ocorre se  $A$  e  $B$  ocorrem;
- (iii)  $\bar{A} \rightarrow$  é o evento que ocorre se  $A$  não ocorre ( $\bar{A}$  é chamado complementar de  $A$ ).

### Exemplo

Seja o experimento  $E$ : "lançar um dado e observar o número de cima". Então:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Seja o evento  $A$ : "ocorrer número ímpar". Então:

$$A = \{1, 3, 5\}$$

Seja o evento  $B$ : "ocorrer um número menor ou igual a 3". Então:

$$B = \{1, 2, 3\}$$

Assim:

- $A \cup B = \{1, 2, 3, 5\}$
- $A \cap B = \{1, 3\}$
- $\bar{A} = \{2, 4, 6\}$
- $\bar{B} = \{4, 5, 6\}$

### Eventos mutuamente exclusivos

Dois eventos  $A$  e  $B$  são denominados mutuamente exclusivos, se eles não puderem ocorrer simultaneamente, isto é,  $A \cap B = \emptyset$ .

## Exemplo

$E$ : "jogar um dado e observar o resultado"

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Sejam os eventos:  $A$ : "ocorrer nº par" e  $B$ : "ocorrer nº ímpar". Então:

$$A = \{2, 4, 6\} \text{ e } B = \{1, 3, 5\}, \text{ assim, } A \cap B = \emptyset.$$

$A$  e  $B$  são mutuamente exclusivos, pois a ocorrência de um número par impede a ocorrência de um número ímpar e vice-versa.

## Definição de Probabilidade

Dado um experimento aleatório  $E$  e  $\Omega$  o espaço amostral, probabilidade de um evento  $A$ , denotada por  $P(A)$ , é uma função definida em  $\Omega$  que associa a cada evento um número real, satisfazendo os seguintes axiomas:

- (i)  $0 \leq P(A) \leq 1$ ;
- (ii)  $P(\Omega) = 1$ ;
- (iii) Se  $A$  e  $B$  forem eventos mutuamente exclusivos, ( $A \cap B = \emptyset$ ), então  $P(A \cup B) = P(A) + P(B)$ .

## Principais Teoremas

1. Se  $\emptyset$  é o conjunto vazio, então  $P(\emptyset) = 0$ .
2. Se  $\bar{A}$  é o complemento do evento  $A$ , então  $P(\bar{A}) = 1 - P(A)$ .
3. Se  $A \subset B$ , então  $P(A) \leq P(B)$ .
4. Teorema da soma: Se  $A$  e  $B$  são dois eventos quaisquer, então:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

## Espaços Amostrais Finitos Equiprováveis

Quando se associa a cada elemento do espaço amostral a mesma probabilidade, o espaço amostral chama-se equiprovável ou uniforme. Em particular se  $\Omega$  contém  $n$  elementos, então, a probabilidade de cada ponto será  $1/n$ .

Por outro lado, se um evento  $A$  contém  $r$  pontos, então:

$$P(A) = r \cdot \left(\frac{1}{n}\right) = \frac{r}{n}.$$

Este método de avaliar  $P(A)$  é frequentemente enunciado da seguinte maneira:

$$P(A) = \frac{\text{nº de elementos do evento } A}{\text{nº de elementos do espaço amostral } \Omega}$$

ou, ainda,

$$P(A) = \frac{\text{nº de casos favoráveis}}{\text{nº total de casos}}.$$

### Exemplo

Escolha aleatoriamente uma carta de um baralho com 52 cartas.

Sejam:

$A$ : "a carta é de ouros"

$B$ : "a carta é uma figura"

Calcular  $P(A)$  e  $P(B)$ .

$$P(A) = \frac{\text{nº de cartas de ouros}}{\text{nº total de cartas}} = \frac{13}{52} = 0,25.$$

$$P(B) = \frac{\text{nº de figuras}}{\text{nº total de cartas}} = \frac{12}{52} = 0,2308.$$

Como se observa, o cálculo da probabilidade de um evento se reduz a um problema de contagem. Assim, a Análise Combinatória (Teoria da Contagem) tem fundamental importância para se contar o número de casos favoráveis e o total de casos. Na maioria dos problemas tratados neste curso, a combinação é a técnica que pode ser aplicada.

Combinação de  $n$  elementos tomados (combinados)  $r$  a  $r$  ( $r \leq n$ ). Calcula-se por:

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

onde:

- $n! = n \cdot (n - 1) \cdot (n - 2) \dots 1$
- $r! = r \cdot (r - 1) \cdot (r - 2) \dots 1$
- admite-se que  $0! = 1$

## Exemplo

Quantas comissões de duas pessoas pode-se formar com um grupo de quatro pessoas?

## Solução

$$\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{4!}{2! 2!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1} = \frac{24}{4} = 6.$$

Podemos conferir esse resultado:

Grupo de 4 pessoas:

A B C D

Comissões de 2 pessoas:

AB AC AD BC BD CD

**Observação:** Nesse caso a ordem não importa, por exemplo,  $AB = BA$ .

## Comandos no Software R para calcular combinações:

```
#Usando fatoriais (comando factorial()):
factorial(4) / (factorial(2) * factorial(4-2))

#Carregando o pacote combinat:
library(combinat)

#Mostrando todas as combinações:
combn(c("A", "B", "C", "D"), 2)

#Número de combinações:
dim(combn(4, 2)) [2]
```



## Exemplo

Quantas comissões de três pessoas pode-se formar com um grupo de dez pessoas?

## Solução

$$\binom{10}{3} = \frac{10!}{3!(10-3)!} = \frac{10!}{3! 7!} = \frac{10 \cdot 9 \cdot 8 \cdot \cancel{7!}}{3! \cdot \cancel{7!}} = \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = 120.$$

### Comandos no Software R para calcular combinações:

```
#Usando fatoriais:
factorial(10) / (factorial(3) * factorial(10-3))

#Carregando o pacote combinat:
library(combinat)

#Número de combinações:
dim(combn(10, 3)) [2]
```

## Exemplo

Num lote de 12 peças, 4 são defeituosas; duas peças são retiradas aleatoriamente. Calcule:

- A probabilidade de ambas serem defeituosas;
- A probabilidade de ambas serem boas;
- A probabilidade de ao menos uma ser defeituosa.

## Solução

- A: "ambas são defeituosas".

$$A \text{ contém } \binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{4 \cdot 3 \cdot \cancel{2!}}{2 \cdot \cancel{2!}} = 6 \text{ elementos.}$$

$$\Omega \text{ contém } \binom{12}{2} = \frac{12!}{2!(12-2)!} = \frac{12 \cdot 11 \cdot \cancel{10!}}{2 \cdot \cancel{10!}} = 66 \text{ elementos.}$$

$$\text{Logo, } P(A) = \frac{\text{nº de casos favoráveis}}{\text{nº total de casos}} = \frac{6}{66} = 0,0909.$$

b)  $B$ : "ambas são boas".

$$B \text{ contém } \binom{8}{2} = \frac{8!}{2!(8-2)!} = \frac{8 \cdot 7 \cdot \cancel{6!}}{2 \cdot \cancel{6!}} = 28 \text{ elementos.}$$

$$\text{Logo, } P(B) = \frac{\text{nº de casos favoráveis}}{\text{nº total de casos}} = \frac{28}{66} = 0,4242.$$

c)  $C$ : "ao menos uma é defeituosa".

Observe que  $C$  (uma é defeituosa ou as duas são defeituosas) é o complemento de  $B$  (as duas são boas), ou seja,  $C = \bar{B}$ .

Possibilidades:

$$\left\{ \underbrace{(\text{duas boas})}_B, \underbrace{(\text{uma defeituosa e uma boa}), (\text{duas defeituosas})}_C \right\}$$

Logo,  $C = \bar{B}$ .

Portanto:

$$P(C) = 1 - P(B) = 1 - 0,4242 = 0,5758.$$

## 1.2 Probabilidade condicional, Teorema de Bayes e independência de eventos

### Probabilidade Condicional

#### Exemplo

Seja  $E$ : "lançar um dado", e o evento  $A$ : "sair o nº 3". Então  $\Omega = \{1, 2, 3, 4, 5, 6\}$ ,  $A = \{3\}$  e, portanto:

$$P(A) = \frac{1}{6}.$$

Considere agora o evento  $B$ : "sair um nº ímpar", ou seja,  $B = \{1, 3, 5\}$ . Pode-se estar interessado em avaliar a probabilidade do evento  $A$  condicionada à ocorrência do evento  $B$ . Em símbolos, designa-se por  $P(A|B)$ ; lê-se "probabilidade do evento  $A$  condicionada à ocorrência de  $B$ ", ou ainda, "probabilidade de  $A$  dado  $B$ ".

Dada a informação da ocorrência do evento  $B$  (ocorreu um nº ímpar), temos a redução do espaço amostral  $\Omega = \{1, 2, 3, 4, 5, 6\}$  para  $\Omega^* = \{1, 3, 5\}$  (pois já sabemos que o nº que ocorreu é ímpar), e é nesse espaço amostral reduzido que se avalia a probabilidade do evento.

- $A = \{3\} \Rightarrow$  nº de elementos de  $A = 1$ ;
- $\Omega^* = \{1, 3, 5\} \Rightarrow$  nº de elementos de  $\Omega^* = 3$ ;

Portanto:

$$P(A|B) = \frac{1}{3} = 0,3333.$$

Podemos, também, calcular a probabilidade condicional de  $A$  dado  $B$  sem ter que considerar o espaço amostral reduzido.

**Definição:** Dados dois eventos  $A$  e  $B$ , denota-se  $P(A|B)$  a probabilidade condicional do evento  $A$ , quando  $B$  tiver ocorrido, por:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

com  $P(B) \neq 0$ , pois  $B$  já ocorreu.

## Exemplo

Seja  $E$ : "lançar um dado", então  $\Omega = \{1, 2, 3, 4, 5, 6\}$ . Sejam:

- $A$ : "sair o nº 3" =  $\{3\} \Rightarrow P(A) = \frac{1}{6}$
- $B$ : "sair um nº ímpar" =  $\{1, 3, 5\} \Rightarrow P(B) = \frac{3}{6}$ .

Então,  $A \cap B = \{3\} \Rightarrow P(A \cap B) = \frac{1}{6}$ .

Assim, podemos calcular  $P(A|B)$  como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{3/6} = \frac{1}{3} = 0,3333.$$

Podemos, também, calcular a probabilidade de  $B$  dado  $A$  como:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

### Teorema do Produto

A partir da definição de probabilidade condicional pode-se enunciar o teorema do produto:

"A probabilidade da ocorrência simultânea de dois eventos,  $A$  e  $B$ , do mesmo espaço amostral, é igual ao produto da probabilidade de um deles pela probabilidade condicional do outro, dado o primeiro."

Assim:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(B) \cdot P(A|B)$$

ou

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(A \cap B) = P(A) \cdot P(B|A)$$

Usamos o Teorema do Produto quando já conhecemos a probabilidade condicional (quando não precisamos calculá-la usando fórmula) e queremos encontrar a probabilidade da interseção de  $A$  e  $B$ , como será visto no exemplo a seguir.

### Exemplo

Em um lote de 12 peças, 4 são defeituosas. Duas são retiradas uma após a outra sem reposição. Qual a probabilidade de que ambas sejam boas?

$D_1$	$D_2$	$D_3$	$D_4$	$B_1$	$B_2$	$B_3$	$B_4$	$B_5$	$B_6$	$B_7$	$B_8$
-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------

### Solução

- $A$ : "a primeira peça é boa"
- $B$ : "a segunda peça é boa"

$$P(A \cap B) = P(A) \cdot P(B|A) = \frac{8}{12} \cdot \frac{7}{11} = \frac{56}{132} = 0,4242.$$

## Regra Geral da Multiplicação

Seja  $A_1, A_2, \dots, A_n$  uma sequência de eventos de um espaço amostral  $\Omega$ . Então

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdots P(A_n|A_1 \cap A_2 \cap \dots \cap A_{n-1})$$

## Exemplo

Suponhamos uma extração de 3 cartas sem reposição, de um baralho, onde estamos interessados no evento “nenhuma copas”.

Seja  $C = \{\text{carta de copas}\}$ . Logo  $P(C) = 13/52$  e seu complementar será  $P(\bar{C}) = 39/52$ . Então:

$$P(\bar{C}_1 \cap \bar{C}_2 \cap \bar{C}_3) = P(\bar{C}_1) \cdot P(\bar{C}_2|\bar{C}_1) \cdot P(\bar{C}_3|\bar{C}_1 \cap \bar{C}_2) = \frac{39}{52} \cdot \frac{38}{51} \cdot \frac{37}{50} = 0,4135$$

## Independência Estatística

Um evento  $A$  é considerado independente de um outro evento  $B$  se a probabilidade de  $A$  é igual à probabilidade condicional de  $A$  dado  $B$ , isto é, se

$$P(A) = P(A|B)$$

ou seja,  $A$  é independente de  $B$  se o fato de  $B$  ter ocorrido não afeta em nada a probabilidade da ocorrência de  $A$ .

É evidente que, se  $A$  é independente de  $B$ ,  $B$  é independente de  $A$ ; assim:

$$P(B) = P(B|A).$$

Considerando o teorema do produto, pode-se afirmar que: se  $A$  e  $B$  são independentes, então:

$$P(A \cap B) = P(A) \cdot P(B).$$

## Exemplo

Em uma caixa de 10 peças, 4 são defeituosas. São retiradas duas peças, uma após a outra, com reposição. Calcular a probabilidade de ambas serem boas.

$D_1$	$D_2$	$D_3$	$D_4$	$B_1$	$B_2$	$B_3$	$B_4$	$B_5$	$B_6$
-------	-------	-------	-------	-------	-------	-------	-------	-------	-------

### Solução

Sejam os eventos:

$A$ : "a primeira peça é boa",

$B$ : "a segunda peça é boa".

Note que  $A$  e  $B$  são independentes, pois,  $P(B) = P(B|A)$ . Ou seja, o fato de a 1ª peça ter sido boa não alterou a probabilidade de a 2ª peça ser boa (pois a 1ª peça foi colocada de volta na caixa). Logo:

$$P(A \cap B) = P(A) \cdot P(B) = \frac{6}{10} \cdot \frac{6}{10} = \frac{36}{100} = 0,36.$$

### Regra Geral da Multiplicação (Eventos Independentes)

Seja  $A_1, A_2, \dots, A_n$  uma sequência de eventos independentes de um espaço amostral  $\Omega$ . Então

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2) \cdots P(A_n)$$

### Exemplo

As probabilidades de 3 jogadores marcarem um penalty são respectivamente  $2/3$ ,  $4/5$  e  $7/10$ . Se cada um "cobrar" uma única vez, qual a probabilidade de todos acertarem?

### Solução

Os eventos são  $A_1$ : {o 1º jogador acertar},  $A_2$ : {o 2º jogador acertar} e  $A_3$ : {o 3º jogador acertar}. Observe que os eventos são independentes (a ocorrência de um evento não altera a probabilidade de ocorrência do outro), então:

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2) \cdot P(A_3) = \frac{2}{3} \cdot \frac{4}{5} \cdot \frac{7}{10} = 0,3733$$

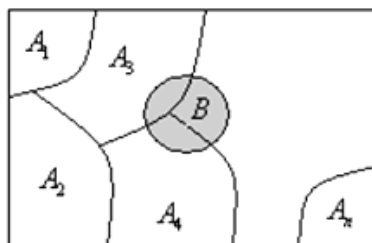
## Teorema da Probabilidade Total

Sejam  $A_1, A_2, \dots, A_n$ ,  $n$  eventos mutuamente exclusivos tais que

$$A_1 \cup A_2 \cup \dots \cup A_n = \Omega.$$

Sejam  $P(A_i)$  as probabilidades conhecidas dos vários eventos, e  $B$  um evento qualquer de  $\Omega$  tal que são conhecidas todas as probabilidades condicionais  $P(B|A_i)$ . Então

$$P(B) = P(A_1).P(B|A_1) + P(A_2).P(B|A_2) + \dots + P(A_n).P(B|A_n).$$



## Exemplo

Considere um lote contendo 100 peças, das quais 20 são defeituosas e 80 são não-defeituosas. São extraídas duas peças, sem reposição. Definindo-se:

$A$ : {a primeira peça extraída é defeituosa},

$B$ : {a segunda peça extraída é defeituosa},

determine  $P(B)$ .

## Solução

Na primeira extração pode ocorrer apenas um dos seguintes eventos:

$A_1$ : {a primeira peça extraída é defeituosa}, ou

$A_2$ : {a primeira peça extraída é não-defeituosa}.

então, a probabilidade de a segunda peça extraída ser defeituosa é calculada por:

$$\begin{aligned} P(B) &= P(A_1).P(B|A_1) + P(A_2).P(B|A_2) \\ &= \frac{20}{100} \cdot \frac{19}{99} + \frac{80}{100} \cdot \frac{20}{99} \\ &= 0,2 \end{aligned}$$

## Teorema de Bayes

Sejam  $A_1, A_2, \dots, A_n$ ,  $n$  eventos mutuamente exclusivos tais que

$$A_1 \cup A_2 \cup \dots \cup A_n = \Omega.$$

Sejam  $P(A_i)$  as probabilidades conhecidas dos vários eventos, e  $B$  um evento qualquer de  $\Omega$  tal que são conhecidas todas as probabilidades condicionais  $P(B|A_i)$ . Então, para cada índice " $i$ ", tem-se:

$$P(A_i|B) = \frac{P(A_i) \cdot P(B|A_i)}{P(A_1) \cdot P(B|A_1) + P(A_2) \cdot P(B|A_2) + \dots + P(A_n) \cdot P(B|A_n)}$$

## Exemplo

Considere três urnas, cada uma contendo bolas pretas, brancas e vermelhas.

Admita a seguinte configuração:

Urnas Cores	$u_1$	$u_2$	$u_3$
Pretas	3	4	2
Brancas	1	3	3
Vermelhas	5	2	3

Escolheu-se uma urna ao acaso e dela extraiu-se uma bola ao acaso, verificando-se que a bola é branca. Qual a probabilidade da bola ter vindo da urna 2?

## Solução

- Como temos três urnas então a probabilidade de se escolher uma urna ao acaso é de  $1/3$ .
- Dado que a urna escolhida foi  $u_1$ , então a probabilidade de se escolher uma bola branca é de  $1/9$ .



Urnas Cores	$P(u_1) = 1/3$	$u_2$	$u_3$
Pretas	$P(p u_1) = 3/9$	4	2
Branças	<b><math>P(b u_1) = 1/9</math></b>	3	3
Vermelhas	$P(v u_1) = 5/9$	2	3

- Dado que a urna escolhida foi  $u_2$ , então a probabilidade de se escolher uma bola branca é de  $3/9$ .

Urnas Cores	$u_1$	$P(u_2) = 1/3$	$u_3$
Pretas	3	$P(p u_2) = 4/9$	2
Branças	1	<b><math>P(b u_2) = 3/9</math></b>	3
Vermelhas	5	$P(v u_2) = 2/9$	3

- Dado que a urna escolhida foi  $u_3$ , então a probabilidade de se escolher uma bola branca é de  $3/8$ .

Urnas Cores	$u_1$	$u_2$	$P(u_3) = 1/3$
Pretas	3	4	$P(p u_3) = 2/8$
Branças	1	3	<b><math>P(b u_3) = 3/8</math></b>
Vermelhas	5	2	$P(v u_3) = 3/8$

Assim:

$$P(u_1) = \frac{1}{3}, \quad P(u_2) = \frac{1}{3}, \quad P(u_3) = \frac{1}{3}.$$

$$P(b|u_1) = \frac{1}{9}, \quad P(b|u_2) = \frac{3}{9}, \quad P(b|u_3) = \frac{3}{8}$$

deseja-se calcular  $P(u_2|b)$ .

Aplicando-se o teorema de Bayes, tem-se:

$$P(A_i|B) = \frac{P(A_i).P(B|A_i)}{P(A_1).P(B|A_1) + P(A_2).P(B|A_2) + \dots + P(A_n).P(B|A_n)}$$

$$P(u_2|br) = \frac{P(u_2).P(br|u_2)}{P(u_1).P(br|u_1) + P(u_2).P(br|u_2) + P(u_3).P(br|u_3)}$$

$$= \frac{\frac{1}{3} \cdot \frac{1}{3}}{\frac{1}{3} \cdot \frac{1}{9} + \frac{1}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{3}{8}} = 0,4068.$$

Observe que a probabilidade a priori de  $u_2$  era de  $1/3$ . Dada a informação que saiu uma bola branca, a probabilidade a posteriori de  $u_2$  será 0,4068.

## Exemplo

Uma doença ataca 3% de uma população. Um determinado teste rápido de sangue consegue identificar corretamente 98% das pessoas que possuem a doença. Contudo, o teste identifica como positivo para a doença 8% das pessoas que na realidade não possuem a doença (falsos positivos). Qual é a probabilidade de um indivíduo que foi classificado como positivo no teste ter efetivamente a doença?

## Solução

Sejam:

- $I$ : Infectado (o indivíduo possui a doença);
- $S$ : Saudável (o indivíduo não possui a doença);
- $CP$ : Classificado como positivo no teste;
- $CN$ : Classificado como negativo no teste.

Desejamos encontrar a probabilidade de um indivíduo que, foi classificado como positivo no teste, realmente ter a doença, ou seja, dado que o indivíduo foi classificado como positivo, qual a probabilidade dele realmente ter a doença? Podemos escrever isso como:

$$P(I|CP) = ?$$

Pelo Teorema de Bayes temos que:

$$P(A_i|B) = \frac{P(A_i).P(B|A_i)}{P(A_1).P(B|A_1) + P(A_2).P(B|A_2)}$$

$$P(I|CP) = \frac{P(I).P(CP|I)}{P(I).P(CP|I) + P(S).P(CP|S)}$$

O enunciado nos dá as seguintes informações:

- $P(I) = 0,03$  (a doença ataca 3% da população)
- $P(S) = 0,97$  (logo, a doença não ataca 97% da população)
- $P(CP|I) = 0,98$  (o teste identifica como infectados 98% das pessoas doentes)
- $P(CP|S) = 0,08$  (o teste identifica como infectados 8% das pessoas saudáveis)

Portanto:

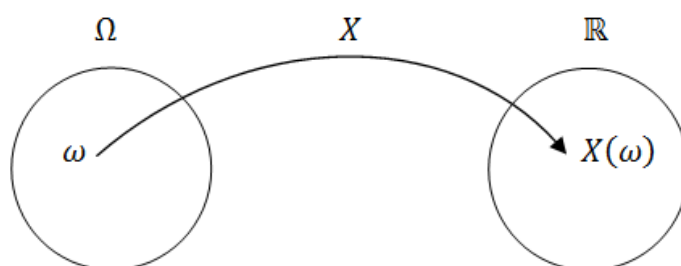
$$\begin{aligned} P(I|CP) &= \frac{P(I).P(CP|I)}{P(I).P(CP|I) + P(S).P(CP|S)} \\ &= \frac{0,03 \cdot 0,98}{0,03 \cdot 0,98 + 0,97 \cdot 0,08} \\ &= 0,2747. \end{aligned}$$

---

## 2. Variáveis aleatórias discretas

### 2.1 Conceito. Valor esperado e variância de uma variável aleatória

**Definição:** Sejam  $E$  um experimento e  $\Omega$  o espaço amostral associado ao experimento. Uma função  $X$ , que associe a cada elemento  $\omega \in \Omega$  um número real  $X(\omega)$  é denominada variável aleatória. Veja a ilustração.



#### Exemplo

Sejam:  $E$ : "lançamento de duas moedas";

$X$ : "nº de caras obtidas nas duas moedas".

Então:  $\Omega = \{(c, c), (c, k), (k, c), (k, k)\}$ , em que:  $c$  = cara e  $k$  = coroa.

- $X = 0 \rightarrow$  corresponde ao elemento:  $(k, k)$
- $X = 1 \rightarrow$  corresponde aos elementos:  $(c, k)$  e  $(k, c)$
- $X = 2 \rightarrow$  corresponde ao elemento:  $(c, c)$

#### Observações

1. "Variável Aleatória" é uma função cujo domínio é  $\Omega$  e o contradomínio é  $\mathbb{R}$ ;
2. Nas aplicações é conveniente trabalhar com números e não com eventos, daí o uso da variável aleatória;
3. Se  $\Omega$  é numérico, então  $X(\omega) = \omega$ ;

4. Uma variável aleatória  $X$  será discreta se o número de valores possíveis de  $X$  (seu contradomínio) for finito ou infinito numerável. Caso seu contradomínio seja um intervalo ou uma coleção de intervalos, ela será uma variável aleatória contínua.

## Função de probabilidades

Seja  $X$  uma variável aleatória discreta. Portanto, o contradomínio de  $X$  será formado no máximo por um número infinito numerável de valores  $x_1, x_2, \dots$ . A cada possível resultado  $x_i$  associaremos um número  $p(x_i) = P(X = x_i)$ , denominado probabilidade de  $x_i$ . Os números  $p(x_i)$ ,  $i = 1, 2, \dots$  devem satisfazer às seguintes condições:

- i.  $p(x_i) \geq 0$  para todo  $i$ ,
- ii.  $\sum_{i=1}^n p(x_i) = 1$ .

A função  $p$ , definida acima, é denominada **Função de Probabilidades** da variável aleatória  $X$ . A probabilidade de um determinado valor da variável aleatória é igual a probabilidade do evento associado ao valor da variável.

## Exemplo

Sejam:  $E$ : "lançamento de duas moedas";  
 $X$ : "nº de caras obtidas nas duas moedas".

Então:

- $P(X = 0) = P[(k, k)] = \frac{1}{4}$ ;
- $P(X = 1) = P[(c, k), (k, c)] = \frac{2}{4} = \frac{1}{2}$ ;
- $P(X = 2) = P[(c, c)] = \frac{1}{4}$ .

A coleção de pares  $[x_i, p(x_i)]$ ,  $i = 1, 2, \dots$ , é algumas vezes denominada **Distribuição de Probabilidades** de  $X$ .

## Exemplo

Sejam:  $E$ : "lançamento de duas moedas";

$X$ : "nº de caras obtidas".

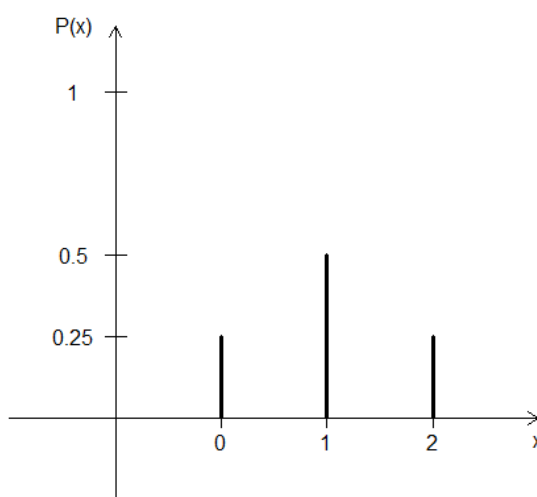
Eis as várias expressões para  $P(x)$ :

### 1. Tabela

**Tabela 2.1.** Distribuição de probabilidades da variável aleatória  $X$ : "nº de caras obtidas"

$x$	$P(x)$
0	$1/4$
1	$1/2$
2	$1/4$

### 2. Gráfico



### 3. Fórmula

$$P(x) = \frac{1}{4} \binom{2}{x}, \quad \text{para } x = 0, 1, 2.$$

Verificando a fórmula:

- $P(X = 0) = \frac{1}{4} \binom{2}{0} = \frac{1}{4} \frac{2!}{0!(2-0)!} = \frac{1}{4} \cdot \frac{2!}{1! \cdot 2!} = \frac{1}{4} \cdot 1 = \frac{1}{4}$
- $P(X = 1) = \frac{1}{4} \binom{2}{1} = \frac{1}{4} \frac{2!}{1!(2-1)!} = \frac{1}{4} \cdot \frac{2!}{1! \cdot 1!} = \frac{1}{4} \cdot 2 = \frac{1}{2}$
- $P(X = 2) = \frac{1}{4} \binom{2}{2} = \frac{1}{4} \frac{2!}{2!(2-2)!} = \frac{1}{4} \cdot \frac{2!}{2! \cdot 0!} = \frac{1}{4} \cdot 1 = \frac{1}{4}$

Note que

$$\sum_{i=1}^3 P(x_i) = P(0) + P(1) + P(2) = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1.$$

## Função de Distribuição Acumulada

Seja  $X$  uma variável aleatória discreta. Define-se função de distribuição acumulada (ou função de repartição) da variável aleatória  $X$ , no ponto  $x$ , como sendo a probabilidade de que  $X$  assumo um valor menor ou igual a  $x$ , isto é:

$$F(x) = P(X \leq x).$$

## Propriedades

1.  $F(x) = \sum_{x_i \leq x} P(x_i)$
2.  $F(-\infty) = 0$
3.  $F(+\infty) = 1$
4.  $P(a < X \leq b) = F(b) - F(a)$
5.  $P(a \leq X \leq b) = F(b) - F(a) + P(X = a)$
6.  $P(a < X < b) = F(b) - F(a) - P(X = b)$
7.  $F(x)$  é contínua à direita  $\rightarrow \lim_{x \rightarrow x_0} F(x) = F(x_0)$
8.  $F(x)$  é descontínua à esquerda, nos pontos em que a probabilidade é diferente de zero.
9. A função é não decrescente, isto é,  $F(b) \geq F(a)$ , para  $b > a$ .

## Exemplo

Sejam:  $E$ : "lançamento de duas moedas";  
 $X$ : "nº de caras obtidas".

Vimos que a distribuição de probabilidades de  $X$  é dada por:

Recordando a Tabela 2.1

$x$	$P(x)$
0	$1/4$
1	$1/2$
2	$1/4$

Então:

$$F(x) = 0 \quad \text{se} \quad x < 0$$

$$(F(x) = 0)$$

$$F(x) = \frac{1}{4} \quad \text{se} \quad 0 \leq x < 1$$

$$(F(x) = \frac{1}{4} = 0,25)$$

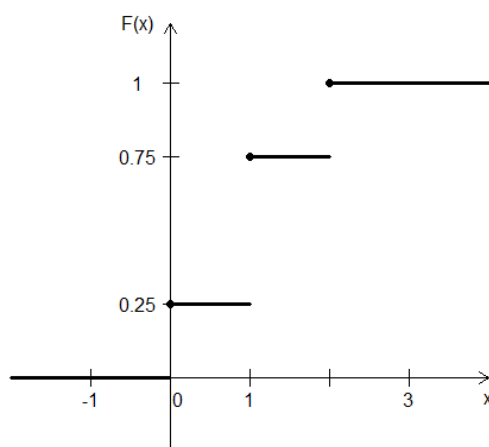
$$F(x) = \frac{3}{4} \quad \text{se} \quad 1 \leq x < 2$$

$$(F(x) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4} = 0,75)$$

$$F(x) = 1 \quad \text{se} \quad x \geq 2$$

$$(F(x) = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1)$$

Eis o gráfico de  $F(x)$ :



## Média ou Esperança Matemática

Define-se Esperança Matemática ou Média de uma variável aleatória discreta como:

$$E(X) = \mu = \sum_{i=1}^n x_i P(x_i).$$



## Exemplo

Seja o experimento  $E$ : "lançamento de duas moedas", e seja a variável aleatória  $X$ : "número de caras". Então:

$$\Omega = \{(c, c), (c, k), (k, c), (k, k)\}.$$

A distribuição de probabilidades de  $X$  é dada por:

**Tabela 2.2.** Distribuição de probabilidades da variável aleatória  $X$ : "nº de caras"

$x$	$P(x)$
0	1/4
1	1/2
2	1/4
Total	1

Então,

$$\begin{aligned}
 E(X) &= \sum_{i=1}^n x_i P(x_i) \\
 &= 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} \\
 &= 1.
 \end{aligned}$$

## Exemplo

Seja  $X$  uma variável aleatória com função de probabilidade dada por:

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \dots, n$$

Onde  $n$  é um número inteiro positivo,  $0 \leq p \leq 1$ , e para cada par fixo  $n$  e  $p$  a função de probabilidade soma 1. O valor esperado desta variável aleatória é dado por:

$$\begin{aligned}
 E(X) &= \sum_{x=0}^n x \binom{n}{x} p^x (1-p)^{n-x} \\
 &= \sum_{x=1}^n x \binom{n}{x} p^x (1-p)^{n-x} \quad (\text{em } x = 0 \text{ o termo é } 0)
 \end{aligned}$$

Utilizando a identidade  $x \binom{n}{x} = n \binom{n-1}{x-1}$ , temos

$$\begin{aligned}
E(X) &= \sum_{x=1}^n n \binom{n-1}{x-1} p^x (1-p)^{n-x} \\
&= \sum_{y=0}^{n-1} n \binom{n-1}{y} p^{y+1} (1-p)^{n-(y+1)} \quad (\text{substituindo } y = x - 1) \\
&= \sum_{y=0}^{n-1} n \binom{n-1}{y} p^y \cdot p \cdot (1-p)^{(n-1)-y} \\
&= np \underbrace{\sum_{y=0}^{n-1} \binom{n-1}{y} p^y (1-p)^{(n-1)-y}}_1 \\
&= np
\end{aligned}$$

uma vez que a última soma deve ser 1, sendo a soma de todos os valores possíveis de uma função de probabilidade.

### Propriedades da Média

1.  $E(k) = k$ , onde  $k$  é uma constante.
2.  $E(kX) = k \cdot E(X)$
3.  $E(X \pm Y) = E(X) \pm E(Y)$
4.  $E(X \pm k) = E(X) \pm k$ , onde  $k$  é uma constante.
5.  $E(X - \mu) = 0$ , onde  $\mu$  é a média de  $X$
6. A média do produto de duas variáveis aleatórias independentes é o produto das médias:

$$E(XY) = E(X) \cdot E(Y).$$

### Variância

Define-se variância de uma variável aleatória como sendo:

$$Var(X) = \sigma^2 = E[(X - \mu)^2].$$

A raiz quadrada positiva de  $V(X)$  é o desvio padrão de  $X$ , ou seja,

$$dp(X) = \sqrt{V(X)}.$$

Podemos reescrever a fórmula da variância como:

$$\begin{aligned}
 V(X) &= E(X - E(X))^2 \\
 &= E(X^2 - 2XE(X) + E^2(X)) \\
 &= E(X^2) - 2E(X)E(X) + E^2(X) \\
 &= E(X^2) - 2E^2(X) + E^2(X) \\
 &= E(X^2) - E^2(X)
 \end{aligned}$$

ou seja,

$$V(X) = E(X^2) - E^2(X)$$

A variância da uma medida do grau de dispersão de uma distribuição ao redor de sua média. O **desvio padrão** é mais fácil de ser interpretado, no sentido de que a unidade de medida no **desvio padrão** é a mesma que para a variável original  $X$ . A unidade de medida na variância é o quadrado da unidade original.

Se  $X$  é uma variável aleatória discreta, então:

$$\sigma^2 = \sum (x_i - \mu)^2 P(x_i).$$

## Exemplo

Seja o experimento  $E$ : "lançamento de duas moedas", e seja a variável aleatória  $X$ : "número de caras". Vimos a distribuição de probabilidades de  $X$  na Tabela 2.2.

Recordando a Tabela 2.2

$x$	$P(x)$
0	1/4
1	1/2
2	1/4
Total	1

Já vimos, também, que  $E(X) = 1$ . Logo:

$$\begin{aligned}
 Var(X) &= \sum (x_i - \mu)^2 P(x_i) \\
 &= (0 - 1)^2 \cdot \frac{1}{4} + (1 - 1)^2 \cdot \frac{1}{2} + (2 - 1)^2 \cdot \frac{1}{4} \\
 &= 0,5.
 \end{aligned}$$

Usando a fórmula alternativa da variância:

$$\begin{aligned} E(X) &= \sum_{i=1}^n x_i P(x_i) \\ &= 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} \\ &= 1,0. \end{aligned}$$

e

$$E(X^2) = \sum_{i=1}^n x_i^2 P(x_i) = 0^2 \cdot \frac{1}{4} + 1^2 \cdot \frac{1}{2} + 2^2 \cdot \frac{1}{4} = 1,5.$$

Portanto:

$$\begin{aligned} V(X) &= E(X^2) - [E(X)]^2 \\ &= 1,5 - 1,0^2 \\ &= 1,5 - 1,0 \\ &= 0,5. \end{aligned}$$

O desvio padrão da variável aleatória  $X$  é obtido por:

$$dp(X) = \sqrt{V(X)} = \sqrt{0,5} \approx 0,7071.$$

## 2.2 Covariância e Correlação

Anteriormente discutimos a ausência ou presença de uma relação entre duas variáveis aleatórias, independência ou não independência. Mas se houve uma relação, esta poderá ser forte ou fraca. Uma das medidas para medir este grau de dependência é a *covariância*, como já definimos anteriormente. Uma outra medida, mais utilizada, e de melhor interpretação, pode ser obtida através do *Coeficiente de Correlação*.

### Definição

A covariância mede o grau de dependência entre duas variáveis aleatórias  $X$  e  $Y$ .

$$\begin{aligned}
COV(X, Y) &= E\{[X - E(X)] \cdot [Y - E(Y)]\} \\
&= E\{[XY - XE(Y) - YE(X) + E(X)E(Y)]\} \\
&= E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y) \\
&= E(XY) - E(X)E(Y)
\end{aligned}$$

Se  $X$  e  $Y$  forem independentes temos  $E(XY) = E(X)E(Y)$  então  $COV(X, Y) = 0$ .

### Exemplo

Considere a distribuição conjunta de  $X$  e  $Y$  dada na tabela a seguir:

$Y \backslash X$	0	1	2	$p(y)$
1	3/20	3/20	2/20	8/20
2	1/20	1/20	2/20	4/20
3	4/20	1/20	3/20	8/20
$p(x)$	8/20	5/20	7/20	1,00

Temos que:

$$E(X) = 0 \times \frac{8}{20} + 1 \times \frac{5}{20} + 2 \times \frac{7}{20} = 0,95$$

$$E(Y) = 1 \times \frac{8}{20} + 2 \times \frac{4}{20} + 3 \times \frac{8}{20} = 2,00$$

$$\begin{aligned}
E(XY) &= (0 \times 1) \times \frac{3}{20} + (0 \times 2) \times \frac{1}{20} + (0 \times 3) \times \frac{4}{40} + \\
&\quad + (1 \times 1) \times \frac{3}{20} + (1 \times 2) \times \frac{1}{20} + (1 \times 3) \times \frac{1}{20} + \\
&\quad + (2 \times 1) \times \frac{2}{20} + (2 \times 2) \times \frac{2}{20} + (2 \times 3) \times \frac{3}{20} \\
&= 1,90
\end{aligned}$$

Obtemos, então:

$$\begin{aligned}
COV(X, Y) &= E(XY) - E(X)E(Y) \\
&= 1,90 - (0,95)(2,00) \\
&= 0.
\end{aligned}$$

Portanto, as variáveis aleatórias  $X$  e  $Y$  desse exemplo são independentes.

## Propriedades da Variância

- a.  $V(k) = 0$ ;
- b.  $V(kX) = k^2 V(X)$ , sendo  $k$  uma constante;
- c.  $V(X \pm Y) = V(X) + V(Y) \pm 2COV(X, Y)$ ;
- d.  $V(aX \pm bY) = a^2 V(X) + b^2 V(Y) \pm 2ab COV(X, Y)$  em que  $a$  e  $b$  são constantes.

Sabemos que se  $X$  e  $Y$  são **independentes** então  $COV(X, Y) = 0$ . Logo,

$$V(X \pm Y) = V(X) + V(Y).$$

## Definição

A correlação (*Coeficiente de Correlação*) de  $X$  e  $Y$  é o número definido por

$$\rho_{XY} = \frac{COV(X, Y)}{\sigma_X \sigma_Y}$$

A correlação é sempre entre -1 e 1, com os valores  $-1$  e  $1$  indicando uma perfeita relação linear entre  $X$  e  $Y$ , isto é,  $-1 \leq \rho_{XY} \leq 1$ .

- a) Quando  $\rho_{XY} > 0$  existe uma correlação positiva entre as duas variáveis;
- b) Quando  $\rho_{XY} < 0$  existe uma correlação negativa entre as duas variáveis;
- c) Quando  $\rho_{XY} = 0$  não existe uma correlação entre as duas variáveis.

## Exemplo

Duas variáveis aleatórias  $X$  e  $Y$  resultaram  $COV(X, Y) = 0,25$ ,  $Var(X) = 0,75$  e  $Var(Y) = 0,25$ , logo:

$$\begin{aligned} \rho(X, Y) &= \frac{COV(X, Y)}{\sigma(X)\sigma(Y)} \\ &= \frac{0,25}{\sqrt{0,75} \cdot \sqrt{0,25}} \\ &= 0,58. \end{aligned}$$

## 2.3 Distribuição de Bernoulli

Nos experimentos de Bernoulli o espaço amostral é composto por apenas dois resultados possíveis: "sucesso" (resultado de interesse) ou "fracasso" (resultado pelo qual não estamos interessados).

### Exemplos

- (i)  $E_1$ : "Lançar uma moeda": pode sair cara ou coroa;
- (ii)  $E_2$ : "Plantar uma semente": pode germinar ou não.

Seja  $X$  a variável aleatória: sucesso ou fracasso. Seja  $p$  a probabilidade de ocorrer sucesso e  $1 - p$  a probabilidade de fracasso. Podemos ver, pela Tabela 2.3, as possibilidades, as probabilidades e os valores da variável aleatória de Bernoulli. Note que nos valores da variável codificamos sucesso como  $x_i = 1$  e fracasso como  $x_i = 0$ .

**Tabela 2.3.** Resultados possíveis, probabilidades e valores ( $x_i$ ) da variável aleatória  $X$  de Bernoulli

Resultados Possíveis	Probabilidades	$x_i$
Sucesso	$p$	1
Fracasso	$1 - p$	0

A distribuição de probabilidade de  $X$  com distribuição de Bernoulli, com parâmetro  $p$  é apresentada na Tabela 2.4.

**Tabela 2.4.** Distribuição de probabilidade da variável aleatória  $X$  de Bernoulli

$x_i$	$P(X = x_i)$
0	$1 - p$
1	$p$
Total	1

Pode-se calcular a média desta distribuição:

$$\begin{aligned}
 \mu = E(X) &= \sum_{i=1}^n x_i P(x_i) \\
 &= 0 \cdot (1 - p) + 1 \cdot p \\
 &= p.
 \end{aligned}$$

Pode-se calcular a variância desta distribuição:

$$\begin{aligned}
 \text{Var}(X) &= \sum_{i=1}^n [x_i - \mu]^2 P(x_i) \\
 &= (0 - \mu)^2 \cdot (1 - p) + (1 - \mu)^2 \cdot p \\
 &= (0 - p)^2 \cdot (1 - p) + (1 - p)^2 \cdot p \\
 &= p^2(1 - p) + p(1 - p)^2 \\
 &= p^2(1 - p) + p(1 - 2p + p^2) \\
 &= p^2 - p^3 + p - 2p^2 + p^3 \\
 &= p - p^2 \\
 &= p(1 - p).
 \end{aligned}$$

Portanto:

$$E(X) = p \quad \text{e} \quad \text{Var}(X) = p(1 - p).$$

**Notação:**  $X \sim \text{Bernoulli}(p)$ .

## Função de probabilidades

A função de probabilidades de uma distribuição de Bernoulli é dada por:

$$P(X = x_i) = p^{x_i}(1 - p)^{1-x_i}, \quad x_i = 0, 1.$$

## 2.4 Distribuição Binomial

Trata-se de uma distribuição de probabilidade adequada aos experimentos que apresentam apenas dois resultados (sucesso ou fracasso). Este modelo fundamenta-se nas seguintes hipóteses:

1. São realizados  $n$  ensaios **independentes** (o resultado de um experimento não é afetado pelo resultado dos outros) de Bernoulli;
2. A probabilidade de sucesso em cada ensaio é  $p$  e de fracasso é  $q = 1 - p$ ;
3. O número observado de sucessos é um número inteiro entre 0 e  $n$ .



Diz-se que a variável  $X$ : "número de sucessos em  $n$  ensaios" tem **distribuição binomial** com parâmetros  $n$  e  $p$ , onde  $n$  é o número de ensaios e  $p$  é a probabilidade de sucesso em cada ensaio.

## Função de Probabilidades

A função de probabilidades de uma variável  $X$  com distribuição Binomial,  $X \sim \text{Bin}(n, p)$ , é dada por:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, \dots, n$$

ou

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, \dots, n$$

em que  $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ .

**Notação:**  $X \sim \text{Bin}(n, p)$ .

## Exemplo

Suponha que em uma maternidade irão nascer três bebês. Vamos estudar as possibilidades do sexo dos bebês nos três nascimentos. Considerando a variável  $X$ : "número de meninos em três nascimentos", e considerando que a probabilidade de nascer menino é  $p = 0,5$  (então  $q = 0,5$ , pois  $q = 1 - p$ ), temos:

- $P(X = 0) = \binom{3}{0} 0,5^0 \cdot 0,5^{3-0} = 1 \cdot 1 \cdot 0,5^3 = 0,125$
- $P(X = 1) = \binom{3}{1} 0,5^1 \cdot 0,5^{3-1} = 3 \cdot 0,5 \cdot 0,5^2 = 0,375$
- $P(X = 2) = \binom{3}{2} 0,5^2 \cdot 0,5^{3-2} = 3 \cdot 0,5^2 \cdot 0,5 = 0,375$
- $P(X = 3) = \binom{3}{3} 0,5^3 \cdot 0,5^{3-3} = 1 \cdot 0,5^3 \cdot 1 = 0,125$

**Comandos no Software R para calcular as probabilidades da distribuição Binomial:**

```
dbinom(0, 3, 0.5) #x=0, n=3, p=0.5
dbinom(1, 3, 0.5) #x=1, n=3, p=0.5
dbinom(2, 3, 0.5) #x=2, n=3, p=0.5
dbinom(3, 3, 0.5) #x=3, n=3, p=0.5
```

## Exemplo

Considere a variável  $X$ : "número de meninos em três nascimentos", do exemplo anterior, porém, considere agora que a probabilidade de nascer menino é  $p = 0,6$  (então  $q = 0,4$ , pois  $q = 1 - p$ ). Assim, temos:

- $P(X = 0) = \binom{3}{0} 0,6^0 \cdot 0,4^{3-0} = 1 \cdot 1 \cdot 0,4^3 = 0,064$
- $P(X = 1) = \binom{3}{1} 0,6^1 \cdot 0,4^{3-1} = 3 \cdot 0,6 \cdot 0,4^2 = 0,288$
- $P(X = 2) = \binom{3}{2} 0,6^2 \cdot 0,4^{3-2} = 3 \cdot 0,6^2 \cdot 0,4 = 0,432$
- $P(X = 3) = \binom{3}{3} 0,6^3 \cdot 0,4^{3-3} = 1 \cdot 0,6^3 \cdot 1 = 0,216$

**Comandos no Software R para calcular as probabilidades da distribuição Binomial:**

```
dbinom(0, 3, 0.6) #x=0, n=3, p=0.6
dbinom(1, 3, 0.6) #x=1, n=3, p=0.6
dbinom(2, 3, 0.6) #x=2, n=3, p=0.6
dbinom(3, 3, 0.6) #x=3, n=3, p=0.6
```

## Média e variância na Distribuição Binomial

A média  $\mu$  de uma variável  $X \sim \text{Bin}(n, p)$  é dada por:

$$\mu = np$$

e a variância  $\sigma^2$  de uma variável  $X \sim \text{Bin}(n, p)$  é dada por:

$$\sigma^2 = npq.$$

## Exemplo

A probabilidade de nascer um menino é  $p = 0,5$  (ignorando nascimento de gêmeos e nascimentos múltiplos). Calcule a média e a variância do número de meninos em 1.000 nascimentos.

## Solução

A média é

$$\mu = np = 1000 \cdot 0,5 = 500 \text{ meninos,}$$

e a variância é

$$\sigma^2 = npq = 1000 \cdot 0,5 \cdot 0,5 = 250.$$

## 2.5 Distribuição de Poisson

A distribuição de Poisson é largamente empregada quando se deseja contar o número de ocorrências de um evento de interesse, por unidade de tempo, comprimento, área ou volume.

### Exemplos

- a) Número de insetos de uma espécie coletados por armadilha por dia;
- b) Número de chamadas recebidas por um telefone durante cinco minutos;
- c) Número de falhas de um computador num dia de operação;
- d) Número de bactérias por *ml* de urina;
- e) Número de pacientes que chegam a um pronto atendimento de uma pequena cidade durante a madrugada.

Note que os possíveis valores que as variáveis descritas podem assumir são: 0, 1, . . . , . O comportamento dessas variáveis pode ser descrito pela chamada distribuição de Poisson.

### Função de Probabilidades

A função de probabilidades de uma variável  $X$  com distribuição Poisson é dada por:

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, \dots$$

em que  $e = 2,7182 \dots$  é o número de Euler, e  $\lambda$  é o número médio de ocorrências do evento de interesse por unidade de tempo, distância ou área.

**Notação:**  $X \sim \text{Pois}(\lambda)$

## Média e variância na distribuição de Poisson

A esperança e a variância de uma variável aleatória  $X \sim \text{Pois}(\lambda)$  são dadas, respectivamente, por:

$$\mu = \lambda \quad \text{e} \quad \sigma^2 = \lambda.$$

### Exemplo

Uma telefonista recebe, em média, 2 chamadas por hora. Calcular a probabilidade de se receber:

- a) nenhuma chamada em 1 hora;
- b) 1 chamada em 1 hora;
- c) 2 chamadas em 1 hora;
- d) no máximo 2 chamadas em 1 hora.
- e) 3 chamadas em 2 horas;

### Solução

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

$$\text{a) } P(X = 0) = \frac{e^{-2} 2^0}{0!} = 0,1353$$

$$\text{b) } P(X = 1) = \frac{e^{-2} 2^1}{1!} = 0,2707$$

$$\text{c) } P(X = 2) = \frac{e^{-2} 2^2}{2!} = 0,2707$$

$$\text{d) } P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

$$= 0,1353 + 0,2707 + 0,2707 = 0,6767$$

- e) Como  $\lambda = 2$  chamadas em 1 hora, então, em 2 horas, teremos  $\lambda^* = 4$  chamadas. Logo:

$$P(X = 3) = \frac{e^{-(\lambda^*)} (\lambda^*)^x}{x!} = \frac{e^{-4} 4^3}{3!} = 0,1954$$

**Comandos no Software R para calcular as probabilidades da distribuição de Poisson:**

```
dpois(0, 2) #x=0, lambda=2
dpois(1, 2) #x=1, lambda=2
dpois(2, 2) #x=2, lambda=2
sum(dpois(0:2, 2)) #x=0:2, lambda=2
dpois(3, 4) #x=3, lambda=4
```

**Exemplo**

Um pesquisador está interessado no número de ovos depositados por uma espécie de pássaro. Na primavera, ele procura e acha 80 ninhos. O número médio de ovos por ninho foi 3,8 e a variância foi 3,1. Porque a variância é aproximadamente igual à média, ele acha que pode ser razoável descrever o número de ovos por ninho como tendo uma distribuição Poisson com média  $\lambda = 3,8$ .

- a) Se esta realmente representa a distribuição populacional, qual seria a probabilidade de encontrar um ninho com nenhum ovo?
- b) Qual seria a probabilidade de encontrar um ninho com no máximo 4 ovos?
- c) Qual seria a probabilidade de encontrar um ninho com mais de 4 ovos?

**Solução**

$$\text{a) } P(X = 0) = \frac{e^{-3,8} 3,8^0}{0!} = 0,0224$$

$$\text{b) } P(X \leq 4) = P(0) + P(1) + P(2) + P(3) + P(4)$$

$$\begin{aligned} &= \frac{e^{-3,8} 3,8^0}{0!} + \frac{e^{-3,8} 3,8^1}{1!} + \frac{e^{-3,8} 3,8^2}{2!} + \frac{e^{-3,8} 3,8^3}{3!} + \frac{e^{-3,8} 3,8^4}{4!} \\ &= 0,0224 + 0,0850 + 0,1615 + 0,2046 + 0,1944 \\ &= 0,6679 \end{aligned}$$

$$\text{c) } P(X > 4) = 1 - P(X \leq 4) = 1 - 0,6679 = 0,3321$$

**Comandos no Software R para calcular as probabilidades da distribuição de Poisson:**

```
dpois(0, 3.8) #x=0, lambda=3.8
sum(dpois(0:4, 3.8)) #x=0:4, lambda=3.8
1-sum(dpois(0:4, 3.8)) #x=5:Inf, lambda=3.8
```

## 2.6 Distribuição Geométrica

Considere uma sequência de ensaios de Bernoulli independentes (com probabilidade de sucesso  $p$ ,  $0 < p < 1$ ). Defina  $X$  como o número de fracassos anteriores ao primeiro sucesso ou, em outras palavras, o tempo de espera (em termos de ensaios anteriores) para o primeiro sucesso. A variável  $X$  segue o modelo Geométrico com parâmetro  $p$ .

### Função de Probabilidades

A função de probabilidade do modelo Geométrico é dada por:

$$P(X = x) = (1 - p)^x \cdot p, \quad x = 0, 1, 2, \dots$$

**Notação:**  $X \sim Geo(p)$ .

### Média e variância da Distribuição Geométrica

$$E(X) = \frac{1}{p} \quad \text{e} \quad V(X) = \frac{1 - p}{p^2}.$$

### Exemplo

Uma linha de fabricação de um equipamento de precisão é interrompida na primeira ocorrência de um defeito. A partir da manutenção, o equipamento tem probabilidade de 0,01 de apresentar defeito em um dia qualquer. Seja  $X$  a variável que conta o número de dias que antecedem a interrupção. Admitindo que o desempenho, nos dias sucessivos, sejam independentes, temos que  $X \sim Geo(p = 0,01)$ . Dessa forma,

$$P(X = x) = 0,99^x \cdot 0,01, \quad x = 0, 1, 2, \dots$$

Por exemplo, para uma interrupção no sexto dia temos:

$$P(X = 5) = 0,99^5 \cdot 0,01 = 0,0095.$$

**Observação:** Para que a interrupção ocorra no sexto dia então o defeito deve ocorrer no sexto dia, assim, o número de dias que antecedem ao defeito é  $x = 5$  dias.

Considerando ainda, o mesmo exemplo, qual seria o intervalo de tempo ideal para uma manutenção preventiva, se desejamos uma probabilidade de, pelo menos, 0,90 de que o defeito não ocorrerá?

Para isso precisamos saber quantos dias são necessários para acumular uma probabilidade de ocorrer defeito próxima de 0,10 (que é análogo a acumular uma probabilidade de não ocorrer defeito próxima de 0,90).

**Tabela.** Probabilidades antes de ocorrer defeito

Tempo de espera (dias) para ocorrer defeito	Probabilidade	Prob. acumulada de <b>Ocorrer defeito</b>	Prob. acumulada de <b>Não ocorrer defeito</b>
$x$	$P(X = x)$	$P(X \leq x)$	$1 - P(X \leq x)$
0	0,0100	0,0100	0,9900
1	0,0099	0,0199	0,9801
2	0,0098	0,0297	0,9703
3	0,0097	0,0394	0,9606
4	0,0096	0,0490	0,9510
5	0,0095	0,0585	0,9415
6	0,0094	0,0679	0,9321
7	0,0093	0,0773	0,9228
8	0,0092	0,0865	0,9135
9	0,0091	<b>0,0956</b>	<b>0,9044</b>
10	0,0090	0,1047	0,8953

Observe que obtivemos  $P(X \leq 9) = 0,0956$ , ou seja, a probabilidade (acumulada) do tempo de espera antes de ocorrer defeito está próxima de 0,10. Portanto, a manutenção preventiva deve ser feita à partir de 9 dias de operação.

## 2.7 Distribuição Hipergeométrica

Considere um conjunto com  $n$  objetos dos quais  $m$  são do tipo  $I$  e  $n - m$  são do tipo  $II$ . Uma amostra é escolhida, ao acaso e sem reposição, com tamanho  $r$  ( $r < n$ ) e definimos  $X$  como o número de objetos com a característica  $I$ , na amostra. Nesse caso, diremos que a variável aleatória  $X$  segue o modelo Hipergeométrico de parâmetros  $m$ ,  $n$  e  $r$ .

## Função de Probabilidades

A função de probabilidade do modelo Hipergeométrico é dada por:

$$P(X = x) = \frac{\binom{m}{x} \binom{n-m}{r-x}}{\binom{n}{r}},$$

sendo  $x$  inteiro e tal que  $\max\{0, r - (n - m)\} \leq x \leq \min\{r, m\}$ . Estes limites para  $x$  garantem que situações absurdas sejam evitadas.

**Notação:**  $X \sim Hgeo(m, n, r)$ .

## Exemplo

Considere que, num lote de 20 peças, existam 4 defeituosas. Seleccionando-se 5 peças, sem reposição, qual seria a probabilidade de 2 defeituosas serem escolhidas?

## Solução

A variável aleatória que indica o número de peças defeituosas na amostra seleccionada segue o modelo Hipergeométrico. Denominando por  $X$  essa variável, temos:

$$n = 20, \quad m = 4, \quad r = 5, \quad x = 2, \quad \text{e} \quad P(X = x) = \frac{\binom{m}{x} \binom{n-m}{r-x}}{\binom{n}{r}},$$

então:

$$P(X = 2) = \frac{\binom{4}{2} \binom{20-4}{5-2}}{\binom{20}{5}} = \frac{\binom{4}{2} \binom{16}{3}}{\binom{20}{5}} = 0,2167.$$

## Exemplo

Pequenos motores elétricos são expedidos em lotes de 50 unidades. Antes que uma remessa seja aprovada, um inspetor escolhe 5 desses motores e os inspeciona. Se nenhum dos motores inspecionados for defeituoso o lote é aprovado. Se um ou mais forem verificados defeituosos, todos os motores da remessa são inspecionados.



Suponha que existam, de fato, três motores defeituosos no lote. Qual é a probabilidade de que a inspeção de 100% do lote seja necessária?

### Solução

A inspeção de 100% do lote será necessária se, e somente se,  $X \geq 1$ . Logo,

$$P(X \geq 1) = 1 - P(X = 0)$$

Temos, que:

$$n = 50, \quad m = 3, \quad r = 5 \quad x = 0 \quad \text{e} \quad P(X = x) = \frac{\binom{m}{x} \binom{n-m}{r-x}}{\binom{n}{r}}$$

Portanto:

$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) \\ &= 1 - \frac{\binom{3}{0} \binom{50-3}{5-0}}{\binom{50}{5}} \\ &= 1 - \frac{\binom{3}{0} \binom{47}{5}}{\binom{50}{5}} \\ &= 1 - 0,7240 \\ &= 0,2760. \end{aligned}$$

### 3. Variáveis Aleatórias Contínuas

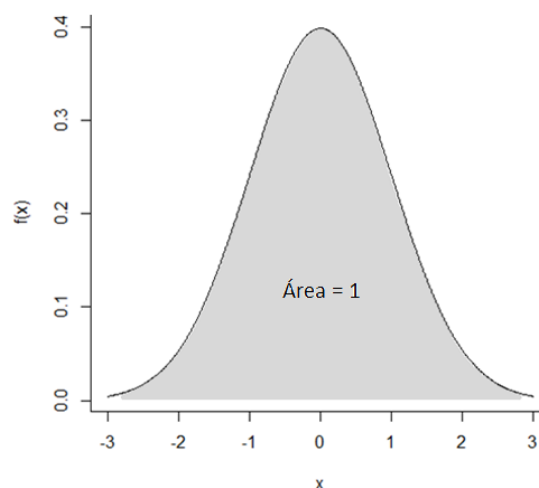
#### 3.1 Conceito. Noções básicas de esperança matemática e variância

Uma variável aleatória contínua pode tomar um número infinito de valores, e esse valores podem ser associados a mensurações em uma escala contínua e as probabilidades necessárias ao seu estudo são calculadas como a área abaixo da curva da distribuição, chamada de função densidade de probabilidade.

**Definição:** Uma variável aleatória contínua  $X$  é contínua em  $\mathbb{R}$ , se existir uma função  $f(x)$ , tal que:

- i)  $f(x) \geq 0, \forall x \in \mathbb{R}$ ;
- ii) Área entre o gráfico da função  $f$  e o eixo  $x$  é igual a 1, ou seja,

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$



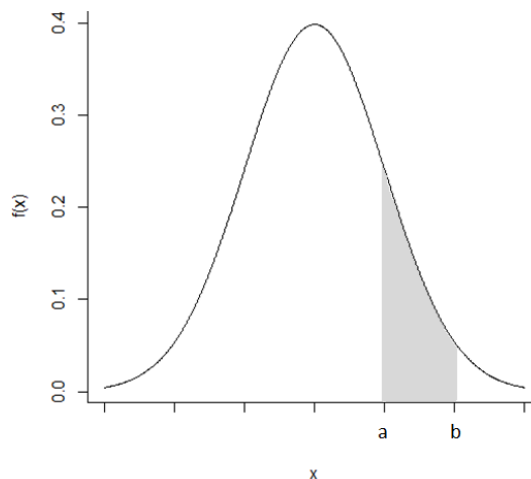
A função  $f(x)$  é chamada **função densidade de probabilidade** (f.d.p.).

## Cálculo de probabilidades para uma variável aleatória contínua

Para qualquer  $a < b$  pertencente ao contradomínio de  $X$ , temos:

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

Graficamente, a probabilidade acima corresponde à área limitada pela função  $f(x)$ , eixo  $x$  e pelas retas  $x = a$  e  $x = b$ :



Da relação entre a probabilidade e a área sob a função, a inclusão ou não dos extremos  $a$  e  $b$  não afetará os resultados. Assim, será admitido que

$$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b).$$

### Exemplo

Considere a variável aleatória contínua  $X$  cuja fdp é dada por:

$$f(x) = \begin{cases} 2x, & \text{se } 0 < x < 1, \\ 0, & \text{se } x \leq 0 \text{ ou } x \geq 1 \end{cases}$$

Evidentemente,  $f(x) \geq 0$ . Além disso,

$$\int_{-\infty}^{+\infty} f(x) dx = \int_{-\infty}^0 0 dx + \int_0^1 2x dx + \int_1^{+\infty} 0 dx = \int_0^1 2x dx = x^2 \Big|_0^1 = 1^2 - 0^2 = 1.$$

Logo,  $f(x)$  é realmente uma fdp.

Para exemplificar o cálculo de uma probabilidade, vamos calcular, por exemplo,  $P(0 < X < 1/2)$ :

$$P\left(0 < X < \frac{1}{2}\right) = \int_0^{\frac{1}{2}} 2x \, dx = x^2 \Big|_0^{\frac{1}{2}} = \left(\frac{1}{2}\right)^2 - 0^2 = \frac{1}{4}.$$

### Função de distribuição acumulada

Se  $X$  é uma variável aleatória contínua com função densidade de probabilidade  $f(y)$  define-se a sua função de distribuição acumulada  $F(x)$  como:

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt$$

Se  $a$  e  $b$  forem dois números reais quaisquer, tem-se que:

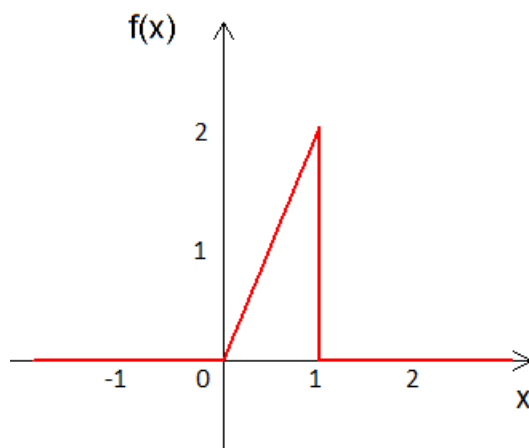
$$P(a < X < b) = F(b) - F(a)$$

### Exemplo

Considere a variável aleatória contínua  $X$  cuja fdp é dada por:

$$f(x) = \begin{cases} 0, & \text{se } x \leq 0, \\ 2x, & \text{se } 0 < x < 1, \\ 0, & \text{se } x \geq 1. \end{cases}$$

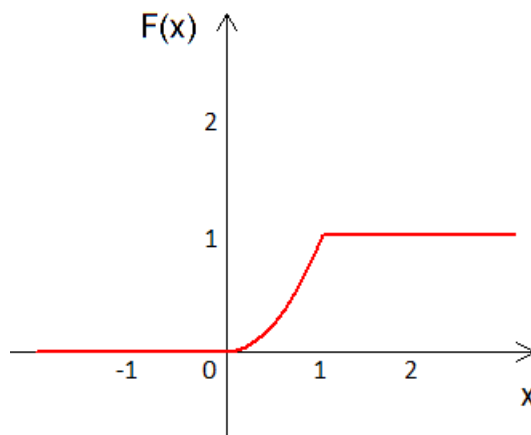
Eis o gráfico de  $f(x)$ :



Portanto, a função de distribuição é dada por:

$$F(x) = \begin{cases} 0 & \text{se } x \leq 0, \\ \int_0^x 2t \, dt = x^2 & \text{se } 0 < x \leq 1, \\ 1 & \text{se } x > 1. \end{cases}$$

Eis o gráfico de  $F(x)$ :



### Teorema

Seja  $F$  a função de distribuição acumulada de uma variável aleatória contínua, com fdp  $f$ . Então,

$$f(x) = \frac{d}{dx}F(x),$$

para todo  $x$  no qual  $F$  seja derivável.

### Demonstração

$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) \, dt$ . Aplicando-se o Teorema Fundamental do Cálculo, obteremos  $F'(x) = f(x)$ .

### Exemplo

Seja  $X$  uma variável aleatória contínua com função de distribuição acumulada  $F$  dada por

$$F(x) = \begin{cases} 0, & \text{se } x \leq 0 \\ 1 - e^{-x}, & \text{se } x > 0 \end{cases}$$

Neste caso,  $F'(x) = e^{-x}$  para  $x > 0$ , e, por isso, a fdp será dada por

$$f(x) = \begin{cases} 0, & \text{se } x \leq 0 \\ e^{-x}, & \text{se } x > 0 \end{cases}$$

## Média e Variância

A esperança matemática (ou média,  $\mu$ ) e a variância ( $\sigma^2$ ) de uma variável aleatória contínua  $X$ , são dadas, respectivamente, por:

$$\mu = E(X) = \int_{-\infty}^{+\infty} xf(x)dx,$$

e

$$\sigma^2 = E(X^2) - [E(X)]^2,$$

em que:

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x)dx.$$

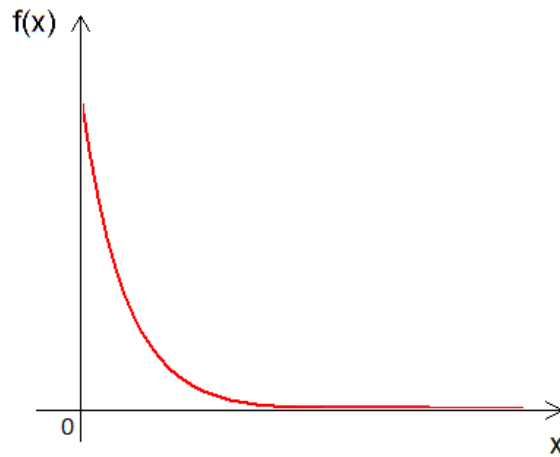
## Exemplo

Seja  $X$  uma variável aleatória com função de densidade de probabilidade dada por:

$$f(x) = \lambda e^{-\lambda x}, \quad \lambda > 0, \quad x > 0$$

Primeiramente, vamos verificar se  $f(x)$  é uma fdp de  $X$ . Temos que:

i)  $f(x) \geq 0$  para todo  $x$ , como pode-se observar pelo gráfico de  $f$ ;



ii)

$$\begin{aligned}
 \int_0^{\infty} \lambda e^{-\lambda x} dx &= -e^{-\lambda x} \Big|_0^{\infty} \\
 &= \lim_{x \rightarrow \infty} (-e^{-\lambda x}) - (-e^{-0}) \\
 &= \lim_{x \rightarrow \infty} \left( -\frac{1}{e^{\lambda x}} \right) + e^0 \\
 &= 0 + 1 \\
 &= 1
 \end{aligned}$$

ou seja,  $f(x)$  é realmente uma *fdp* da variável  $X$ .

Vamos, agora, determinar a esperança matemática de  $X$ :

$$E(X) = \int_0^{\infty} x \lambda e^{-\lambda x} dx$$

Resolveremos usando integração por partes:

$$\int u dv = uv - \int v du$$

então  $u = x$  e  $dv = \lambda e^{-\lambda x} dx$  teremos  $du = dx$  e  $v = -e^{-\lambda x}$ . Logo

$$\begin{aligned}
 E(X) &= \int_0^{\infty} x \lambda e^{-\lambda x} dx = -x e^{-\lambda x} \Big|_0^{\infty} - \int_0^{\infty} -e^{-\lambda x} dx \\
 &= -\frac{x}{e^{\lambda x}} \Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx
 \end{aligned}$$

$$\begin{aligned}
&= -\frac{x}{e^{\lambda x}} \Big|_0^{\infty} + \frac{-e^{-\lambda x}}{\lambda} \Big|_0^{\infty} \\
&= -\frac{x}{e^{\lambda x}} \Big|_0^{\infty} + \frac{-\left(\frac{1}{e^{\lambda x}}\right)}{\lambda} \Big|_0^{\infty} \\
&= [-0 - (-0)] + \left[-0 - \left(-\frac{1}{\lambda}\right)\right] \\
&= \frac{1}{\lambda}.
\end{aligned}$$

**Observação:** Para determinar a variância da variável aleatória acima devemos obter  $E(X^2)$ , que é calculada por:

$$E(X^2) = \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx,$$

e, então, calcular a variância como:

$$Var(X) = E(X^2) - E^2(X).$$

O cálculo da variância desta variável aleatória fica como exercício.

### 3.2 Distribuição exponencial

A variável aleatória  $X$ , que mede a "distância" entre contagens sucessivas de um processo de Poisson, com média  $\lambda > 0$ , é uma variável aleatória com distribuição **Exponencial** com parâmetro  $\lambda$ . **Notação:**  $X \sim Exp(\lambda)$ .

#### Função densidade de probabilidade

$$f(x) = \lambda e^{-\lambda x}, \quad \text{para } 0 \leq x < \infty.$$

#### Cálculo de probabilidades

$$P(a \leq X \leq b) = \int_a^b \lambda e^{-\lambda x} dx.$$



## Função de distribuição

$$F(x) = P(X \leq x) = \begin{cases} 1 - e^{-\lambda x}, & \text{se } x \geq 0 \\ 0, & \text{se } x < 0 \end{cases}$$

## Demonstração

$$\begin{aligned} \int_{-\infty}^x \lambda e^{-\lambda t} dt &= \int_0^x \lambda e^{-\lambda t} dt = (-e^{-\lambda t}) \Big|_0^x = (-e^{-\lambda x}) - (-e^{-\lambda(0)}) = -e^{-\lambda x} + e^0 \\ &= -e^{-\lambda x} + 1 = 1 - e^{-\lambda x} \end{aligned}$$

## Média e variância

Se  $X \sim \text{Exp}(\lambda)$ , então:

$$\mu = \frac{1}{\lambda} \quad \text{e} \quad \sigma^2 = \frac{1}{\lambda^2}$$

## Exemplo

Uma média de cinco chamadas por hora são recebidas em um departamento de manutenção. Considerando que o tempo ( $X$ ) entre as chamadas é uma variável aleatória com distribuição Exponencial com  $\lambda = 5$  e, começando a observação em qualquer ponto no tempo, determine a probabilidade de que a primeira chamada de serviço chegue dentro de meia hora.

## Solução

A probabilidade de ocorrer uma chamada dentro de meia hora (0,5 horas) é determinada por:

$$\begin{aligned} P(X \leq 0,5) &= 1 - e^{-5(0,5)} \\ &= 1 - e^{-2,5} \\ &= 1 - 0,0821 \\ &= \mathbf{0,9179.} \end{aligned}$$

## Exemplo

Em uma grande rede de computadores, as conexões dos usuários do sistema podem ser modeladas como um processo de Poisson, com média de 25 conexões por hora. Qual é a probabilidade de :

- a) o tempo até a próxima conexão estar entre 2 e 3 minutos?
- b) não haver conexão nos próximos 6 minutos, ou, analogamente, o tempo até a próxima conexão ser maior que 6 minutos?

## Solução

a)  $\lambda = 25$  conexões por hora, então, convertendo minutos para horas temos:

- 2 minutos  $= \frac{2}{60} = 0,033$  horas;
- 3 minutos  $= \frac{3}{60} = 0,05$  horas.

Logo,

$$\begin{aligned}
 P(0,033 < X < 0,05) &= \int_{0,033}^{0,05} 25e^{-25x} dx \\
 &= (-e^{-25x}) \Big|_{0,033}^{0,05} \\
 &= (-e^{-25(0,05)}) - (-e^{-25(0,033)}) \\
 &= \mathbf{0,1517}
 \end{aligned}$$

b)  $\lambda = 25$  conexões por hora, então, convertendo minutos para horas temos:

- 6 minutos  $= \frac{6}{60} = 0,1$  horas.

Logo,

$$\begin{aligned}
 P(X > 0,1) &= 1 - P(X \leq 0,1) \\
 &= 1 - [1 - e^{-25(0,1)}] \\
 &= e^{-25(0,1)} \\
 &= \mathbf{0,0821}
 \end{aligned}$$

### 3.3 Distribuição normal: características; distribuição normal padronizada

Dentre todas as distribuições de probabilidades, sejam discretas ou contínuas, a mais estudada e mais utilizada é a distribuição Normal. As principais razões que fazem a distribuição Normal o modelo mais importante na estatística são:

- 1) Muitas variáveis biométricas tendem a ter distribuição Normal. Isto ocorre principalmente quando a variável é influenciada por um grande número de fatores que atuam de modo independente e aditivo;
- 2) A distribuição das médias amostrais de uma variável qualquer tendem a ter distribuição Normal, mesmo que a variável em si não tenha distribuição Normal;
- 3) Muitos testes e modelos estatísticos têm como pressuposição a normalidade dos dados isto é, que os dados possuem distribuição Normal.

A distribuição Normal é também conhecida como distribuição Gaussiana em homenagem a Karl F. Gauss (1777-1855), brilhante matemático e físico alemão, que a desenvolveu no início do século XIX.

#### Função Densidade da Normal

A função densidade de probabilidade de uma variável aleatória contínua  $X$ , é dada por:

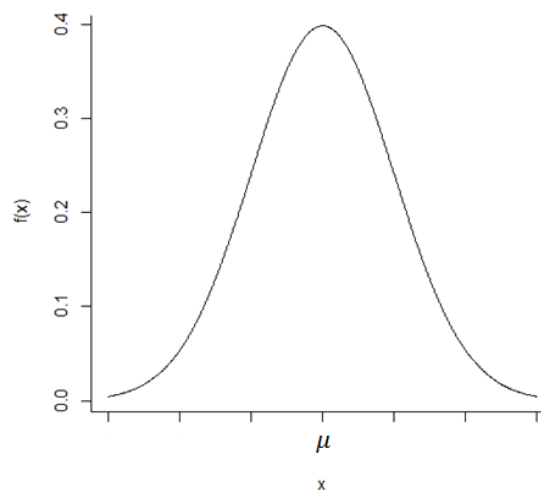
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad \text{para } -\infty < x < \infty.$$

em que:

- $\mu$  é a posição central da distribuição (média);
- $\sigma^2$  é a dispersão da distribuição (variância);
- $x$  são os valores que a variável aleatória em estudo  $X$  assume.

**Notação:**  $X \sim N(\mu, \sigma^2)$ .

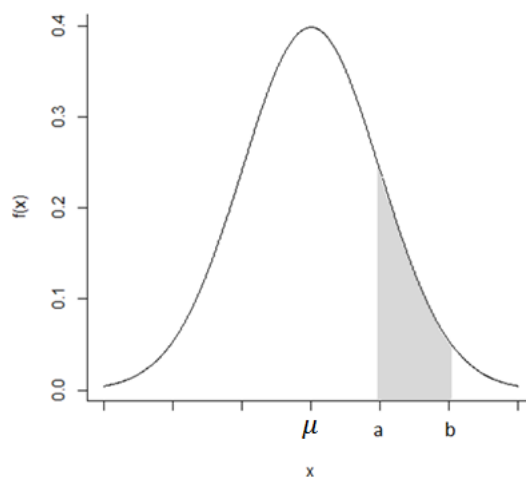
O gráfico da função densidade da distribuição Normal é apresentado a seguir:



Para se calcular a probabilidade da variável aleatória  $X$  assumir valores entre  $a$  e  $b$  basta calcular a área compreendida entre estes intervalos usando a fórmula:

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

cuja área é mostrada na figura a seguir.

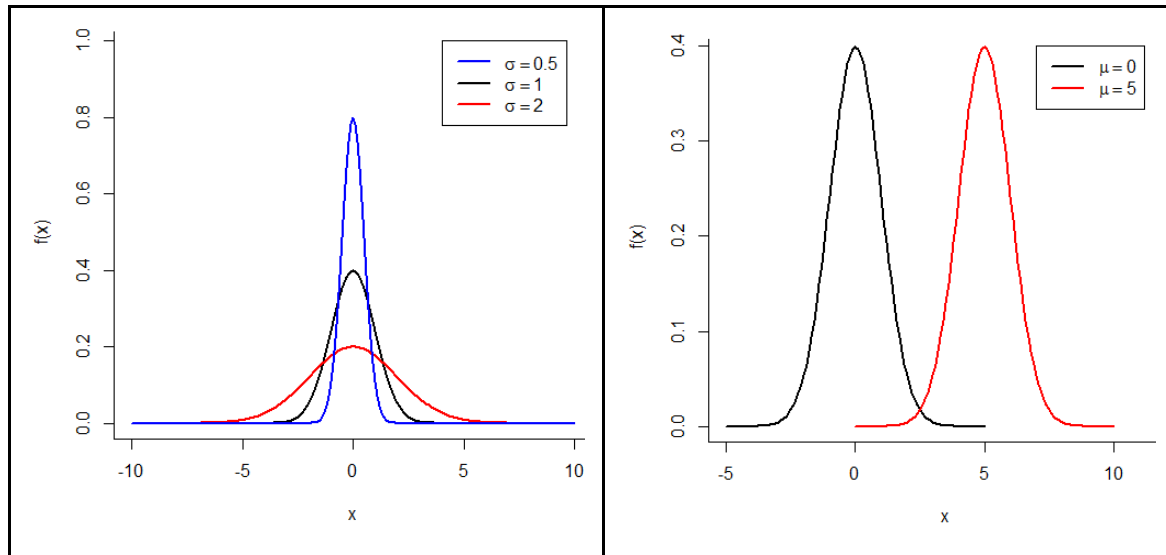


## Propriedades da Distribuição Normal

As principais características dessa função são:

- 1) A função gera um gráfico em forma de sino, sendo unimodal e simétrica;
- 2) é definida por dois parâmetros: a média ( $\mu$ ) e o desvio padrão ( $\sigma$ ), sendo que a média controla a localização do centro da distribuição (é o ponto de simetria), já o desvio padrão controla a dispersão da curva ao redor da média;
- 3) O ponto de máximo de  $f(x)$  ocorre no ponto de abscissa  $x = \mu$ ;

4) Não possui limite inferior ou superior;



O cálculo direto de probabilidades envolvendo a distribuição normal exige recursos do cálculo avançado e, mesmo assim, dada a forma da função densidade, não é um processo muito elementar. Por isso, elas foram tabeladas, permitindo-nos obter diretamente o valor da probabilidade desejada. Devido as dificuldades de cálculo e da dificuldade em se construir tabelas da função dependendo de dois parâmetros ( $\mu$  e  $\sigma^2$ ), recorre-se a uma mudança de variável, transformando a variável aleatória  $X$  na variável aleatória  $Z$ . Essa nova variável chama-se **variável normal padronizada**, ou reduzida.

### Distribuição Normal Padrão

Denomina-se distribuição normal padrão, a distribuição normal de média zero e variância 1, ou seja:

$$Z \sim N(\mu = 0, \sigma^2 = 1), \text{ ou simplesmente, } Z \sim N(0,1).$$

As probabilidades associadas a distribuição normal reduzida são facilmente obtidas em tabelas.

## Uso da Tabela da Distribuição Normal Padrão

### Exemplo

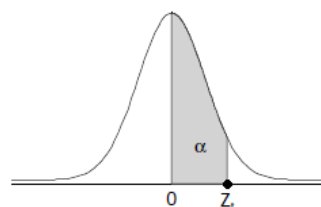
Seja  $Z \sim N(0, 1)$ . Usando a tabela da distribuição normal padrão, calcular

$$P(0 < Z < 1,57).$$

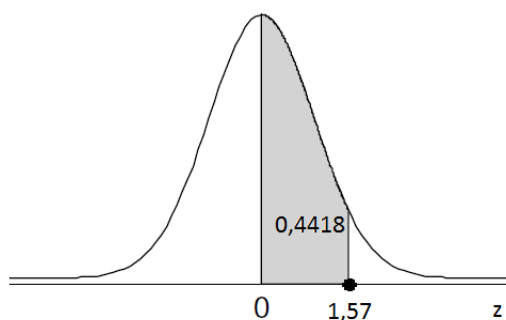
### Solução

#### Tabela da distribuição Normal Padrão

Probabilidades ( $\alpha$ ) da distribuição Normal Padrão  $N(0,1)$  para valores do quantil  $Z_t$  padronizado de acordo com o seguinte evento:  $P(0 < Z < Z_t) = \alpha$ .



$Z_t$	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
⋮										
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545



Portanto,

$$P(0 < Z < 1,57) = 0,4418.$$

### Exercício

Seja  $Z \sim N(0, 1)$ . Usando a tabela da distribuição normal padrão, calcular:

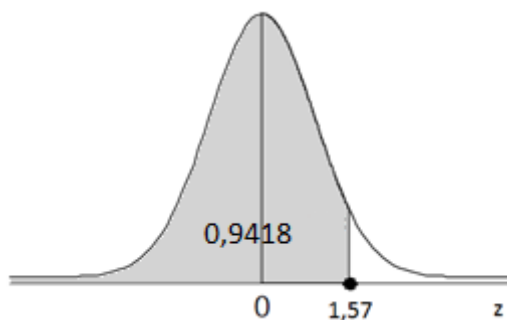
- a)  $P(Z > 0)$
- b)  $P(Z < 0)$
- c)  $P(0 < Z < 1,08)$
- d)  $P(-1,89 < Z < 0)$
- e)  $P(-1,23 < Z < 1,05)$
- f)  $P(Z < 1)$
- g)  $P(Z > 1)$
- h)  $P(0,5 < Z < 1)$

### Distribuição Normal usando o software R

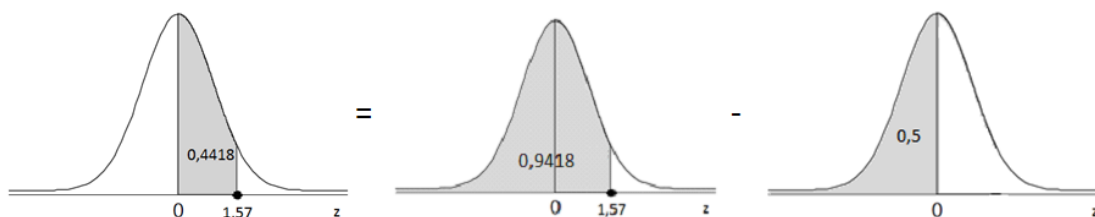
Na tabela que utilizamos acima as probabilidades são calculadas de 0 até  $Z_t$ , ou seja,  $P(0 < Z < Z_t)$ . Porém, no software R as probabilidades são calculadas de  $-\infty$  até  $Z_t$ , ou seja,  $P(-\infty < Z < Z_t)$ . Assim, quando usamos o comando:

```
pnorm(1.57, mean=0, sd=1)
```

o R calcula  $P(-\infty < Z < 1,57)$ :



Logo, para obter  $P(0 < Z < 1,57)$ , precisamos subtrair  $P(-\infty < Z < 0)$  de  $P(-\infty < Z < 1,57)$ :



$$\begin{aligned}
 P(0 < Z < 1,57) &= P(-\infty < Z < 1,57) - P(-\infty < Z < 0) \\
 &= \text{pnorm}(1.57, \text{mean}=0, \text{sd}=1) - \text{pnorm}(0, \text{mean}=0, \text{sd}=1)
 \end{aligned}$$

**Comandos no Software R para calcular as probabilidades da distribuição Normal:**

```
pnorm(1.57, mean=0, sd=1)           #P(-Inf < Z < 1.57)
pnorm(1.57, mean=0, sd=1) - pnorm(0, mean=0, sd=1)  #P(0 < Z < 1.57)
```

**Padronização de uma variável  $X$  com distribuição Normal**

Os problemas da vida real, entretanto, não se apresentam na forma reduzida da variável  $Z$ , ao contrário, são formulados em termos da variável normal original  $X$ , com média  $\mu$  e desvio-padrão  $\sigma$ . É preciso então, antes de passarmos à sua resolução, padronizar ou reduzir a variável aleatória normal  $X$ , transformando-a na variável aleatória  $Z$ :

$$Z = \frac{X - \mu}{\sigma}.$$

**Exemplo 1**

Seja  $X \sim N(4, 1)$ . Determine  $P(X \leq 4)$ .

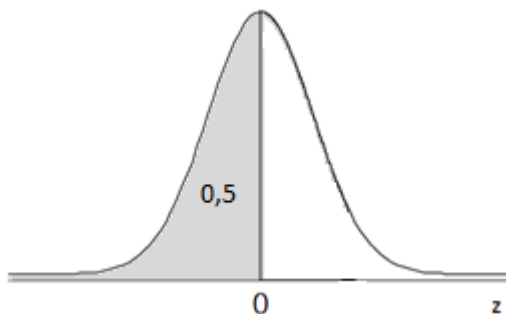
**Solução**

Como  $X \sim N(4, 1)$ , então:

$$\mu = 4 \quad \text{e} \quad \sigma = \sqrt{\sigma^2} = \sqrt{1} = 1$$

Logo:

$$\begin{aligned} P(X \leq 4) &= P\left(\frac{X - \mu}{\sigma} \leq \frac{4 - \mu}{\sigma}\right) = P\left(Z \leq \frac{4 - \mu}{\sigma}\right) = P\left(Z \leq \frac{4 - 4}{1}\right) \\ &= P(Z \leq 0) = 0,5 \end{aligned}$$



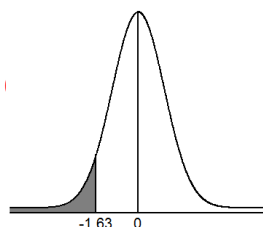


### Comandos no Software R para calcular as probabilidades da distribuição Normal:

```
pnorm(4, mean=4, sd=1) #Usando a variável  $X \sim N(4, 1)$ 
```

### Exemplo 2

A duração da gravidez humana, da concepção ao parto, varia segundo uma distribuição aproximadamente normal com média 266 dias e desvio padrão de 16 dias. Qual a probabilidade de uma gravidez durar menos de 240 dias?



$$\begin{aligned}
 P(X < 240) &= P\left(\frac{X - \mu}{\sigma} < \frac{240 - \mu}{\sigma}\right) \\
 &= P\left(Z < \frac{240 - \mu}{\sigma}\right) \\
 &= P\left(Z < \frac{240 - 266}{16}\right) \\
 &= P(Z < -1,63) \\
 &= 0,5 - 0,4484 \\
 &= \mathbf{0,0516}
 \end{aligned}$$

### Exercício 1

Seja  $X \sim N(4, 1)$ . Determine:

- a)  $P(4 < X < 5)$
- b)  $P(2 < X < 5)$
- c)  $P(5 < X < 7)$
- d)  $P(X \leq 1)$
- e)  $P(0 \leq X \leq 2)$

### Exercício 2

Seja  $X \sim N(3, 16)$ , ou seja, a variável  $X$  tem distribuição Normal com média  $\mu = 3$  e variância  $\sigma^2 = 16$ . Faça o gráfico da distribuição e determine  $P(3 < X < 8)$ .

### Exercício 3

A estatura média dos alunos da UFRRJ é  $\mu = 1,75m$  e desvio padrão  $\sigma = 0,15m$ . Assumindo-se que a variável estatura ( $X$ ) seja normalmente distribuída, calcule a probabilidade de um aluno aleatoriamente selecionado ter estatura entre  $1,70m$  e  $1,80m$ .

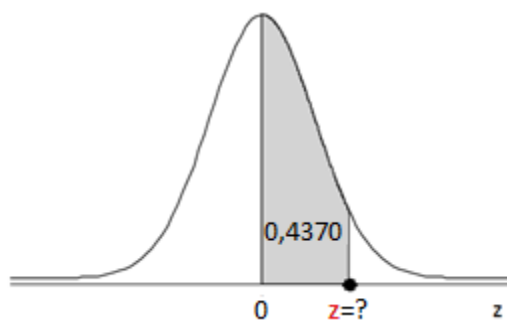
### Mais exemplos sobre o uso da tabela da Distribuição Normal Padrão

#### Exemplo

Sabendo-se que  $Z \sim N(0, 1)$  e usando a tabela da Distribuição Normal Padrão, obter  $z$  tal que  $P(0 < Z < z) = 0,4370$ .

#### Solução

Foi dado  $P(0 < Z < z) = 0,4370$ , ou seja, temos a probabilidade e queremos encontrar o valor de  $z$  tal que a área do gráfico, compreendida entre 0 e  $z$ , é igual a 0,4370. Graficamente, temos:



Observe que este é exatamente o tipo de probabilidade encontrada na tabela da Distribuição Normal Padrão, ou seja,  $P(0 < Z < Z_t)$ . Então, devemos localizar a probabilidade 0,4370 no interior da tabela da Distribuição Normal Padrão:

1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545

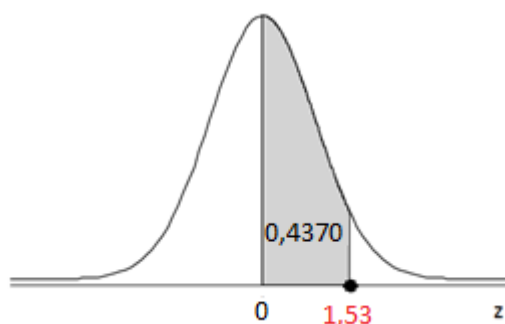
e depois, obtemos, nas bordas da tabela, o valor de  $z$  correspondente a esta probabilidade:

Tabela da distribuição Normal Padrão										
Probabilidades ( $\alpha$ ) da distribuição Normal Padrão $N(0,1)$ para valores do quantil $Z_t$ padronizado de acordo com o seguinte evento: $P(0 < Z < Z_t) = \alpha$ .										
$Z_t$	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545

Portanto, o valor de  $z$  tal que  $P(0 < Z < z) = 0,4370$  é  $z = 1,53$ , ou seja:

$$P(0 < Z < 1,53) = 0,4370.$$

Graficamente, temos:

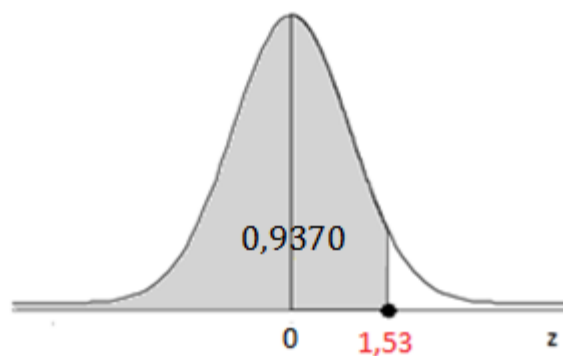


**Comandos no Software R para calcular os quantis da distribuição Normal:**

```
qnorm(0.9370, mean=0, sd=1)
```

**Observação:** Note que usamos 0,9370 ao invés de 0,4370 porque o R considera a probabilidade de  $-\infty$  até  $z$  (e não de 0 até  $z$ ), ou seja, no R usamos:

$$P(Z < z) = P(Z < 0) + P(0 < Z < z) = 0,5 + 0,4370 = 0,9370$$



### Exercício 1

Sabendo-se que  $Z \sim N(0, 1)$  e usando a tabela da Distribuição Normal Padrão, obter  $z$  tal que:

- a)  $P(0 < Z < z) = 0,475$
- b)  $P(-z < Z < 0) = 0,3531$
- c)  $P(-z < Z < 0) = 0,4949$
- d)  $P(-z < Z < z) = 0,95$
- e)  $P(-z < Z < z) = 0,97$
- f)  $P(Z < z) = 0,8212$
- g)  $P(Z < z) = 0,3023$
- h)  $P(Z > z) = 0,9525$
- i)  $P(Z > z) = 0,0749$
- j)  $P(Z < z) = 0,5$

**Observação:** Se acaso a tabela não contiver a probabilidade desejada, utilize a probabilidade mais próxima.

## III - Inferência Estatística

### 1 Introdução à inferência estatística.

#### 1.1 Conceitos básicos. Amostra e população.

Até o momento vimos a *Estatística Descritiva*, que mostra como relatar os dados que temos em mãos. A interpretação do material coletado é feita por meio de tabelas e gráficos e da apresentação de estatísticas como média e desvio-padrão. Por exemplo, se você medir o peso e a altura de 100 estudantes da UFRRJ, saberá apresentar e resumir os dados, ou seja, descrever os resultados que você encontrou nesse grupo de estudantes.

É possível generalizar os resultados obtidos à partir das observações feitas nesses estudantes (uma amostra) para todos os estudantes da UFRRJ (a população). Mas para isso é preciso usar um conjunto de técnicas de Estatística que permitem, com base em uma amostra, fazer *inferência* para a população de onde foi retirada.

#### População e Amostra

**População:** é o conjunto de unidades sobre o qual desejamos informação.

**Amostra:** é qualquer subconjunto de unidades retiradas da população para obter a informação desejada.

A chave para o bom entendimento da Estatística é saber distinguir entre os dados observados (amostra) e a vasta quantidade de dados que poderiam ter sido observados (população). O uso de amostras permite obter respostas para a questão estudada, com *margens de erro* conhecidas.

O termo *população* não se restringe, porém a um conjunto de pessoas, referindo-se, sim, a qualquer conjunto "grande" de unidades que têm algo em comum, como, por exemplo:

- radiografias feitas pelos alunos de uma faculdade em determinado curso;
-

- prontuários de pacientes atendidos pelo SUS durante todo o ano;
- auditorias das contas hospitalares de uma maternidade;
- certidões de óbito registradas numa cidade em determinado período.

### **Parâmetro, estimador e estimativa**

**Parâmetro:** é um valor em geral desconhecido (e, portanto, que precisa ser estimado) que representa determinada característica da população.

São exemplos de parâmetros a média populacional  $\mu$ , a variância populacional  $\sigma^2$  e o desvio padrão populacional  $\sigma$ .

**Estimador:** é uma função dos elementos da amostra. É usado para estimar o parâmetro correspondente, da população de onde a amostra foi retirada.

Assim, por exemplo, a média amostral  $\bar{X}$  é um estimador da média populacional  $\mu$ , a variância amostral  $S^2$  é um estimador da variância populacional  $\sigma^2$  e o desvio padrão amostral  $S$  é um estimador do desvio padrão populacional  $\sigma$ .

**Estimativa:** O valor numérico de um estimador é conhecido como estimativa. Assim  $\bar{x} = 17,8$  é uma estimativa da média populacional  $\mu$ .

Algumas razões para se tomar uma amostra ao invés de usar a população toda são as seguintes:

- custo alto para obter informação da população toda;
- tempo muito longo para obter informação da população toda;
- algumas vezes é impossível, por exemplo, estudo de poluição atmosférica;
- algumas vezes é logicamente impossível, por exemplo, em ensaios destrutivos.

## **1.2 Amostragem aleatória simples: obtenção de uma amostra aleatória**

É compreensível que o estudo de todos os elementos da população possibilita preciso conhecimento das variáveis que estão sendo pesquisadas; todavia, nem

---

sempre é possível obter as informações de todos os elementos da população. Torna-se claro que a representatividade da amostra dependerá de seu tamanho (quanto maior, melhor) e de outras considerações de ordem metodológica. Isto é, o pesquisador procurará acercar-se de cuidados, visando a obtenção de uma amostra que de fato represente "o melhor possível" toda a população.

A amostragem aleatória simples é o processo mais elementar e frequentemente utilizado. Atribui-se a cada elemento da população um número distinto de 1 a  $N$  (tamanho da população). Efetuam-se sucessivos sorteios até completar-se o tamanho da amostra:  $n$ .

## Exemplo

Um dentista quer obter uma amostra de 2% dos quinhentos pacientes de sua clínica para entrevistá-los sobre a qualidade de atendimento da secretária. Para obter uma amostra aleatória de 2% dos quinhentos pacientes, é preciso sortear 10 pacientes.

Isso pode ser feito de maneira mais antiga e mais conhecida (e também mais trabalhosa): atribuem-se números de 1 a 500 a cada um dos pacientes e escrevem-se esses números em pedaços de papel. Coloca-se todos os pedaços de papel em uma caixa, misturando-os bem, e retira-se um papel. O procedimento é repetido até serem retirados 10 papéis referentes a 10 pacientes.

O procedimento pode ser feito mais facilmente utilizando-se um gerador de números aleatórios. Na calculadora isso pode ser feito utilizando a tecla #Ran (em laranja em cima da tecla do ponto "."). Para sortear 10 números entre 1 e 500 digitamos:

500 Shift #Ran =

Para sortear outro número basta teclar "=" novamente. Considera-se apenas a parte inteira do número. Se um número repetir ignore e sorteie outro.

---

### 1.3 Conceito de Distribuições amostrais

No início do curso vimos medidas que caracterizam uma amostra, como, por exemplo, a média, a variância, o desvio padrão, ... Nas seções anteriores vimos os principais modelos de distribuição de probabilidade, particularmente o modelo da distribuição Normal.

Neste capítulo juntam-se os modelos de distribuição de probabilidade e as medidas descritivas obtendo-se as distribuições amostrais dos principais estimadores. O conceito de distribuição amostral é fundamental na inferência estatística.

Considere todas as possíveis amostras de tamanho  $n$  que podem ser extraídas de determinada população. Se para cada uma delas se calcular um valor do estimador (por exemplo da média amostral  $\bar{X}$ ), tem-se uma distribuição amostral desse estimador. Como o estimador é uma variável aleatória, pode-se determinar suas características, isto é, encontrar sua média, variância, desvio-padrão...

### 1.4 Distribuição amostral da média

A distribuição amostral da média refere-se à distribuição gerada pelas estimativas da média, obtidas a partir de todas as amostras, de tamanho  $n$ , retiradas de uma população de referência.

#### Exemplo

Considere uma população fictícia, de tamanho  $N = 3$ ,  $X = \{1, 2, 3\}$ , cuja média é  $\mu = 2$  e a variância é  $\sigma^2 = 2/3$ . Obter a distribuição amostral da média  $\bar{X}$  para amostras de tamanho  $n = 2$  retiradas com reposição dessa população.

As amostras *com reposição* de tamanho  $n = 2$  juntamente com a média amostral são apresentadas a seguir:

---

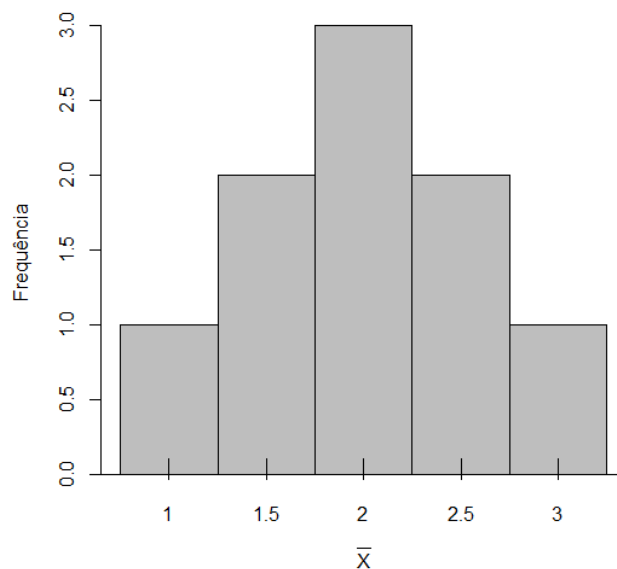


Amostras com reposição	$\bar{X}$
(1, 1)	1,0
(1, 2)	1,5
(1, 3)	2,0
(2, 1)	1,5
(2, 2)	2,0
(2, 3)	2,5
(3, 1)	2,0
(3, 2)	2,5
(3, 3)	3,0

Agrupando as médias comuns e computando suas frequências, obtêm-se os seguintes resultados.

$\bar{X}$	Frequência
1,0	1
1,5	2
2,0	3
2,5	2
3,0	1

O gráfico da distribuição amostral das médias da amostra é apresentado a seguir:



Calculando-se, agora a média e a variância de  $\bar{X}$  para todas as 9 médias amostrais (população de médias amostrais) obtidas, têm-se:

$$\mu_{\bar{X}} = \frac{\sum \bar{X}_i}{9} = \frac{1 + 1,5 + 1,5 + \dots + 3,0}{9} = 2$$

e

$$\sigma_{\bar{X}}^2 = \frac{1}{9} \left[ \sum \bar{X}_i^2 - \frac{(\sum \bar{X}_i)^2}{9} \right] = \frac{1}{9} \left[ (1^2 + 1,5^2 + \dots + 3^2) - \frac{18^2}{9} \right] = \frac{1}{3}$$

### Teorema

Seja  $X$  uma variável aleatória com média  $\mu$  e variância  $\sigma^2$ , e seja  $(X_1, X_2, \dots, X_n)$  uma amostra aleatória de  $X$ . Seja  $\bar{X} = (X_1 + X_2 + \dots + X_n)/n$  a média amostral, então:

$$\mu_{\bar{X}} = E(\bar{X}) = \mu \quad \text{e} \quad \sigma_{\bar{X}}^2 = Var(\bar{X}) = \frac{\sigma^2}{n}.$$

Podemos verificar numericamente este resultado a partir do exemplo anterior:

- Médias:
  - $\mu = 2$  (média populacional)
  - $\mu_{\bar{X}} = 2$  (média das médias amostrais)
- Variâncias:
  - $\sigma^2 = \frac{2}{3}$  (variância populacional)
  - $\sigma_{\bar{X}}^2 = \frac{1}{3}$  (variância das médias amostrais)

Temos que:

$$\frac{\sigma^2}{n} = \frac{\left(\frac{2}{3}\right)}{2} = \frac{1}{3} = \sigma_{\bar{X}}^2$$

ou seja, verificamos numericamente que  $\mu_{\bar{X}} = \mu$  e  $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$ .

### Teorema

Seja  $X \sim N(\mu, \sigma^2)$ , ou seja,  $X$  é uma variável aleatória com distribuição Normal, com média  $\mu$  e variância  $\sigma^2$ , então:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

ou seja, a média amostral ( $\bar{X}$ ) tem distribuição Normal com média  $\mu$  e variância  $\sigma^2/n$ .

## Corolário

Seja  $X \sim N(\mu, \sigma^2)$ , então:

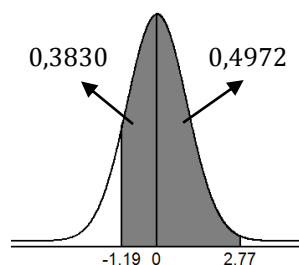
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

## Exemplo

A duração da gravidez humana, da concepção ao parto, varia segundo uma distribuição aproximadamente normal com média  $\mu = 266$  dias e desvio padrão de  $\sigma = 16$  dias. Qual a probabilidade de em uma amostra de  $n = 10$  mulheres grávidas a duração média da gravidez durar entre 260 e 280 dias?

## Solução

$$\begin{aligned} P(260 < \bar{X} < 280) &= P\left(\frac{260 - \mu}{\sigma/\sqrt{n}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{280 - \mu}{\sigma/\sqrt{n}}\right) = P\left(\frac{260 - \mu}{\sigma/\sqrt{n}} < Z < \frac{280 - \mu}{\sigma/\sqrt{n}}\right) \\ &= P\left(\frac{260 - 266}{16/\sqrt{10}} < Z < \frac{280 - 266}{16/\sqrt{10}}\right) = P(-1,19 < Z < 2,77) \\ &= 0,3830 + 0,4972 = \mathbf{0,8802} \end{aligned}$$



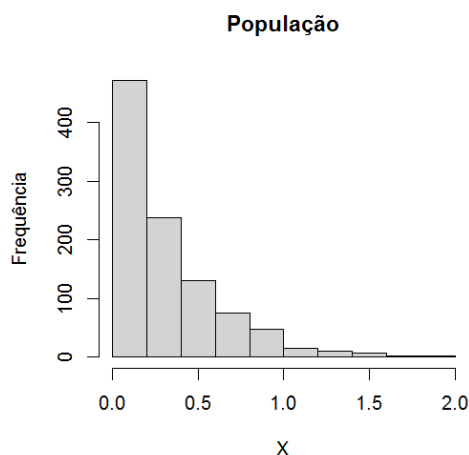
## Teorema Central do Limite

Seja  $(X_1, X_2, \dots, X_n)$  uma amostra de variáveis aleatórias com *distribuição de probabilidades qualquer* (não precisa ser normal), com média  $\mu$  e variância  $\sigma^2$ . A distribuição amostral da média  $\bar{X}$  **aproxima-se, para  $n$  grande ( $n \geq 30$ )**, de uma distribuição normal, com média  $\mu$  e variância  $\sigma^2/n$ , ou seja:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

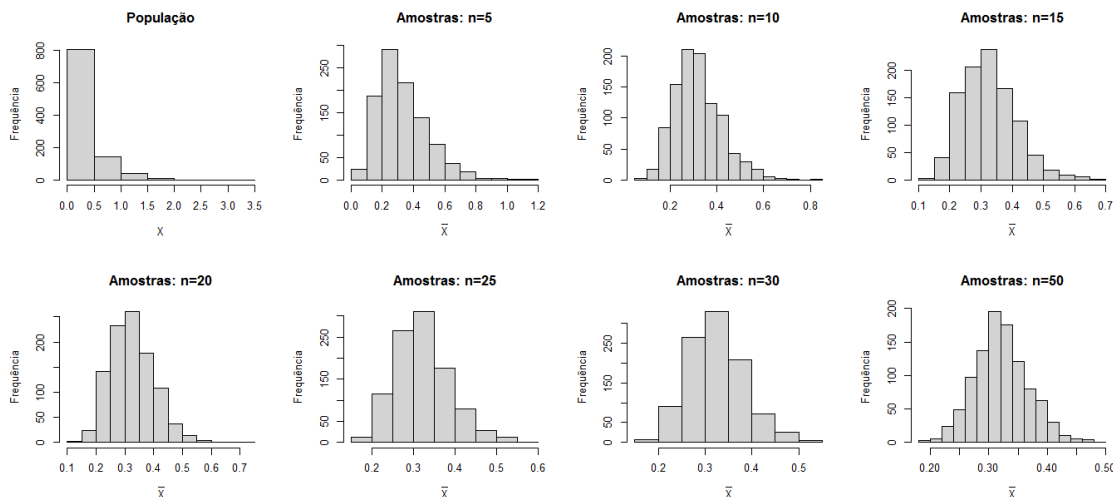
## Exemplo

Para ilustrar o Teorema Central do Limite considere uma população com distribuição Exponencial, cujo histograma é apresentado a seguir:



- São extraídas, desta população, 1000 amostras aleatórias de tamanho  $n = 5$ , outras 1000 amostras aleatórias de tamanho  $n = 10$ , ...,  $n = 15$ , 20, 30, 40 e 50;
- Para cada uma das amostras foi calculada a média amostral  $\bar{X}$ ;
- Foi feito um histograma para o conjunto de todas as médias obtidas de amostras de tamanho  $n = 5$ ; também foi feito um histograma para o conjunto de todas as médias obtidas de amostras de tamanho  $n = 10$ , ..., também foi feito um histograma para o conjunto de todas as médias obtidas de amostras de tamanho  $n = 50$ .

São apresentados, a seguir, o histograma dos dados da população ( $X$ ) e os histogramas das médias amostrais ( $\bar{X}$ ) para as médias obtidas de amostras de tamanho  $n = 5, 10, 15, 20, 30$  e 50:



Observe que, mesmo a população ( $X$ ) não tendo distribuição Normal (histograma muito assimétrico), para amostras de tamanho  $n = 30$  ou maior, a distribuição da média amostral ( $\bar{X}$ ) tem uma distribuição aproximadamente Normal (histograma quase simétrico).

### Corolário (do Teorema Central do Limite)

Seja  $(X_1, X_2, \dots, X_n)$  uma amostra de variáveis aleatórias com *distribuição de probabilidades qualquer* (não precisa ser normal), com média  $\mu$  e variância  $\sigma^2$ , e  $\bar{X} = (X_1 + X_2 + \dots + X_n)/n$ , então, **para  $n$  grande** ( $n \geq 30$ ):

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Utilizando o Teorema Central do Limite podemos calcular probabilidades relacionadas à média amostral independente de conhecermos a distribuição de probabilidades da população da qual a amostra é proveniente.

### Exemplo

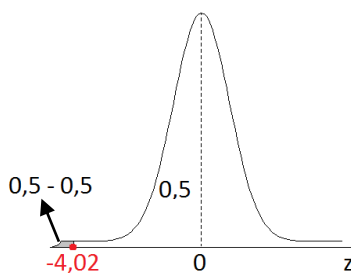
Suponha que numa população de camarões machos adultos a média dos comprimentos seja igual a  $\mu = 27,3 \text{ mm}$  e o desvio padrão seja  $\sigma = 7,8 \text{ mm}$ . Qual a probabilidade de que numa amostra de  $n = 35$  camarões, obtenha-se uma média menor que 22 milímetros (ou seja,  $\bar{X} < 22 \text{ mm}$ )?

## Solução

Deseja-se obter  $P(\bar{X} < 22)$ .

Note que nada foi dito sobre a distribuição de probabilidades dos comprimentos dos camarões, mas como  $n > 30$  ( $n$  grande) podemos utilizar a distribuição Normal para calcular probabilidades sobre a média amostral ( $\bar{X}$ ). Então:

$$\begin{aligned} P(\bar{X} < 22) &= P\left(\frac{\bar{X} - \mu}{(\sigma/\sqrt{n})} < \frac{22 - \mu}{(\sigma/\sqrt{n})}\right) = P\left(Z < \frac{22 - \mu}{(\sigma/\sqrt{n})}\right) = P\left(Z < \frac{22 - 27,3}{(7,8/\sqrt{35})}\right) \\ &= P(Z < -4,02) = 0,5 - 0,5 = 0 \end{aligned}$$



**Observação:** Na tabela da distribuição Normal Padrão, para  $z = 3,9$  temos que  $P(0 < Z < 3,9) = 0,5$ , ou seja,  $z = 3,9$  é um valor tão distante nesta distribuição que a área entre 0 e 3,9 já ocupa toda a metade da área do gráfico. Então, para qualquer valor " $z_{maior}$ " maior do que 3,9, temos  $P(0 < Z < z_{maior}) = 0,5$ .

## 2 Estimação

### 2.1 Conceitos básicos. Estimadores não viciados

Vimos que a Inferência Estatística tem por objetivo fazer generalizações sobre uma população, com base nos dados de uma amostra. Salientamos que dois problemas básicos nesse processo são:

- a) Estimação de parâmetros;
- b) Teste de hipóteses sobre parâmetros.

Lembremos que parâmetros são funções de valores populacionais, por exemplo a média populacional  $\mu$ , enquanto estatísticas são funções de valores amostrais, por exemplo a média amostral  $\bar{X}$ .

### Estimação Pontual

É usada quando a partir da amostra procura-se obter um único valor para "tentar adivinhar" o verdadeiro valor de certo parâmetro populacional, ou seja, obter estimativas a partir dos valores amostrais.

Principais estimadores:

- Média Amostral ( $\bar{X}$ );
- Proporção Amostral ( $\hat{p}$ );
- Variância Amostral ( $S^2$ ).

### Estimadores não viciados

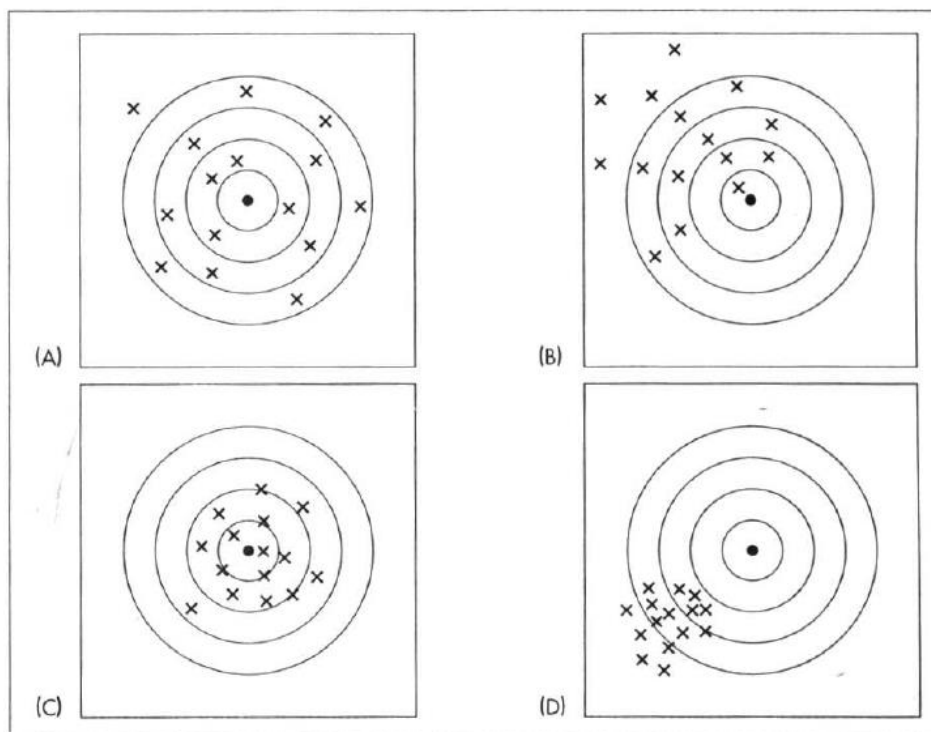
#### Definição

O estimador amostral  $\hat{\theta}$  do parâmetro populacional  $\theta$  é não viciado (ou não viesado) para o parâmetro  $\theta$  se  $E(\hat{\theta}) = \theta$ .

---

## Exemplo

Desejamos comprar um rifle e, após algumas seleções, restaram quatro alternativas, que chamaremos de rifles A, B, C e D. Foi feito um teste com cada rifle, que consistiu em fixá-lo num cavalete, mirar o centro do alvo e disparar 15 tiros. Os resultados estão ilustrados na figura abaixo:



Para analisar qual a melhor arma, podemos fixar critérios. Por exemplo, segundo o critério de "em média acertar o alvo" (não viciado), escolheríamos as armas A e C. Segundo o critério de "não ser muito dispersivo" (variância pequena), a escolha recairia nas armas C e D. Note que a arma C é aquela que reúne as duas propriedades e, segundo esses critérios, seria a melhor arma.

## Exemplo

Considere uma população com  $N$  elementos e a média populacional:

$$\mu = \frac{\sum X_i}{N}.$$

Sejam  $X_1, X_2, \dots, X_n$  uma amostra aleatória de tamanho  $n$  desta população. A média amostral:



$$\bar{X} = \frac{\sum X_i}{n}$$

é um estimador **não viciado** da média populacional  $\mu$  pois  $E(\bar{X}) = \mu$ .

### Demonstração

Cada  $X_i$  tem média  $E(X_i) = \mu$ , pois  $X_1, X_2, \dots, X_n$  são uma amostra aleatória de uma população com média  $\mu$ . Então:

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{\sum X_i}{n}\right) = \frac{1}{n} E\left(\sum X_i\right) = \frac{1}{n} \sum [E(X_i)] = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \left(\underbrace{\mu + \mu + \dots + \mu}_{n \text{ vezes}}\right) \\ &= \frac{1}{n} \cdot n\mu = \mu \end{aligned}$$

### Exemplo

Considere uma população com  $N$  elementos e a variância populacional

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2.$$

Um possível estimador para  $\sigma^2$ , baseado numa amostra aleatória simples de tamanho  $n$  extraída dessa população, pode ser:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Pode-se mostrar, de modo semelhante ao do exemplo anterior (porém mais trabalhoso), que:

$$E(\hat{\sigma}^2) = \left(\frac{n-1}{n}\right) \sigma^2,$$

ou seja:

$$E(\hat{\sigma}^2) \neq \sigma^2.$$

Portanto,  $\hat{\sigma}^2$  é um estimador **viciado** para  $\sigma^2$ .

### Estimador não viciado para $\sigma^2$

Observe que, se multiplicarmos  $\hat{\sigma}^2$  por  $\left(\frac{n}{n-1}\right)$  teremos:

$$E\left(\frac{n}{n-1} \hat{\sigma}^2\right) = \frac{n}{n-1} E(\hat{\sigma}^2) = \frac{n}{n-1} \left(\frac{n-1}{n}\right) \sigma^2 = \sigma^2$$

Então, podemos obter um estimador não viciado para a variância populacional como:

$$S^2 = \frac{n}{n-1} \cdot \hat{\sigma}^2 = \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

### Observação

O motivo pelo qual dividimos por  $(n-1)$  na variância amostral ao invés de dividir por  $n$  é para que o estimador da variância amostral seja não viciado, ou seja:

$$E(S^2) = E\left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right] = \sigma^2.$$

Assim, se retirarmos várias amostras da população e calcularmos a variância amostral ( $S^2$ ) de cada uma das amostras, em média, estaremos "acertando" a verdadeira variância da população ( $\sigma^2$ ).

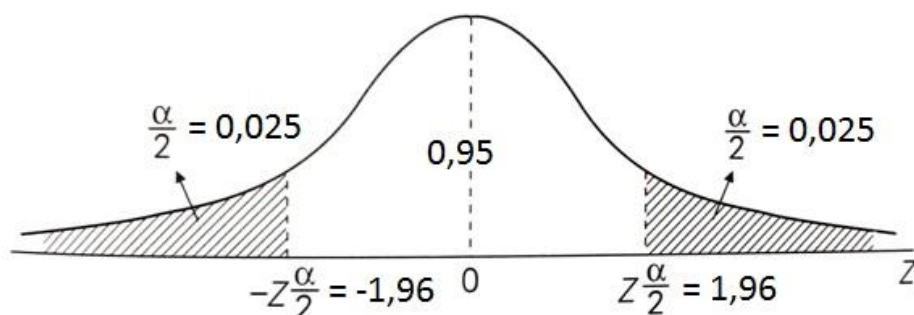
## 2.2 Intervalo de confiança para média de uma população Normal com variância populacional conhecida.

Os estimadores pontuais especificam um único valor para o parâmetro. Esse procedimento não permite julgar qual a possível magnitude do erro que se está cometendo. Assim, surge a idéia de construir intervalos de confiança, que serão baseados na distribuição amostral do estimador pontual.

Foi visto anteriormente que se  $X \sim N(\mu, \sigma^2)$ , então  $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ . Assim, a variável padronizada para  $\bar{X}$  será:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Foi visto também  $P(-1,96 \leq Z \leq 1,96) = 0,95$ .



Substituindo  $Z$  por  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ , temos:

$$P\left(-1,96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1,96\right) = 0,95.$$

Resolvendo-se as inequações para  $\mu$ , temos:

$$P\left(\bar{X} - 1,96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1,96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0,95.$$

Assim, o intervalo de confiança de 95% para a média populacional ( $\mu$ ) é dado por:

$$IC_{(0,95)}(\mu) = \left[ \bar{x} - 1,96 \cdot \frac{\sigma}{\sqrt{n}} ; \bar{x} + 1,96 \cdot \frac{\sigma}{\sqrt{n}} \right].$$

## Interpretação

Com 95% de confiança (ou probabilidade) o intervalo

$$\left[ \bar{x} - 1,96 \cdot \frac{\sigma}{\sqrt{n}} ; \bar{x} + 1,96 \cdot \frac{\sigma}{\sqrt{n}} \right]$$

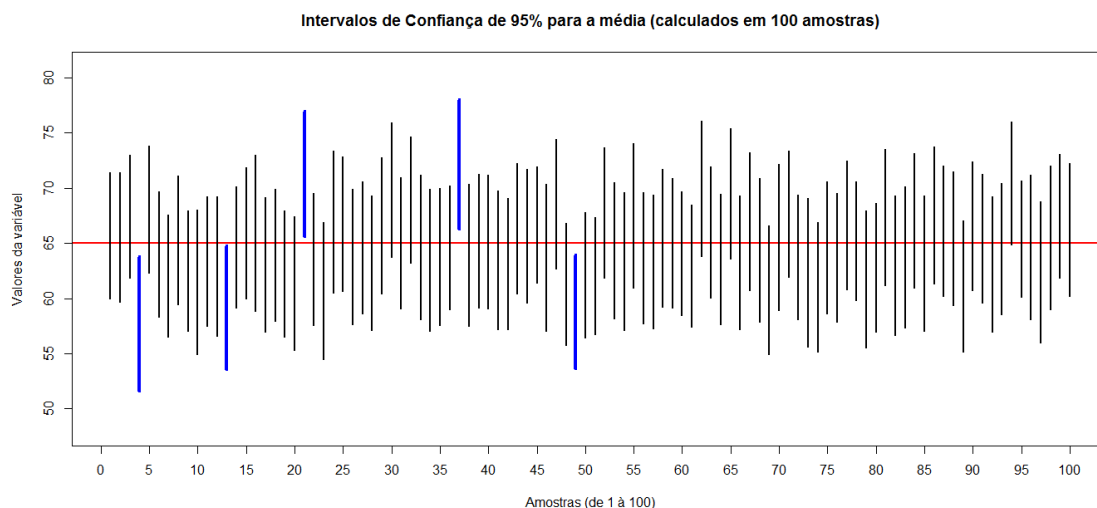
contém o verdadeiro valor da média populacional  $\mu$ .

## Exemplo

Suponhamos uma população cujo verdadeiro valor da média é  $\mu = 65$  (lembre-se que na prática  $\mu$  é desconhecido). Se retirarmos uma amostra dessa população e, a partir dessa amostra, calcularmos o intervalo de confiança de 95% para a média populacional, este intervalo tem 95% de chances (probabilidade) de conter a verdadeira média populacional  $\mu = 65$ .

Dessa forma, se retirarmos 100 amostras desta população e, para cada amostra, obtermos o respectivo intervalo de confiança de 95% para a média, espera-se

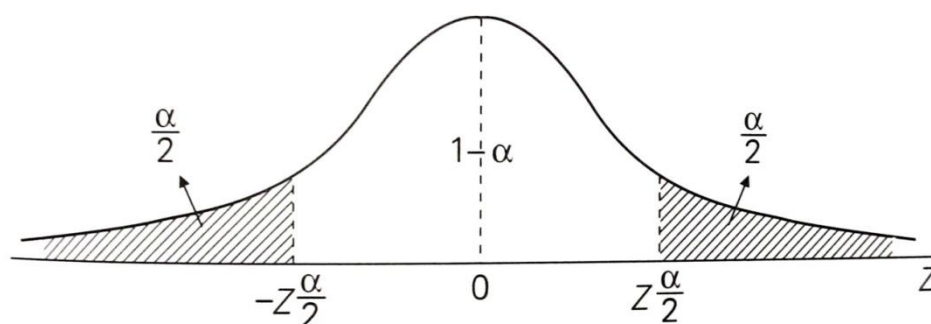
que aproximadamente 95 desses intervalos contenham a verdadeira média populacional  $\mu = 65$ . Observe na figura abaixo que para 100 amostras retiradas desta população, 95 dos 100 intervalos de confiança contém o valor da verdadeira média populacional ( $\mu = 65$ ), enquanto apenas 5 intervalos (destacados em azul) não contém a verdadeira média populacional.



Fixando-se um nível de confiança qualquer:  $1 - \alpha$  temos:

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha.$$

Graficamente, temos:



Substituindo  $Z$  por  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ , temos:

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

Resolvendo-se as inequações para  $\mu$ , temos:

$$P\left(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Assim, o intervalo de confiança de  $100(1 - \alpha)\%$  para a média populacional ( $\mu$ ) é dado por:

$$IC_{(1-\alpha)}(\mu) = \left[ \bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} ; \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right].$$

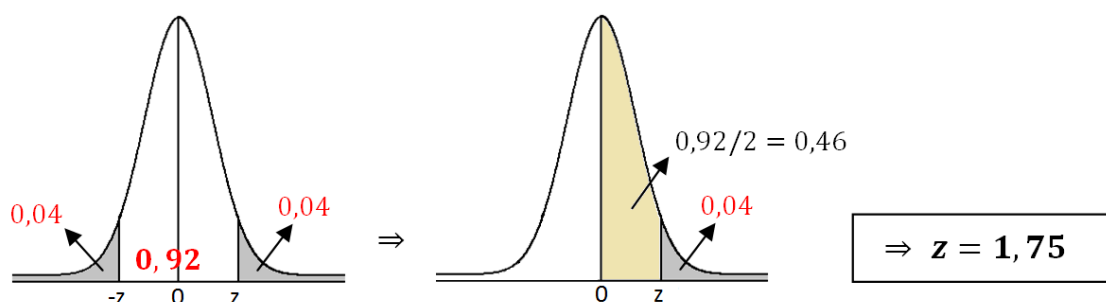
### Exemplo

Sabe-se que o consumo mensal *per capita* de um determinado produto tem distribuição normal com desvio padrão de 2 kg. Foi realizada uma pesquisa de mercado, tomando-se uma amostra de 25 indivíduos e obteve-se um consumo médio *per capita*  $\bar{x} = 7,2$  kg para esta amostra. Pede-se:

- Obtenha o valor de  $z$  tal que  $P(-z \leq Z \leq z) = 0,92$ ;
- Estabeleça um intervalo de 92% de confiança para o consumo médio *per capita* deste produto. Interprete.

### Solução

a)



$Z_t$	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
⋮										
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706

b) Temos  $(1 - \alpha) = 0,92$ , e vimos no item (a) que  $z_{\alpha/2} = 1,75$ . Substituindo em:

$$IC_{(1-\alpha)}(\mu) = \left[ \bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} ; \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

segue que:

$$\begin{aligned} IC_{0,92}(\mu) &= \left[ \bar{x} - 1,75 \frac{\sigma}{\sqrt{n}} , \bar{x} + 1,75 \frac{\sigma}{\sqrt{n}} \right] \\ &= \left[ 7,2 - 1,75 \frac{2}{\sqrt{25}} , 7,2 + 1,75 \frac{2}{\sqrt{25}} \right] \\ &= [6,50 , 7,90] \end{aligned}$$

**Interpretação:** O intervalo que vai de 6,50 kg até 7,90 kg contém o consumo médio *per capita* deste produto (de toda a população) com 92% de confiança (ou probabilidade).

**Comandos no Software R para calcular o Intervalo de Confiança:**

```
require(asbio) #Precisa instalar o pacote asbio
ci.mu.z(conf=0.92, sigma=2, xbar=7.2, n=25, summarized=TRUE)
```

## 2.3 Determinação do tamanho de uma amostra

Podemos reescrever a fórmula do intervalo de confiança apresentado na seção anterior como:

$$IC_{(1-\alpha)}(\mu) = \bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

ou, simplesmente

$$IC_{(1-\alpha)}(\mu) = \bar{X} \pm e$$

em que  $e = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$  é o erro de estimação.

Surge, então, a seguinte questão: "qual deve ser o tamanho da amostra ( $n$ ) para se ter determinada precisão na estimação da média populacional?", ou seja, qual deve ser o  $n$  para que se tenha, no máximo, um determinado erro  $e$ ?

Isolando o  $n$  na equação  $e = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ , tem-se:

$$n = \left( \frac{\sigma \cdot z_{\alpha/2}}{e} \right)^2.$$

### Exemplo

Sabe-se que o consumo mensal *per capita* de um determinado produto tem distribuição normal com desvio padrão de 2 kg. Deseja-se realizar uma pesquisa de mercado para estimar o consumo médio *per capita* deste produto. Qual deve ser o tamanho da amostra ( $n$ ) para que se tenha um erro ( $e$ ) de no máximo 0,5 kg, com um nível de confiança de 95%?

### Solução

Sabe-se que  $\sigma = 2$  e portanto  $\sigma^2 = 4$ . Além disso foi fixado um nível de confiança  $1 - \alpha = 0,95$ , e, portanto  $z_{\alpha/2} = 1,96$ . Como se deseja um erro de estimação ( $e$ ) de no máximo 0,5 kg, segue que:

$$n = \frac{\sigma^2 z_{\alpha/2}^2}{e^2} = \frac{4 \times (1,96)^2}{(0,5)^2} \approx 62$$

ou seja, para que se tenha um erro de no máximo 0,5 kg para mais ou para menos na estimação da média populacional é necessário que a amostra seja composta por 62 indivíduos.

## 2.4 Intervalo de confiança para a média de uma população Normal com variância populacional desconhecida

A pressuposição de normalidade para a média amostral ( $\bar{X}$ ) é garantida para amostras grandes ( $n \geq 30$ ), sendo que para amostras pequenas ( $n < 30$ ) esta pressuposição é válida apenas se sua população for normalmente distribuída e  $\sigma$  for conhecido.

Nesta seção será considerado o caso em que a amostra é pequena ( $n < 30$ ) e a população é normalmente distribuída, porém  $\sigma$  é desconhecido. Neste caso, o desvio padrão populacional ( $\sigma$ ) deve ser substituído pelo desvio padrão amostral ( $S$ ) e a distribuição normal ( $Z$ ) deve ser substituída pela distribuição  $t$  de *Student*.

### Distribuição $t$ de *Student*

Seja  $X_1, X_2, \dots, X_n$  uma amostra aleatória proveniente de uma distribuição normal, com média  $\mu$  com variância  $\sigma^2$  desconhecidas. A variável aleatória

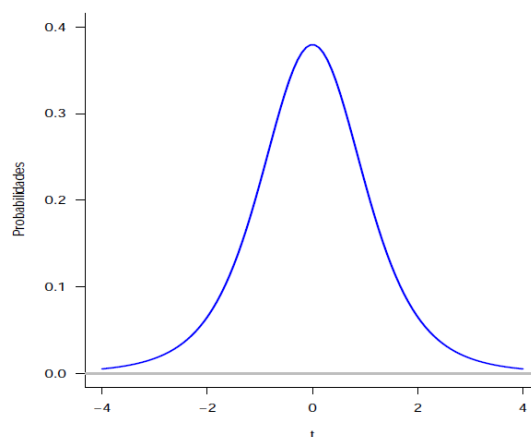
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

tem distribuição  $t$ , com  $n - 1$  graus de liberdade.

A função densidade de probabilidade da distribuição  $t$  de *Student* é

$$f(x) = \frac{\Gamma\left(\frac{\phi+1}{2}\right)}{\sqrt{\pi\phi} \Gamma\left(\frac{\phi}{2}\right)} \cdot \frac{1}{\left(1 + \frac{x^2}{\phi}\right)^{\frac{\phi+1}{2}}}, \quad -\infty < x < \infty,$$

em que  $\phi > 0$  é o número de graus de liberdade.



**Figura.** Gráfico da função densidade de probabilidade da distribuição  $t$  de *Student*. Quando (na prática)  $n > 30$  a distribuição  $t$  tende para a normal padrão.

### Tabela da distribuição $t$ de *Student*

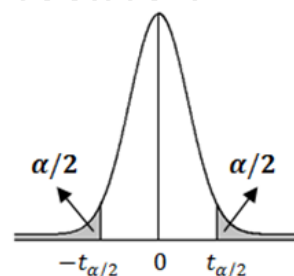
Neste material os quantis da distribuição  $t$  de *Student* são apresentados em uma tabela (bilateral) em função do número de graus de liberdade ( $\phi$ ) e da probabilidade (bilateral)  $\alpha$  tal que  $P(t \leq -t_{\alpha/2}) + P(t \geq t_{\alpha/2}) = \alpha$ .



## Apêndice 2. Tabela (Bilateral) da distribuição $t$ de Student

Quantis da distribuição  $t$  de Student, com  $\phi$  graus de liberdade, e probabilidades  $\alpha$ , de acordo com o seguinte evento:

$$P(t \leq -t_{\alpha/2}) + P(t \geq t_{\alpha/2}) = \alpha$$

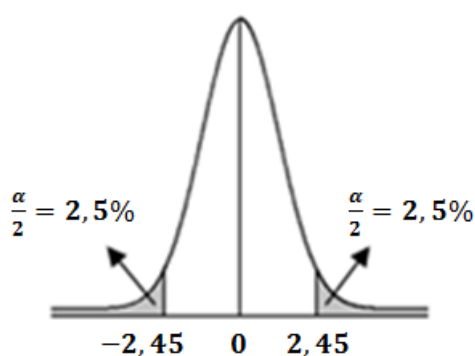


Graus de Liberdade	$\alpha$					
	10%	5%	2%	1%	0,5%	0,1%
1	6,31	12,71	31,82	63,66	127,32	636,62
2	2,92	4,30	6,96	9,92	14,09	31,60
3	2,35	3,18	4,54	5,84	7,45	12,92
4	2,13	2,78	3,75	4,60	5,60	8,61
5	2,02	2,57	3,36	4,03	4,77	6,87
6	1,94	2,45	3,14	3,71	4,32	5,96
7	1,89	2,36	3,00	3,50	4,03	5,41

Por exemplo, para  $\phi = 6$  graus de liberdade, se considerarmos  $\alpha = 5\%$  de probabilidade, então o quantil  $t_{\alpha/2} = t_{0,025} = 2,45$  é tal que:

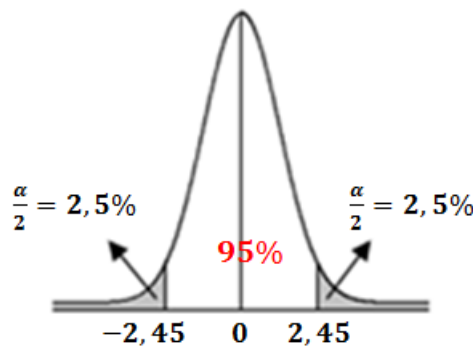
$$P(t \leq -2,45) + P(t \geq 2,45) = 5\%$$

Graus de Liberdade	$\alpha$					
	10%	5%	2%	1%	0,5%	0,1%
1	6,31	12,71	31,82	63,66	127,32	636,62
2	2,92	4,30	6,96	9,92	14,09	31,60
3	2,35	3,18	4,54	5,84	7,45	12,92
4	2,13	2,78	3,75	4,60	5,60	8,61
5	2,02	2,57	3,36	4,03	4,77	6,87
6	1,94	2,45	3,14	3,71	4,32	5,96
7	1,89	2,36	3,00	3,50	4,03	5,41



ou, ainda,  $t_{0,025} = 2,45$  é o quantil da distribuição  $t$  de *Student* tal que:

$$P(-2,45 \leq t \leq 2,45) = 95\%$$



### Intervalo de Confiança

O intervalo de confiança para a média ( $\mu$ ) de uma população com variância populacional ( $\sigma^2$ ) desconhecida é dado por:

$$IC_{(1-\alpha)}(\mu) = \left[ \bar{X} - t_{\alpha/2} \cdot \frac{S}{\sqrt{n}} , \quad \bar{X} + t_{\alpha/2} \cdot \frac{S}{\sqrt{n}} \right],$$

em que  $t_{\alpha/2}$  é o quantil da distribuição  $t$  de *Student* com  $n - 1$  graus de liberdade.

### Exemplo (Levine, et al. 2012, pg. 271)

O Departamento de Transportes dos EUA exige que fabricantes de pneus forneçam informações sobre o desempenho dos pneus na banda lateral, de modo a que os potenciais consumidores possam ser mais bem informados ao tomar uma decisão de compra. Uma medida de grande importância do desempenho do pneu é o índice de desgaste da banda de rodagem, que indica a resistência do pneu em relação ao desgaste da banda de rodagem comparada a um pneu graduado com uma base de 100. Isso significa que um pneu com graduação de 200 deve durar duas vezes mais, em média, do que um pneu graduado com base de 100. Uma organização de defesa do consumidor deseja estimar o verdadeiro índice de desgaste da banda de rodagem de uma determinada marca de pneus, que declara "graduação 200" na banda lateral de seu pneu. Uma amostra aleatória de tamanho  $n = 18$  indica uma média aritmética

$\bar{x} = 195,3$  para o índice de desgaste de banda de rodagem, com desvio padrão da amostra  $s = 21,4$ .

Pressupondo que a população dos índices de desgaste de banda de rodagem seja distribuída conforme uma distribuição normal, construa um intervalo de confiança de 95% para a média do índice de desgaste das bandas de rodagem de todos os pneus produzidos por esse fabricante (de toda a população de pneus).

\* LEVINE et. al. **Estatística Teoria e Aplicações: Usando o Microsoft Excel em Português**. 6 ed. Rio de Janeiro: LTC, 2012.

### Solução

Temos:  $n = 18$ ,  $\bar{x} = 195,3$  e  $s = 21,4$ . Para construir o IC de 95% precisamos encontrar  $t_{\alpha/2} = t_{0,025}$  com  $n - 1 = 17$  graus de liberdade. Como a tabela é bilateral devemos procurar diretamente em  $\alpha = 5\%$ :

Graus de Liberdade	$\alpha$					
	10%	5%	2%	1%	0,5%	0,1%
1	6,31	12,71	31,82	63,66	127,32	636,62
2	2,92	4,30	6,96	9,92	14,09	31,60
3	2,35	3,18	4,54	5,84	7,45	12,92
⋮	⋮	⋮	⋮	⋮	⋮	⋮
16	1,75	2,12	2,58	2,92	3,25	4,01
17	1,74	2,11	2,57	2,90	3,22	3,97
18	1,73	2,10	2,55	2,88	3,20	3,92

Assim,  $t_{0,025} = 2,11$ . Portanto,

$$\begin{aligned}
 IC_{0,95}(\mu) &= \left[ \bar{X} - 2,11 \cdot \frac{S}{\sqrt{n}} , \bar{X} + 2,11 \cdot \frac{S}{\sqrt{n}} \right] \\
 &= \left[ 195,3 - 2,11 \cdot \frac{21,4}{\sqrt{18}} , 195,3 + 2,11 \cdot \frac{21,4}{\sqrt{18}} \right] \\
 &= [184,66 , 205,94 ]
 \end{aligned}$$

**Interpretação:** O intervalo que vai de 184,66 até 205,94 contém a "graduação" média de todos os pneus produzidos por esse fabricante (ou seja, de toda a população de pneus) com 95% de confiança.

## Pergunta

Com base no resultado encontrado no exemplo anterior, você acredita que a organização de defesa do consumidor deveria acusar o fabricante de fabricar pneus que não atendem às informações de desempenho apresentadas na banda lateral do pneu?

## Resposta

Como o intervalo de confiança de 95% é [184,66 , 205,94 ], então temos 95% de confiança de que este intervalo contém a verdadeira "graduação" média  $\mu$  (de toda esta população de pneus), e, como a "graduação 200" está dentro deste intervalo então **não existe motivo** para suspeitar que a verdadeira "graduação" média, de toda esta população de pneus, seja diferente de 200. Assim, a organização de defesa do consumidor **não deve** acusar o fabricante de fabricar pneus que não atendem às informações de desempenho apresentadas na banda lateral do pneu.

### Comandos no Software R para calcular o Intervalo de Confiança:

```
#Carregando o pacote BSDA
library(BSDA) #Precisa instalar o pacote antes

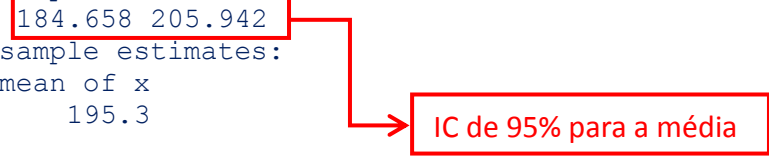
#Intervalo de confiança:
tsum.test(mean.x=195.3, s.x=21.4, n.x=18, conf.level = 0.95,
           alternative="two.sided")

#Saída do R:

One-sample t-Test

data: Summarized x
t = 38.719, df = 17, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
184.658 205.942
sample estimates:
mean of x
195.3

Warning message:
In tsum.test(mean.x = 195.3, s.x = 21.4, n.x = 18, alternative =
"two.sided") :
argument 'var.equal' ignored for one-sample test.
```



## Obtenção do intervalo de confiança no R usando dados brutos

Considerando o exemplo anterior (“graduação de pneus”), suponha que ao invés das estatísticas descritivas (média, desvio padrão e tamanho da amostra) tivéssemos os dados brutos das graduações dos 18 pneus da amostra:

196.9	207.2	168.9	193.2	223.8	218.9
179.2	192.9	220.3	207.5	210.2	183.0
190.0	199.8	206.6	133.3	199.9	183.2

Neste caso, poderíamos obter o intervalo de confiança diretamente no R usando o comando `t.test()`.

### Comandos no Software R para calcular o Intervalo de Confiança:

```
#Entrando com os dados brutos (não foram apresentados no exemplo):
dados <- c(196.9, 179.2, 190.0, 207.2, 192.9, 199.8, 168.9, 220.3,
206.6, 193.2, 207.5, 133.3, 223.8, 210.2, 199.9, 218.9, 183.0,
183.2)

#Medidas descritivas (média, desvio padrão e tamanho da amostra):
mean(dados);sd(dados); length(dados)

#Saída do R (medidas descritivas):
[1] 195.2667
[1] 21.43998
[1] 18

#Observe que se a média e o desvio padrão forem arredondados para
uma casa decimal seus valores ficariam, respectivamente, 195.3 e
21.4, como no exemplo anterior.

#Intervalo de confiança:
t.test(dados, conf.level=0.95)

#Saída do R:
One Sample t-test

data: dados
t = 38.64, df = 17, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
184.6048 205.9285
sample estimates:
mean of x
195.2667
```

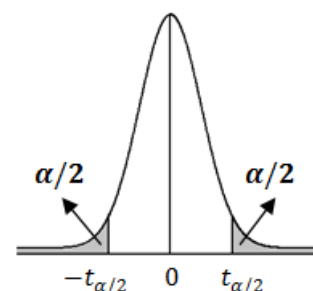
IC de 95% para a média



## Apêndice 2. Tabela (Bilateral) da distribuição $t$ de Student

Quantis da distribuição  $t$  de Student, com  $\phi$  graus de liberdade, e probabilidades  $\alpha$ , de acordo com o seguinte evento:

$$P(t \leq -t_{\alpha/2}) + P(t \geq t_{\alpha/2}) = \alpha$$



Graus de Liberdade	$\alpha$					
	10%	5%	2%	1%	0,5%	0,1%
1	6,31	12,71	31,82	63,66	127,32	636,62
2	2,92	4,30	6,96	9,92	14,09	31,60
3	2,35	3,18	4,54	5,84	7,45	12,92
4	2,13	2,78	3,75	4,60	5,60	8,61
5	2,02	2,57	3,36	4,03	4,77	6,87
6	1,94	2,45	3,14	3,71	4,32	5,96
7	1,89	2,36	3,00	3,50	4,03	5,41
8	1,86	2,31	2,90	3,36	3,83	5,04
9	1,83	2,26	2,82	3,25	3,69	4,78
10	1,81	2,23	2,76	3,17	3,58	4,59
11	1,80	2,20	2,72	3,11	3,50	4,44
12	1,78	2,18	2,68	3,05	3,43	4,32
13	1,77	2,16	2,65	3,01	3,37	4,22
14	1,76	2,14	2,62	2,98	3,33	4,14
15	1,75	2,13	2,60	2,95	3,29	4,07
16	1,75	2,12	2,58	2,92	3,25	4,01
17	1,74	2,11	2,57	2,90	3,22	3,97
18	1,73	2,10	2,55	2,88	3,20	3,92
19	1,73	2,09	2,54	2,86	3,17	3,88
20	1,72	2,09	2,53	2,85	3,15	3,85
21	1,72	2,08	2,52	2,83	3,14	3,82
22	1,72	2,07	2,51	2,82	3,12	3,79
23	1,71	2,07	2,50	2,81	3,10	3,77
24	1,71	2,06	2,49	2,80	3,09	3,75
25	1,71	2,06	2,49	2,79	3,08	3,73
26	1,71	2,06	2,48	2,78	3,07	3,71
27	1,70	2,05	2,47	2,77	3,06	3,69
28	1,70	2,05	2,47	2,76	3,05	3,67
29	1,70	2,05	2,46	2,76	3,04	3,66
30	1,70	2,04	2,46	2,75	3,03	3,65
40	1,68	2,02	2,42	2,70	2,97	3,55
60	1,67	2,00	2,39	2,66	2,91	3,46
120	1,66	1,98	2,36	2,62	2,86	3,37
$\infty$	1,65	1,96	2,33	2,59	2,82	3,31

## Bibliografia

BUSSAB, W. O., MORETTIN, P. A. **Estatística Básica**. 5. ed. São Paulo: Saraiva, 2002.

COSTA, S. C. **Estatística Aplicada à Veterinária**. Londrina: UEL, [ca. 2012]. (Apostila).

FERREIRA, D. F. **Estatística Básica**. Lavras: Editora UFLA, 2005.

FONSECA, J. S., MARTINS, G. A. **Curso de estatística**. 6. ed. São Paulo: Atlas, 1996.

LEVINE et. al. **Estatística Teoria e Aplicações: Usando o Microsoft Excel em Português**. 6 ed. Rio de Janeiro: LTC, 2012.

MONTGOMERY, D. C., RUNGER, G. C. **Estatística Aplicada e Probabilidade para Engenheiros**. 5. ed. Rio de Janeiro: LTC, 2012.

VIEIRA, S. **Introdução à Bioestatística**. 3. ed. rev. Rio de Janeiro: Campus, 1998.

---