# Capítulo 1

# Oganização e Apresentação de Dados

## 1.1 Tabelas e Gráficos

Uma vez os dados observados, sejam de uma amostra ou de uma população, passamos a interpretá-los de acordo com a finalidade da pesquisa. Para isso, necessitamos de algumas ferramentas, como tabelas e gráficos, que permitam organizar, resumir e apresentar estes dados. E esse tipo de tratamento aos dados é chamado de **Estatística Descritiva**.

## 1.2 Tabela

**Def.: Tabela** é usada para organizar e apresentar um conjunto de observações.

A apresentação de tabelas em um relatório é regida por normas específicas elaboradas pelo Instituto Brasileiro de Geografia e Estatística (IBGE) e adotadas pela Associação Brasileira de Normas Técnicas (ABNT). Toda tabela deve ser auto explicativa sendo necessário:

- **Título:** deve responder às perguntas:O quê ?, Quando ?, Onde ?, localizado no topo da tabela.
- **Cabeçalho:** específica o conteúdo das colunas.
- **Corpo:** contém as informações dos dados coletados.

- **Rodapé:** reservado para a fonte dos dados e observações.

Tabela 1.1: *Produção de automóveis - Brasil - 2010 / 2014* → **Título**

| Anos | Quantidade | → **Cabeçalho** |
|---|---|---|
| 2010 | 37.989 | |
| 2011 | 41.223 | |
| 2012 | 48.098 | → **Corpo** |
| 2013 | 54.876 | |
| 2014 | 56.999 | |

Fonte: Dados Fictícios. → **Rodapé**

Podemos trabalhar com tabelas que representam as variáveis qualitativas (séries estatísticas) e as variáveis quantitativas (distribuição de frequência).

**Def.: Séries estatísticas** são as tabelas que apresentam a distribuição de conjuntos de dados estatísticos, em função da época, do local ou da espécie.

### 1.2.1 Série Histórica, Cronológica ou Temporais

Descreve os valores de uma variável qualitativa discriminados segundo intervalos de tempo ou época.

Tabela 1.2: *Rendimento Diário da Poupança - Brasil - Julho/17*

| Dias | Rendimento% |
|---|---|
| 01/07 | 0,6822 |
| 02/07 | 0,6651 |
| 03/07 | 0,6702 |
| 04/07 | 0,6638 |
| 05/07 | 0,6573 |
| 06/07 | 0,6422 |

Fonte:Banco Central

### 1.2.2 Série Geográfica, Territorial ou de Localização

Descreve os valores da variável qualitativa discriminados segundo localidades.

Tabela 1.3: *Temperaturas de cidades do Rio de Janeiro, Brasil - 27/07/17*

| Cidades | Min. | Max. |
|---|---|---|
| Rio de Janeiro | 13 | 26 |
| Petrópoles | 05 | 15 |
| Niteroi | 10 | 22 |
| Nova Iguaçu | 08 | 23 |
| Itaguai | 09 | 21 |

Fonte: Dados fictícios.

### 1.2.3 Série Categórica ou Específica

Descreve os valores da variável qualitativa discriminados segundo alguma especificação ou categoria.

Tabela 1.4: *Percentual das marcas de carros mais vendidas no Rio de Janeiro/2017.*

| Marca | Percentual % |
|---|---|
| Fiat | 30 |
| Wolksvagem | 20 |
| Ford | 10 |
| Honda | 23 |
| Toyota | 27 |

Fonte: Dados fictícios.

### 1.2.4 Tabelas de Distribuição de Frequências

Quando lidamos com grandes conjuntos de dados, formados por valores de uma variável quantitativa, podemos obter boa vizualização e todas as informações necessárias, agrupando esses dados em um certo número de classes ou intervalos.

Um conjunto de dados sem qualquer tipo de tratamento é chamado de **Dados Brutos**. Quando o conjunto de dados é ordenado (crescente ou decrescente) temos um **Rol**.

Para construirmos uma distribuição de frequências devemos seguir os seguintes passos:

**1- Classes:** são intervalos de variação da variável em estudo. K = número de classes.

$$K = 1 + 3,3logn \quad \Rightarrow \quad \text{Regra de Sturges}$$

ou , $K = \sqrt{n}$, onde $n$ = tamanho da amostra.

**2- Amplitude Total de Distribuição:** é a diferença entre o maior e o menor elemento do conjunto.

$$A_T = X_{max} - X_{min}.$$

**3- Amplitude de classe:** é a medida do intervalo que define a classe.

$$A_c = \frac{\text{Amplitude Total}}{\text{Número de classes}} = \frac{A_T}{K}.$$

**4- Frequência absoluta ($f_i$) :** é o número de observações correspondentes a cada classe.

## COMPLEMENTOS DA DISTRIBUIÇÃO DE FREQUÊNCIAS

**5- Ponto médio da classe:** é o ponto que divide a classe em duas partes iguais, sendo indicado por $x_i$.

**6- Frequência relativa ($fr_i$):** são os valores das razões entre as frequências absolutas e a frequência total, isto é:

$$fr_i = \frac{f_i}{\sum_{i=1}^{n} f_i}.$$

**7- Frequência acumulada ($fac_i$):** é o total de frequências de todos os valores inferiores ao limite superior do intervalo de uma dada classe:

$$fac_i = \sum_{j=1}^{i} f_j = f_1 + f_2 + ... + f_i$$

**Exemplo:** Com base no rol das alturas (cm) de 40 alunos da UFRRJ,Out/2017, construir a tabela de distribuição de frequências.

150,151,151,153,154,154,154,155,156,157,

157,157,157,158,159,159,160,160,160,161,

161,161,161,161,162,162,162,163,163,165,

165,165,166,166,168,168,168,170,172,173.

**Solução:**

1- Número de classes : $K = \sqrt{40} \approx 6$.

2- Amplitude Total: $A_T = 173 - 150 = 23$.

3- Amplitude da classe: $A_c = 23/6 = 3,8 \approx 4$.

Baseado nestes valores temos a seguinte distribuição de frequências:

Tabela 1.5: *Altura (cm) de 40 alunos da UFRRJ, Outubro/2017.*

| Classes | $f_i$ | $x_i$ | $fr_i(\%)$ | $fac_i$ |
|---|---|---|---|---|
| 150⊢154 | 4 | 152 | 10 | 4 |
| 154⊢158 | 9 | 156 | 22,5 | 13 |
| 158⊢162 | 11 | 160 | 27,5 | 24 |
| 162⊢166 | 8 | 164 | 20 | 32 |
| 166⊢170 | 5 | 168 | 12,5 | 37 |
| 170⊢174 | 3 | 172 | 7,5 | 40 |
| Total | 40 | — | 100 | — |

Fonte: Dados fictícios.

## 1.3 Gráficos

As tabelas tem a função de resumir um conjunto de dados com grande precisão. Porém, são os gráficos que fornecem uma visualização mais rápida do comportamento dos dados em estudo.

**Def.:** **Gráfico** é uma alternativa para apresentação de dados estatísticos, cujo objetivo é o de produzir uma visualizção mais rápida das informações contidas em uma tabela.

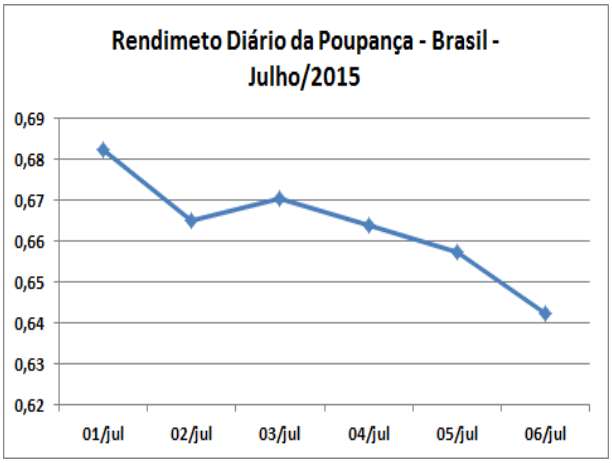
Estudaremos, nesta seção, o gráfico de linhas, o gráfico de barras e de colunas e o gráfico de setores que apresentam os valores das séries estatística. Os valores das tabelas de distribuição de frequências são apresentados pelo histograma e pelo

polígono de frequências. E veremos outra alternativa de resumir dados que é o Ramo-e-Folhas e a apresentação do Box-Plot.

### 1.3.1 Gráfico de Linhas

Este gráfico representa as séries temporais utilizando linhas poligonais. Sua contrução utiliza o plano cartesiano, onde no eixo $X$ colocamos a variavél em estudo e no eixo $Y$ colocamos os valores numéricos, respeitando uma escala numérica previamente estabelecida.

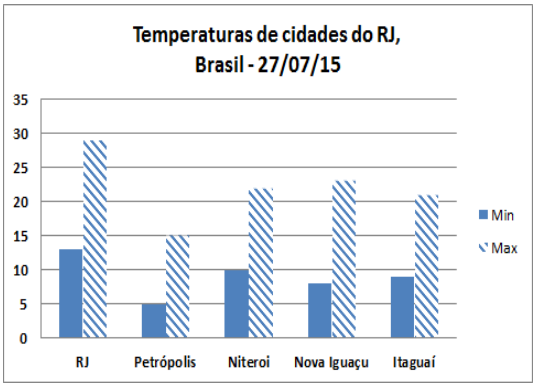
A gráfico de linhas a seguir representa a série estatística temporal (Tabela 1.2) apresentada na seção anterior.




### 1.3.2 Gráfico de Barras e Colunas

Estes gráficos podem ser utilizados para representar todas as séries estatísticas, por meio de retângulos dispostos horizontalmente (Grafico de Barras) e verticalmente (gráfico de colunas). A contrução do gráfico de colunas utiliza o plano cartesiano, onde no eixo $X$ colocamos a variavél em estudo e no eixo $Y$ colocamos os valores numéricos, respeitando uma escala numérica previamente estabelecida. No caso de gráfico em Barras no eixo $Y$ colocamos a variavél em estudo e no eixo $X$ colocamos os valores numéricos, respeitando uma escala numérica previamente estabelecida.

O gráfico de colunas abaixo representa a série estatística geográfica (Tabela 1.3) apresentada na seção anterior.



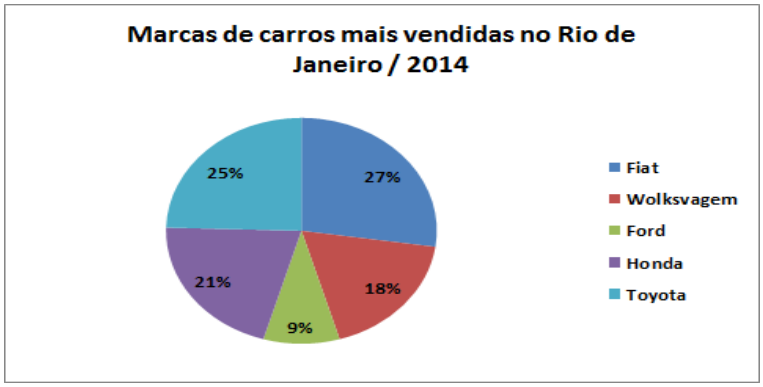

### 1.3.3 Gráfico de Setores

Este gráfico é constituido com base em um círculo. É empregado sempre que desejamos ressaltar a participação do um determinado dado no total. Cada setor é obtido por meio de uma regra de três simples e direta, lembrando que o total da série corresponde a $360^o$.

Por exemplo, o cálculo de um setor feito abaixo:

$$\begin{array}{ll} 100\% & \rightarrow 360^o \\ 55\% & \rightarrow x \end{array} \Rightarrow x = \frac{360 \times 55}{100} = 198^o.$$

O gráfico de setores abaixo representa a série estatística categórica (Tabela 1.4) apresentada na seção anterior.



### 1.3.4 Histograma e Polígono de Frequências

O **Histograma** é formado por retângulos justapostos, feitos sobre as classes da variável quantitativa em estudo, sua construção é feita no plano cartesiano sendo que no eixo $X$ ficam as classes e no eicho $Y$ colocamos a frequencia absoluta, ou frequencia relativa ou uma densidade, respeitando uma escala previamente estabelecida.

O **Polígono de Frequências** é formado por linhas que unem os pares ordenados $(x_i, f_i)$ formados pelos pontos médios $x_i$ e as frequências absolutas $f_i$ das classes. Sua construção é feita no plano cartesiano sendo que no eixo $X$ ficam os pontos médios das classes e no eixo $Y$ colocamos a frequencia absoluta, ou frequencia relativa ou uma densidade,respeitando uma escala previamente estabelecida.

Considerando a Tabela 1.5, que mostra a distribuição de frequências das alturas (cm) de 40 alunos da UFRRJ, temos o histograma e polígono de frequências.

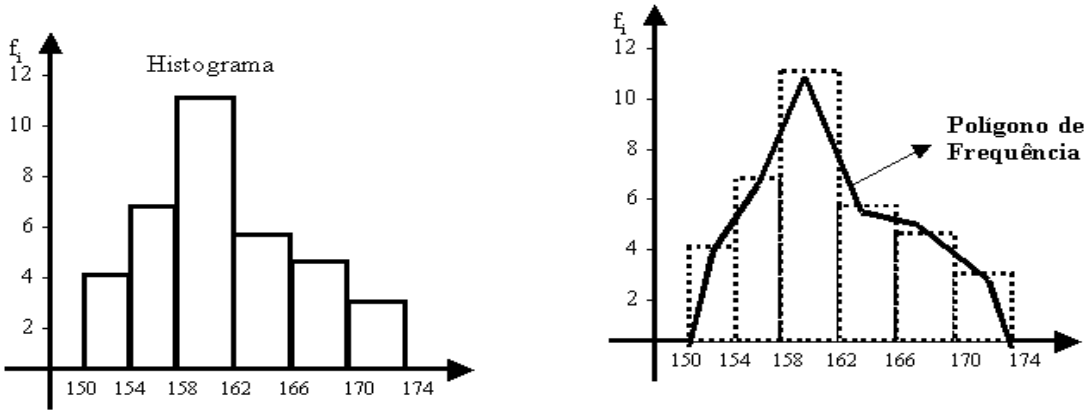



Figura 1.1: *Esquerdo: Histograma. Direito: Polígono de Frequências.*

### 1.3.5 Ramo-e-Folhas

Esta é uma técnica alternativa para resumir um conjunto de dados, com o objetivo de se obter uma idéia da forma de sua distribuição. Uma vantagem deste diagrama sobre o histograma é que não perdemos ( ou perdemos pouca) informação

sobre os dados em si. Num **ramo-e-folhas** os dados ficam ordenados crescentemente, o que facilita a obtenção de algumas medidas descritivas, que serão vistas no próximo capítulo.

Voltemos a considerar o rol das alturas (cm) de 40 alunos da UFRRJ.

150,151,151,153,154,154,154,155,156,157,

157,157,157,158,159,159,160,160,160,161,

161,161,161,161,162,162,162,163,163,165,

165,165,166,166,168,168,168,170,172,173.

Para construir o *ramo-e-folhas* devemos seguir o seguinte processo: para cada valor, o primeiro algarismo é colocado do lado esquerdo do traço vertical, formando os *ramos*. O segundo algarismo é colocado do lado direito do traço formando as *folhas*, as folhas dos ramos tem que ficar uma embaixo da outra para não distorcer as informações quanto ao modelo da distribuição. Porém, na construção de um *ramo-e-folhas*, a escolha dos algarismos mais relevantes depende do conjunto de dados em análise.

Em nosso exemplo, vamos considerar a dezena inicial para formar os ramos e o último algarismo formará as folhas.

| | |
|---|---|
| 15 | 0 1 1 3 4 4 4 5 6 7 7 7 7 8 9 9 |
| 16 | 0 0 0 1 1 1 1 1 2 2 2 3 3 5 5 5 6 6 8 8 8 |
| 17 | 0 2 3 |

**Exemplo:** Consideremos as taxas de mortalidade infantil de 34 municípios da Microrregião Oeste Catarinense, ano de 1982.

32 62 10 22 13 09 11 20 36 23 18 22 20 38 19 27 28

18 27 21 23 13 36 32 29 25 23 15 17 39 22 29 18 33

Apesar dos dados não estarem em ordem, podemos construir o ramo-e-folhas e depois ordenar as folhas.

| | |
|---|---|
| 0 | 9 |
| 1 | 0 3 1 8 9 8 3 5 7 8 |
| 2 | 2 0 3 2 0 7 8 7 1 3 9 5 3 2 9 |
| 3 | 2 6 8 6 2 9 3 |
| 4 | |
| 5 | |
| 6 | 2 |

Ordenando as folhas temos:

| | |
|---|---|
| 0 | 9 |
| 1 | 0 1 3 3 5 7 8 8 8 9 |
| 2 | 0 0 1 2 2 2 3 3 3 5 7 7 8 9 9 |
| 3 | 2 2 3 6 6 8 9 |
| 4 | |
| 5 | |
| 6 | 2 |

Notamos que o valor "62"está distante dos demais valores. É o que chamamos de **valor discrepante**. Então, podemos estudá-lo separadamente.

### 1.3.6 Box Plots

Uma maneira de representar características relevantes de um conjunto de dados é através do Box Plot, também chamado de diagrama em caixas, termo usado por Barbetta (2001). Murteira (1993) usa o termo "caixa-de-bigodes". Para construir o Box Plots precisamos de 5 medidas: $Q_1$, $Me$, $Q_3$, $L_I$ ($Limite\ Inferior$) e o $L_S$ ($Limite\ Superior$). Além dessas medidas é necessário calcular as seguintes quantidades:

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

sendo o $L_i$ = limite inferior, $L_S$ = limite superior e $d_q$ = distância ou intervalo interquartílico.

Utilizando como base o plano cartesiano, normalmente, no eixo $Y$ colocamos uma escala numérica, previamente determinada, que contenha as cinco medidas $Q_1$, $(Me)$, $Q_3$, $L_I$ e $L_S$ . É necessario o estabelecimento de uma escala para que

o gráfico não apresente informações distorcidas da realidade dos dados. O boxplot nos fornece uma análise visual da posição, dispersão, simetria, caudas e valores discrepantes (outliers) do conjunto de dados.

1- **Posição** – Em relação à posição dos dados, observa-se a linha central do retângulo (a mediana ou segundo quartil).

2- **Dispersão** – A dispersão dos dados pode ser representada pelo intervalo interquatilico que é a diferença entre o terceiro quartil e o primeiro quartil (tamanho da caixa), ou ainda pela amplitude que é calculada da seguinte maneira: valor máximo – valor mínimo. Embora a amplitude seja de fácil entendimento, o intervalo interqualitico é uma estatística mais robusta para medir variabilidade uma vez que não sofre influência de outliers.

3- **Caudas** – As linhas que vão do retângulo até aos outliers podem fornecer o comprimento das caudas da distribuição.

4- **Outliers** – Já os outliers indicam possíveis valores discrepantes. No boxplot, as observações são consideradas outliers quando estão abaixo ou acima do limite de detecção de outliers.

5- **Simetria** – Temos três situações possiveis quanto a simetria:

a- **Simetria**, quando a Mediana encontra-se centralizada entre o 1º e 3º quartil;

b- **Assimetria Positiva**, quando a Mediana encontra-se mais próxima do 1º quartil;

c- **Assimetria Negativa**, quando a Mediana encontra-se mais próxima do 3º quartil.

Vale ressaltar que a mediana é a medida de tendência central mais indicada quando os dados possuem distribuição assimétrica, uma vez que a média aritmética é influenciada pelos valores extremos.

Este gráfico é muito útil quando desejamos comparar várias distribuições simultaneamente. No modelo de box-plot abaixo, verificamos uma distribuição assimétrica positiva, pois a Mediana está mais proxima do 1º Quartil e com um valor

discrepante por estar fora dos limites inferior e superior.

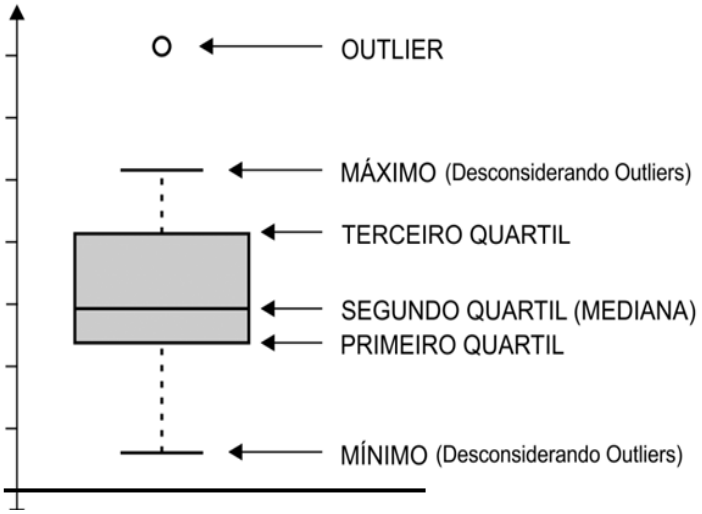



Figura 1.2: *Fonte: https://operdata.com.br*

## 1.4 Exercícios

1- Construir a tabela de distribuição de frequência completa: frequência relativa, frequência acumulada e ponto médio para cada um dos conjuntos de dados abaixo:

a) Peso (kg) de 40 adultos:

70 71 72 68 65 57 55 53 61 60 62 72 83 83 50 51 50 62 63 78

68 67 73 75 84 50 49 51 54 55 63 63 64 63 75 75 70 70 61 61

b) Medida (mm) de 30 peças:

3,5 3,2 3,2 3,2 3,8 2,1 2,1 2,5 2,5 2,4 3,8 3,7 3,5 3,5 3,7

3,1 3,1 3,0 3,0 2,4 2,8 2,8 3,2 3,2 3,5 3,7 3,0 3,0 3,1 3,2

c) Notas de 36 alunos em estatística:

8,0 8,5 7,0 7,0 7,5 6,0 6,5 6,0 5,0 3,0 3,5 3,0 3,5 4,0 4,0 4,0 4,5 3,5

5,5 5,5 9,5 9,5 9,5 9,0 9,0 10 9,0 4,5 4,5 7,0 7,5 7,5 6,0 6,5 6,0 6,0

2- Construir o histograma e o polígono de frequências para os dados do exercício 1.

3- Construir o ramo-e-folhas para os dados do exercício 1.

4- Utilizando a Tabela 1, sobre os 40 funcionários da UFRRJ, construir:

a) as tabelas para cada uma das variáveis;

b) os gráficos para as tabelas construidas em a).

5- Idealizar 3 (três) situações de pesquisa e construir:

a) Uma série temporal e o gráfico de linha.

b) Uma série geográfica e o gráfico de barras(colunas).

c) Uma série específica e o gráfico de setores.