

3

Estatística Descritiva

ESQUEMA DO CAPÍTULO

3.1 INTRODUÇÃO

3.2 CONCEITOS BÁSICOS

3.3 ORGANIZAÇÃO E APRESENTAÇÃO DE DADOS

3.4 SÍNTESE NUMÉRICA

3.5 OUTROS TÓPICOS

3.1 Introdução

- Síntese de dados:
 - É a *primeira etapa* na análise e interpretação dos dados obtidos dos participantes de estudos na área da saúde;
 - Feita por meio de *métodos estatísticos* que recebem genericamente a denominação de estatística descritiva;
 - Consiste basicamente na organização e descrição dos *dados*;
 - Tem como *elementos* básicos:
 - Tabelas;
 - Gráficos ou
 - Medidas ou síntese numéricas.

3.2 Conceitos Básicos

3.2.1 Variável

- É a quantificação da *característica* de interesse do estudo;
- Corresponde ao termo '*parâmetro*', utilizado na área da saúde;
- Produzidas por *perguntas* tais como:
 - Qual a sua idade?
 - Qual é o número de pessoas da sua família?
 - Qual a renda total de sua família?
 - Qual é o seu estado civil?
 - Você tem emprego fixo?

3.2 Conceitos Básicos (cont.)

3.2.1 Variável (cont.)

- Em geral, a mera **identificação** da característica de interesse leva a uma quantificação adequada;
- Entretanto, há situações em que a quantificação da característica de interesse precisa passar por um procedimento de **padronização** (*e.g.*, como quantificar a dor; falta de ar (dispneia); estado geral do indivíduo etc.)
- Em outras situações, a precisa ser obtida pela **manipulação** de outras (*e.g.*, para avaliar a obesidade, é comum utilizar o índice de massa corpora (IMC), que é o peso (em *kg*) dividido pelo quadrado da altura (em m^2) .

3.2 Conceitos Básicos (cont.)

3.2.2 Tipos de Variáveis

Variáveis Categóricas

- Nominal (por exemplo, sexo);
- Ordinal (caso exista uma ordem entre as classes; por exemplo, estágio de uma doença);
- Níveis:
 - Binárias (sexo);
 - Vários (tipo sanguíneo);
 - Dependente do interesse do estudo.

3.2 Conceitos Básicos (cont.)

3.2.2 Tipos de Variáveis

Variáveis Quantitativas

- Discretas:
 - são aquelas provenientes de contagens (por exemplo, o tamanho da família);
- Contínuas:
 - são aquelas provenientes de medições (por exemplo, a pressão arterial);

3.2 Conceitos Básicos (cont.)

3.2.2 Tipos de Variáveis

Note que há uma *diferença* entre:

- variáveis *categóricas ordinais*;
 - por exemplo, estágio de câncer de mama, I, II, III, IV;
- variáveis *quantitativas discretas*;
 - por exemplo, número de crianças na família, 0, 1, 2, 3, 4, ...

3.2 Conceitos Básicos (cont.)

3.2.3 Dados Brutos

- São aqueles dados na forma como foram obtidos *diretamente* da pesquisa;
- Não sofreram *qualquer* processo de síntese ou análise;
- São, em geral, apresentados em *planilhas*;
- Esse *conjunto* de dados é denominado *amostra*;
- O *tamanho* da amostra é denotado por n ;
- Dados coletados são *organizados* em planilhas, em que:
 - *cada linha corresponde a um elemento da amostra* (unidade amostral) e
 - *cada coluna corresponde a uma variável* (atributo);

3.2 Conceitos Básicos (cont.)

3.2.3 Dados Brutos (cont.)

- Planilhas mais conhecidas: Excel, Access;
- *Softwares* estatísticos: R, MINITAB, SPSS, SAS, PRISMA *etc.*
- Um exemplo pode ser visto a seguir, em que cada **linha representa um indivíduo** e cada **coluna representa um atributo** desse indivíduo:

Paciente	Idade	Sexo	Localização do Melanoma	Tratamento
1	48	F	Cabeça/pescoço	Quimioterapia
2	60	M	Membros superior	Cirurgia
3	36	F	Cabeça/pescoço	Cirurgia + Quimioterapia
...				
100	59	F	Membros inferiores	Cirurgia

3.2 Conceitos Básicos (cont.)

3.2.3 Dados Brutos (cont.)

- **Exemplo 3.0:**

Este exemplo usa os dados do livro “Kume, H. (1994). *Métodos Estatísticos para Melhoria da Qualidade*, Ed. Gente”. Numa padaria, dois padeiros, A e B, assam pão francês usando duas máquinas (1 e 2). Os pesos dos pães produzidos foram registrados durante 10 dias, como mostra a Tabela 3.0. Cada dia, quatro pães foram retirados ao acaso de cada máquina e pesados. O peso especificado é de 200 a 225 g.

Com relação aos dados da Tabela 3.0, *como planejar um arquivo de dados apropriado*, com definição das variáveis necessárias e escolha de nomes descritivos?

Tabela 3.0: Dados de uma padaria

Dia	Padeiro	Máquina 1				Máquina 2			
1	A	209,2	209,5	210,2	212,0	214,3	221,8	214,6	214,4
2	A	208,5	208,7	206,2	207,8	215,3	216,7	212,3	212,0
3	A	204,2	210,2	210,5	205,9	215,7	213,8	215,2	207,2
4	B	204,0	203,3	198,2	199,9	212,5	210,2	211,3	210,4
5	B	209,6	203,7	213,2	209,6	208,4	214,9	212,8	214,8
6	A	208,1	207,9	211,0	206,2	212,3	216,2	208,4	210,8
7	A	205,2	204,8	198,7	205,8	208,1	211,9	212,9	209,0
8	B	199,0	197,7	202,0	213,1	207,5	209,9	210,6	212,3
9	B	197,2	210,6	199,5	215,3	206,9	207,1	213,6	212,2
10	B	199,1	207,2	200,8	201,2	209,6	209,5	206,8	214,2

3.2 Conceitos Básicos (cont.)

3.2.3 Dados Brutos (cont.)

Exemplo 3.1: Teor de gordura fecal em crianças

Embora os gastroenterologistas infantis reconhecessem a utilidade diagnóstica do teor de gordura fecal, até 1984 não existia um padrão de referência desta medida para crianças brasileiras. A Tabela 3.1 apresenta os valores obtidos em 43 crianças sadias em *g/24 horas*.

Tabela 3.1: Teor de gordura (*g/24 horas*) em 43 crianças

3,7	1,6	2,5	3,0	3,9	1,9	3,8	1,5	1,1
1,8	1,4	2,7	2,1	3,3	3,2	2,3	2,3	2,4
0,8	3,1	1,8	1,0	2,0	2,0	2,9	3,2	1,9
1,6	2,9	2,0	1,0	2,7	3,0	1,3	1,5	4,6
2,4	2,1	1,3	2,7	2,1	2,8	1,9		

3.2 Conceitos Básicos (cont.)

3.2.3 Dados Brutos (cont.)

Exemplo 3.2: Nível de colesterol

Os valores da Tabela 3.2 são as medidas em *mg/dl* do colesterol referentes a um segundo exame em 80 pacientes. Aqueles que *não* compareceram estão representados pelo número **9**.

Tabela 3.2: Taxa de colesterol (*mg/dl*) em 80 indivíduos

278	182	247	227	277	194	196	276	244	192
118	219	255	201	9	209	219	228	209	209
171	213	233	226	209	200	200	363	209	200
179	167	192	277	317	146	217	292	217	255
212	233	250	243	150	209	184	199	250	479
175	194	221	233	9	184	217	150	167	265
242	180	255	170	209	161	196	165	234	179
248	184	291	185	242	276	243	229	242	250

3.3 Organização e Apresentação de Dados (cont.)

3.3.1 Tabelas de Frequências

Exemplo 3.5: Tentativas de suicídio

A Tabela 3.3 mostra a distribuição da profissão de pacientes potencialmente suicidas. A maior incidência foi observada entre profissões mal remuneradas e com sobrecarga de trabalho, sem perspectiva de ascensão social.

Tabela 3.3: Distribuição de profissões entre pacientes potencialmente suicidas

Profissão	Frequência	Proporção
Serviços gerais*	75	0,248
Doméstica**	55	0,182
Do lar	53	0,175
Indeterminada	29	0,096
Emprego especializado***	23	0,076
Menor	20	0,066
Desempregado	15	0,050
Estudante	14	0,046
Lavrador	12	0,040
Autônomo	4	0,013
Aposentado	2	0,007
Total	302	1,000
*garçon, encanador, pedreiro, frentista, operário, padeiro, açougueiro, borracheiro etc.		
**copeira, faxineira, costureira e bordadeira		
***enfermeira, modelo, protético, escrivão, professor, digitador e vendedor		

3.3 Organização e Apresentação de Dados (cont.)

3.3.1 Tabelas de Frequências (cont.)

Exemplo 3.5: Tentativas de suicídio (final)

A Tabela 3.4 mostra a distribuição de tentativas de suicídio segundo faixa etária. A faixa etária mais prevalente é de 20 a 29 anos, tanto no geral (37,67%), como por sexo (36%, masculino, e 38%, feminino).

Tabela 3.4: Distribuição de idades dos pacientes potencialmente suicidas

Idade (anos)	Frequência absoluta	Frequência relativa
10 ┤ 20	57	18,87%
20 ┤ 30	113	37,42%
30 ┤ 40	59	19,54%
40 ┤ 50	32	10,60%
50 ┤ 60	19	6,29%
60 ┤ 70	7	2,32%
≥ 70	2	0,66%
Indeterminada	13	4,30%
Total	302	100,00%

3.3 Organização e Apresentação de Dados

3.3.1 Tabelas de Frequências

- É uma tabela construída a partir dos dados brutos que leva em conta a frequência com que cada observação ocorre;
- O interesse é conhecer o comportamento da variável, analisando a ocorrência de suas possíveis realizações;

Exemplo 3.8: Melanoma

Melanoma é uma neoplasia de grande potencial maligno responsável por grande parte das mortes por câncer de pele. Na análise de 35.695 biópsias realizadas, encontraram-se os 30 casos de melanoma primitivos de pele. A Tabela 3.8 mostra a distribuição da localização anatômica dos melanomas, sendo a cabeça/pescoço mais comum.

Tabela 3.8: Distribuição de melanomas por localização anatômica

Localização anatômica	Nº de casos	Percentual
Cabeça/pescoço	10	33,3%
Tronco	7	23,3%
Membros superiores	6	20,0%
Membros inferiores	2	6,7%
Acral	5	16,7%
Total	30	100,0%

3.3 Organização e Apresentação de Dados (cont.)

3.3.1 Tabelas de Frequências (cont.)

Etapas para a construção de tabelas de frequência para dados agrupados:

1. Encontrar o menor e o maior valores (denominados *mínimo* e *máximo*) do conjunto de dados;
2. Escolher um número de subintervalos (*classes*), preferencialmente de igual tamanho (*amplitude*), que englobem todos os dados, sem haver superposição dos intervalos; os extremos são conhecidos com *limites de classes*; (obs.: para número de classes, usar $\sqrt{n}$ ou $1+\log_2 n$, para uma amostra de tamanho n .)
3. Contar o número de elementos que pertencem a cada classe; este número é denominado *frequência absoluta*, usualmente denotado por f ;
4. Determinar a *frequência relativa* (denotada por fr) de cada classe, dividindo a frequência da classe pelo número total de observações, n .

3.3 Organização e Apresentação de Dados (cont.)

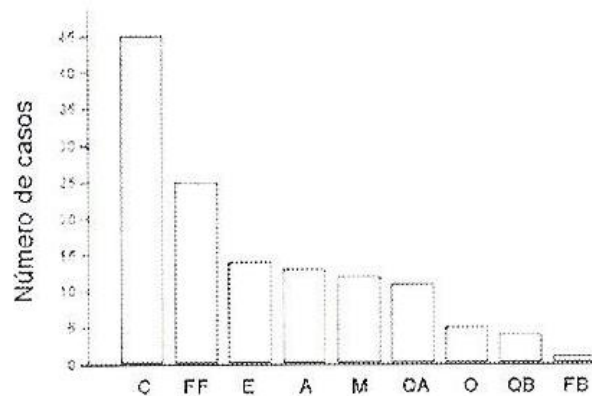
3.3.2 Gráficos

Representação gráfica para variáveis categóricas

Exemplo 3.9: Fraturas na face

Em uma amostra de 130 casos de fraturas na face foram encontrados uma grande maioria de vítimas do sexo masculino (78,4%) e na faixa etária entre 20 e 40 anos. A Figura 3.1 apresenta o diagrama de barras com os agentes etiológicos.

Figura 3.1: Diagrama de barras dos seguintes agentes etiológicos: carro (C), moto (M), queda de altura (QA), queda de bicicleta (QB), ferimentos de arma branca (FB) e outros (O)



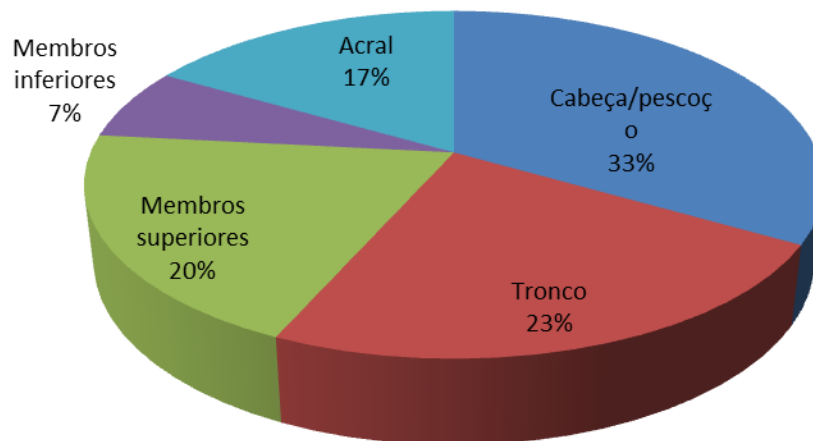
3.3 Organização e Apresentação de Dados (cont.)

3.3.2 Gráficos (cont.)

Representação gráfica para variáveis categóricas (cont.)

Exemplo: Gráfico de setores/pizza

É adequado para representar variáveis categóricas e consiste em repartir um disco em setores circulares correspondentes às porcentagens de cada valor.



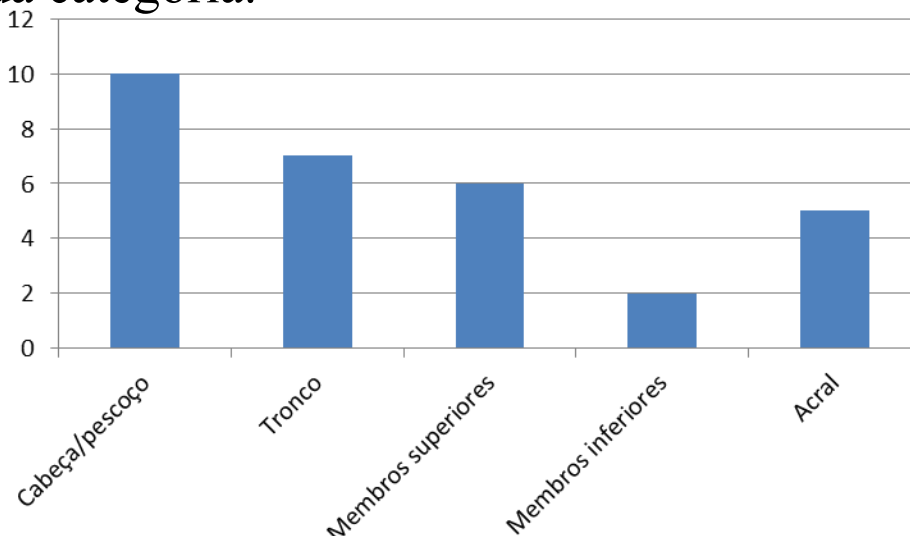
3.3 Organização e Apresentação de Dados (cont.)

3.3.2 Gráficos (cont.)

Representação gráfica para variáveis categóricas (cont.)

Exemplo: Gráfico de barras

Neste tipo de gráfico, o tamanho de cada barra é proporcional ao número de indivíduos da categoria.



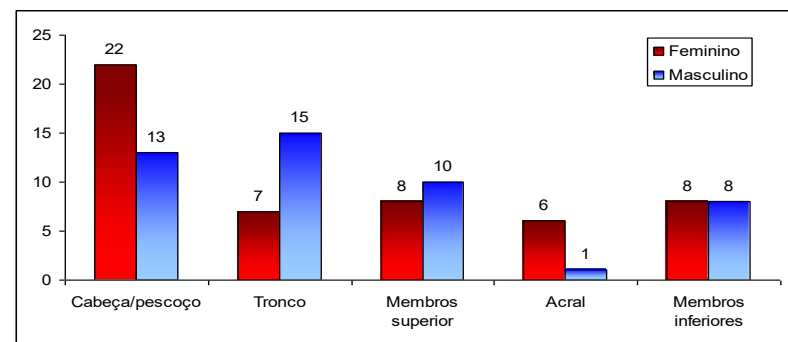
3.3 Organização e Apresentação de Dados (cont.)

3.3.2 Gráficos (cont.)

Representação gráfica para variáveis categóricas (cont.)

Exemplo: Gráfico de barras estratificado

Localização anatômica do melanoma	Feminino	Masculino	Total
Cabeça/pescoço	22	13	35
Tronco	7	15	22
Membros superior	8	10	18
Acral	6	1	7
Membros inferiores	8	8	16
Total	51	47	98



3.3 Organização e Apresentação de Dados (cont.)

3.3.2 Gráficos (cont.)

Representação gráfica para variáveis quantitativas

Exemplo 3.10: Nível de colesterol (cont.)

Para os dados do Exemplo 3.2, $n = 78$, o maior valor é 479 e o menor, 118. Escolhas razoáveis para o número e o tamanho de classes são 8 e 50, aproximação de $(479-118)/8=45,125$. Escolhendo o início da primeira classe 100, obtemos a Tabela 3.10.

Tabela 3.10: Distribuição do nível de colesterol

Nível de colesterol	Frequência absoluta		Frequência relativa	
	simples	acumulada	simples	acumulada
100 † 150	2	2	0,03	0,03
150 † 200	24	26	0,31	0,33
200 † 250	35	61	0,45	0,78
250 † 300	14	75	0,18	0,96
300 † 350	1	76	0,01	0,97
350 † 400	1	77	0,01	0,99
400 † 450	0	77	0,00	0,99
450 † 500	1	78	0,01	1,00
Total	78	-	1,00	-

3.3 Organização e Apresentação de Dados (cont.)

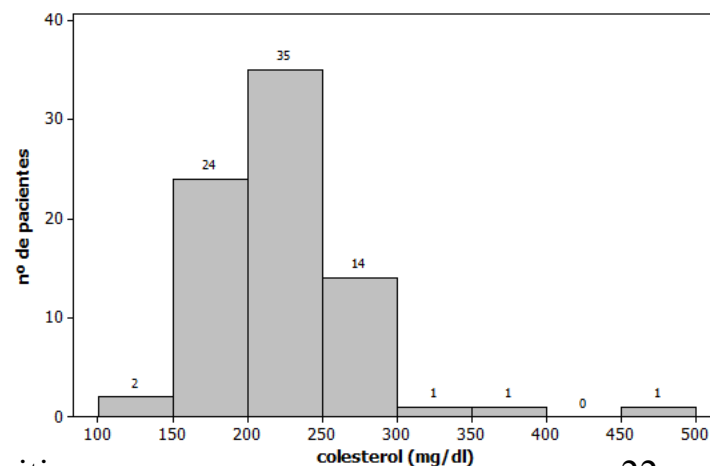
3.3.2 Gráficos (cont.)

Representação gráfica para variáveis quantitativas (cont.)

Exemplo 3.10: Nível de colesterol (cont.)

Com uma simples inspeção do histograma da Figura 3.2 vemos que a grande maioria dos indivíduos tem colesterol em torno de 225. Temos também uma boa descrição de como os diferentes níveis se distribuem em torno deste valor.

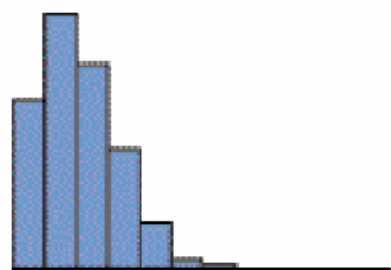
Figura 3.2: Histograma para o nível de colesterol



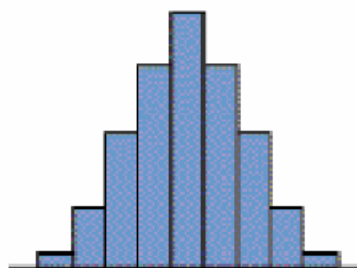
3.3 Organização e Apresentação de Dados (cont.)

3.3.2 Gráficos (cont.)

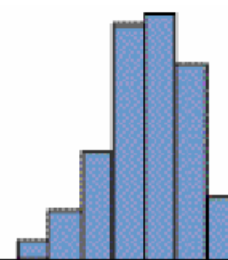
Formas da distribuição de uma variável quantitativa



**Assimétrica
(concentração
à esquerda)**



Simétrica



**Assimétrica
(concentração à
direita)**

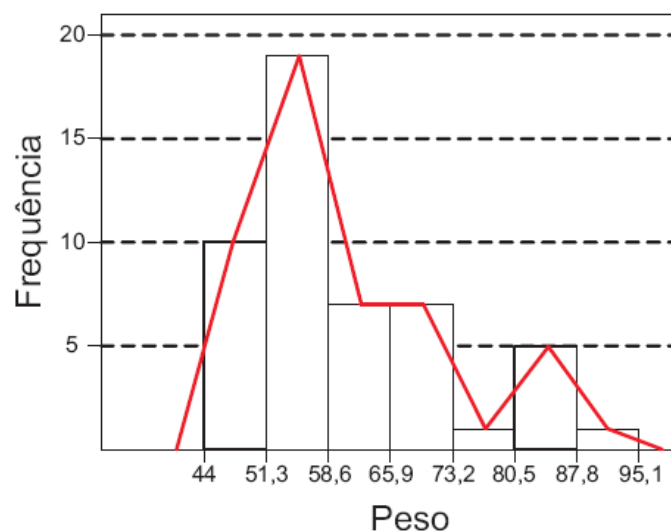


3.3 Organização e Apresentação de Dados (cont.)

3.3.2 Gráficos (cont.)

Polígono de frequências

Obtido unindo-se os pontos correspondentes às frequências (relativas ou absolutas) das diversas classes centradas nos respectivos pontos médios.



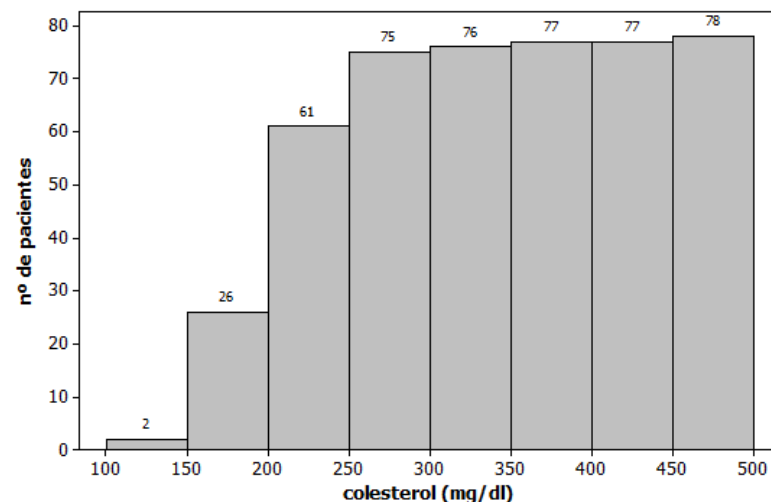
3.3 Organização e Apresentação de Dados (cont.)

3.3.2 Gráficos (cont.)

Diagrama de frequências acumuladas

É uma variação do histograma, em que a altura de cada barra é o número total de observações que é menor que o limite superior da classe.

Figura 3.2: Diagrama de frequências acumuladas para o nível de colesterol

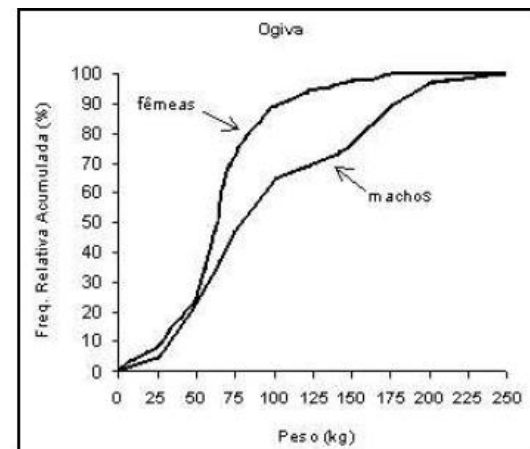
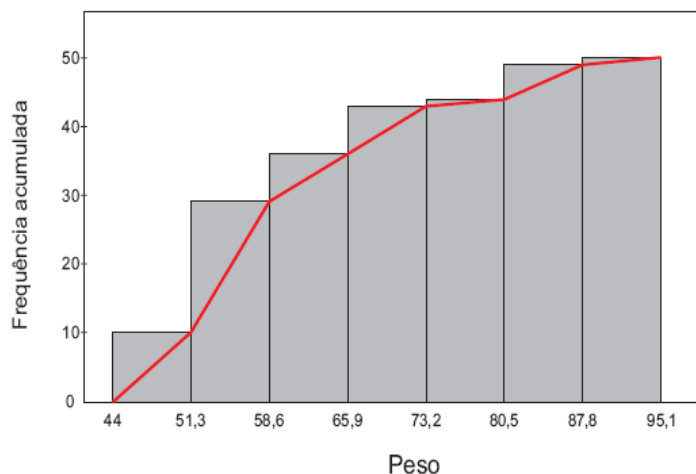


3.3 Organização e Apresentação de Dados (cont.)

3.3.2 Gráficos (cont.)

Ogiva

É construída a partir do gráfico de frequências acumuladas e definida pela diagonal formada por segmentos de reta unindo o ponto inicial inferior da primeira barra e os pontos finais de cada classe. São úteis para comparar diferentes grupos.



3.3 Organização e Apresentação de Dados (cont.)

3.3.2 Gráficos (cont.)

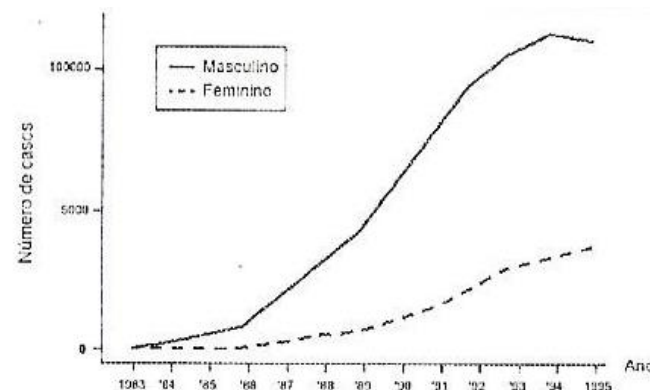
Representação gráfica de dados temporais

Dados coletados ao longo do tempo são muito comuns. São denominados dados temporais ou longitudinais, ou ainda de medidas repetidas.

Exemplo 3.12: AIDS no Brasil no período de 1908 a 1992

O numero de casos de AIDS notificados ao Ministério da Saúde é periodicamente publicado. Na Figura 3.4 vemos o número anual de casos notificados no Brasil, de 1983 a 1995, segundo o sexo.

Figura 3.4: Número de casos notificados de AIDS



3.4 Síntese Numérica

A distribuição de **variáveis quantitativas** pode ter suas características descritas e resumidas em medidas de:

Tendência central: (ou localização)	"qual o tamanho dos números envolvidos?"	moda, mediana, média.
---	--	-----------------------

Variabilidade: (ou dispersão)	"quanta variação existe?"	amplitude total, variância / desvio-padrão, coeficiente de variação.
---	---------------------------	--

Posição:	"qual o k-ésimo valor?"	Percentil, escore padronizado
-----------------	-------------------------	-------------------------------

3.4 Síntese Numérica (cont.)

3.4.1 Medidas de Tendência Central

Média aritmética simples

A média aritmética simples de n observações $(x_1, x_2, \dots, x_n)$ é denotada por $\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$

Exemplo 3.14: Peso de recém-nascidos

Para exemplificar os cálculos, vamos utilizar os seguintes pesos em *kg* de 10 recém-nascidos:

3,2 3,2 2,8 2,1 2,9 3,1 3,2 3,0 3,5 4,0

A média dos pesos do recém-nascidos é:

$$\bar{x} = \frac{3,2 + 3,2 + 2,8 + 2,1 + 2,9 + 3,1 + 3,2 + 3,0 + 3,5 + 4,0}{10} = 3,10$$

Note que a média é normalmente apresentada com um dígito a mais que os valores individuais.

3.4 Síntese Numérica (cont.)

3.4.1 Medidas de Tendência Central (cont.)

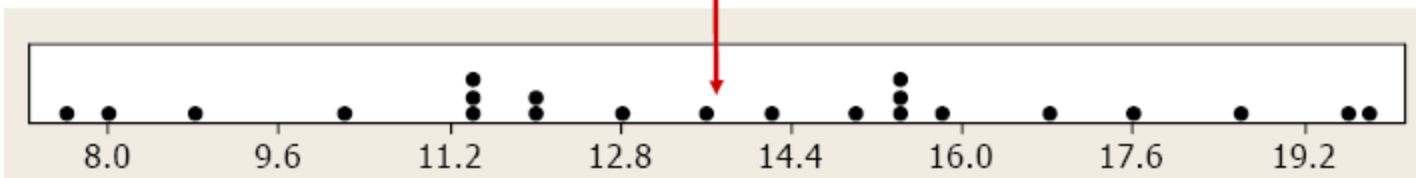
Média aritmética simples (cont.)

Exemplo peso de 22 crianças:

19.8 15.4 11.9 16.7 11.9 10.1 15.4 11.4 15.8 11.4 14.9
18.5 11.4 19.5 7.5 7.9 14.1 8.8 15.4 12.7 17.6 13.6

$x_1 = 19.8$ $x_2 = 15.4$ $x_3 = 11.9$ $x_{21} = 17.6$ $x_{22} = 13.6$

$$\bar{x} = \frac{301.7}{22} = 13.71$$



3.4 Síntese Numérica (cont.)

3.4.1 Medidas de Tendência Central (cont.)

Mediana

A mediana, denotada por $\tilde{x}$, é apropriada para representar variáveis com distribuições assimétricas. É aquele valor que divide os valores observados ao meio, quando colocados em ordem crescente.

n é ímpar

Existe um só
"valor do
meio"

Dados: { 3.3 ; 2.5 ; 5.6 ; 4.3 ; 3.2 }.

Ordenando os valores :

{ 2.5 ; 3.2 ; 3.3 ; 4.3 ; 5.6 }.

A mediana é o valor 3.3.

n é par

Existem dois
"valores do
meio"

Dados (3 ; 4.5 ; 5.5 ; 2.5 ; 1.3 ; 6).

Ordenando os valores:

(1.3 ; 2.5 ; 3 ; 4.5 ; 5.5 ; 6)

A mediana é $(3 + 4.5)/2 = 3.75$.

3.4 Síntese Numérica (cont.)

3.4.1 Medidas de Tendência Central (cont.) Mediana (cont.)

Exemplo 3.14: Peso de recém-nascidos

Para exemplificar os cálculos, vamos utilizar os seguintes pesos em *kg* de 10 recém-nascidos:

3,2 3,2 2,8 2,1 2,9 3,1 3,2 3,0 3,5 4,0

Ordenando esses valores, temos:

2,1 2,8 2,9 3,0 **3,1** **3,2** 3,2 3,2 3,5 4,0

Portanto, como temos um número par de observações, há dois valores centrais (o **3,1** e o **3,2**). A mediana é a média aritmética simples entre esses dois valores:

$$\tilde{x} = \frac{3,1 + 3,2}{2} = 3,15$$

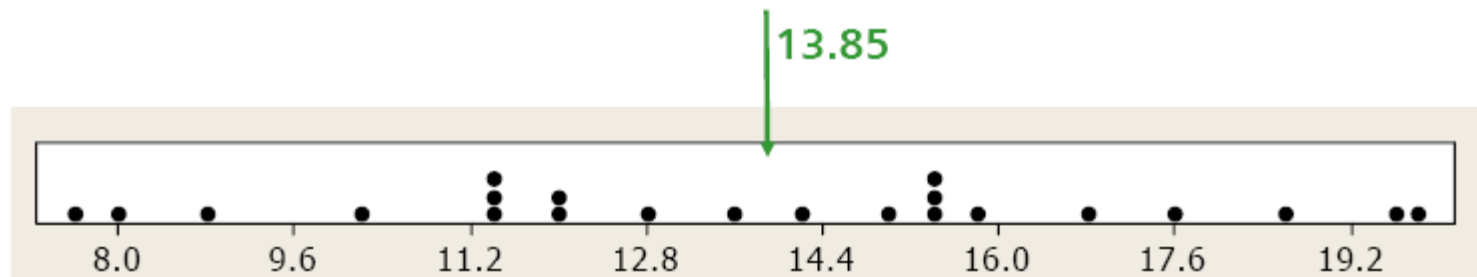
3.4 Síntese Numérica (cont.)

3.4.1 Medidas de Tendência Central (cont.) Mediana (cont.)

Exemplo peso de 22 crianças ($n = 22$, par):

7.5 7.9 8.8 10.1 11.4 11.4 11.4 11.9 11.9 12.7 13.6
14.1 14.9 15.4 15.4 15.4 15.8 16.7 17.6 18.5 19.5 19.8

A mediana é $(13.6 + 14.1)/2 = 13.85$



3.4 Síntese Numérica (cont.)

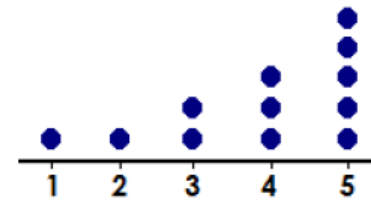
3.4.1 Medidas de Tendência Central (cont.)

Moda

A moda é o valor mais frequente.

Moda = 5

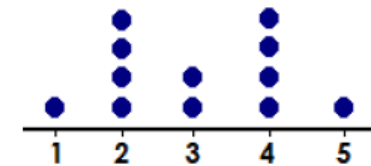
Unimodal



1ª Moda = 2

2ª Moda = 4

Bimodal



Nenhuma moda

Amodal



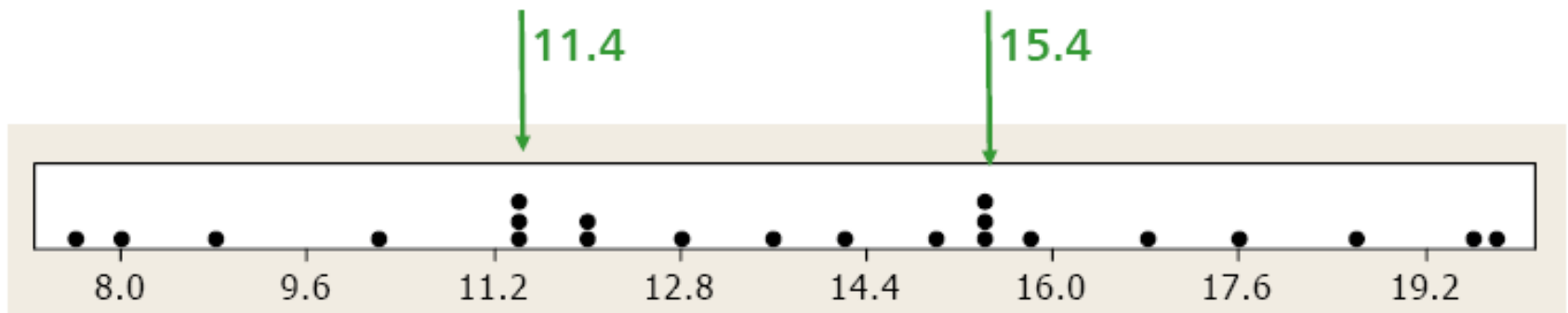
3.4 Síntese Numérica (cont.)

3.4.1 Medidas de Tendência Central (cont.)

Moda (cont.)

Exemplo peso de 22 crianças:

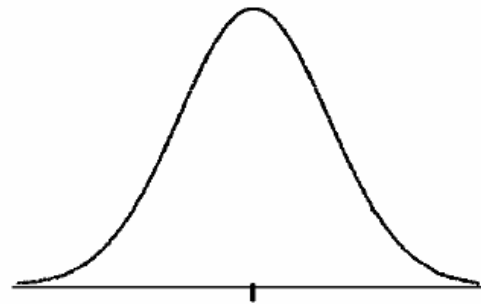
Há duas modas.



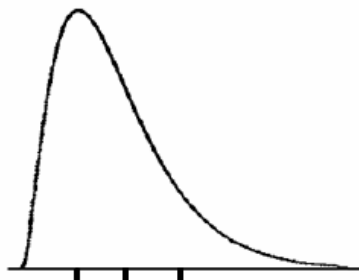
3.4 Síntese Numérica (cont.)

3.4.1 Medidas de Tendência Central (cont.)

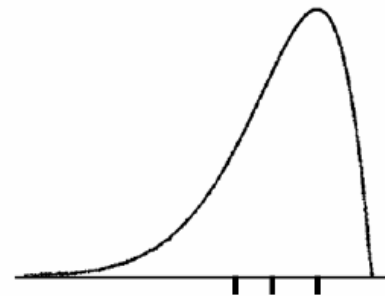
Forma da Distribuição X Medidas de Tendência Central



$\text{moda} = \text{mediana} = \text{média}$



$\text{moda} < \text{mediana} < \text{média}$



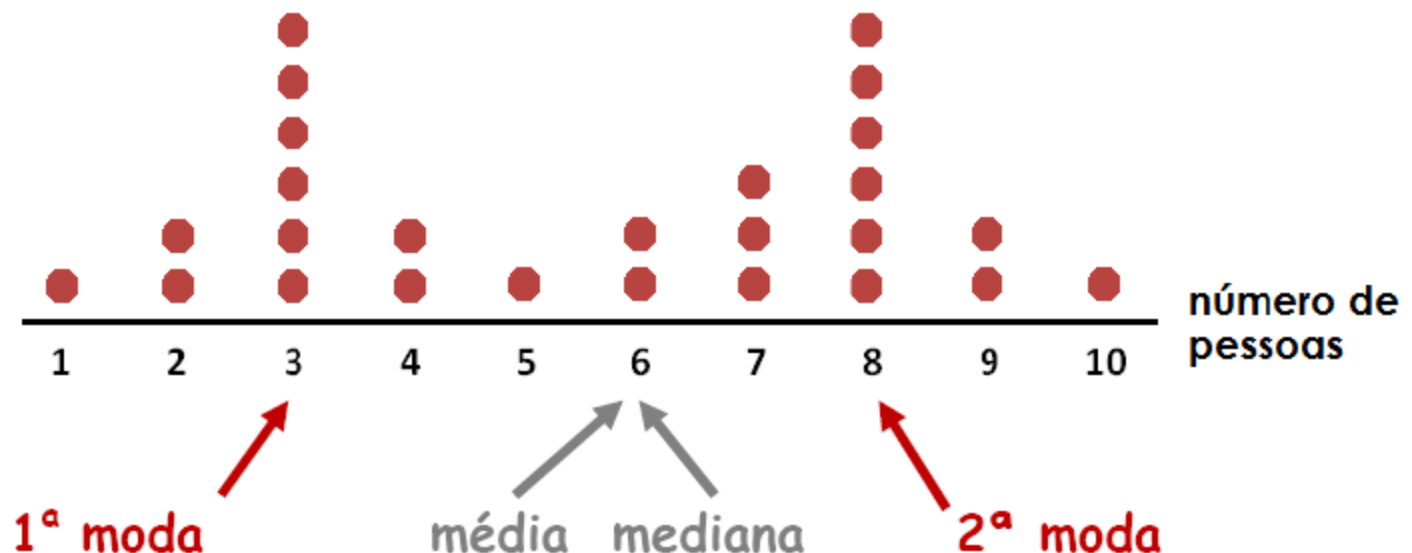
$\text{média} < \text{mediana} < \text{moda}$

3.4 Síntese Numérica (cont.)

3.4.1 Medidas de Tendência Central (cont.)

A moda é mais adequada para variáveis com distribuições multimodais.

Exemplo tamanho de família ideal:



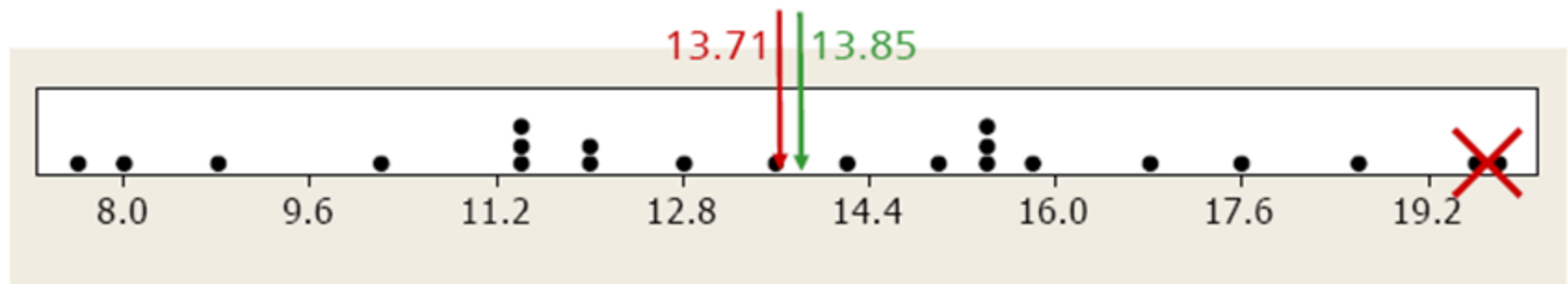
3.4 Síntese Numérica (cont.)

3.4.1 Medidas de Tendência Central (cont.)

A mediana é menos sensível a valores extremos dos dados (muito baixos ou muito altos).

Exemplo peso de 22 crianças:

Se forem retirados os dois maiores valores, a mediana do novo conjunto de dados será 13,60 (antes era 13,85). A média do novo conjunto de dados será 13,42 (antes era 13,71).



A mediana do novo conjunto de dados será 13.60.

A média do novo conjunto de dados será 13.42.

3.4 Síntese Numérica (cont.)

3.4.1 Medidas de Tendência Central (cont.)

A mediana pode ser calculada se é conhecida apenas a ordem de algum dos dados.

Exemplo tempo (em min.) até um medicamento fazer efeito:

<20 23 34 45 56 **67** 70 87 90 >96 >120

3.4 Síntese Numérica (cont.)

3.4.2 Medidas de Variabilidade

- Identificar o elemento típico de um conjunto de dados *não* é suficiente para caracterizá-lo;
- Dois conjuntos de dados podem ter o mesmo elemento típico, mas serem bem *diferentes* um do outro;
- É necessário quantificar a dispersão em torno do elemento típico, ou seja, quantificar a *variabilidade* de um conjunto de dados

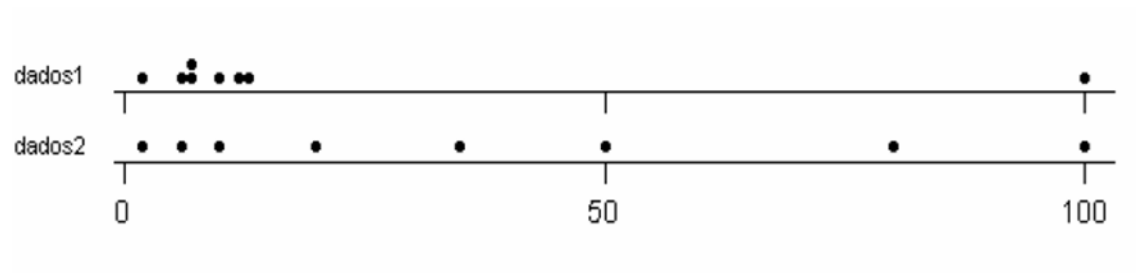
3.4 Síntese Numérica (cont.)

3.4.2 Medidas de Variabilidade (cont.)

Amplitude

A amplitude é uma medida simples de variabilidade, porém, é muito *grosseira*, pois só considera os valores extremos do conjunto de dados.

Exemplo:



$$AT_1 = 100 - 2 = 98$$

$$AT_2 = 100 - 2 = 98$$

3.4 Síntese Numérica (cont.)

3.4.2 Medidas de Variabilidade (cont.)

Melhor seria uma medida de variabilidade (em relação ao valor típico) que considerasse **todos** os dados. Por exemplo:

- $(x_i - \bar{x})$, é a **distância** do i -ésimo valor em relação à média dos dados;
- $\frac{\sum_{i=1}^n (x_i - \bar{x})}{n}$, é a distância **típica** dos valores em relação à média;

O problema é que o somatório abaixo é sempre **zero**, qualquer que seja a média:

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

3.4 Síntese Numérica (cont.)

3.4.2 Medidas de Variabilidade (cont.)

Exemplo numérico:

i	x_i	$(x_i - \bar{x})$
1	1,0	(1,0 - 4,0)= -3,00
2	1,5	(1,5 - 4,0)= -2,50
3	2,0	(2,0 - 4,0)= -2,00
4	3,5	(3,5 - 4,0)= -0,50
5	4,0	(4,0 - 4,0)= 0,00
6	4,0	(4,0 - 4,0)= 0,00
7	4,5	(4,5 - 4,0)= 0,50
8	6,0	(6,0 - 4,0)= 2,00
9	6,5	(6,5 - 4,0)= 2,50
10	7,0	(7,0 - 4,0)= 3,00
Soma	40,0	0,00
Média	4,00	0,00

Os valores se anulam
(sinais positivos e
negativos).

Como resolver esse
problema mantendo a
ideia de distância até a
média?

3.4 Síntese Numérica (cont.)

3.4.2 Medidas de Variabilidade (cont.)

Variância

É a medida de variabilidade do dados, definida como o desvio *quadrático* médio em torno da média:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}, \quad (3.2)$$

que pode ser *reescrita* como

$$s^2 = \frac{\left[\sum_{i=1}^n (x_i^2) \right] - n(\bar{x})^2}{n-1}, \quad (3.3)$$

ou ainda (forma *preferida* para uso em calculadoras)

$$s^2 = \frac{\left[\sum_{i=1}^n (x_i^2) \right] - \frac{\left[\sum_{i=1}^n (x_i) \right]^2}{n}}{n-1}. \quad (3.4)$$

3.4 Síntese Numérica (cont.)

3.4.2 Medidas de Variabilidade (cont.)

Desvio-padrão

- Note que a unidade de medida de s^2 é a unidade de medida das observações elevada ao **quadrado**;
- Por exemplo, se as **observações** foram medidas em cm , a unidade de s^2 é cm^2 ;
- Logo, para obter uma medida de variabilidade com a **mesma unidade** das observações, extrai-se a raiz quadrada:

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

3.4 Síntese Numérica (cont.)

3.4.2 Medidas de Variabilidade (cont.)

Exemplo 3.15: Peso de recém-nascidos (cont.)

Para o cálculo da variância dos pesos de recém-nascidos do Exemplo 3.15, organizam-se os dados na Tabela 3.12. A Tabela 3.13 apresenta os cálculos da variância através das três formulas.

Tabela 3.12: Cálculos para obtenção do desvio-padrão

i	x_i	x_i^2	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	3,2	10,24	(3,20 - 3,10)= 0,10	0,010
2	3,2	10,24	(3,20 - 3,10)= 0,10	0,010
3	2,8	7,84	(2,80 - 3,10)= -0,30	0,090
4	2,1	4,41	(2,10 - 3,10)= -1,00	1,000
5	2,9	8,41	(2,90 - 3,10)= -0,20	0,040
6	3,1	9,61	(3,10 - 3,10)= 0,00	0,000
7	3,2	10,24	(3,20 - 3,10)= 0,10	0,010
8	3,0	9,00	(3,00 - 3,10)= -0,10	0,010
9	3,5	12,25	(3,50 - 3,10)= 0,40	0,160
10	4,0	16,00	(4,00 - 3,10)= 0,90	0,810
Soma	31,00	98,24	0,00	2,140
Média	3,10			

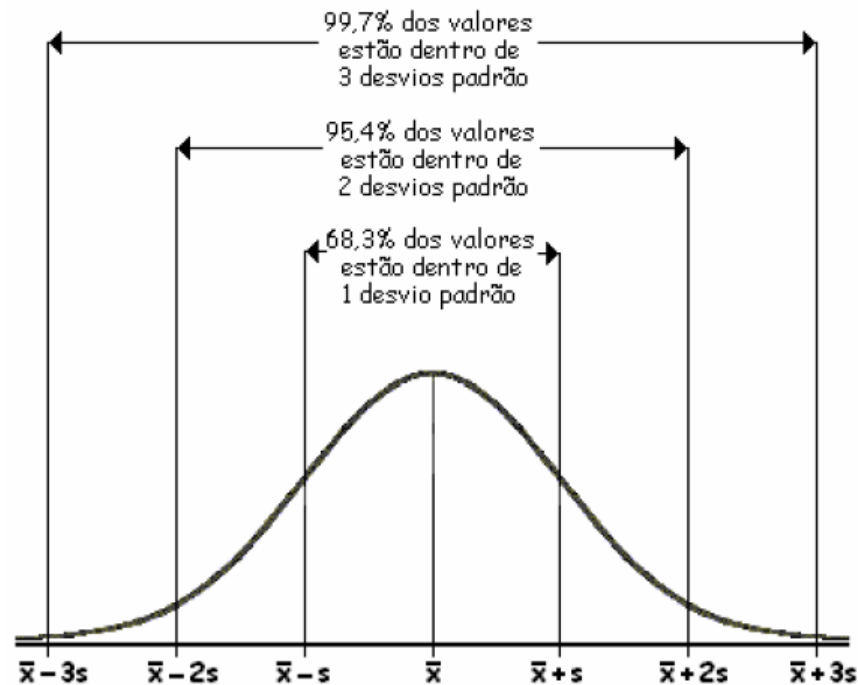
Tabela 3.13: Cálculos da variância e desvio-padrão

Fórmula	Variância	Desvio-padrão
(3.2) $s^2 = \frac{2,14}{9} = 0,24$		$s = \sqrt{0,24} = 0,49$
(3.3) $s^2 = \frac{98,24 - 10 \times (3,10)^2}{9} = 0,24$		$s = \sqrt{0,24} = 0,49$
(3.4) $s^2 = \frac{98,24 - (31)^2/10}{9} = 0,24$		$s = \sqrt{0,24} = 0,49$

3.4 Síntese Numérica (cont.)

3.4.2 Medidas de Variabilidade (cont.)

Regra do Desvio-Padrão para Dados com Distribuição de Frequências Simétrica



3.4 Síntese Numérica (cont.)

3.4.2 Medidas de Variabilidade (cont.)

Um desvio-padrão igual a 10 é grande ou pequeno ?

$s=10$ significa **muita** dispersão se $\bar{X}=100$

$s=10$ significa **pouca** dispersão se $\bar{X}=1000$

$$\frac{10}{100} = 0.1 \text{ (10\%)}$$

$$\frac{10}{1000} = 0.01 \text{ (1\%)}$$

Para termos idéia da magnitude do valor do desvio-padrão, é necessário verificar o quanto ele “ocupa” da escala de medida, representada pela média.

3.4 Síntese Numérica (cont.)

3.4.3 Coeficiente de Variação

- Uma pergunta que pode surgir é se um desvio-padrão é grande ou pequeno, que é uma questão relevante, por exemplo, na avaliação da precisão de métodos;
- Pode-se obter um índice relativo de dispersão comparando-se o desvio-padrão (s) com a média ($\bar{x}$);
- A medida utilizada é denominada coeficiente de variação (CV) e é definida por:

$$CV = \frac{s}{\bar{x}};$$

- O coeficiente de variação é adimensional, isto é, é um número puro, usualmente expresso em porcentagem.

3.4 Síntese Numérica (cont.)

3.4.3 Coeficiente de Variação (cont.)

Exemplo 3.16: Comparação do colesterol em dois grupos

Em um grupo de jovens médicos residentes, obteve-se, ao medir o colesterol, a média de 205 *mg/dl* e um desvio-padrão de 22 *mg/dl*.

Para o grupo de médicos especialistas, entretanto, a média obtida foi de 244 *mg/dl* e desvio-padrão de 45 *mg/dl*.

O grupo de médicos mais idosos apresenta não só a média mais alta como também a maior variabilidade em torno da média.

O coeficiente de variação capta essa diferença. Nesse caso, o coeficiente de variação é 10,7% para os residentes e 18,4%, para os especialistas.

3.4 Síntese Numérica (cont.)

3.4.3 Coeficiente de Variação (cont.)

Exemplo: Comparação de homogeneidade entre grupos e variáveis

Idade (em anos) de motoristas e cobradores.

Grupo	n	Média	DP	CV
Motoristas	150	35.6	5.08	0.143 (14.3%)
Cobradores	50	22.6	3.11	0.137 (13.7%)

Idade, tempo de profissão e salário de motoristas.

Variável	Média	DP	CV
Idade	35.6 anos	5.08 anos	0.143 (14.3%)
Tempo de profissão	6.5 anos	2.98 anos	0.458 (45.8%)
Salário	537.52 reais	25.34 reais	0.047 (4.7%)

3.4 Síntese Numérica (cont.)

3.4.4 Escore Padronizado

- Na seção anterior foi visto como relacionar a média e o desvio-padrão, para caracterizar a homogeneidade de um grupo;
- Esta seção também trata de uma forma de relacionar essas duas estatísticas (média e desvio-padrão), mas para cada indivíduo;
- A ideia é que, na comparação dos resultados de dois indivíduos, é importante a padronização em relação ao grupo;
- Por exemplo, um estudante que
 - tira 7 em uma prova cuja média da turma foi 5
 - apresentou melhor rendimento do que
 - numa prova em que tirou 8, mas a média da classe foi 9.

3.4 Síntese Numérica (cont.)

3.4.4 Escore Padronizado (cont.)

Exemplo 3.14: Peso de recém-nascidos (cont.)

Considere como referência os mesmos 10 bebês do Exemplo 3.14 e suponha que nasça um bebe de 4,1 kg. Este peso está praticamente dois desvios-padrão acima da média porque:

$$\frac{x_i - \bar{x}}{s} = \frac{4,1 - 3,1}{s} \cong 2,04$$

Definição:

Sejam $x_1, x_1, \dots, x_n$, os dados observados de uma amostra de tamanho n , $\bar{x}$ e s , a média e o desvio-padrão, respectivamente. Define-se z_i o *escore padronizado* da observação x_i :

$$z_i = \frac{x_i - \bar{x}}{s}$$

3.4 Síntese Numérica (cont.)

3.4.4 Escore Padronizado (cont.)

Exemplo 3.14: Peso de recém-nascidos (cont.)

- O escore padronizado fornece o número de desvios-padrão que a observação dista da média da amostra;
- Quando um escore padronizado deve ser considerado grande?
- Para a maioria dos conjuntos de dados:
 - 80% dos dados estão no intervalo centrado na média, com amplitude de 4 desvios-padrão;
 - poucas observações estão além de dois desvios-padrão da média e
 - raramente há uma observação além de 3 desvios-padrão.
- Diante disso:
 - um bebê de 4,1 kg tem um peso muito diferente dos outros dez;
 - não se espera, a priori, um valor tão grande como o que ocorreu;
 - devem-se procurar razões substantivas para este fato.

3.4 Síntese Numérica (cont.)

3.4.4 Escore Padronizado (cont.)

Exemplo 3.17: Dosagens laboratoriais

A Tabela 3.14 apresenta os resultados de exames laboratoriais solicitados a duas pacientes, mãe (A) e filha (B), com respectivamente 60 a 40 anos de idade. Também são apresentados os resultados padronizados pelo grupo de adultos do sexo feminino. A paciente A apresentou um resultado de colesterol bastante alto (2 desvios-padrão acima da média), mas o paciente B não apresentou nenhum resultado preocupante.

Tabela 3.14: Média ($\bar{x}$) e desvio-padrão (s) para adultos do sexo feminino e resultados de exames laboratoriais de mãe (A) e filha (B)

Exame	$\bar{x}$	s	Resultado original		Escore padronizado	
			A	B	A	B
Glicemia em jejum	85	12,5	90	79	0,40	-0,48
Ácido úrico	4,2	0,9	3,5	3,1	-0,78	-1,22
Triglicérides	105	30	97	66	-0,27	1,30
Colesterol total	200	25	251	185	2,04	-0,60

3.4 Síntese Numérica (cont.)

3.4.5 Percentis

O percentil de ordem K ($0 < K < 100$),
denotado por P_k ,
é o valor da variável tal que
 $K\%$ dos valores dessa variável nos dados
são menores ou iguais a ele.

Exemplo: No conjunto de notas na 1ª Fase do Vestibular,
o percentil de ordem 10 foi igual a 25 pontos:
 $P_{10} = 25$ pontos;
ou seja,
10% das notas foram menores ou iguais a 25 pontos.

3.4 Síntese Numérica (cont.)

3.4.5 Percentis (cont.)

Exemplo: o percentil de ordem 10, o P_{10} , é o valor da variável tal que 10% dos valores são menores ou iguais a ele.

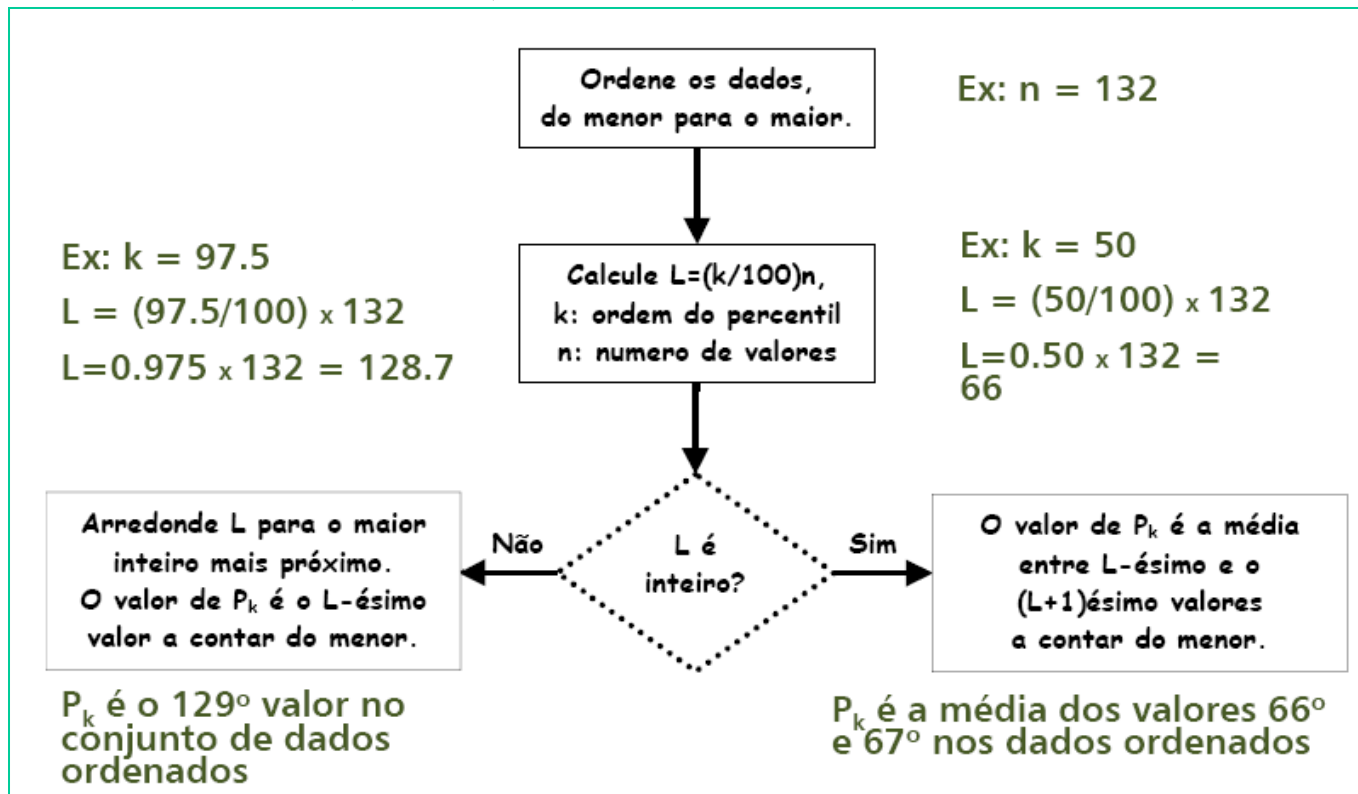
Os percentis de ordem 10, 20, 30, ... 90 dividem o conjunto de dados em **dez** partes com mesmo número de observações e são chamados de *decis*.

Os percentis de ordem 25, 50 e 75 dividem o conjunto de dados em **quatro** partes com o mesmo número de observações. Estes três percentis recebem o nome de *quartis*

Primeiro Quartil (Q1), Segundo quartil (Q2) ou mediana e Terceiro quartil (Q3).

3.4 Síntese Numérica (cont.)

3.4.5 Percentis (cont.)



Fonte: Introdução à Estatística, Mario F. Triola

3.4 Síntese Numérica (cont.)

3.4.5 Percentis (cont.)

Exemplo nota de 40 alunos:

40 41 42 42 44 47 48 48 49 49 51 52 53 58 59 62 63 64 65 66
67 68 69 70 75 76 83 83 85 86 86 87 87 88 92 93 94 95 97 98

P_{10} : 10% de 40 = 4. P_{10} = média(4º e 5º valores) = $(42 + 44)/2 = 43$.

P_{95} : 95% de 40 = 38. P_{95} = média(38º e 39º valores) = $(95 + 97)/2 = 96$.

Primeiro Quartil: 25% de 40 = 10.

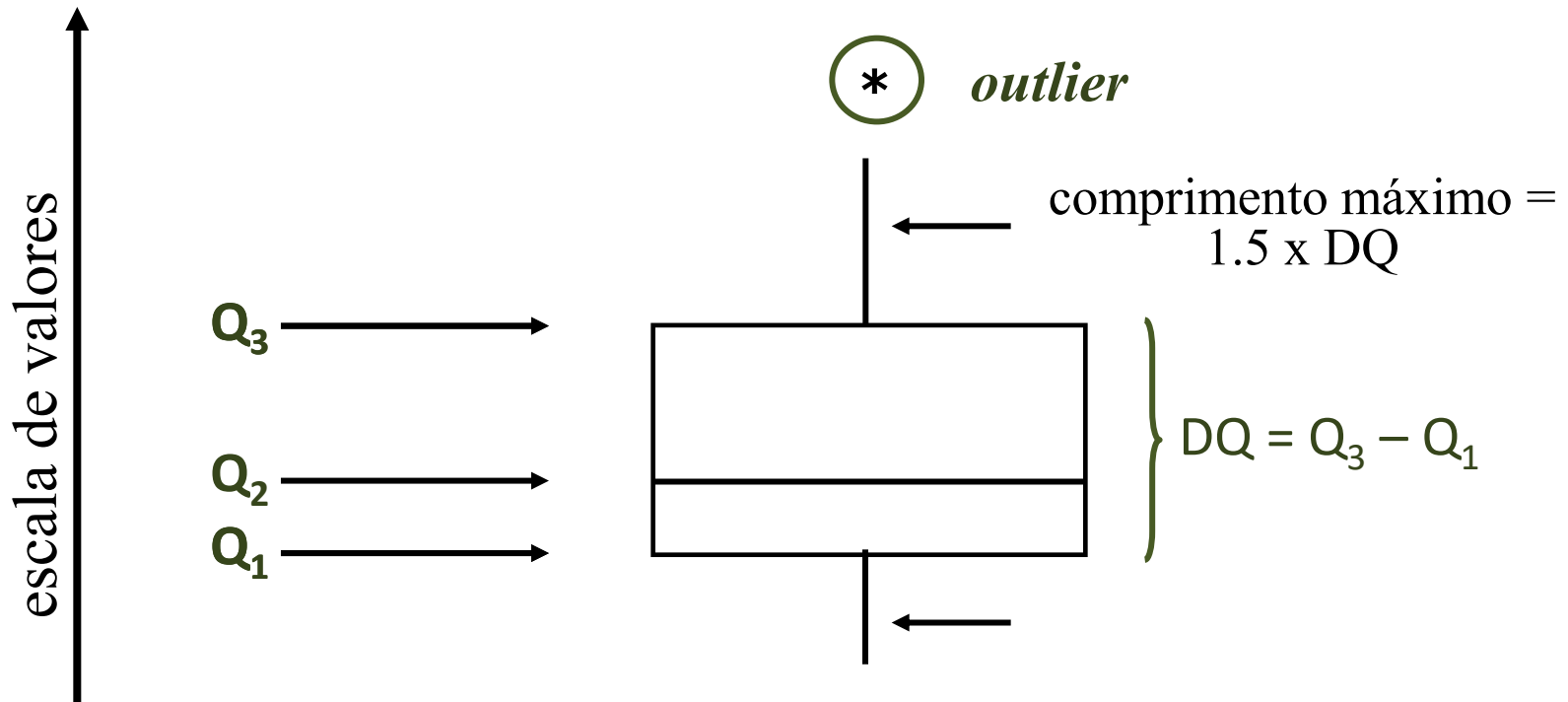
Q_1 = média(10º e 11º valores) = $(49 + 51)/2 = 50$.

Terceiro Quartil: 75% de 40 = 30.

Q_3 = média(30º e 31º valores) = $(86 + 86)/2 = 86$.

3.5 Outros Tópicos

3.5.4 *Boxplot*

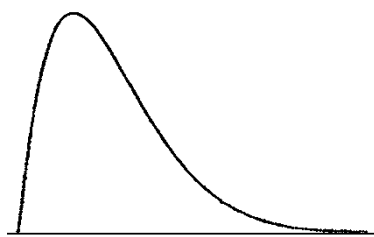


3.5 Outros Tópicos

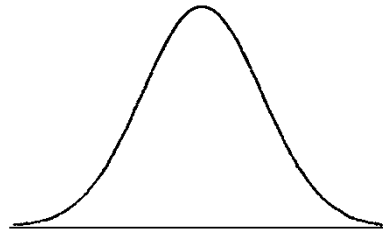
3.5.4 *Boxplot* (cont.)

Exemplo:

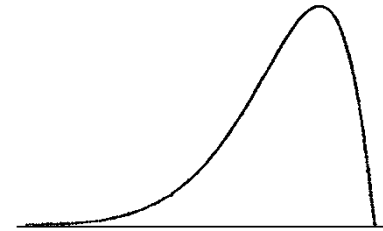
O *boxplot* e as formas básicas das distribuições de frequências:



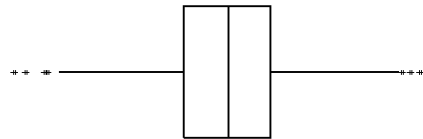
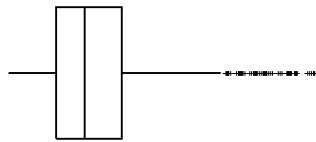
assimétrica (concentração à esquerda)



simétrica



assimétrica (concentração à direita)



3.5 Outros Tópicos

3.5.4 *Boxplot* (cont.)

Exemplo 3.22: Doença de Chagas

Mota (1992) faz um estudo comparativo de crianças filhas de mães chagásicas com filhas de mães não chagásicas (grupo controle).

Para o grupo de mães chagásicas, os dados já ordenados da dosagem de bilirrubina (*mg/dl*) são apresentados na Tabela 3.19.

Tabela 3.19: Dosagem de bilirrubina (*mg/dl*) de 46 crianças filhas de mães chagásicas

1,3	1,5	1,9	2,0	2,1	2,8	2,8	2,9	2,9	3,1
3,1	3,2	3,2	3,2	3,2	3,2	3,3	3,7	3,7	3,8
3,9	4,0	4,3	4,3	4,4	4,9	6,0	6,1	6,2	6,3
6,3	6,6	6,7	6,8	6,8	7,8	8,2	8,3	8,8	9,5
9,7	10,1	10,7	11,2	11,3	12,5				

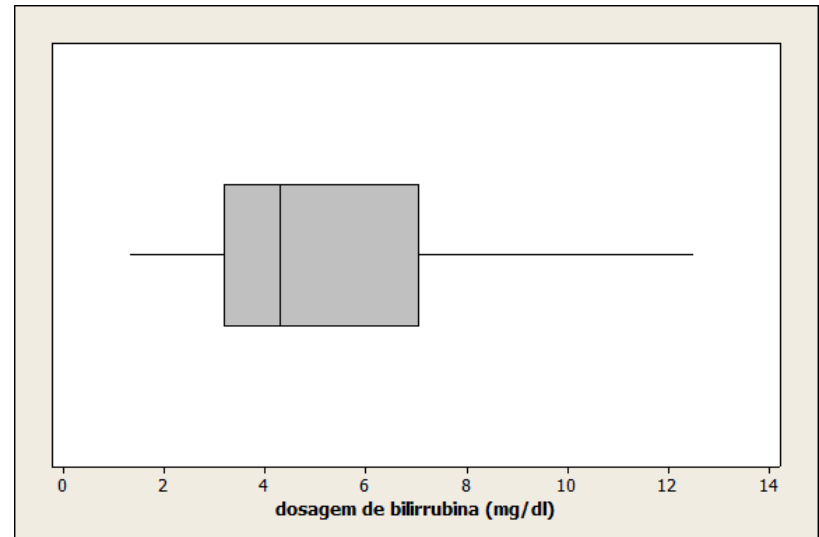
3.5 Outros Tópicos

3.5.4 *Boxplot* (cont.)

Exemplo 3.22: Doença de Chagas (cont.)

As estatísticas descritivas são $\bar{x} = 5,404$, $s = 2,967$, $x_{\min} = 1,300$, $x_{\max} = 12,500$, $Q_1 = 3,175$, $Q_2 = 4,300$, $Q_3 = 7,050$. O *boxplot* correspondente é apresentado na Figura 3.8.

Figura 3.8: *Boxplot* da dosagem de bilirrubina (mg/dl)



3.5 Outros Tópicos

3.5.5 Observações Atípicas (*outliers*)

- Observações atípicas são valores muito grandes ou muito pequenos em relação aos demais;
- Afetam muito a média e as medidas de variabilidade;
- Possíveis causas para o aparecimento:
 - Incorreta leitura, anotação ou transcrição dos dados;
 - erro na execução do experimento ou na tomada da medida;
 - mudanças não controláveis nas condições experimentais ou dos pacientes;
 - característica inerente à variável estudada.

3.5 Outros Tópicos

3.5.5 Observações Atípicas (*outliers*) (cont.)

- Como detectar?
 - gráficos (como o *boxplot*);
 - valores maiores que $Q_3 + 1,5 \times DQ$;
 - valores menores que $Q_1 - 1,5 \times DQ$.

3.5 Outros Tópicos

3.5.5 Observações Atípicas (*outliers*) (cont.)

- Outliers* podem ser “pontos externos” ou “pontos soltos”;

