Estatística com Apoio Computacional

Análise Exploratória

Universidade Estadual Vale do Acaraú – UVA

Paulo Regis Menezes Sousa

paulo_regis@uvanet.br

- Resumir, de forma eficiente, a informação contida nos dados.
- Identificar padrões, características e revelar tendências.
- Fazer previsões e tomar decisões.

Análise Exploratória

Conjunto de técnicas de tratamento de dados que nos ajuda a tirar conclusões preliminares sobre a informação disponível, sem implicar em uma fundamentação matemática mais rigorosa.

- A questão mais importante para um cientista em atividade – talvez a pergunta mais útil que alguém pode fazer – é: "**O que esta acontecendo aqui?**"
- Responder a esta pergunta requer o uso criativo de diferentes maneiras de visualizar **conjuntos de dados**, para resumi-los e expor qualquer estrutura que possa estar lá.
- Esta é uma atividade que é conhecida como **estatística descritiva**.

- É muito comum ter os dados tabulados em um arquivo-texto ou em outros formatos que permitem a conversão para dados texto.
- Geralmente, os dados estão estruturados de forma que cada linha corresponda a cada registro (instância), com elementos separados por espaços ou vírgulas.
- Usamos o comando `read.table()` para ler esses dados e armazená-los em um objeto.

```
1 dados <- read.table(          #lê dados de um arquivo texto
2             "dados-01.txt", #caminho e nome do arquivo
3             header = TRUE,  #primeira é cabeçalho
4             sep = ";")      #ponto-e-vírgula como separador
5 dados                       #exibe o objeto dados
```

Script: Lê dados de um arquivo `.txt`

- O formato de arquivo CSV ("Comma Separated Value") é geralmente usado para trocar dados entre aplicativos diferentes. O formato de arquivo, como é usado no Microsoft Excel, tornou-se um pseudo-padrão em todo o setor, mesmo entre plataformas não-Microsoft.
- Usamos o comando `read.csv()` para ler esses dados nesse tipo de arquivo e armazená-los em um objeto.

```
1 dados <- read.csv(           #lê dados de um arquivo csv
2             "dados-01.csv", #caminho e nome do arquivo
3             header = TRUE,  #primeira é cabeçalho
4             sep = ";")      #ponto-e-vírgula como separador
5 dados                       #exibe o objeto dados
```

Script: Lê dados de um arquivo `.csv`

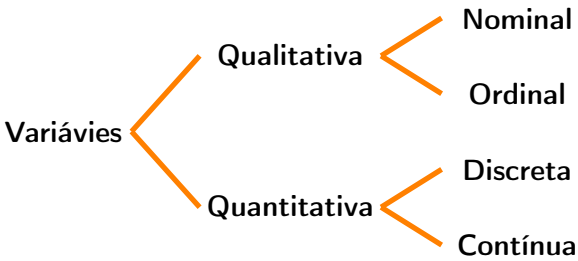
- Existem funções para extrair dados de planilhas de cálculo a partir do R sem a necessidade de conversão manual para `csv` . Uma dessas função é `read.xls()` do pacote **gdata**:

```
1 library(gdata)
2 bx <- read.xls("planilha.xls")
```

- A função `read.xls()` poderá falhar no Linux se o arquivo `xls` contiver acentos nos nomes das colunas e se a codificação de caracteres não for UTF-8. Nesse caso, pode-se tentar acrescentar o argumento `encoding` com o valor `"latin1"`:

```
1 bx <- read.xls("planilha.xls", encoding = "latin1")
```

- As variáveis, de uma forma geral, podem ser classificadas em tipos, conforme o seguinte:



Variável qualitativa nominal ou categórica

Seus valores possíveis são diferentes categorias não-ordenadas, em que cada observação pode ser classificada. *Exemplos: raça, nacionalidade, área de atividade*.

Variável qualitativa ordinal

Seus valores possíveis são diferentes categorias ordenadas, em que cada observação pode ser classificada. *Exemplos: classe social, nível de instrução*.

Variável quantitativa discreta

Seus valores possíveis são em geral resultados de um processo de contagem. *Exemplos: número de filhos, número de séries escolares cursadas com aprovação*.

Variável quantitativa contínua

Seus valores possíveis podem ser expressos através de números reais e varrem uma escala contínua de medição. *Exemplos: renda mensal, peso, altura.*

- A Tabela a seguir apresenta um conjunto de dados brutos de uma pesquisa antropométrica realizada com mulheres cuja idade está acima de 60 anos.

Ident	*Categ.*	*Idade*	*Peso (kg)*	*Altura (cm)*	*IMC (kg/m^2)*	*Classe IMC*	*Cintura (cm)*	*Quadril (cm)*	*RCQ*	*Classe RCQ*
ID1	A	61	58,2	154,0	24,5	normal	87	109	0,80	MR
ID2	S	69	63,0	152,0	27,3	sobrepeso	89	104	0,86	GR
ID3	S	61	70,1	158,0	28,1	sobrepeso	106	123	0,86	GR
ID4	S	71	73,2	156,0	30,1	sobrepeso	110	122	0,90	GR
ID5	A	63	58,6	152,0	25,4	sobrepeso	99	121	0,82	MR
ID6	S	71	77,0	160,0	30,1	sobrepeso	125	132	0,95	GR
ID7	S	72	76,2	165,0	28,0	sobrepeso	115	125	0,92	GR
ID8	S	68	59,8	160,0	23,4	normal	85	103	0,83	MR
ID9	A	66	64,3	155,0	26,8	sobrepeso	100	120	0,83	MR
ID10	S	69	52,1	151,0	22,8	normal	74	83	0,89	GR
ID11	S	72	62,0	156,0	25,5	sobrepeso	90	111	0,81	MR
ID12	S	67	52,1	151,0	22,8	normal	76	90	0,84	MR
ID13	S	63	58,0	157,0	23,5	normal	80	102	0,78	MR

- Neste exemplo cada observação (ou indivíduo) é uma mulher acima de 60 anos, e as variáveis (ou características) são:
  - Categoria, sendo A = ativa e S = sedentária
  - Idade, em anos
  - Peso, medido em kg
  - Altura, medida em cm
  - Índice de Massa Corporal (IMC), que é a seguinte razão: $peso/(altura)^2$
  - Classe segundo o IMC: normal ou sobrepeso
  - Circunferência da cintura, medida em cm
  - Circunferência do quadril, medida em cm
  - Relação cintura/quadril (RCQ), adimensional
  - Classe segundo a RCQ, sendo PR = pequeno risco, MR = médio risco e GR = grande risco.

- Para descrever o comportamento de uma variável é comum apresentar os valores que ela assume organizados sob a forma de tabelas de frequências e gráficos.
- Em uma tabela de frequências para uma variável qualitativa:
  - cada linha corresponde a um valor possível da variável;
  - através de um processo de contagem são obtidos os valores que constam na coluna de frequências da tabela. O resultado dessa contagem é a chamada **frequência absoluta**.
  - a partir das frequências absolutas podem ser também calculadas frequências relativas, usualmente apresentadas sob a forma de percentuais.

Exemplos de tabelas de frequências para as variáveis Categoria e Classe relação cintura-quadril (RCQ).

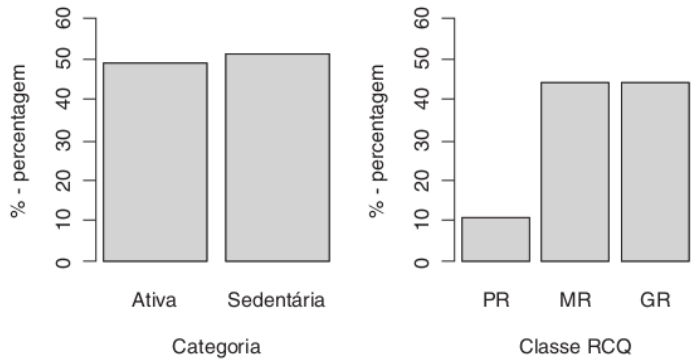
Categoria	*Freqüência*	*Percentuais*
Ativa	22	48,89
Sedentária	23	51,11
Total	45	100,00

Tabela: Frequência e percentual das 45 idosas, segundo a categoria

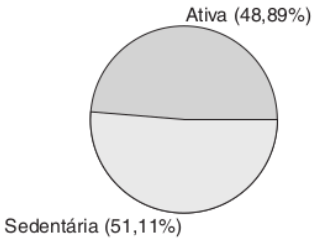
Classe RCQ	*Freqüência*	*Percentuais*
Pequeno Risco	5	11,11
Médio Risco	20	44,44
Grande Risco	20	44,44
Total	45	100,00

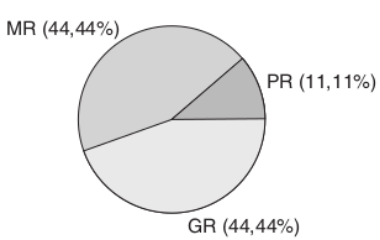
Tabela: Frequência e percentual das 45 idosas, segundo a classe de RCQ

- Com base em uma tabela de frequências podem ser construídos gráficos da distribuição de frequências, entre os quais os mais comuns são o **gráfico de barras** e o **gráfico de pizza**.



- Com base em uma tabela de frequências podem ser construídos gráficos da distribuição de frequências, entre os quais os mais comuns são o **gráfico de barras** e o **gráfico de pizza**.



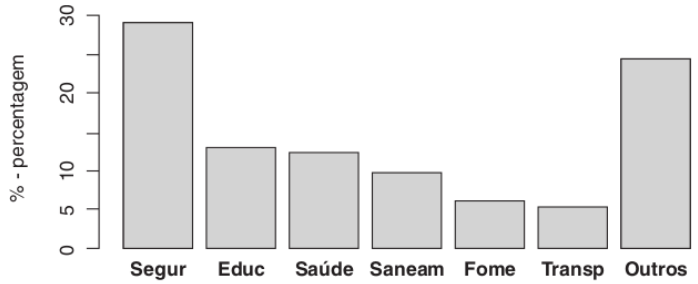


```r
tabela <- c(48.89, 51.11) # vetor de dados

# atribuindo nomes aos elementos do vetor
names(tabela) <- c("Ativa","Sedentária")

# gráfico de barras
barplot(tabela,                    # dados
        xlab = "Categorias",       # rótulo do eixo x
        ylab = "% - percentagem", # rótulo do eixo y
        ylim = c(0, 60))           # limites do eixo y

# gráfico de pizza/setores
pie(tabela,                      # dados
      col = c("red","green")) # cores
```

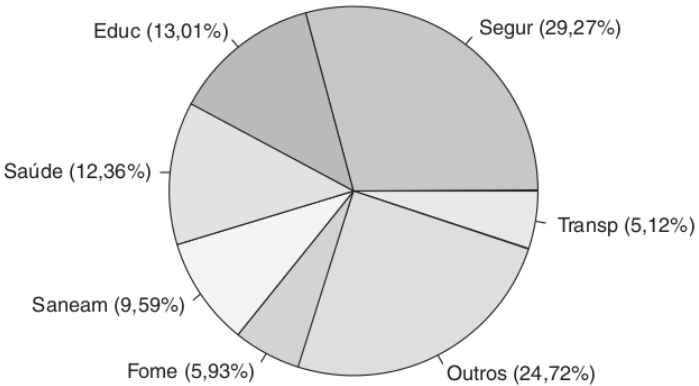
- Consideremos agora uma pesquisa por amostragem feita em 1986 junto à população do Estado do Rio de Janeiro.
- Foram ouvidas 1.230 pessoas que, entre outras coisas, apontaram qual era, na sua opinião, o problema mais grave do Estado naquele momento.

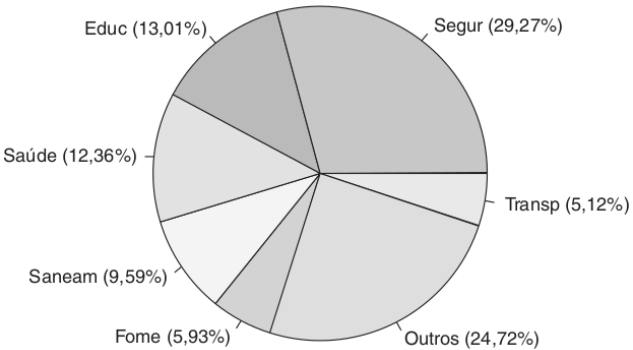
Problema mais grave do Estado	*Freqüências*	*Percentuais*
Segurança / Violência	360	29,27
Educação	160	13,01
Saúde	152	12,36
Saneamento	118	9,59
Alimentação/Fome/Pobreza	73	5,93
Transporte	63	5,12
Outros	304	24,72
Total	1230	100,00

- Foram construídos os gráficos a seguir com base nos dados.

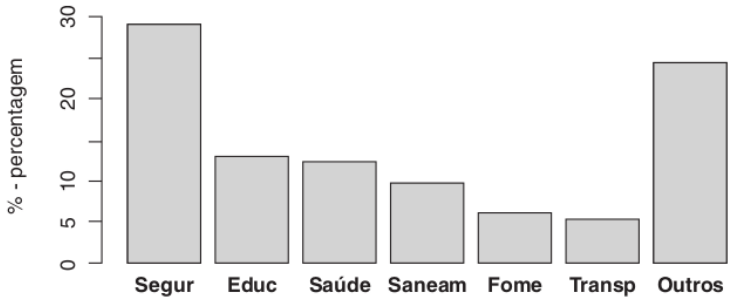


- Foram construídos os gráficos a seguir com base nos dados.





- O gráfico de setores (pizza), por não implicar uma ordenação das categorias, é mais apropria do para as **variáveis qualitativas nominais**.



- O gráfico de barras, em que as categorias estão naturalmente ordenadas, é mais apropriado para as **variáveis qualitativas ordinais**.

- No caso de **variável quantitativa discreta** com um **pequeno número de valores** possíveis (por exemplo: número de filhos), a construção de uma tabela de frequência segue os mesmos moldes do que foi visto para variáveis qualitativas.
- Para uma **variável quantitativa discreta** com um **grande número de valores** possíveis ou com uma **variável quantitativa contínua**, é preciso dividir o seu intervalo de variação em vários subintervalos com a mesma amplitude.

- A Tabela a seguir apresenta um conjunto de dados brutos de uma pesquisa antropométrica realizada com mulheres cuja idade está acima de 60 anos.

Ident	*Categ.*	*Idade*	*Peso (kg)*	*Altura (cm)*	*IMC (kg/m^2)*	*Classe IMC*	*Cintura (cm)*	*Quadril (cm)*	*RCQ*	*Classe RCQ*
ID1	A	61	58,2	154,0	24,5	normal	87	109	0,80	MR
ID2	S	69	63,0	152,0	27,3	sobrepeso	89	104	0,86	GR
ID3	S	61	70,1	158,0	28,1	sobrepeso	106	123	0,86	GR
ID4	S	71	73,2	156,0	30,1	sobrepeso	110	122	0,90	GR
ID5	A	63	58,6	152,0	25,4	sobrepeso	99	121	0,82	MR
ID6	S	71	77,0	160,0	30,1	sobrepeso	125	132	0,95	GR
ID7	S	72	76,2	165,0	28,0	sobrepeso	115	125	0,92	GR
ID8	S	68	59,8	160,0	23,4	normal	85	103	0,83	MR
ID9	A	66	64,3	155,0	26,8	sobrepeso	100	120	0,83	MR
ID10	S	69	52,1	151,0	22,8	normal	74	83	0,89	GR
ID11	S	72	62,0	156,0	25,5	sobrepeso	90	111	0,81	MR
ID12	S	67	52,1	151,0	22,8	normal	76	90	0,84	MR
ID13	S	63	58,0	157,0	23,5	normal	80	102	0,78	MR

- Vamos construir a tabela de frequências para as variáveis Idade e Índice de Massa Corporal.
- No caso da variável Idade, o menor valor é 60, e o maior é 79. Portanto, vamos dividir o intervalo [60, 80], que contém todos os valores observados da variável considerada, em subintervalos de amplitude 5 (fechados à esquerda e abertos à direita) e contar o número de ocorrências em cada um deles.

Faixa Etária	*Freqüência*	*Percentuais*
60 |— 65	16	35,56
65 |— 70	16	35,56
70 |— 75	12	26,67
75 |— 80	1	2,22
Total	45	100,00

- Para a variável IMC, o menor valor é 20,3, e o maior é 30,1. Portanto, vamos dividir o intervalo [20,0 ; 32,5], que contém todos os valores observados da variável considerada, em subintervalos de amplitude 2,5 (fechados à esquerda e abertos à direita) e contar o número de ocorrências em cada um deles.

IMC	*Freqüência*	*Percentuais*
20,0 |— 22,5	7	15,56
22,5 |— 25,0	20	44,44
25,0 |— 27,5	11	24,44
27,5 |— 30,0	5	11,11
30,0 |— 32,5	2	4,44
Total	45	100,00

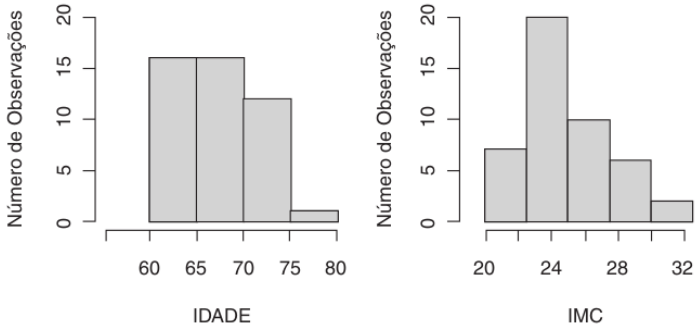
- A Tabela a baixo de dados brutos reporta o número de linhas telefônicas por mil habitantes em cada estado do Brasil, em 2001.

Acre	183,8	Maranhão	86,1	Rio de Janeiro	347,5
Alagoas	125,4	M. Grosso	199,6	R. G. Norte	150,1
Amapá	193,3	M. G. Sul	235,3	R. G. Sul	236,9
Amazonas	162,0	M. Gerais	218,6	Rondônia	214,6
Bahia	142,3	Pará	128,0	Roraima	214,1
Ceará	140,6	Paraíba	125,4	S. Catarina	257,3
D. Federal	456,8	Paraná	244,2	S. Paulo	362,8
E.S.	228,7	Pernambuco	147,8	Sergipe	140,7
Goiás	231,4	Piauí	118,2	Tocantins	113,8

- O menor valor é 86 (MA), e o maior valor é 457 (DF). Portanto, vamos dividir o intervalo [50;500], que contém todos os valores observados da variável considerada, em subintervalos de amplitude 50 (fechados à esquerda e abertos à direita) e contar o número de ocorrências em cada um deles.

Classe	*Freqüência*	*Percentual*
50 |—- 100	1	3,70
100 |—- 150	9	33,33
150 |—- 200	5	18,52
200 |—- 250	8	29,63
250 |—- 300	1	3,70
300 |—- 350	1	3,70
350 |—- 400	1	3,70
400 |—- 450	0	0,00
450 |—- 500	1	3,70
Total	27	100,00

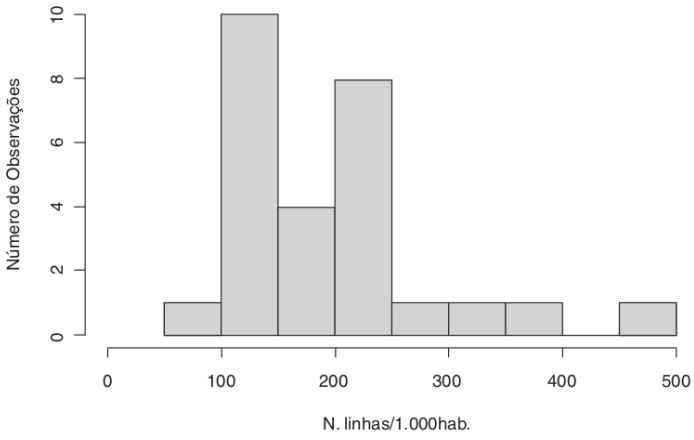
- No histograma os intervalos de classe da variável considerada são marcados em um eixo e as frequências (ou percentuais) no outro eixo.
- A largura das barras é proporcional à amplitude do intervalo, e a altura é proporcional à frequência (ou ao percentual).



- A Tabela a baixo de dados brutos reporta o número de linhas telefônicas por mil habitantes em cada estado do Brasil, em 2001.

Acre	183,8	Maranhão	86,1	Rio de Janeiro	347,5
Alagoas	125,4	M. Grosso	199,6	R. G. Norte	150,1
Amapá	193,3	M. G. Sul	235,3	R. G. Sul	236,9
Amazonas	162,0	M. Gerais	218,6	Rondônia	214,6
Bahia	142,3	Pará	128,0	Roraima	214,1
Ceará	140,6	Paraíba	125,4	S. Catarina	257,3
D. Federal	456,8	Paraná	244,2	S. Paulo	362,8
E.S.	228,7	Pernambuco	147,8	Sergipe	140,7
Goiás	231,4	Piauí	118,2	Tocantins	113,8

- Histograma da variável número de linhas telefônicas por mil habitantes.



```r
# dados para o histograma
dados <- c(96,96,102,102,102,104,104,108,126,126,128,128,140,
           156,160,160,164,170,115,121,118,142,145,145,149,112,
           152,144,122,121,133,134,109,108,107,148,162,96)

# raiz quadrada da quantidade de observações
k <- round(sqrt(length(dados))) # round: arredondamento

# histograma
hist(dados,
     nclass = k,          # número de classes
     col = "blue",        # cor do histograma
     main = "Histogama") # título do gráfico
```

- Para uma dada variável quantitativa, uma medida de centralidade é um "valor típico" em torno do qual se situam os valores daquela variável.
- Há várias formas de se definir uma medida de centralidade:
  - a **média aritmética**,
  - a **mediana**
  - e a **moda**

  são as mais conhecidas entre elas.

Média aritmética

Sejam $x_1, x_2, \ldots, x_n$ os valores observados da variável considerada.

$$x = \frac{x_1, x_2, \ldots, x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Mediana

Sejam $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ os mesmos valores que compõem a amostra, porém dispostos em ordem crescente.

$$\text{Mediana}(x) = \begin{cases} \text{valor da posição central,} & \text{se n é ímpar} \\ \text{média dos valores de posição central,} & \text{se n é par} \end{cases}$$

Moda

A **moda** dos dados é aquele valor da amostra que ocorre com maior frequência.

- Média Aritmética

```
> x <- c(10, 14, 13, 15, 16, 18, 12)
> mean(x)
[1] 14
```

- Mediana

```
> k <- c(1,3,0,0,2,4,1,2,5)
> median(k)
[1] 2
> g <- c(1,3,0,0,2,4,1,3,5,6)
> median(g)
[1] 2.5
```

- Moda

```
> y <- c(7,8,9,10,10,10,11,12)
> table(y)
y
7  8  9 10 11 12
1  1  1  3  1  1
>
subset(table(y),table(y)==max(table(y)))
10
3
```

- Mediana

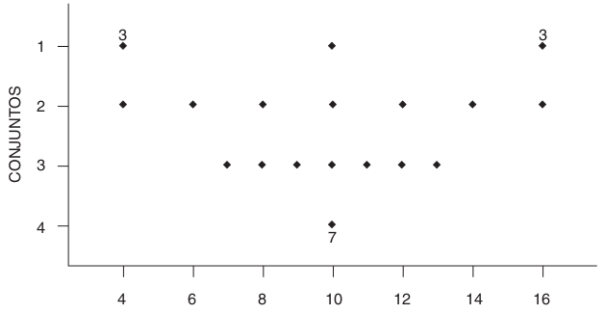
```
k <- c(1,3,0,0,2,4,1,2,5)
median(k)
[1] 2
g <- c(1,3,0,0,2,4,1,3,5,6)
median(g)
[1] 2.5
```

- Quartis

```
> z <- c(5,2,6,9,10,13,15)
> summary(z)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.000   5.500   9.000   8.571  11.500  15.000
> quantile(z)
0%  25%  50%  75% 100%
2.0  5.5  9.0 11.5 15.0
> quantile(z,0.25)
25%
5.5
```

- Uma medida de dispersão para uma variável quantitativa é um indicador do grau de **espalhamento** dos valores da amostra em torno da medida de centralidade.

Conjunto 1:	4	4	4	10	16	16	16
Conjunto 2:	4	6	8	10	12	14	16
Conjunto 3:	7	8	9	10	11	12	13
Conjunto 4:	10	10	10	10	10	10	10



- Pode ser observado que o ponto central dos quatro conjuntos é igual a 10.

$$\text{Conjunto } 1 \rightarrow \quad \bar{x} = 4+4+4+10+16+16+16 = \frac{70}{7} = 10$$

$$\text{Conjunto } 2 \rightarrow \quad \bar{x} = 4+6+8+10+12+14+16 = \frac{70}{7} = 10$$

$$\text{Conjunto } 3 \rightarrow \quad \bar{x} = 7+8+9+10+11+12+13 = \frac{70}{7} = 10$$

$$\text{Conjunto } 4 \rightarrow \quad \bar{x} = 10+10+10+10+10+10+10 = \frac{70}{7} = 10$$

- Observa-se também que:
  - todas as observações do conjunto 4 são exatamente iguais ao ponto central;
  - no conjunto 3 os dados estão um pouco mais dispersos em relação a 10;
  - no conjunto 2, mais ainda;
  - finalmente, o conjunto 1 é aquele em que há a maior dispersão em torno da média.

- Há diferentes formas de se medir a dispersão de uma variável quantitativa. Aqui serão vistas a variância, o desvio-padrão, o coeficiente de variação e a distância interquartil.

Variância

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

Desvio padrão

O desvio-padrão é a raiz quadrada positiva da variância.

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}} \tag{2}$$

Coeficiente de variação

O coeficiente de variação é o quociente entre o desvio-padrão e a média

$$\mathrm{CV} = \frac{S}{\bar{x}} \tag{3}$$

- Intervalo de dados

```
> range(dados)
```

- Variância

```
> var(dados)
```

- Desvio Padrão

```
> sd(dados)
```

- Coeficiente de Variação (CV)

```
> v <-c(10,11,9,10,10,9,11)
> CV = 100*sd(v)/mean(v)
> CV
[1] 8.164966  #em torno de 8%
```

- Os dados a seguir representam o peso em quilograma de cinco mães e de seus respectivos bebês recém-nascidos:

Peso da mãe	52,3	52,5	53	53,5	54
Peso do recém-nascido	2,3	2,5	3	3,5	4

Variável	*Média*	*Variância*	*Desvio-padrão*
Peso da mãe	53,06	0,493	0,702
Peso do recém-nascido	3,06	0,493	0,702

Variável	*Coeficiente de variação*
Peso da mãe	0,009
Peso do recém-nascido	0,161

- A **mediana** é um valor tal que metade dos dados é menor que ele, e metade dos dados é maior que ele.

Quartis

Primeiro quartil	Q1 tem 1/4 dos dados abaixo dele e 3/4 dos dados acima dele.
Segundo quartil	Q2 é a própria mediana.
Terceiro quartil	Q3 tem 3/4 dos dados abaixo dele e 1/4 dos dados acima dele.
Distância interquartil	é dada por DIQ = Q3-Q1.

- Distância interquartil para cada um dos conjuntos fictícios de exemplo.

Conjunto de dados	*Q1*	*Q2*	*Q3*	*DIQ=Q3-Q1*
1	4	10	16	12
2	7	10	13	6
3	8,5	10	11,5	3
4	10	10	10	0

- Algumas observações sobre as medidas de dispersão:
  1. Não é difícil observar que quanto mais dispersos estiverem os dados maiores tendem a ser a variância S^2 , o desvio-padrão S, o coeficiente de variação cv e a distância interquartil DIQ.
  2. A unidade da variância é o quadrado da unidade dos dados. Por exemplo, se os dados forem medidos em metros, a unidade da variância será metro ao quadrado. Consequentemente, a unidade do desvio-padrão é a mesma dos dados originais.
  3. A média e o desvio-padrão são possivelmente as duas medidas mais comumente utilizadas na prática. Porém, enquanto a média aritmética é uma medida normalmente muito conhecida de todos, o mesmo não acontece com o desvio-padrão.

Exercício 1

Um artigo no Journal of Structural Engineering (Vol. 115, 1989) descreve um experimento para testar a resistência resultante em tubos circulares com calotas soldadas nas extremidades. Os primeiros resultados (em kN) são: 96, 96, 102, 102, 102, 104, 104, 108, 126, 126, 128, 128, 140, 156, 160, 160, 164 e 170. Pede-se:

a) Calcule a média da amostra

b) Calcule o mediana e os quartis 1 e 3

c) Calcule a variância e o desvio padrão da amostra

- Diz-se que uma medida de centralidade ou de dispersão é resistente quando ela é pouco afetada pela presença de observações discrepantes. Por exemplo:

Acre	183,8	Maranhão	86,1	Rio de Janeiro	347,5
Alagoas	125,4	M. Grosso	199,6	R. G. Norte	150,1
Amapá	193,3	M. G. Sul	235,3	R. G. Sul	236,9
Amazonas	162,0	M. Gerais	218,6	Rondônia	214,6
Bahia	142,3	Pará	128,0	Roraima	214,1
Ceará	140,6	Paraíba	125,4	S. Catarina	257,3
D. Federal	456,8	Paraná	244,2	S. Paulo	362,8
E.S.	228,7	Pernambuco	147,8	Sergipe	140,7
Goiás	231,4	Piauí	118,2	Tocantins	113,8

Medida	*Amostra Completa*	*Amostra Expurgada*
Nº de observações	27	26
Média	200,20	190,33
Mediana	193,3	188,6
Desvio-padrão	84,44	68,41
Distância Interquartil	92,7	90,2

Exercício 2

Suponha que tivesse havido um erro de digitação de modo que o valor 456,8 correspondente ao Distrito Federal passasse a ser 4568. Verifique qual seria a variação em cada uma dessas quatro medidas, sempre comparando a amostra completa com a amostra expurgada.

- Eventualmente em uma massa de dados há valores que foram coletados em condições anormais (falha de equipamento, queda de energia, erro do operador, erro de leitura, erro de digitação etc.).
- São as chamadas observações discrepantes ou **outliers**.
- Um critério bastante utilizado para a identificação de observações discrepantes que se baseia em medidas pouco resistentes é apontar toda observação que estiver fora do intervalo $(\bar{x} - 3S,\ \bar{x} + 3S)$.
- Um segundo critério também muito usado que se baseia em medidas mais resistentes para a identificação de observações discrepantes é apontar qualquer valor abaixo de $Q1 - \frac{3}{2} \times DIQ$ ou acima de $Q3 + \frac{3}{2} \times DIQ$.

Exercício 3

No caso do arquivo com dados antropométricos, vamos analisar a eventual presença de observações discrepantes no caso das variáveis idade (em anos) e peso (em kg).