Capítulo 2

Resumo de Dados

2.1 Tipos de Variáveis

Para ilustrar o que segue, consideremos o seguinte exemplo.

Exemplo 2.1. Um pesquisador está interessado em fazer um levantamento sobre alguns aspectos socioeconômicos dos empregados da seção de orçamentos da Companhia MB. Usando informações obtidas do departamento pessoal, ele elaborou a Tabela 2.1.

De modo geral, para cada elemento investigado numa pesquisa, tem-se associado um (ou mais de um) resultado correspondendo à realização de uma característica (ou características). No exemplo em questão, considerando-se a característica (variável) *estado civil*, para cada empregado pode-se associar uma das realizações, *solteiro* ou *casado* (note que poderia haver outras possibilidades, como separado, divorciado, mas somente as duas mencionadas foram consideradas no estudo). Podemos atribuir uma letra, digamos X, para representar tal variável. Observamos que o pesquisador colheu informações sobre seis variáveis:

Variável	Representação
Estado civil	X
Grau de instrução	Y
Número de filhos	Z
Salário	S
Idade	U
Região de procedência	V

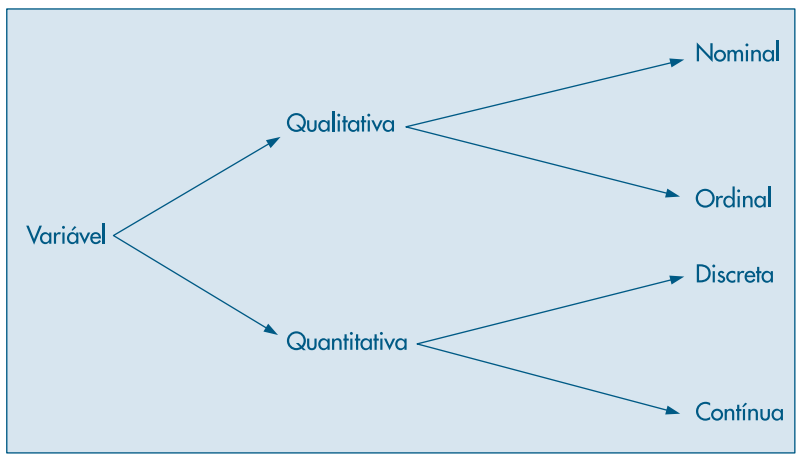
Algumas variáveis, como sexo, educação, estado civil, apresentam como possíveis realizações uma qualidade (ou atributo) do indivíduo pesquisado, ao passo que outras, como número de filhos, salário, idade, apresentam como possíveis realizações números resultantes de uma contagem ou mensuração. As variáveis do primeiro tipo são chamadas *qualitativas*, e as do segundo tipo, *quantitativas*.

Dentre as variáveis qualitativas, ainda podemos fazer uma distinção entre dois tipos: variável qualitativa *nominal*, para a qual não existe nenhuma ordenação nas possíveis realizações, e variável qualitativa *ordinal*, para a qual existe uma ordem nos seus resultados. A região de procedência, do Exemplo 2.1, é um caso de variável nominal, enquanto grau de instrução é um Exemplo de variável ordinal, pois ensinos fundamental, médio e superior correspondem a uma ordenação baseada no número de anos de escolaridade completos. A variável qualitativa *classe social*, com as possíveis realizações alta, média e baixa, é outro exemplo de variável ordinal.

De modo análogo, as variáveis quantitativas podem sofrer uma classificação dicotômica: (a) variáveis quantitativas *discretas*, cujos possíveis valores formam um conjunto finito ou enumerável de números, e que resultam, freqüentemente, de uma contagem, como por exemplo número de filhos (0, 1, 2, ...); (b) variáveis quantitativas *contínuas*, cujos possíveis valores pertencem a um intervalo de números reais e que resultam de uma mensuração, como por exemplo estatura e peso (melhor seria dizer massa) de um indivíduo.

A Figura 2.1 esquematiza as classificações feitas acima.

Figura 2.1: Classificação de uma variável.



Para cada tipo de variável existem técnicas apropriadas para resumir as informações, donde a vantagem de usar uma tipologia de identificação como a da Figura 2.1. Entretanto, verificaremos que técnicas usadas num caso podem ser adaptadas para outros.

Para finalizar, cabe uma observação sobre variáveis qualitativas. Em algumas situações podem-se atribuir valores numéricos às várias qualidades ou atributos (ou, ainda, classes) de uma variável qualitativa e depois proceder-se à análise como se esta fosse quantitativa, desde que o procedimento seja passível de interpretação.

Existe um tipo de variável qualitativa para a qual essa quantificação é muito útil: a chamada variável dicotômica. Para essa variável só podem ocorrer duas realizações, usualmente chamadas *sucesso* e *fracasso*. A variável *estado civil* no exemplo acima estaria nessa situação. Esse tipo de variável aparecerá mais vezes nos próximos capítulos.

Tabela 2.1: Informações sobre estado civil, grau de instrução, número de filhos, salário (expresso como fração do salário mínimo), idade (medida em anos e meses) e procedência de 36 empregados da seção de orçamentos da Companhia MB.

Nº	Estado civil	Grau de instrução	Nº de filhos	Salário (× sal. mín.)	Idade		Região de procedência
					anos	meses	
1	solteiro	ensino fundamental	—	4,00	26	03	interior
2	casado	ensino fundamental	1	4,56	32	10	capital
3	casado	ensino fundamental	2	5,25	36	05	capital
4	solteiro	ensino médio	—	5,73	20	10	outra
5	solteiro	ensino fundamental	—	6,26	40	07	outra
6	casado	ensino fundamental	0	6,66	28	00	interior
7	solteiro	ensino fundamental	—	6,86	41	00	interior
8	solteiro	ensino fundamental	—	7,39	43	04	capital
9	casado	ensino médio	1	7,59	34	10	capital
10	solteiro	ensino médio	—	7,44	23	06	outra
11	casado	ensino médio	2	8,12	33	06	interior
12	solteiro	ensino fundamental	—	8,46	27	11	capital
13	solteiro	ensino médio	—	8,74	37	05	outra
14	casado	ensino fundamental	3	8,95	44	02	outra
15	casado	ensino médio	0	9,13	30	05	interior
16	solteiro	ensino médio	—	9,35	38	08	outra
17	casado	ensino médio	1	9,77	31	07	capital
18	casado	ensino fundamental	2	9,80	39	07	outra
19	solteiro	superior	—	10,53	25	08	interior
20	solteiro	ensino médio	—	10,76	37	04	interior
21	casado	ensino médio	1	11,06	30	09	outra
22	solteiro	ensino médio	—	11,59	34	02	capital
23	solteiro	ensino fundamental	—	12,00	41	00	outra
24	casado	superior	0	12,79	26	01	outra
25	casado	ensino médio	2	13,23	32	05	interior
26	casado	ensino médio	2	13,60	35	00	outra
27	solteiro	ensino fundamental	—	13,85	46	07	outra
28	casado	ensino médio	0	14,69	29	08	interior
29	casado	ensino médio	5	14,71	40	06	interior
30	casado	ensino médio	2	15,99	35	10	capital
31	solteiro	superior	—	16,22	31	05	outra
32	casado	ensino médio	1	16,61	36	04	interior
33	casado	superior	3	17,26	43	07	capital
34	solteiro	superior	—	18,75	33	07	capital
35	casado	ensino médio	2	19,40	48	11	capital
36	casado	superior	3	23,30	42	02	interior

Fonte: Dados hipotéticos.

2.2 Distribuições de Freqüências

Quando se estuda uma variável, o maior interesse do pesquisador é conhecer o *comportamento* dessa variável, analisando a ocorrência de suas possíveis realizações. Nesta seção

veremos uma maneira de se dispor um conjunto de realizações, para se ter uma idéia global sobre elas, ou seja, de sua distribuição.

Exemplo 2.2. A Tabela 2.2 apresenta a *distribuição de freqüências* da variável grau de instrução, usando os dados da Tabela 2.1.

Tabela 2.2: Freqüências e porcentagens dos 36 empregados da seção de orçamentos da Companhia MB segundo o grau de instrução.

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

Fonte: Tabela 2.1.

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

Uma medida bastante útil na interpretação de tabelas de freqüências é a proporção de cada realização em relação ao total. Assim, 6/36 = 0,1667 dos empregados da companhia MB (seção de orçamentos) têm instrução superior. Na última coluna da Tabela 2.2 são apresentadas as porcentagens para cada realização da variável grau de instrução. Usaremos a notação n_i para indicar a freqüência (absoluta) de cada classe, ou categoria, da variável, e a notação $f_i = n_i/n$ para indicar a *proporção* (ou *freqüência relativa*) de cada classe, sendo n o número total de observações. As proporções são muito úteis quando se quer comparar resultados de duas pesquisas distintas. Por exemplo, suponhamos que se queira comparar a variável grau de instrução para empregados da seção de orçamentos com a mesma variável para todos os empregados da Companhia MB. Digamos que a empresa tenha 2.000 empregados e que a distribuição de freqüências seja a da Tabela 2.3.

Tabela 2.3: Freqüências e porcentagens dos 2.000 empregados da Companhia MB, segundo o grau de instrução.

Grau de instrução	Freqüência n_i	Porcentagem $100\ f_i$
Fundamental	650	32,50
Médio	1.020	51,00
Superior	330	16,50
Total	2.000	100,00

Fonte: Dados hipotéticos.

Não podemos comparar diretamente as colunas das freqüências das Tabelas 2.2 e 2.3, pois os totais de empregados são diferentes nos dois casos. Mas as colunas das porcentagens são comparáveis, pois reduzimos as freqüências a um mesmo total (no caso 100).

A construção de tabelas de freqüências para variáveis contínuas necessita de certo cuidado. Por exemplo, a construção da tabela de freqüências para a variável salário, usando o mesmo procedimento acima, não resumirá as 36 observações num grupo menor, pois não existem observações iguais. A solução empregada é agrupar os dados por faixas de salário.

Exemplo 2.3. A Tabela 2.4 dá a distribuição de freqüências dos salários dos 36 empregados da seção de orçamentos da Companhia MB por faixa de salários.

Tabela 2.4: Freqüências e porcentagens dos 36 empregados da seção de orçamentos da Companhia MB por faixa de salário.

Classe de salários	Freqüência n_i	Porcentagem $100\, f_i$
4,00 ⊢ 8,00	10	27,78
8,00 ⊢ 12,00	12	33,33
12,00 ⊢ 16,00	8	22,22
16,00 ⊢ 20,00	5	13,89
20,00 ⊢ 24,00	1	2,78
Total	36	100,00

Fonte: Tabela 2.1.

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 a tabela original (Tabela 2.1). Sem perda de muita precisão, poderíamos supor que todos os oito salários daquela classe fossem iguais ao ponto médio da referida classe, isto é, 14 (o leitor pode verificar qual o erro cometido, comparando-os com os dados originais da Tabela 2.1). Voltaremos a este assunto no Capítulo 3. 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 $a \vdash b$.

A escolha dos intervalos é arbitrária e a familiaridade do pesquisador com os dados é que lhe indicará quantas e quais classes (intervalos) devem ser usadas. Entretanto, deve-se observar que, com um pequeno número de classes, perde-se informação, e com um número grande de classes, o objetivo de resumir os dados fica prejudicado. Estes dois extremos têm a ver, também, com o grau de suavidade da representação gráfica dos dados, a ser tratada a seguir, baseada nestas tabelas. Normalmente, sugere-se o uso de 5 a 15 classes com a mesma amplitude. O caso de classes com amplitudes diferentes é tratado no Problema 10.

Problemas

1. Escalas de medidas. A seguir descrevemos outros possíveis critérios para classificar variáveis, em função da escala adotada. Observe a similaridade com a classificação apresentada anteriormente. Nossas observações são resultados de medidas feitas sobre os elementos de uma população. Existem quatro escalas de medidas que podem ser consideradas:

   Escala nominal. Nesta escala somente podemos afirmar que uma medida é diferente ou não de outra, e ela é usada para categorizar indivíduos de uma população. Um exemplo é o sexo de um indivíduo. Para cada categoria associamos um numeral diferente (letra ou número). Por exemplo, no caso de sexo: podemos associar as letras M (masculino) e F (feminino) ou 1 (masculino) e 2 (feminino). Não podemos realizar operações aritméticas aqui e uma medida de posição apropriada é a moda. (As medidas citadas nesse problema, como a média, mediana e moda, são definidas no Capítulo 3.)

   Escala ordinal. Aqui podemos dizer que uma medida é diferente e maior do que outra. Temos a situação anterior, mas as categorias são ordenadas, e a ordem dos numerais associados ordena as categorias. Por exemplo, a classe socioeconômica de um indivíduo pode ser baixa (1 ou X), média (2 ou Y) e alta (3 ou Z). Transformações que preservam a ordem não alteram a estrutura de uma escala ordinal. No exemplo acima, podemos representar as categorias por 1, 10 e 100 ou A, L e Z. Medidas de posição apropriadas são a mediana e a moda.

   Escala intervalar. Nesta escala podemos afirmar que uma medida é igual ou diferente, maior e quanto maior do que outra. Podemos quantificar a diferença entre as categorias da escala ordinal. Necessitamos de uma origem arbitrária e de uma unidade de medida. Por exemplo, considere a temperatura de um indivíduo, na escala Fahrenheit. A origem é 0^{o} F e a unidade é 1^{o} F. Transformações que preservam a estrutura dessa escala são do tipo $y = ax + b$, $a > 0$. Por exemplo, a transformação $y = 5/9\ (x - 32)$ transforma graus Fahrenheit em centígrados. Para essa escala, podemos fazer operações aritméticas, e média, mediana e moda são medidas de posição apropriadas.

   Escala razão. Dadas duas medidas nessa escala, podemos dizer se são iguais, ou se uma é diferente, maior, quanto maior e quantas vezes a outra. A diferença com a escala intervalar é que agora existe um zero absoluto. A altura de um indivíduo é um exemplo de medida nessa escala. Se ela for medida em centímetros (cm), 0 cm é a origem e 1 cm é a unidade de medida. Um indivíduo com 190 cm é duas vezes mais alto do que um indivíduo com 95 cm, e esta relação continua a valer se usarmos 1 m como unidade. Ou seja, a estrutura da escala razão não é alterada por transformações da forma $y = cx$, $c > 0$. Por exemplo, $y = x/100$ transforma cm em m. As estatísticas apropriadas para a escala intervalar são também apropriadas para a escala razão.

   Para cada uma das variáveis abaixo, indique a escala usualmente adotada para resumir os dados em tabelas de freqüências:

   (a) Salários dos empregados de uma indústria.
   (b) Opinião de consumidores sobre determinado produto.
   (c) Número de respostas certas de alunos num teste com dez itens.
   (d) Temperatura diária da cidade de Manaus.
   (e) Porcentagem da receita de municípios aplicada em educação.
   (f) Opinião dos empregados da Companhia MB sobre a realização ou não de cursos obrigatórios de treinamento.
   (g) QI de um indivíduo.

2. Usando os dados da Tabela 2.1, construa a distribuição de freqüências das variáveis:
   (a) Estado civil.
   (b) Região de procedência.
   (c) Número de filhos dos empregados casados.
   (d) Idade.
3. Para o Conjunto de Dados 1 (CD-Brasil), construa a distribuição de freqüências para as variáveis população urbana e densidade populacional.

2.3 Gráficos

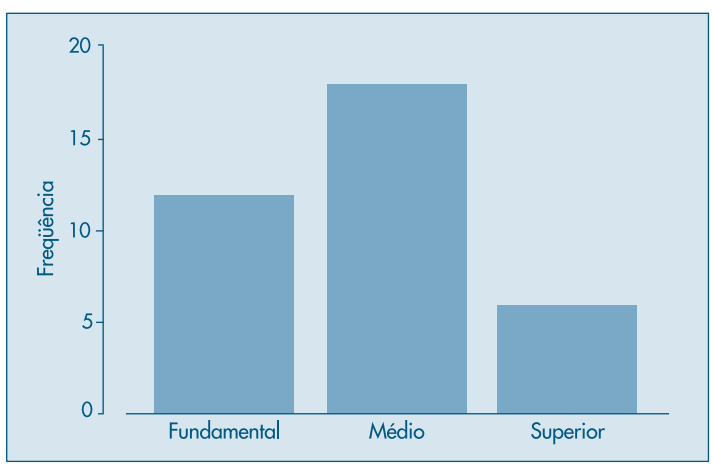
Como já salientamos no Capítulo 1, a representação gráfica da distribuição de uma variável tem a vantagem de, rápida e concisamente, informar sobre sua variabilidade. Existem vários gráficos que podem ser utilizados e abordaremos aqui os mais simples para variáveis quantitativas. No Capítulo 3, voltaremos a tratar deste assunto, em conexão com medidas associadas à distribuição de uma variável.

2.3.1 Gráficos para Variáveis Qualitativas

Existem vários tipos de gráficos para representar variáveis qualitativas. Vários são versões diferentes do mesmo princípio, logo nos limitaremos a apresentar dois deles: gráficos em barras e de composição em setores ("pizza" ou retângulos).

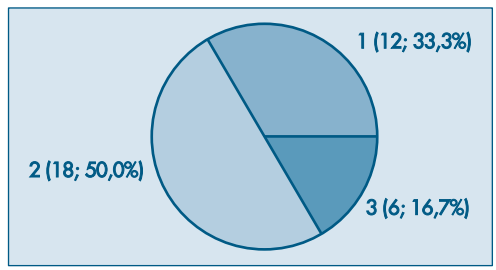
Exemplo 2.4. Tomemos como ilustração a variável Y: grau de instrução, exemplificada nas Tabelas 2.2 e 2.3. O gráfico em barras consiste em construir retângulos ou barras, em que uma das dimensões é proporcional à magnitude a ser representada (n_i ou f_i), sendo a outra arbitrária, porém igual para todas as barras. Essas barras são dispostas paralelamente umas às outras, horizontal ou verticalmente. Na Figura 2.2 temos o gráfico em barras (verticais) para a variável Y.

Figura 2.2: Gráfico em barras para a variável Y: grau de instrução.



Já o gráfico de composição em setores, sendo em forma de "pizza" o mais conhecido, destina-se a representar a composição, usualmente em porcentagem, de partes de um todo. Consiste num círculo de raio arbitrário, representando o todo, dividido em setores, que correspondem às partes de maneira proporcional. A Figura 2.3 mostra esse tipo de gráfico para a variável *Y*. Muitas vezes é usado um retângulo no lugar do círculo, para indicar o todo.

Figura 2.3: Gráfico em setores para a variável *Y*: grau de instrução.



1 = Fundamental, 2 = Médio e 3 = Superior

2.3.2 Gráficos para Variáveis Quantitativas

Para variáveis quantitativas podemos considerar uma variedade maior de representações gráficas.

Exemplo 2.5. Considere a distribuição da variável *Z*, número de filhos dos empregados casados da seção de orçamentos da Companhia MB (Tabela 2.1). Na Tabela 2.5 temos as freqüências e porcentagens.

Além dos gráficos usados para as variáveis qualitativas, como ilustrado na Figura 2.4, podemos considerar um gráfico chamado *gráfico de dispersão unidimensional*, como o da Figura 2.5 (a), em que os valores são representados por pontos ao longo da reta (provida de uma escala). Valores repetidos são acompanhados por um número que indica as repetições. Outra possibilidade é considerar um gráfico em que os valores repetidos são "empilhados", um em cima do outro, como na Figura 2.5 (b). Pode-se também apresentar o ponto mais alto da pilha, como aparece na Figura 2.5 (c).

Figura 2.4: Gráfico em barras para a variável *Z*: número de filhos.

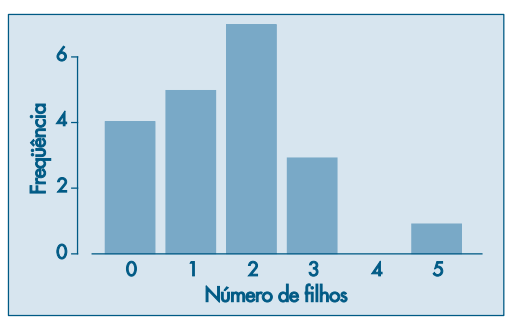
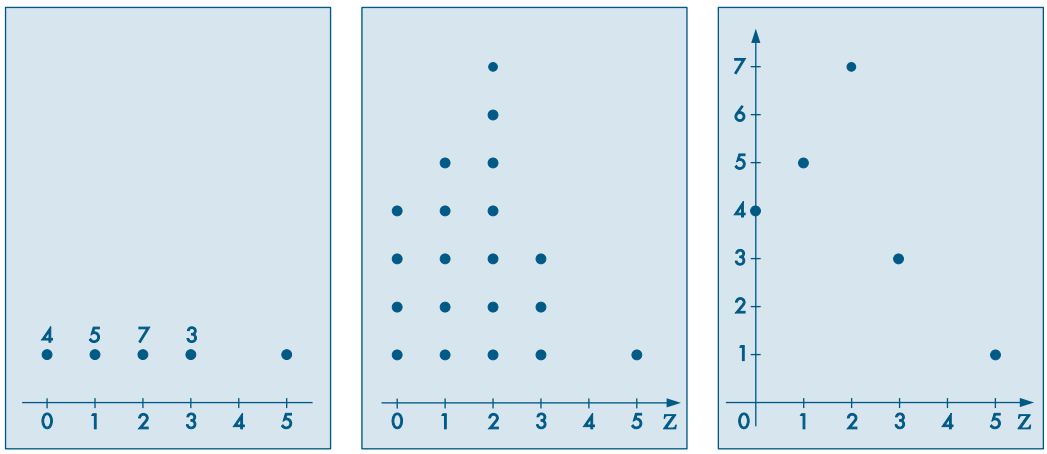


Figura 2.5: Gráficos de dispersão unidimensionais para a variável Z: número de filhos.



Para variáveis quantitativas contínuas, necessita-se de alguma adaptação, como no exemplo a seguir.

Tabela 2.5: Freqüências e porcentagens dos empregados da seção de orçamentos da Companhia MB, segundo o número de filhos.

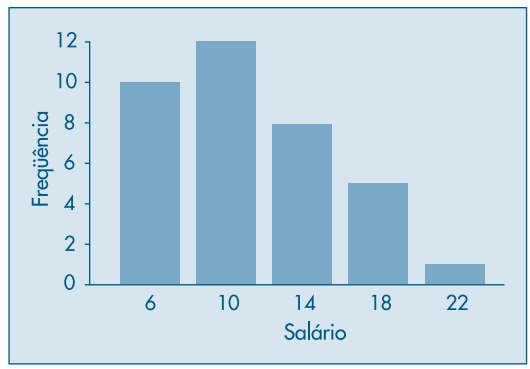
Nº de filhos z_i	Freqüência n_i	Porcentagem $100\, f_i$
0	4	20
1	5	25
2	7	35
3	3	15
5	1	5
Total	20	100

Fonte: Tabela 2.1.

Exemplo 2.6. Queremos representar graficamente a distribuição da variável S, salário dos empregados da seção de orçamentos da Companhia MB. A Tabela 2.4 fornece a distribuição de freqüências de S. Para fazer uma representação similar às apresentadas anteriormente, devemos usar o artifício de aproximar a variável contínua por uma variável discreta, sem perder muita informação. Isto pode ser feito supondo-se que todos os salários em determinada classe são iguais ao ponto médio desta classe. Assim, os dez salários pertencentes à primeira classe (de quatro a oito salários) serão admitidos iguais a 6,00, os 12 salários da segunda classe (oito a doze salários) serão admitidos iguais a 10,00 e assim por diante. Então, podemos reescrever a Tabela 2.4 introduzindo os pontos médios das classes. Estes pontos estão na segunda coluna da Tabela 2.6.

Com a tabela assim construída podemos representar os pares (s_i, n_i) ou (s_i, f_i), por um gráfico em barras, setores ou de dispersão unidimensional. Veja a Figura 2.6.

Figura 2.6: Gráfico em barras para a variável S: salários.



O artifício usado acima para representar uma variável contínua faz com que se perca muito das informações nela contidas. Uma alternativa a ser usada nestes casos é o gráfico conhecido como *histograma*.

Tabela 2.6: Distribuição de freqüências da variável S, salário dos empregados da seção de orçamentos da Companhia MB.

Classes de salários	Ponto médio s_i	Freqüência n_i	Porcentagem $100\, f_i$
4,00 ⊢ 8,00	6,00	10	27,78
8,00 ⊢ 12,00	10,00	12	33,33
12,00 ⊢ 16,00	14,00	8	22,22
16,00 ⊢ 20,00	18,00	5	13,89
20,00 ⊢ 24,00	22,00	1	2,78
Total	—	36	100,00

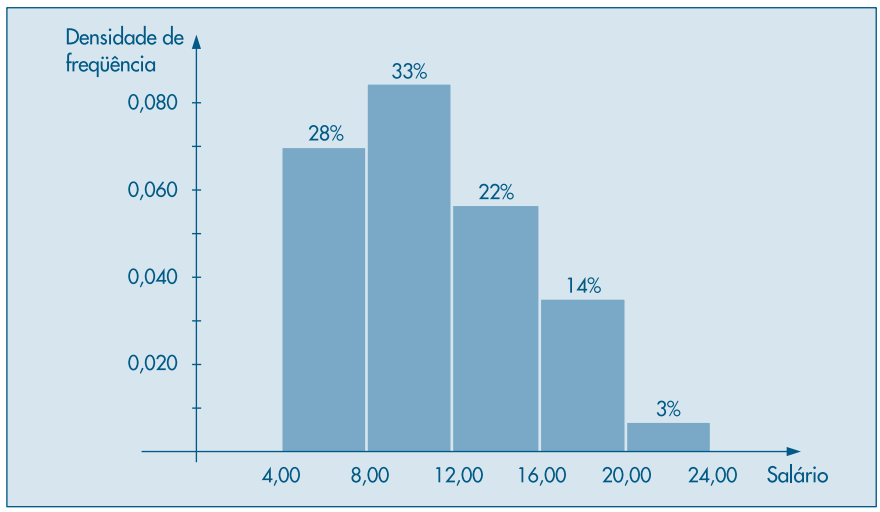
Fonte: Tabela 2.4.

Exemplo 2.7. Usando ainda a variável S do Exemplo 2.4, apresentamos na Figura 2.7 o histograma de sua distribuição.

O histograma é um gráfico de barras contíguas, com as bases proporcionais aos intervalos das classes e a área de cada retângulo proporcional à respectiva freqüência. Pode-se usar tanto a freqüência absoluta, n_i, como a relativa, f_i. Indiquemos a amplitude do i-ésimo intervalo por Δ_i. Para que a área do retângulo respectivo seja proporcional a f_i, a sua altura deve ser proporcional a f_i/Δ_i (ou a n_i/Δ_i), que é chamada *densidade de freqüência* da i-ésima classe. Quanto mais dados tivermos em cada classe, mais alto deve ser o retângulo. Com essa convenção, a área total do histograma será igual a um.

Quando os intervalos das classes forem todos iguais a Δ, a densidade de freqüência da i-ésima classe passa a ser f_i/Δ (ou n_i/Δ). É claro que marcar no eixo das ordenadas os valores n_i, f_i, n_i/Δ ou f_i/Δ leva a obter histogramas com a mesma forma; somente as áreas é que serão diferentes. O Problema 10 traz mais informações sobre a construção de histogramas.

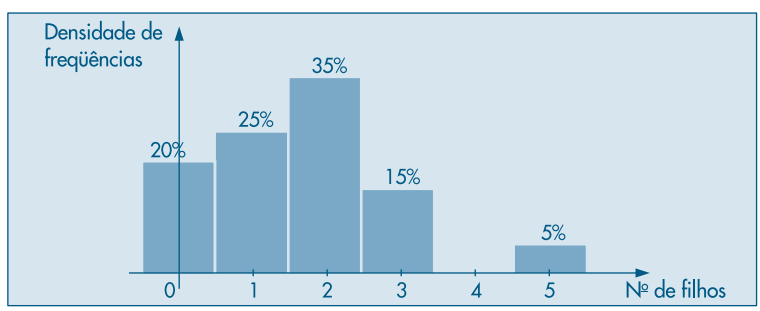
Figura 2.7: Histograma da variável *S*: salários.



Para facilitar o entendimento, foi colocada acima de cada setor (retângulo) a respectiva porcentagem das observações (arredondada). Assim, por meio da figura, podemos dizer que 61% dos empregados têm salário inferior a 12 salários mínimos, ou 17% possuem salário superior a 16 salários mínimos.

Do mesmo modo que usamos um artifício para representar uma variável contínua como uma variável discreta, podemos usar um artifício para construir um histograma para variáveis discretas. A Figura 2.8 é um exemplo de como ficaria o histograma da variável *Z*, número de filhos dos empregados casados da seção de orçamentos da Companhia MB, segundo os dados da Tabela 2.5. O gráfico é suficientemente auto-explicativo, de modo que omitimos detalhes sobre sua construção.

Figura 2.8: Histograma da variável *Z*: número de filhos.



2.4 Ramo-e-Folhas

Tanto o histograma como os gráficos em barras dão uma idéia da *forma da distribuição* da variável sob consideração. Veremos, no Capítulo 3, outras características da distribuição de uma variável, como medidas de posição e dispersão. Mas a forma da distribuição é tão importante quanto estas medidas. Por exemplo, saber que a renda *per capita* de um país é de tantos dóla-res pode ser um dado interessante, mas saber como esta renda se distribui é mais importante.

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 deste diagrama sobre o histograma é que não perdemos (ou perdemos pouca) informação sobre os dados em si.

Exemplo 2.8. Na Figura 2.9 construímos o ramo-e-folhas dos salários de 36 empregados da Companhia MB (Tabela 2.1). 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 (a *folha*) é colocada à direita. Assim, para os salários 4,00 e 4,56, o 4 é o ramo e 00 e 56 são as folhas.

Um ramo com muitas folhas significa maior incidência daquele ramo (realização).

Figura 2.9: Ramo-e-folhas para a variável *S*: salários.

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			

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

(a) Há um destaque grande para o valor 23,30.
(b) Os demais valores estão razoavelmente concentrados entre 4,00 e 19,40.
(c) Um valor mais ou menos típico para este conjunto de dados poderia ser, por exemplo, 10,00.
(d) Há uma leve assimetria em direção aos valores grandes; a suposição de que estes dados possam ser considerados como amostra de uma população com distribuição simétrica, em forma de sino (a chamada distribuição normal), pode ser questionada.

A escolha do número de linhas do ramo-e-folhas é equivalente à escolha do número de classes de um histograma. Um número pequeno de linhas (ou de classes) enfatiza a parte M da relação (1.1), enquanto um número grande de linhas (ou de classes) enfatiza a parte R.

Exemplo 2.9. Os dados abaixo referem-se à dureza de 30 peças de alumínio (Hoaglin, Mosteller e Tukey, 1983, pág. 13).

53,0	70,2	84,3	69,5	77,8	87,5	53,4	82,5	67,3	54,1
70,5	71,4	95,4	51,1	74,4	55,7	63,5	85,8	53,5	64,3
82,7	78,5	55,7	69,1	72,3	59,5	55,3	73,0	52,4	50,7

Na Figura 2.10 temos o ramo-e-folhas correspondente. Aqui, optamos por truncar cada valor, omitindo os décimos, de modo que 69,1 e 69,5, por exemplo, tornam-se 69 e 69 e aparecem como 9 na linha que corresponde ao ramo 6.

Figura 2.10: Ramo-e-folhas para os dados de dureza de peças de alumínio.

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

Este é um exemplo em que temos muitas folhas em cada ramo. Uma maneira alternativa é duplicar os ramos. Criamos os ramos 5* e 5•, 6* e 6• etc., onde colocamos folhas de 0 a 4 na linha * e folhas de 5 a 9 na linha •. Obtemos o ramo-e-folhas da Figura 2.11.

Um ramo-e-folhas pode ser "adornado" com outras informações, como o número de observações em cada ramo. Para outros exemplos, veja o Problema 19.

Figura 2.11: Ramo-e-folhas para os dados de dureza, com ramos divididos.

5*	0 1 2 3 3 3 4
5•	5 5 5 9
6*	3 4
6•	7 9 9
7*	0 0 1 2 3 4
7•	7 8
8*	2 2 4
8•	5 7
9*	
9•	5

Problemas

4. Contou-se o número de erros de impressão da primeira página de um jornal durante 50 dias, obtendo-se os resultados abaixo:

8	11	8	12	14	13	11	14	14	15
6	10	14	19	6	12	7	5	8	8
10	16	10	12	12	8	11	6	7	12
7	10	14	5	12	7	9	12	11	9
14	8	14	8	12	10	12	22	7	15

   (a) Represente os dados graficamente.
   (b) Faça um histograma e um ramo-e-folhas.

5. Usando os resultados do Problema 2 e da Tabela 2.3:
   (a) construa um histograma para a variável idade; e
   (b) proponha uma representação gráfica para a variável grau de instrução.

6. As taxas médias geométricas de incremento anual (por 100 habitantes) dos 30 maiores municípios do Brasil estão dadas abaixo.

3,67	1,82	3,73	4,10	4,30
1,28	8,14	2,43	4,17	5,36
3,96	6,54	5,84	7,35	3,63
2,93	2,82	8,45	5,28	5,41
7,77	4,65	1,88	2,12	4,26
2,78	5,54	0,90	5,09	4,07

   (a) Construa um histograma.
   (b) Construa um gráfico de dispersão unidimensional.

7. Você foi convidado para chefiar a seção de orçamentos ou a seção técnica da Companhia MB. Após analisar o tipo de serviço que cada seção executa, você ficou indeciso e resolveu tomar a decisão baseado em dados fornecidos para as duas seções. O departamento pessoal forneceu as dados da Tabela 2.1 para os funcionários da seção de orçamentos, ao passo que para a seção técnica os dados vieram agrupados segundo as tabelas abaixo, que apresentam as freqüências dos 50 empregados dessa seção, segundo as variáveis grau de instrução e salário. Baseado nesses dados, qual seria a sua decisão? Justifique.

Instrução	Freqüência
Fundamental	15
Médio	30
Superior	5
Total	50

Classe de Salários	Freqüência
7,50 ⊢ 10,50	14
10,50 ⊢ 13,50	17
13,50 ⊢ 16,50	11
16,50 ⊢ 19,50	8
Total	50

8. Construa um histograma, um ramo-e-folhas e um gráfico de dispersão unidimensional para o conjunto de dados 2 (CD-Municípios).

2.5 Exemplos Computacionais

Nesta seção vamos analisar dois dos conjuntos de dados apresentados no final do livro, utilizando técnicas vistas neste capítulo e programas computacionais.

Exemplo 2.10. Considere o conjunto de notas em Estatística de 100 alunos de um curso de Economia (conjunto de dados 3, CD-Notas). O histograma dos dados está na Figura 2.12, que mostra que a distribuição dos dados é razoavelmente simétrica. O gráfico de dispersão unidimensional e o ramo-e-folhas correspondentes estão nas Figuras 2.13 e 2.14, respectivamente, e ambos contêm informação semelhante à dada pelo histograma.

Figura 2.12: Histograma para o CD-Notas. SPlus.

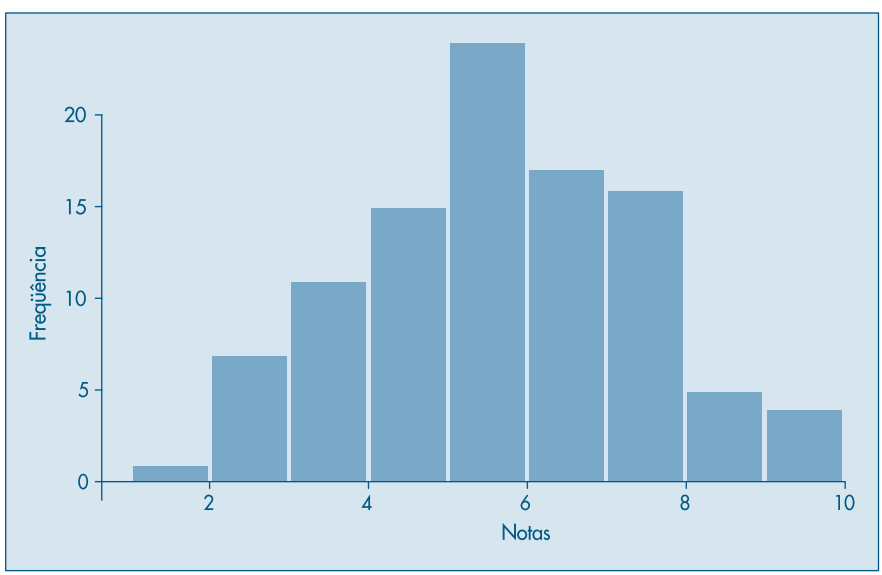


Figura 2.13: Gráfico de dispersão unidimensional para o CD-Notas. Minitab.

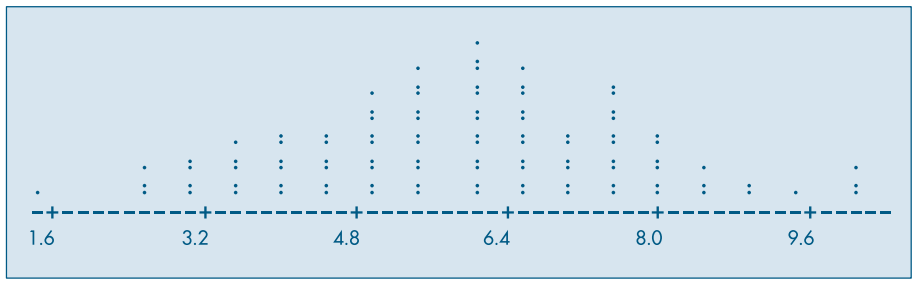
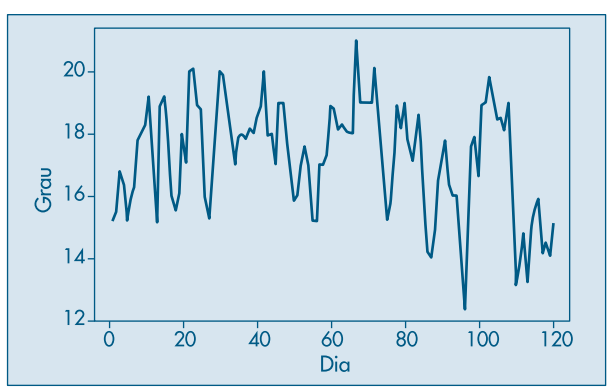


Figura 2.14: Ramo-e-folhas para o CD-Notas. Minitab.

1	*5*
2	*555*
3	*000055555*
4	*000000555555*
5	*00000000055555555555*
6	*000000000000055555555555*
7	*0000005555555555*
8	*000000555*
9	*005*
10	*000*

Exemplo 2.11. O conjunto de dados 4 (CD-Poluição) traz dados sobre a poluição na cidade de São Paulo. Tomemos os dados de temperatura, de 1º de janeiro a 30 de abril de 1991 (120 dados). Essas observações constituem o que se chama *série temporal*, ou seja, os dados são observados em instantes ordenados do tempo. Espera-se que exista relação entre as observações em instantes de tempo diferentes, o que não acontece com os dados do exemplo anterior: a nota de um aluno, em princípio, é independente da nota de outro aluno qualquer. O gráfico dessa série temporal está na Figura 2.15. Observa-se uma variação da temperatura no decorrer do tempo, entre 12 e 22 °C.

Figura 2.15: Dados de temperatura de São Paulo. SPlus.



O histograma e o gráfico de dispersão unidimensional estão nas Figuras 2.16 e 2.17, respectivamente, mostrando que a distribuição dos dados não é simétrica. O ramo-e-folhas da Figura 2.18 ilustra o mesmo comportamento.

Figura 2.16: Histograma dos dados de temperatura de São Paulo. SPlus.

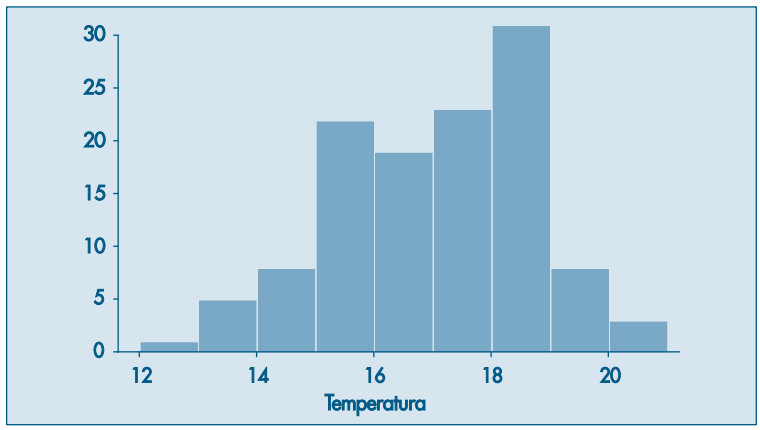


Figura 2.17: Gráfico de dispersão unidimensional para os dados de temperatura de São Paulo. Minitab.

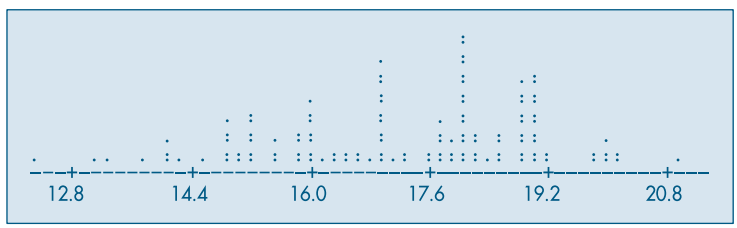


Figura 2.18: Ramo-e-folhas para os dados de temperatura de São Paulo. Minitab.

12	3
13	128
14	0012588899
15	112222225558899
16	000000013344678999
17	0000000012366888888999
18	00000000001111233345566889999999
19	00000000012289
20	00011
21	0

Em cada figura está indicado o pacote computacional que foi utilizado, com as devidas adaptações.

2.6 Problemas e Complementos

9. A MB Indústria e Comércio, desejando melhorar o nível de seus funcionários em cargos de chefia, montou um curso experimental e indicou 25 funcionários para a primeira turma. Os dados referentes à seção a que pertencem, notas e graus obtidos no curso estão na tabela a seguir. Como havia dúvidas quanto à adoção de um único critério de avaliação, cada instrutor adotou seu próprio sistema de aferição. Usando dados daquela tabela, responda às questões:
   (a) Após observar atentamente cada variável, e com o intuito de resumi-las, como você identificaria (qualitativa ordinal ou nominal e quantitativa discreta ou contínua) cada uma das 9 variáveis listadas?
   (b) Compare e indique as diferenças existentes entre as distribuições das variáveis Direito, Política e Estatística.
   (c) Construa o histograma para as notas da variável Redação.
   (d) Construa a distribuição de freqüências da variável Metodologia e faça um gráfico para indicar essa distribuição.
   (e) Sorteado ao acaso um dos 25 funcionários, qual a probabilidade de que ele tenha obtido grau A em Metodologia?
   (f) Se, em vez de um, sorteássemos dois, a probabilidade de que ambos tivessem tido A em Metodologia é maior ou menor do que a resposta dada em (e)?
   (g) Como é o aproveitamento dos funcionários na disciplina Estatística, segundo a seção a que eles pertencem?

Func.	Seção (*)	Administr.	Direito	Redação	Estatíst.	Inglês	Metodologia	Política	Economia
1	P	8,0	9,0	8,6	9,0	B	A	9,0	8,5
2	P	8,0	9,0	7,0	9,0	B	C	6,5	8 0
3	P	8,0	9,0	8,0	8,0	D	B	9,0	8,5
4	P	6,0	9,0	8,6	8,0	D	C	6,0	8,5
5	P	8,0	9,0	8,0	9,0	A	A	6,5	9,0
6	P	8,0	9,0	8,5	10,0	B	A	6,5	9,5
7	P	8,0	9,0	8,2	8,0	D	C	9,0	7,0
8	T	10,0	9,0	7,5	8,0	B	C	6,0	8,5
9	T	8,0	9,0	9,4	9,0	B	B	10,0	8,0
10	T	10,0	9,0	7,9	8,0	B	C	9,0	7,5
11	T	8,0	9,0	8,6	10,0	C	B	10,0	8,5
12	T	8,0	9,0	8,3	7,0	D	B	6,5	8,0
13	T	6,0	9,0	7,0	7,0	B	C	6,0	8,5
14	T	10,0	9,0	8,6	9,0	A	B	10,0	7,5
15	V	8,0	9,0	8,6	9,0	C	B	10,0	7,0
16	V	8,0	9,0	9,5	7,0	A	A	9,0	7,5
17	V	8,0	9,0	6,3	8,0	D	C	10,0	7,5
18	V	6,0	9,0	7,6	9,0	C	C	6,0	8,5
19	V	6,0	9,0	6,8	4,0	D	C	6,0	9,5
20	V	6,0	9,0	7,5	7,0	C	B	6,0	8,5
21	V	8,0	9,0	7,7	7,0	D	B	6,5	8,0
22	V	6,0	9,0	8,7	8,0	C	A	6,0	9,0
23	V	8,0	9,0	7,3	10,0	C	C	9,0	7,0
24	V	8,0	9,0	8,5	9,0	A	A	6,5	9,0
25	V	8,0	9,0	7,0	9,0	B	A	9,0	8,5

(*) (P = departamento pessoal, T = seção técnica e V = seção de vendas)

10. Intervalos de classes desiguais. É muito comum o uso de classes com tamanhos desiguais no agrupamento dos dados em tabelas de freqüências. Nestes casos deve-se tomar alguns cuidados especiais quanto à análise e construção do histograma.

A tabela abaixo fornece a distribuição de 250 empresas classificadas segundo o número de empregados. Uma análise superficial pode levar à conclusão de que a concentração vem aumentando até atingir um máximo na classe $40 \vdash 60$, voltando a diminuir depois, mas não tão acentuadamente. Porém, um estudo mais detalhado revela que a amplitude da classe $40 \vdash 60$ é o dobro da amplitude das classes anteriores. Assim, espera-se que mais elementos caiam nessa classe, mesmo que a concentração seja levemente inferior. Então, um primeiro cuidado é construir a coluna que indica as amplitudes Δ_i de cada classe. Estes valores estão representados na terceira coluna da tabela.

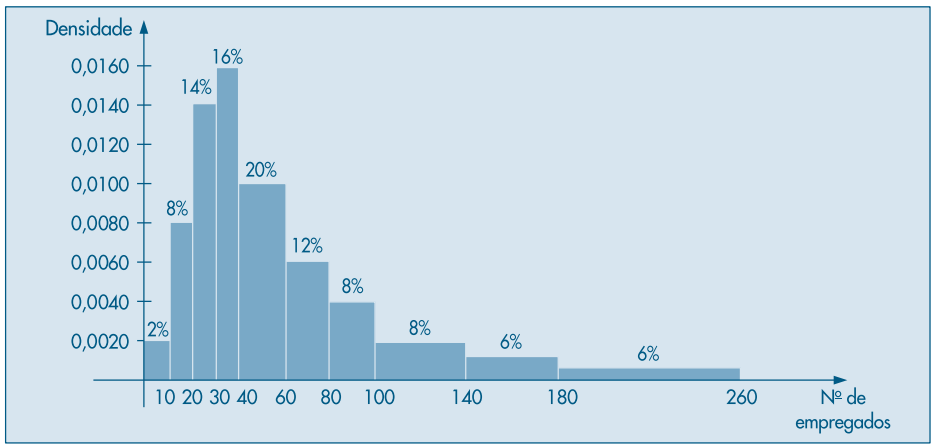
Número de empregados	Freqüência n_i	Amplitude Δ_i	Densidade n_i/Δ_i	Proporção f_i	Densidade f_i/Δ_i
$0 \vdash 10$	5	10	0,50	0,02	0,0020
$10 \vdash 20$	20	10	2,00	0,08	0,0080
$20 \vdash 30$	35	10	3,50	0,14	0,0140
$30 \vdash 40$	40	10	4,00	0,16	0,0160
$40 \vdash 60$	50	20	2,50	0,20	0,0100
$60 \vdash 80$	30	20	1,50	0,12	0,0060
$80 \vdash 100$	20	20	1,00	0,08	0,0040
$100 \vdash 140$	20	40	0,50	0,08	0,0020
$140 \vdash 180$	15	40	0,38	0,06	0,0015
$180 \vdash 260$	15	80	0,19	0,06	0,0008
Total	250	—	—	1,00	—

Um segundo passo é a construção da coluna das densidades de freqüências em cada classe, que é obtida dividindo as freqüências n_i pelas amplitudes Δ_i, ou seja, a medida que indica qual a concentração por unidade da variável. Assim, observando-se os números da quarta coluna, vê-se que a classe de maior concentração passa a ser a $30 \vdash 40$, enquanto a última é a de menor concentração. Para compreender a distribuição, estes dados são muito mais informativos do que as freqüências absolutas simplesmente.

De modo análogo, pode-se construir a densidade da proporção (ou porcentagem) por unidade da variável (verifique a construção através da 5ª e da 6ª colunas). A interpretação para f_i/Δ_i é muito semelhante àquela dada para n_i/Δ_i.

Para a construção do histograma, basta lembrar que a área total deve ser igual a 1 (ou 100%), o que sugere usar no eixo das ordenadas os valores de f_i/Δ_i. O histograma para estes dados está na Figura 2.19.

Figura 2.19: Histograma dos dados do Problema 10.



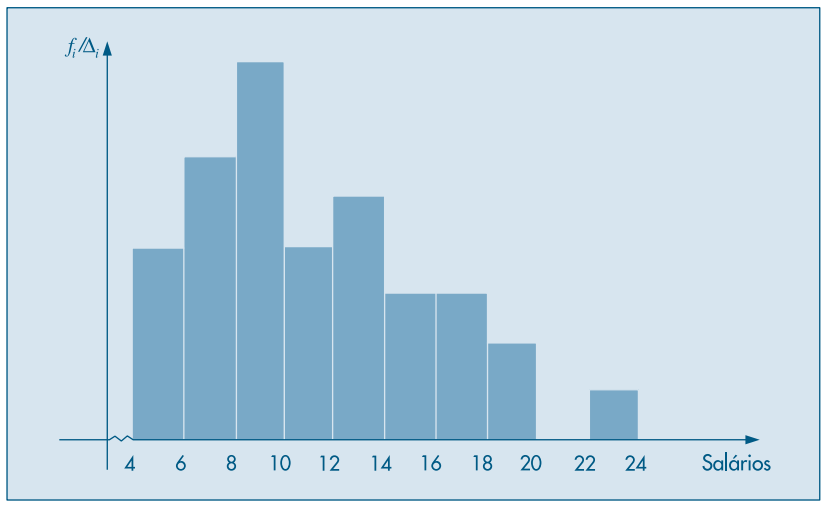
11. Dispomos de uma relação de 200 aluguéis de imóveis urbanos e uma relação de 100 aluguéis rurais.
   (a) Construa os histogramas das duas distribuições.
   (b) Com base nos histogramas, discuta e compare as duas distribuições.

Classes de aluguéis (codificados)	Zona urbana	Zona rural
2 ⊢ 3	10	30
3 ⊢ 5	40	50
5 ⊢ 7	80	15
7 ⊢ 10	50	5
10 ⊢ 15	20	0
Total	200	100

12. Histograma alisado. Na Tabela 2.4 tem-se a distribuição de freqüências dos salários de 36 funcionários, agrupados em classes de amplitude 4. Na Figura 2.7 tem-se o respectivo histograma. Reagrupando-se os dados em classes de amplitude 2, obter-se-ia a seguinte tabela de freqüências e o correspondente histograma (Fig. 2.20 (a)).

Classe de salários	Freqüências n_i
4,00 ⊢ 6,00	4
6,00 ⊢ 8,00	6
8,00 ⊢ 10,00	8
10,00 ⊢ 12,00	4
12,00 ⊢ 14,00	5
14,00 ⊢ 16,00	3
16,00 ⊢ 18,00	3
18,00 ⊢ 20,00	2
20,00 ⊢ 22,00	0
22,00 ⊢ 24,00	1
Total	36

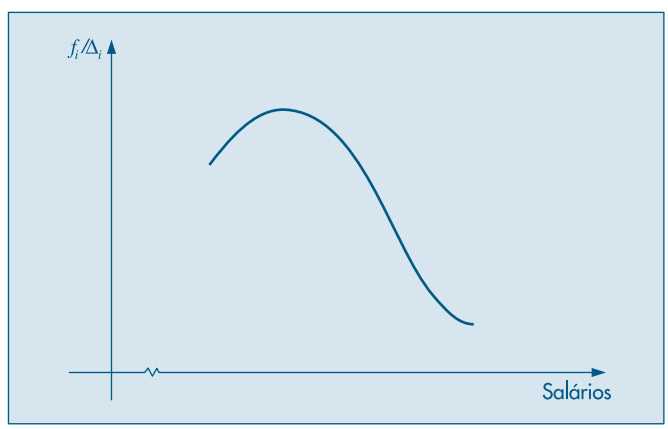
Figura 2.20 (a): Histograma para a variável S: salário, $\Delta = 2$.



Se houvesse um número suficientemente grande de observações, poder-se-ia ir diminuindo os intervalos de classe, e o histograma iria ficando cada vez menos irregular, até atingir um caso limite com uma curva bem mais suave. Por exemplo, o comportamento da distribuição dos salários poderia ter a representação da Figura 2.20 (b). Esse histograma alisado é muito útil para ilustrar rapidamente qual o tipo de comportamento que se espera para a distribuição de uma dada variável. No capítulo referente a variáveis aleatórias contínuas, voltaremos a estudar este histograma sob um ponto de vista mais matemático.

A interpretação desse gráfico é a mesma do histograma. Assim, nas regiões onde a curva é mais alta, significa uma maior densidade de observações. No exemplo acima, conforme se aumenta o salário, observa-se que a densidade de freqüência vai diminuindo.

Figura 2.20 (b): Histograma alisado para a variável S: salário.



13. Esboce o histograma alisado para cada uma das situações descritas abaixo:
    (a) Distribuição dos salários registrados em carteira de trabalho de moradores da cidade de São Paulo.
    (b) Distribuição das idades de alunos de uma Faculdade de Economia e Administração.
    (c) Distribuição das idades dos alunos de uma classe da Faculdade do item anterior. Compare as duas distribuições.
    (d) Distribuição do número de óbitos segundo a faixa etária.
    (e) Distribuição do número de divórcios segundo o número de anos de casado.
    (f) Distribuição do número formado pelos dois últimos algarismos do primeiro prêmio da Loteria Federal, durante os dez últimos anos.

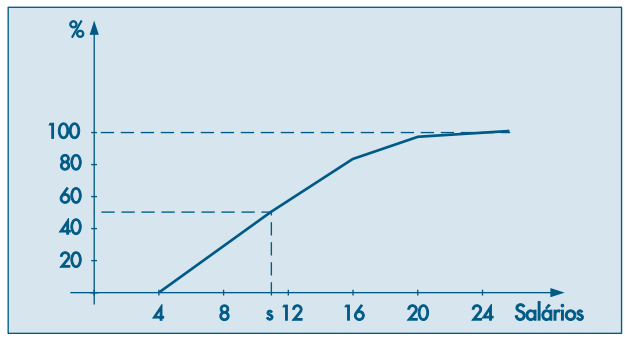
14. Faça no mesmo gráfico um esboço das três distribuições descritas abaixo:
    (a) Distribuição das alturas dos brasileiros adultos.
    (b) Distribuição das alturas dos suecos adultos.
    (c) Distribuição das alturas dos japoneses adultos.

15. Freqüências acumuladas. Uma outra medida muito usada para descrever dados quantitativos é a freqüência acumulada, que indica quantos elementos, ou que porcentagem deles, estão abaixo de um certo valor. Na tabela a seguir, a terceira e a quinta colunas indicam respectivamente a freqüência absoluta acumulada e a proporção (porcentagem) acumulada. Assim, observando a tabela podemos afirmar que 27,78% dos indivíduos ganham até oito salários mínimos; 61,11% ganham até 12 salários mínimos; 83,33% ganham até 16 salários mínimos; 97,22% ganham até 20 salários mínimos e 100% dos funcionários ganham até 24,00 salários.

Classe de salários	Freqüência n_i	Freqüência acumulada N_i	Porcentagem $100\,f_i$	Porcentagem acumulada $100\,F_i$
4,00 ⊢ 8,00	10	10	27,78	27,78
8,00 ⊢ 12,00	12	22	33,33	61,11
12,00 ⊢ 16,00	8	30	22,22	83,33
16,00 ⊢ 20,00	5	35	13,89	97,22
20,00 ⊢ 24,00	1	36	2,78	100,00
Total	36	—	100,00	—

A Figura 2.21 é a ilustração gráfica da porcentagem acumulada.

Figura 2.21: Porcentagens acumuladas para o Problema 15.



Este gráfico pode ser usado para fornecer informações adicionais. Por exemplo, para saber qual o salário s tal que 50% dos funcionários ganhem menos do que s, basta procurar o ponto $(s, 50)$ na curva. Observando as linhas pontilhadas no gráfico, verificamos que a solução é um pouco mais do que 10 salários mínimos.

16. Usando os dados da Tabela 2.1:
    (a) Construa a distribuição de freqüências para a variável idade.
    (b) Faça o gráfico da porcentagem acumulada.
    (c) Usando o gráfico anterior, ache os valores de i correspondentes aos pontos $(i, 25\%)$, $(i, 50\%)$ e $(i, 75\%)$.

17. Freqüências acumuladas (continuação). Para um tratamento estatístico mais rigoroso das variáveis quantitativas, costuma-se usar uma definição mais precisa para a distribuição das freqüências acumuladas. Em capítulos posteriores será vista a sua utilização.

Definição. Dadas n observações de uma variável quantitativa e um número x real qualquer, indicar-se-á por $N(x)$ o número de observações menores ou iguais a x, e chamar-se-á de *função de distribuição empírica* (f.d.e.) a função $F_n(x)$ ou $F_e(x)$

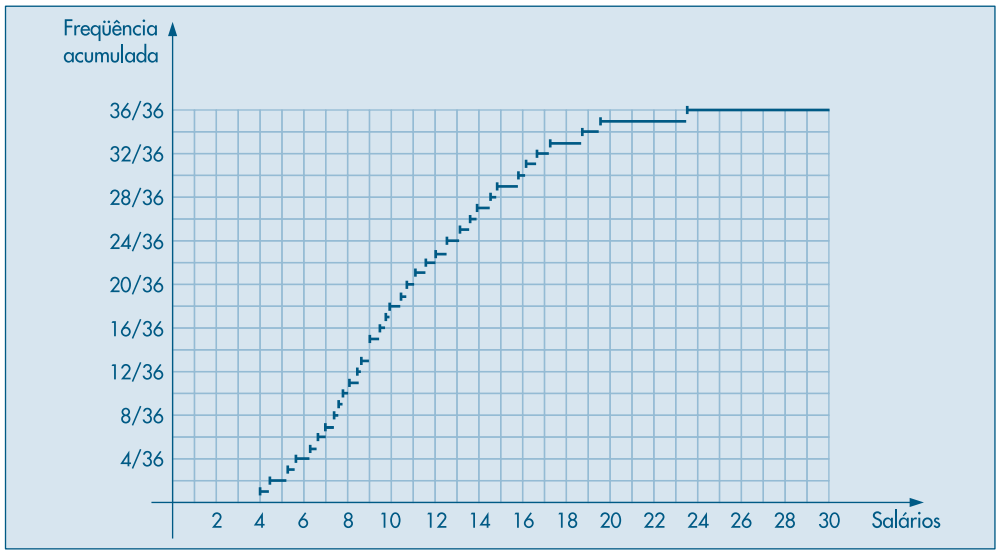
$$F_e(x) = F_n(x) = \frac{N(x)}{n}.$$

Exemplo 2.12. Para a variável S = salário dos 36 funcionários listados na Tabela 2.1, é fácil verificar que:

$$F_{36}(s) = \begin{cases} 0, & \text{se } s < 4{,}00 \\ 1/36\ , & \text{se } 4{,}00 \leqslant s < 4{,}56 \\ 2/36\ , & \text{se } 4{,}56 \leqslant s < 5{,}25 \\ \vdots & \vdots \\ 1, & \text{se } s \geqslant 23{,}30 \end{cases}$$

O gráfico está na Figura 2.22. Àqueles não familiarizados com a representação gráfica de funções, recomenda-se a leitura de Morettin, Hazzan & Bussab (2005).

Figura 2.22: Função de distribuição empírica para o Exemplo 2.12.

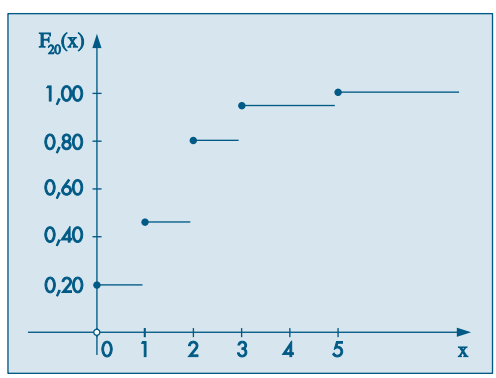


Exemplo 2.13. Esta definição também vale para variáveis quantitativas discretas. Assim, para a variável número de filhos resumida na Tabela 2.5, tem-se a seguinte f.d.e.:

$$F_{20}(x) = \begin{cases} 0{,}00, & \text{se } x < 0 \\ 0{,}20, & \text{se } 0 \leqslant x < 1 \\ 0{,}45, & \text{se } 1 \leqslant x < 2 \\ 0{,}80, & \text{se } 2 \leqslant x < 3 \\ 0{,}95, & \text{se } 3 \leqslant x < 5 \\ 1{,}00, & \text{se } x \geqslant 5 \end{cases}$$

cujo gráfico é o da Figura 2.23.

Figura 2.23: Função de distribuição empírica para o Exemplo 2.13.



18. Construir a f.d.e. para a variável idade referente aos dados da Tabela 2.1.

19. Ramo-e-folhas (continuação). Os dados abaixo referem-se à produção, em toneladas, de dado produto, para 20 companhias químicas (numeradas de 1 a 20).

(1, 50), (2, 280), (3, 560), (4, 170), (5, 180),
(6, 500), (7, 250), (8, 200), (9, 1.050), (10, 240),
(11, 180), (12, 1.000), (13, 1.100), (14, 120), (15, 4.200),
(16, 5.100), (17, 480), (18, 90), (19, 870), (20, 360).

Vemos que os valores estendem-se de 50 a 5.100 e, usando uma representação semelhante à da Figura 2.9, teríamos um grande número de linhas. A Figura 2.24 (a) mostra uma outra forma de ramo-e-folhas, com ramos divididos. A divisão ocorre no ramo, cada vez que se muda por um fator de 10.

Uma economia de 4 linhas poderia ser obtida, representando-se os valores 50 e 90 da Figura 2.24 (a) num ramo denominado 0. Obtemos a Figura 2.24 (b).

Os pacotes computacionais trazem algumas opções adicionais ao construir um ramo-e-folhas. Por exemplo, podemos ter a contagem do número de folhas em cada ramo, como mostra a Figura 2.25 (a). Aqui, temos o ramo-e-folhas dos salários dos empregados da Tabela 2.1. Na Figura 2.25 (b) acrescentamos as contagens de folhas a partir de cada extremo até o ramo que contém a mediana. Esse tipo de opção é chamado *profundidade* (*depth*) nos pacotes.

Figura 2.24: Ramo-e-folhas das produções de companhias químicas.

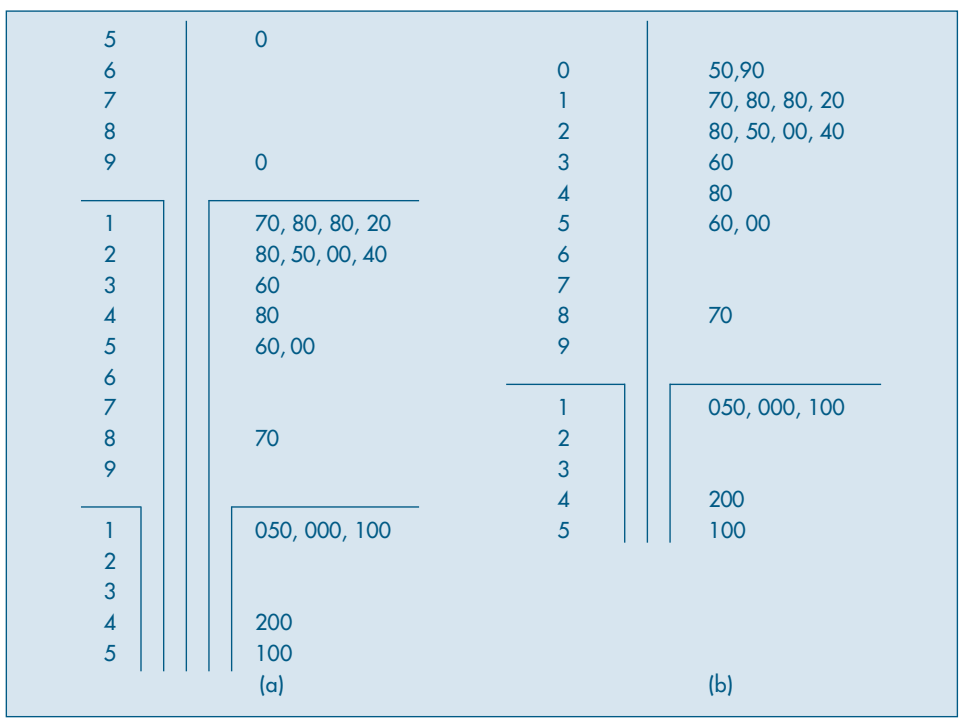


Figura 2.25: Ramo-e-folhas com: (a) freqüências em cada ramo, (b) profundidade.

2	4	00	56			2	4	00	56		
2	5	25	73			4	5	25	73		
3	6	26	66	86		7	6	26	66	86	
3	7	39	44	59		10	7	39	44	59	
4	8	12	46	74	95	14	8	12	46	74	95
4	9	13	35	77	80	18	9	13	35	77	80
2	10	53	76			(2)	10	53	76		
2	11	06	59			16	11	06	59		
2	12	00	79			14	12	00	79		
3	13	23	60	85		12	13	23	60	85	
2	14	69	71			9	14	69	71		
1	15	99				7	15	99			
2	16	22	61			6	16	22	61		
1	17	26				4	17	26			
1	18	75				3	18	75			
1	19	40				2	19	40			
0	20					1	20				
0	21					1	21				
0	22					1	22				
1	23	30				1	23	30			
(a)						(b)					

20. Construa um ramo-e-folhas para a variável CO (monóxido de carbono) do conjunto de dados 4 (CD-Poluição).

Capítulo 3

Medidas-Resumo

3.1 Medidas de Posição

Vimos que o resumo de dados por meio de tabelas de freqüências e ramo-e-folhas fornece muito mais informações sobre o comportamento de uma variável do que a própria tabela original de dados. Muitas vezes, queremos resumir ainda mais estes dados, apresentando um ou alguns valores que sejam *representativos* da série toda. 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.

A *moda* é definida como a realização mais freqüente do conjunto de valores observados. Por exemplo, considere a variável *Z*, número de filhos de cada funcionário casado, resumida na Tabela 2.5 do Capítulo 2. Vemos que a moda é 2, correspondente à realização com maior freqüência, 7. Em alguns casos, pode haver mais de uma moda, ou seja, a distribuição dos valores pode ser bimodal, trimodal etc.

A *mediana* é a realização que ocupa a posição central da série de observações, quando estão ordenadas em ordem crescente. Assim, se as cinco observações de uma variável forem 3, 4, 7, 8 e 8, a mediana é o valor 7, correspondendo à terceira observação. Quando o número de observações for par, usa-se como mediana a média aritmética das duas observações centrais. Acrescentando-se o valor 9 à série acima, a mediana será $(7 + 8)/2 = 7{,}5$.

Finalmente, a *média aritmética*, conceito familiar ao leitor, é a soma das observações dividida pelo número delas. Assim, a média aritmética de 3, 4, 7, 8 e 8 é $(3 + 4 + 7 + 8 + 8)/5 = 6$.

Exemplo 3.1. Usando os dados da Tabela 2.5, já encontramos que a moda da variável *Z* é 2. Para a mediana, constatamos que esta também é 2, média aritmética entre a décima e a décima primeira observações. Finalmente, a média aritmética será

$$\frac{4 \times 0 + 5 \times 1 + 7 \times 2 + 3 \times 3 + 5 \times 1}{20} = \frac{33}{20} = 1{,}65.$$

Neste exemplo, as três medidas têm valores próximos e qualquer uma delas pode ser usada como *representativa* da série toda. A média aritmética é, talvez, a medida mais usada. Contudo, ela pode conduzir a erros de interpretação. Em muitas situações, a mediana é uma medida mais adequada. Voltaremos a este assunto mais adiante.

Vamos formalizar os conceitos introduzidos acima. Se $x_1, \ldots, x_n$ são os n valores (distintos ou não) da variável X, a média aritmética, ou simplesmente média, de X pode ser escrita

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{3.1}$$

Agora, se tivermos n observações da variável X, das quais n_1 são iguais a x_1, n_2 são iguais a x_2 etc., n_k iguais a x_k, então a média de X pode ser escrita

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{n} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i. \tag{3.2}$$

Se $f_i = n_i / n$ representar a freqüência relativa da observação x_i, então (3.2) também pode ser escrita

$$\bar{x} = \sum_{i=1}^{k} f_i x_i. \tag{3.3}$$

Consideremos, agora, as observações ordenadas em ordem crescente. Vamos denotar a menor observação por $x_{(1)}$, a segunda por $x_{(2)}$, e assim por diante, obtendo-se

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n-1)} \leqslant x_{(n)}. \tag{3.4}$$

Por exemplo, se $x_1 = 3$, $x_2 = -2$, $x_3 = 6$, $x_4 = 1$, $x_5 = 3$, então $-2 \leqslant 1 \leqslant 3 \leqslant 3 \leqslant 6$, de modo que $x_{(1)} = -2$, $x_{(2)} = 1$, $x_{(3)} = 3$, $x_{(4)} = 3$ e $x_{(5)} = 6$.

As observações ordenadas como em (3.4) são chamadas *estatísticas de ordem*.

Com esta notação, a mediana da variável X pode ser definida como

$$\text{md}(X) = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ ímpar;} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se } n \text{ par.} \end{cases} \tag{3.5}$$

Exemplo 3.2. A determinação das medidas de posição para uma variável quantitativa contínua, através de sua distribuição de freqüências, exige aproximações, pois perdemos a informação dos valores das observações. Consideremos a variável S: salário dos 36 funcionários da Companhia MB, agrupados em classes de salários, conforme a Tabela 2.6. Uma aproximação razoável é supor que todos os valores dentro de uma classe tenham seus valores iguais ao ponto médio desta classe. Este procedimento nos deixa na mesma situação do caso discreto, onde as medidas são calculadas usando-se os pares (x_i, n_i) ou (x_i, f_i), como em (3.2) e (3.3).

A moda, mediana e média para os dados da Tabela 2.6 são, respectivamente,

$$\text{mo(S)} \simeq 10{,}00,$$
$$\text{md(S)} \simeq 10{,}00,$$
$$\bar{s} \simeq \frac{10 \times 6{,}00 + 12 \times 10{,}00 + 8 \times 14{,}00 + 5 \times 18{,}00 + 1 \times 22{,}00}{36} = 11{,}22.$$

Observe que colocamos o sinal de $\simeq$ e não de igualdade, pois os valores verdadeiros não são os calculados. Por exemplo, a mediana de S é a média entre as duas observações centrais, quando os dados são ordenados, isto é, 9,80 e 10,53, portanto md(S) = 10,16. Quais são, neste exemplo, a média e moda verdadeiras?

Observe que, para calcular a moda de uma variável, precisamos apenas da distribuição de freqüências (contagem). Já para a mediana necessitamos minimamente ordenar as realizações da variável. Finalmente, a média só pode ser calculada para variáveis quantitativas.

Estas condições limitam bastante o cálculo de medidas-resumos para as variáveis qualitativas. Para as variáveis nominais somente podemos trabalhar com a moda. Para as variáveis ordinais, além da moda, podemos usar também a mediana. Devido a esse fato, iremos apresentar daqui em diante medidas-resumo para variáveis quantitativas, que permitem o uso de operações aritméticas com seus valores.

Exemplo 3.2. (continuação) Retomemos os dados da Companhia MB. A moda para a variável V: região de procedência é mo(V) = outra. Para a variável Y: grau de instrução, temos que mo(Y) = ensino médio e md(Y) = ensino médio.

Veremos, na seção 3.3, que a mediana é uma medida resistente, ao passo que a média não o é, em particular para distribuições muito assimétricas ou contendo valores atípicos. Por outro lado, a média é ótima (num sentido que será discutido no Capítulo 10) se a distribuição dos dados for aproximadamente normal.

Uma outra medida de posição também resistente é a média aparada, definida no Problema 39. Esta medida envolve calcular a média das observações centrais, desprezando-se uma porcentagem das iniciais e finais.

3.2 Medidas de Dispersão

O resumo de um conjunto de dados por uma única medida representativa de posição central esconde toda a informação sobre a variabilidade do conjunto de observações. Por exemplo, suponhamos que cinco grupos de alunos submeteram-se a um teste, obtendo-se as seguintes notas:

grupo A (variável X): 3, 4, 5, 6, 7
grupo B (variável Y): 1, 3, 5, 7, 9
grupo C (variável Z): 5, 5, 5, 5, 5
grupo D (variável W): 3, 5, 5, 7
grupo E (variável V): 3, 5, 5, 6, 6

Vemos que $\bar{x} = \bar{y} = \bar{z} = \bar{w} = \bar{v} = 5{,}0$. A identificação de cada uma destas séries por sua média (5, em todos os casos) nada informa sobre suas diferentes variabilidades. Notamos, então, a conveniência de serem criadas medidas que sumarizem a variabilidade de um conjunto de observações e que nos permita, por exemplo, comparar conjuntos diferentes de valores, como os dados acima, segundo algum critério estabelecido.

Um critério freqüentemente usado para tal fim é aquele que mede a dispersão dos dados em torno de sua média, e duas medidas são as mais usadas: desvio médio e variância. O princípio básico é analisar os desvios das observações em relação à média dessas observações.

Para o grupo A acima os desvios $x_i - \bar{x}$ são: –2, –1, 0, 1, 2. É fácil ver (Problema 14 (a)) que, para *qualquer* conjunto de dados, a soma dos desvios é igual a zero. Nestas condições, a soma dos desvios $\sum_{i=1}^{5} (x_i - \bar{x})$ não é uma boa medida de dispersão para o conjunto A. Duas opções são: (a) considerar o total dos desvios em valor absoluto; (b) considerar o total dos quadrados dos desvios. Para o grupo A teríamos, respectivamente,

$$\sum_{i=1}^{5} |x_i - \bar{x}| = 2 + 1 + 0 + 1 + 2 = 6,$$

$$\sum_{i=1}^{5} (x_i - \bar{x})^2 = 4 + 1 + 0 + 1 + 4 = 10.$$

O uso desses totais pode causar dificuldades quando comparamos conjuntos de dados com números diferentes de observações, como os conjuntos A e D acima. Desse modo, é mais conveniente exprimir as medidas como médias, isto é, o *desvio médio* e a *variância* são definidos por

$$\text{dm}(X) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}, \tag{3.6}$$

$$\text{var}(X) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}, \tag{3.7}$$

respectivamente. Para o grupo A temos

$$\text{dm}(X) = 6/5 = 1{,}2,$$
$$\text{var}(X) = 10/5 = 2{,}0,$$

enquanto para o grupo D temos

$$\text{dm}(W) = 4/4 = 1{,}0,$$
$$\text{var}(W) = 8/4 = 2{,}0.$$

Podemos dizer, então, que, segundo o desvio médio, o grupo D é mais homogêneo que A, enquanto ambos são igualmente homogêneos, segundo a variância.

Sendo a variância uma medida de dimensão igual ao quadrado da dimensão dos dados (por exemplo, se os dados são expressos em cm, a variância será expressa em cm^2), pode

causar problemas de interpretação. Costuma-se usar, então, o *desvio padrão*, que é definido como a raiz quadrada positiva da variância. Para o grupo A o desvio padrão é

$$\text{dp}(X) = \sqrt{\text{var}\,(X)} = \sqrt{2} = 1{,}41.$$

Ambas as medidas de dispersão (dm e dp) indicam em média qual será o "erro" (desvio) cometido ao tentar substituir cada observação pela medida resumo do conjunto de dados (no caso, a média).

Exemplo 3.3. Vamos calcular as medidas de dispersão acima para a variável Z = número de filhos, resumida na Tabela 2.5. Como vimos no Exemplo 3.1, $\bar{z} = 1{,}65$. Os desvios são $z_i - \bar{z}$: $-1{,}65$; $-0{,}65$; $0{,}35$; $1{,}35$; $3{,}35$. Segue-se que

$$\text{dm}(Z) = \frac{4 \times (1{,}65) + 5 \times (0{,}65) + 7 \times (0{,}35) + 3 \times (1{,}35) + 1 \times (3{,}35)}{20} = 0{,}98.$$

Também,

$$\text{var}(Z) = \frac{4(-1{,}65)^2 + 5(-0{,}65)^2 + 7(0{,}35)^2 + 3(1{,}35)^2 + 1(3{,}35)^2}{20} = 1{,}528.$$

Conseqüentemente, o desvio padrão de Z é

$$\text{dp}(Z) = \sqrt{1{,}528} = 1{,}24.$$

Suponha que observemos n_1 vezes os valores x_1 etc., n_k vezes o valor x_k da variável X. Então,

$$\text{dm}(X) = \frac{\sum_{i=1}^{k} n_i |x_i - \bar{x}|}{n} = \sum_{i=1}^{k} f_i |x_i - \bar{x}|, \tag{3.8}$$

$$\text{var}(X) = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}{n} = \sum_{i=1}^{k} f_i (x_i - \bar{x})^2, \tag{3.9}$$

$$\text{dp}(X) = \sqrt{\text{var}\,(X)}. \tag{3.10}$$

O cálculo (aproximado) das medidas de dispersão no caso de variáveis contínuas, agrupadas em classes, pode ser feito de modo análogo àquele usado para encontrar a média no Exemplo 2.2.

Exemplo 3.4. Consideremos a variável S = salário. A média encontrada no Exemplo 3.2 foi $s = 11{,}22$. Com os dados da Tabela 2.6 e usando (3.9) encontramos

$$\text{var}(S) \simeq [10(6,\,00 - 11{,}22)^2 + 12(10{,}00 - 11{,}22)^2 + 8(14 - 11{,}22)^2 + 5(18{,}00 - 11{,}22)^2 + 1(22{,}00 - 11{,}22)^2]/36 = 19{,}40$$

e, portanto,

$$\text{dp}(S) \simeq \sqrt{19{,}40} = 4{,}40.$$

É fácil ver que $\text{dm}(S) \simeq 3{,}72$.

Veremos, mais tarde, que a variância de uma amostra será calculada usando-se o denominador $n - 1$, em vez de n. A justificativa será dada naquele capítulo, mas para grandes amostras pouca diferença fará o uso de um ou outro denominador.

Tanto a variância como o desvio médio são medidas de dispersão calculadas em relação à média das observações. Assim como a média, a variância (ou o desvio padrão) é uma boa medida se a distribuição dos dados for aproximadamente normal. O desvio médio é mais resistente que o desvio padrão, no sentido a ser estudado na seção seguinte.

Poderíamos considerar uma medida que seja calculada em relação à mediana. O desvio absoluto mediano é um exemplo e é mais resistente que o desvio padrão. Veja o Problema 41.

Usando o Problema 14 (b), uma maneira computacionalmente mais eficiente de calcular a variância é

$$\text{var}(X) = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2, \tag{3.11}$$

e, no caso de observações repetidas,

$$\text{var}(X) = \sum_{i=1}^{k} f_i x_i^2 - \bar{x}^2. \tag{3.12}$$

Problemas

1. Quer se estudar o número de erros de impressão de um livro. Para isso escolheu-se uma amostra de 50 páginas, encontrando-se o número de erros por página da tabela abaixo.
   (a) Qual o número médio de erros por página?
   (b) E o número mediano?
   (c) Qual é o desvio padrão?
   (d) Faça uma representação gráfica para a distribuição.
   (e) Se o livro tem 500 páginas, qual o número total de erros esperado no livro?

Erros	Freqüência
0	25
1	20
2	3
3	1
4	1

2. As taxas de juros recebidas por 10 ações durante um certo período foram (medidas em porcentagem) 2,59; 2,64; 2,60; 2,62; 2,57; 2,55; 2,61; 2,50; 2,63; 2,64. Calcule a média, a mediana e o desvio padrão.

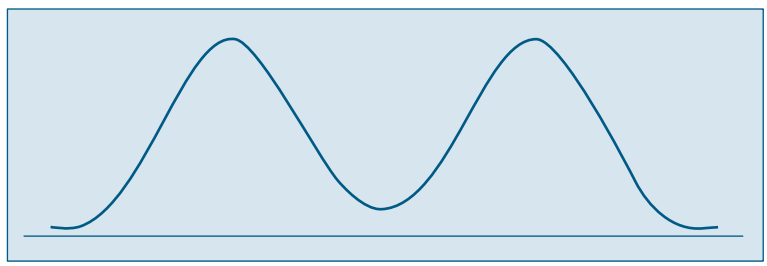
3. Para facilitar um projeto de ampliação da rede de esgoto de uma certa região de uma cidade, as autoridades tomaram uma amostra de tamanho 50 dos 270 quarteirões que compõem a região, e foram encontrados os seguintes números de casas por quarteirão:

2	2	3	10	13	14	15	15	16	16
18	18	20	21	22	22	23	24	25	25
26	27	29	29	30	32	36	42	44	45
45	46	48	52	58	59	61	61	61	65
66	66	68	75	78	80	89	90	92	97

(a) Use cinco intervalos e construa um histograma.
(b) Determine uma medida de posição central e uma medida de dispersão.

4. (a) Dê uma situação prática onde você acha que a mediana é uma medida mais apropriada do que a média.
(b) Esboce um histograma onde a média e a mediana coincidem. Existe alguma classe de histogramas onde isso sempre acontece?
(c) Esboce os histogramas de três variáveis (X, Y e Z) com a mesma média aritmética, mas com as variâncias ordenadas em ordem crescente.

5. Suponha que a variável de interesse tenha a distribuição como na figura abaixo.



Você acha que a média é uma boa medida de posição? E a mediana? Justifique.

6. Numa pesquisa realizada com 100 famílias, levantaram-se as seguintes informações:

Número de filhos	0	1	2	3	4	5	mais que 5
Freqüência de famílias	17	20	28	19	7	4	5

(a) Qual a mediana do número de filhos?
(b) E a moda?
(c) Que problemas você enfrentaria para calcular a média? Faça alguma suposição e encontre-a.

3.3 Quantis Empíricos

Tanto a média como o desvio padrão podem não ser medidas adequadas para representar um conjunto de dados, pois:

(a) são afetados, de forma exagerada, por valores extremos;
(b) apenas com estes dois valores não temos idéia da simetria ou assimetria da distribuição dos dados.

Para contornar esses fatos, outras medidas têm de ser consideradas.

Vimos que a mediana é um valor que deixa metade dos dados abaixo dela e metade acima (ver fórmula (3.5)). De modo geral, podemos definir uma medida, chamada *quantil de ordem p* ou *p-quantil*, indicada por $q(p)$, onde 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.

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

Dependendo do valor de p, há dificuldades ao se calcular os quantis. Isso é ilustrado no exemplo a seguir.

Exemplo 3.5. Suponha que tenhamos os seguintes valores de uma variável X:

$$15,\ 5,\ 3,\ 8,\ 10,\ 2,\ 7,\ 11,\ 12.$$

Ordenando os valores, obtemos as estatísticas de ordem $x_{(1)} = 2$, $x_{(2)} = 3$,..., $x_{(9)} = 15$, ou seja, teremos

$$2 < 3 < 5 < 7 < 8 < 10 < 11 < 12 < 15.$$

Usando a definição de mediana dada, teremos que md $= q(0{,}5) = q_2 = x_{(5)} = 8$. Suponha que queiramos calcular os dois outros quartis, q_1 e q_3. A idéia é dividir os dados em quatro partes:

$$2 \quad \mathbf{3} \quad \mathbf{5} \quad 7 \quad \mathbf{8} \quad 10 \quad \mathbf{11} \quad \mathbf{12} \quad 15$$

Uma possibilidade razoável é, então, considerar a mediana dos primeiros quatro valores para obter q_1, ou seja,

$$q_1 = \frac{3 + 5}{2} = 4,$$

e a mediana dos últimos quatro valores para obter q_3, ou seja,

$$q_3 = \frac{11 + 12}{2} = 11{,}5.$$

Obtemos, então, a sequência

$$2 \quad 3 \quad (\mathbf{4}) \quad 5 \quad 7 \quad (\mathbf{8}) \quad 10 \quad 11 \quad (\mathbf{11{,}5}) \quad \mathbf{12} \quad 15$$

Observe que a média dos $n = 9$ valores é $\overline{x} = 8{,}1$, próximo à mediana.

Exemplo 3.5. (continuação). Acrescentemos, agora, o valor 67 à lista de nove valores do Exemplo 3.5, obtendo-se agora os $n = 10$ valores ordenados:

$$2 < 3 < 5 < 7 < 8 < 10 < 11 < 12 < 15 < 67$$

Agora, $\bar{x} = 14$, enquanto que a mediana fica

$$q_2 = \frac{x_{(5)} + x_{(6)}}{2} = 9,$$

que está próxima da mediana dos nove valores originais, mas ambas (8 e 9) relativamente longes de $\bar{x}$. Dizemos que a mediana é *resistente* (ou *robusta*), no sentido que que ela não é muito afetada pelo valor discrepante (ou atípico) 67.

Para calcular q_1 e q_3 para este novo conjunto de valores, considere-os assim dispostos:

2 3 **5** 7 8 **9** 10 11 **12** 15 67

de modo que $q_1 = 5$ e $q_3 = 12$.

Obtemos, então os dados separados em 4 partes por q_1, q_2 e q_3:

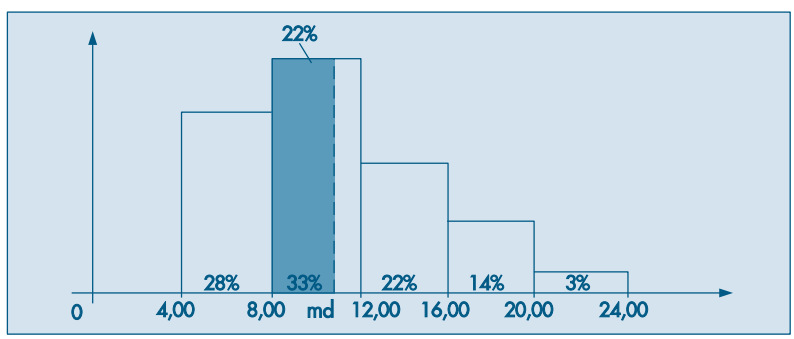
2 3 (**5**) 7 8 (**9**) 10 11 (**12**) 15 67

Suponha, agora, que queiramos calcular $q(0{,}20)$, ou seja, aquele valor que deixa 20% dos dados à sua esquerda, para o conjunto original de $n = 9$ valores de X. Como 20% das observações correspondem a 1,8 observações, qual valor devemos tomar como $q(0, 20)$? O valor 3, que é a segunda observação ordenada, ou 5, ou a média de 3 e 5? Se adotarmos esta última solução, então $q(0, 20) = q(0, 25) = q_1$, o que pode parecer não razoável.

Para responder a esta questão, temos que definir quantil de uma sequência de valores de uma variável de modo apropriado. Isto está feito no Problema 17.

Se os dados estiverem agrupados em classes, podemos obter os quantis usando o histograma. Por exemplo, para obter a mediana, sabemos que ela deve corresponder ao valor da abscissa que divide a área do histograma em duas partes iguais (50% para cada lado). Então, usando argumentos geométricos, podemos encontrar um ponto, satisfazendo essa propriedade. Vejamos como proceder através de um exemplo.

Exemplo 3.6. Vamos repetir abaixo a Figura 2.7, que é o histograma da variável S = salário dos empregados da Companhia MB.



Devemos localizar o ponto das abscissas que divide o histograma ao meio. A área do primeiro retângulo corresponde a 28% do total, os dois primeiros a 61%; portanto, a mediana md é algum número situado entre 8,00 e 12,00. Ou melhor, a mediana irá corresponder ao valor md no segundo retângulo, cuja área do retângulo de base 8,00 ⊢ md é a mesma altura que o retângulo de base 8,00 ⊢ 12,00 seja 22% (28% do primeiro retângulo mais 22% do segundo, perfazendo os 50%). Consulte a figura para melhor compreensão. Pela proporcionalidade entre a área e a base do retângulo, temos:

$$\frac{12{,}00 - 8{,}00}{33\%} = \frac{\text{md} - 8{,}00}{22\%}$$

ou

$$\text{md} - 8{,}00 = \frac{22\%}{33\%} \cdot 4{,}00,$$

logo

$$\text{md} = 8{,}00 + 2{,}67 = 10{,}67,$$

que é uma expressão mais precisa para a mediana do que a mediana bruta encontrada anteriormente.

O cálculo dos quantis pode ser feito de modo análogo ao cálculo da mediana, usando argumentos geométricos no histograma. Vejamos a determinação de alguns quantis, usando os dados do último exemplo.

(a) $q(0{,}25)$: Verificamos que $q(0{,}25)$ deve estar na primeira classe, pois a proporção no primeiro retângulo é 0,28. Logo,

$$\frac{q(0{,}25) - 4{,}00}{25\%} = \frac{8{,}00 - 4{,}00}{28\%},$$

e então

$$q(0{,}25) = 4{,}00 + \frac{25}{28}\ 4{,}00 = 7{,}57.$$

(b) $q(0{,}95)$: Analisando a soma acumulada das proporções, verificamos que este quantil deve pertencer à quarta classe, e que nesse retângulo devemos achar a parte correspondente a 12%, pois a soma acumulada até a classe anterior é 83%, faltando 12% para atingirmos os 95%. Portanto,

$$\frac{q(0{,}95) - 16{,}00}{12\%} = \frac{20{,}00 - 16{,}00}{14\%},$$

logo

$$q(0{,}95) = 16{,}00 + \frac{12}{14} \times 4 = 19{,}43.$$

(c) $q(0{,}75)$: De modo análogo, concluímos que o terceiro quantil deve pertencer ao intervalo $12{,}00 \vdash 16{,}00$, portanto

$$\frac{q(0{,}75) - 12{,}00}{14\%} = \frac{16{,}00 - 12{,}00}{22\%}$$

e

$$q(0{,}75) = 14{,}55.$$

Uma medida de dispersão alternativa ao desvio padrão é a *distância interquartil*, definida como a diferença entre o terceiro e primeiro quartis, ou seja,

$$d_q = q_3 - q_1. \tag{3.13}$$

Para o Exemplo 3.5, temos $q_1 = 4$, $q_3 = 11{,}5$, de modo que $d_q = 7{,}5$. Para um cálculo mais preciso, veja o Problema 17. Lá obtemos $q_1 = 4{,}5$, $q_3 = 11{,}25$, logo $d_q = 6{,}75$.

Os quartis $q(0{,}25) = q_1$, $q(0{,}5) = 92$ e $9(0{,}75) = 93$ são medidas de localização resistentes de uma distribuição.

Dizemos que uma medida de localização ou dispersão é resistente quando for pouco afetada por mudanças de uma pequena porção dos dados. A mediana é uma medida resistente, ao passo que a média não o é. Para ilustrar este fato, considere as populações dos 30 municípios do Brasil, considerados acima. Se descartarmos Rio de Janeiro e São Paulo, a média das populações dos 28 municípios restantes é 100,6 e a mediana é 82,1. Para todos os dados, a média pasa a ser 145,4, ao passo que a mediana será 84,3. Note que a média aumentou bastante, influenciada que foi pelos dois valores maiores, que são muito discrepantes da maioria dos dados. Mas a mediana variou pouco. O desvio padrão também não é uma medida resistente. Verifique como este varia para este exemplo dos municípios.

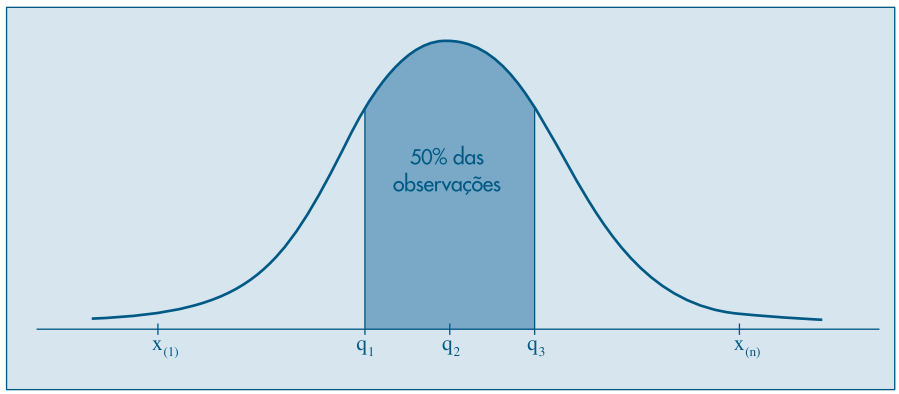
Os cinco valores, $x_{(1)}$, q_1, q_2, q_3 e $x_{(n)}$ são importantes para se ter uma boa idéia da assimetria da distribuição dos dados. Para uma distribuição simétrica ou aproximadamente simétrica, deveríamos ter:

(a) $q_2 - x_{(1)} \simeq x_{(n)} - q_2$;
(b) $q_2 - q_1 \simeq q_3 - q_2$;
(c) $q_1 - x_{(1)} \simeq x_{(n)} - q_3$;
(d) distâncias entre mediana e q_1, q_3 menores do que distâncias entre os extremos e q_1, q_3.

A diferença $q_2 - x_{(1)}$ é chamada *dispersão inferior* e $x_{(n)} - q_2$ é a *dispersão superior*. A condição (a) nos diz que estas duas dispersões devem ser aproximadamente iguais, para uma distribuição aproximadamente simétrica.

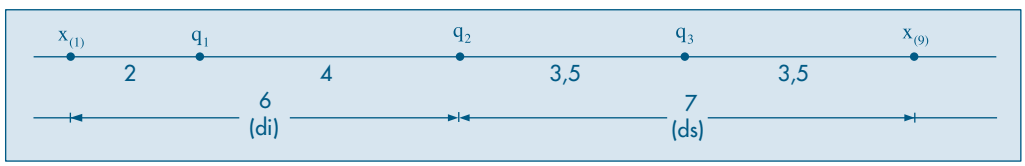
A Figura 3.1 ilustra estes fatos para a chamada *distribuição normal* ou *gaussiana*.

Figura 3.1: Uma distribuição simétrica: normal ou gaussiana.



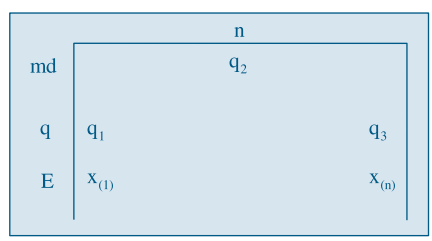
Na Figura 3.2 temos ilustradas estas cinco medidas para os $n = 9$ valores do Exemplo 3.5.

Figura 3.2: Quantis e distâncias para o Exemplo 3.5.



As cinco estatísticas de ordem consideradas acima podem ser representadas esquematicamente como na Figura 3.3, onde também incorporamos o número de observações, *n*. Representamos a mediana por *md*, os quartis por *q* e os extremos por *E*. Podemos ir além, considerando os chamados *oitavos*, ou seja, o primeiro oitavo, que corresponde a $q(0{,}125)$, o sétimo oitavo, que corresponde a $q(0{,}875)$ etc. Teríamos, então, sete números para representar a distribuição dos dados. Em geral, podemos considerar as chamadas *letras-resumos*, descendo aos *dezesseis-avos*, *trinta e dois-avos* etc. Para detalhes, ver Hoaglin, Mosteller and Tukey(1983).

Figura 3.3: Esquema dos cinco números.



Exemplo 3.7. Os aplicativos SPlus e Minitab, assim como a planilha Excel, possuem ferramentas que geram as principais medidas descritas nesse capítulo e outras. Por exemplo, o comando *describe* do Minitab, usado para as populações dos municípios brasileiros produz a saída do Quadro 3.1.

Quadro 3.1. Medidas-resumo para o CD-Municípios. Minitab.

```
MTB > Describe C1.
Descriptive Statistics
Variable    N     Mean   Median  Tr mean   StDev  SE Mean
C1         30    145.4     84.3    104.7   186.6     34.1
Variable   Min     Max       Q1       Q3
C1        46.3   988.8     63.5    139.7
```

Aqui, temos $N = 30$ dados, a média é 145,4, a mediana 84,3, o desvio padrão 186,6, o menor valor 46,3, o maior valor 988,8, o primeiro quartil 63,5 e o terceiro quartil 139,7. Além desses valores, o resumo traz a *média aparada* (*trimmed mean*) e o erro padrão da média, a ser tratado no Capítulo 11. Esse é dado por $S/\sqrt{n} = 186{,}6/\sqrt{30} = 34{,}1$.

O comando *summary* do SPlus produz a saída do Quadro 3.2 para os mesmos dados. Note a diferença no cálculo dos quantis $q(0{,}25)$ e $q(0{,}75)$. Conclui-se que é necessário saber como cada programa efetua o cálculo de determinada estatística, para poder reportá-lo.

Quadro 3.2. Medidas-resumo para o CD-Municípios. SPlus.

```
> summary (munic)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   46.3   64.48    84.3   145.4   134.3   988.8
```

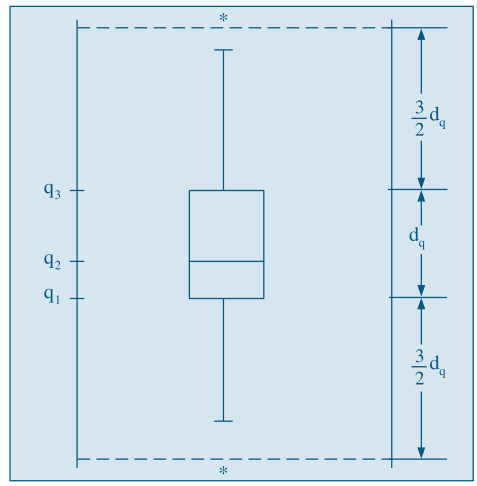
Problemas

7. Obtenha o esquema dos cinco números para os dados do Problema 3. Calcule o intervalo interquartil e as dispersões inferior e superior. Baseado nessas medidas, verifique se a forma da distribuição dos dados é normal.
8. Refaça o problema anterior, utilizando desta vez os dados do Problema 5 do Capítulo 2.
9. Obter os três quartis, $q(0{,}1)$ e $q(0{,}90)$ para os dados do Problema 3.
10. Para a variável *população urbana* do CD-Brasil, obtenha $q(0{,}10)$, $q(0{,}25)$, $q(0{,}50)$, $q(0{,}75)$, $q(0{,}80)$ e $q(0{,}95)$.

3.4 Box Plots

A informação contida no esquema dos cinco números da Figura 3.3 pode ser traduzida graficamente num diagrama, ilustrado na Figura 3.4, que chamaremos de *box plot*. Murteira (1993) usa o termo "caixa-de-bigodes".

Figura 3.4: *Box Plot.*

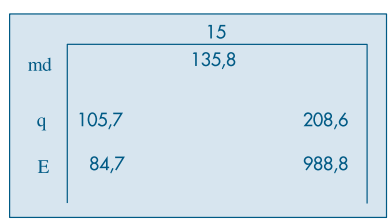


Para construir este diagrama, consideremos um retângulo onde estão representados a mediana e os quartis. A partir do retângulo, para cima, segue uma linha até o ponto mais remoto que não exceda LS = $q_3 + (1{,}5)d_q$, chamado *limite superior*. De modo similar, da parte inferior do retângulo, para baixo, segue uma linha até o ponto mais remoto que não seja menor do que LI = $q_1 - (1{,}5)d_q$, chamado *limite inferior*. Os valores compreendidos entre esses dois limites são chamados *valores adjacentes*. As observações que estiverem acima do limite superior ou abaixo do limite inferior estabelecidos serão chamadas *pontos exteriores* e representadas por asteriscos. Essas são observações destoantes das demais e podem ou não ser o que chamamos de *outliers* ou *valores atípicos*.

O *box plot* 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, q_3 dão uma noção da assimetria da distribuição. Os comprimentos das caudas são dados pelas linhas que vão do retângulo aos valores remotos e pelos valores atípicos.

Exemplo 3.8. Retomemos o exemplo dos 15 maiores municípios do Brasil, ordenados pelas populações. Usando o procedimento do Problema 17 (veja também o Problema 18), obtemos $q_1 = 105{,}7$, $q_2 = 135{,}8$, $q_3 = 208{,}6$. O diagrama para os cinco números $x_{(1)}$, q_1, $q_2 = md$, q_3, $x_{(15)}$ está na Figura 3.5 abaixo.

Figura 3.5: Esquema dos cinco números para o Exemplo 3.8.

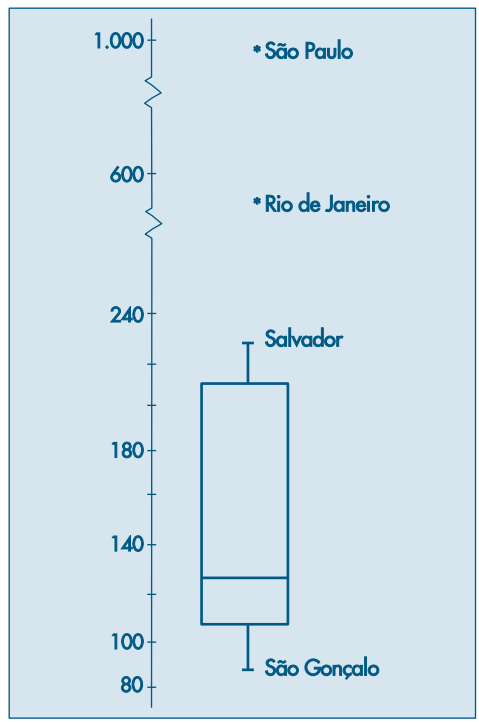


Temos que

$$\text{LI} = q_1 - (1{,}5)d_q = 105{,}7 - (1{,}5)\ (102{,}9) = -48{,}7,$$
$$\text{LS} = q_3 + (1{,}5)d_q = 208{,}6 + (1{,}5)\ (102{,}9) = 362{,}9.$$

Então, as cidades com populações acima de 3.629.000 habitantes são pontos exteriores, ou seja, Rio de Janeiro e São Paulo. O *box plot* correspondente está na Figura 3.6. Vemos que os dados têm uma distribuição assimétrica à direita, com 13 valores concentrados entre 80 e 230 e duas observações discrepantes, bastante afastadas do corpo principal dos dados.

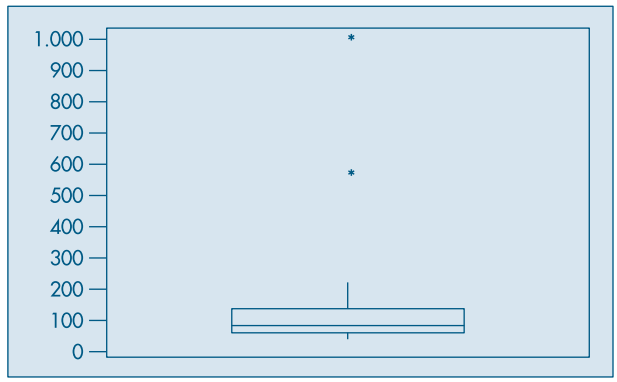
Figura 3.6: *Box plot* para os quinze maiores municípios do Brasil.



Do ponto de vista estatístico, um *outlier* pode ser produto de um erro de observação ou de arredondamento. No exemplo acima, as populações de São Paulo e Rio de Janeiro não são *outliers* neste sentido, pois elas representam dois valores realmente muito diferentes dos demais. Daí, usarmos o nome pontos (ou valores) exteriores. Contudo, na prática, estas duas denominações são freqüentemente usadas com o mesmo significado: observações fora de lugar, discrepantes ou atípicas.

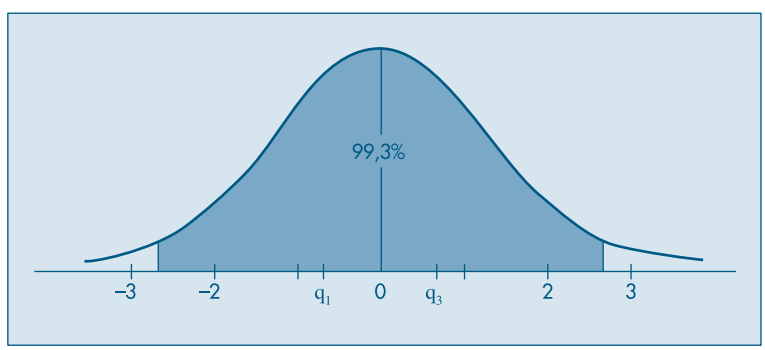
A Figura 3.7 mostra o *box plot* para as populações dos trinta municípios brasileiros, feito com o Minitab.

Figura 3.7: *Box plot* com Minitab.



A justificativa para usarmos os limites acima, LI = $q_1 - (1{,}5)d_q$ e LS = $q_3 + (1{,}5)d_q$, para definir as observações atípicas é a seguinte: considere uma curva normal com média zero e, portanto, com mediana zero. É fácil verificar (veja o Capítulo 7 e Tabela III) que $q_1 = -0{,}6745$, $q_2 = 0$, $q_3 = 0{,}6745$ e portanto $d_q = 1{,}349$. Segue-se que os limites são LI = –2,698 e LS = 2,698. A área entre estes dois valores, embaixo da curva normal, é 0,993, ou seja, 99,3% da distribuição está entre estes dois valores. Isto é, para dados com uma distribuição normal, os pontos exteriores constituirão cerca de 0,7% da distribuição. Veja a Figura 3.8.

Figura 3.8: Área sob a curva normal entre LI e LS.



Problemas

11. Construa o *box plot* para os dados do Exemplo 2.3, Capítulo 2. O que você pode concluir a respeito da distribuição?
12. Refaça a questão anterior com os dados do Problema 3 deste capítulo.
13. Faça um *box plot* para o Problema 10. Comente sobre a simetria, caudas e presença de valores atípicos.

3.5 Gráficos de Simetria

Os quantis podem ser úteis para se verificar se a distribuição dos dados é simétrica (ou aproximadamente simétrica).

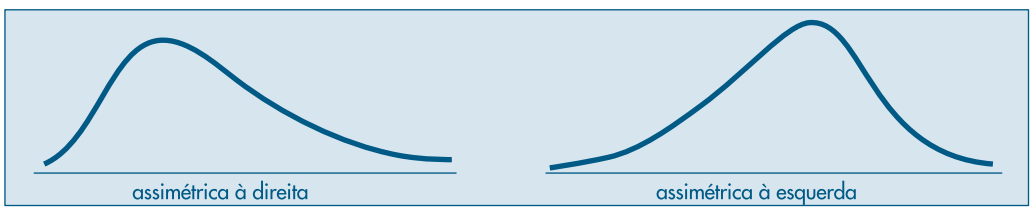
Se um conjunto de observações for perfeitamente simétrico devemos ter

$$q(0{,}5) - x_{(i)} = x_{(n+1-i)} - q(0{,}5), \tag{3.14}$$

onde $i = 1, 2, ..., n/2$, se n for par e $i = 1, 2, ..., (n + 1)/2$, se n for ímpar.

Pela relação (3.14), vemos que, se os quantis da direita estão mais afastados da mediana, do que os da esquerda, os dados serão *assimétricos à direita*. Se ocorrer o contrário, os dados serão *assimétricos à esquerda*. A Figura 3.9 ilustra essas duas situações.

Figura 3.9: Distribuições assimétricas.

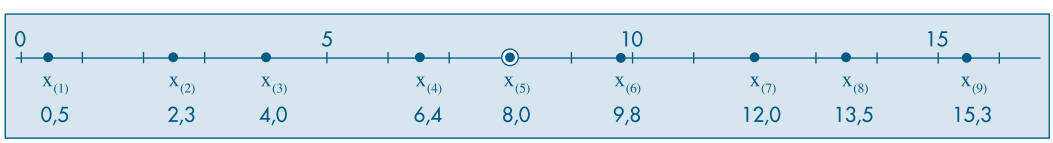


Para os dados do Exemplo 3.8, vemos que as observações são assimétricas à direita. Em geral, esse tipo de situação ocorre com dados positivos.

Podemos fazer um *gráfico de simetria*, usando a identidade (3.14). Chamando de u_i o primeiro membro e de v_i o segundo membro, fazendo-se um gráfico cartesiano, com os u_i's como abscissas e os v_i's como ordenadas, se os dados forem aproximadamente simétricos, os pares (u_i, v_i) estarão dispersos ao redor da reta $v = u$.

Exemplo 3.9. Considere os dados que, dispostos em ordem crescente, ficam representados no eixo real como na Figura 3.10.

Figura 3.10: Dados aproximadamente simétricos.

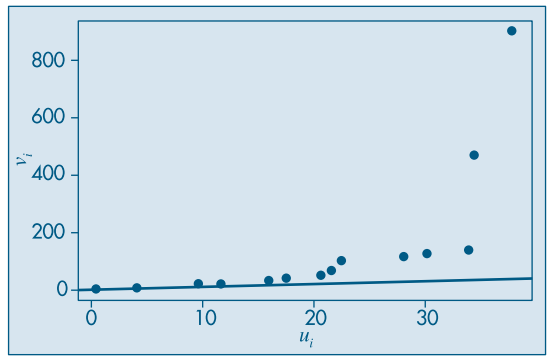


Esses dados são aproximadamente simétricos, pois como $q_2 = 8$, $u_i = q_2 - x_{(i)}$, $v_i = x_{(n+1-i)} - q_2$, teremos:

$$u_1 = 8{,}0 - 0{,}5 = 7{,}5,\ v_1 = 15{,}3 - 8{,}0 = 7{,}3,$$
$$u_2 = 8{,}0 - 2{,}3 = 5{,}7,\ v_2 = 13{,}5 - 8{,}0 = 5{,}5,$$
$$u_3 = 8{,}0 - 4{,}0 = 4{,}0,\ v_3 = 12{,}0 - 8{,}0 = 4{,}0,$$
$$u_4 = 8{,}0 - 6{,}4 = 1{,}6,\ v_4 = 9{,}8 - 8{,}0 = 1{,}8.$$

A Figura 3.11 mostra o gráfico de simetria para as populações dos trinta municípios do Brasil. Vemos que a maioria dos pontos estão acima da reta $v = u$, mostrando a assimetria à direita da distribuição dos valores. Nessa figura, vemos destacados os pontos correspondentes a Rio de Janeiro e São Paulo.

Figura 3.11: Gráfico de simetria para o CD-Municípios.



3.6 Transformações

Vários procedimentos estatísticos são baseados na suposição de que os dados provêm de uma distribuição normal (em forma de sino) ou então mais ou menos simétrica. Mas, em muitas situações de interesse prático, a distribuição dos dados da amostra é assimétrica e pode conter valores atípicos, como vimos em exemplos anteriores.

Se quisermos utilizar tais procedimentos, o que se propõe é efetuar uma transformação das observações, de modo a se obter uma distribuição mais simétrica e próxima da normal. Uma família de transformações freqüentemente utilizada é

$$x^{(p)} = \begin{cases} x^p, & \text{se } p > 0 \\ \ell n(x), & \text{se } p = 0 \\ -x^p, & \text{se } p < 0. \end{cases} \qquad (3.15)$$

Normalmente, o que se faz é experimentar valores de p na seqüência

$$\ldots, -3, -2, -1, -1/2, -1/3, -1/4, 0, 1/4, 1/3, 1/2, 1, 2, 3, \ldots$$

e para cada valor de p obtemos gráficos apropriados (histogramas, desenhos esquemáticos etc.) para os dados originais e transformados, de modo a escolhermos o valor mais adequado de p.

Vimos que, para dados positivos, a distribuição dos dados é usualmente assimétrica à direita. Para essas distribuições, a transformação acima com $0 < p < 1$ é apropriada, pois valores grandes de x decrescem mais, relativamente a valores pequenos. Para distribuições assimétricas à esquerda, tome $p > 1$.

Exemplo 3.10. Consideremos os dados das populações do CD-Municípios e tomemos alguns valores de p: 0, 1/4, 1/3, 1/2. Na Figura 3.12 temos os histogramas para os dados transformados e, na Figura 3.13, os respectivos *box plots*. Vemos que $p = 0$ (transformação logarítmica) e $p = 1/3$ (transformação raiz cúbica) fornecem distribuições mais próximas de uma distribuição simétrica.

Figura 3.12: Histogramas para os dados transformados. CD-Municípios.

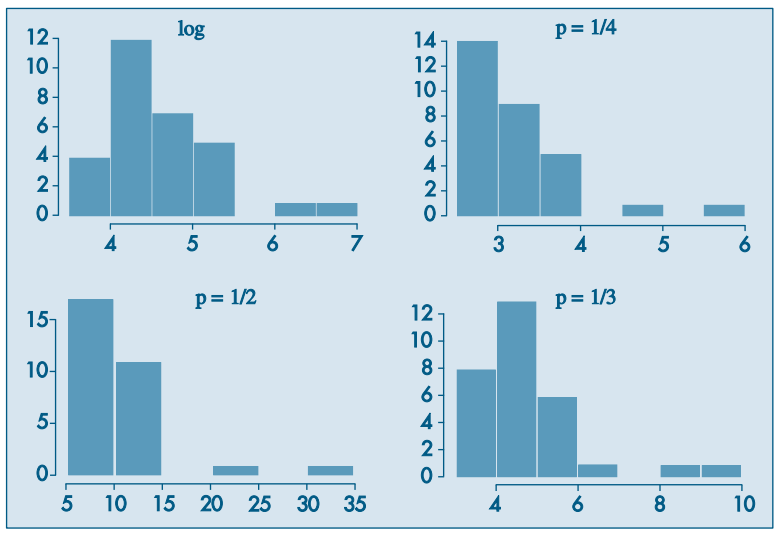
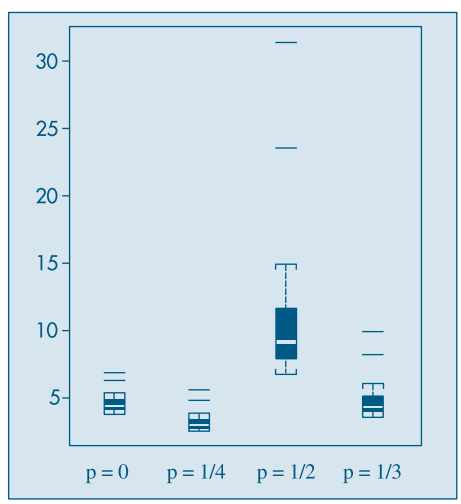


Figura 3.13: *Box plots* para os dados transformados. CD-Municípios. SPlus.



3.7 Exemplos Computacionais

Vamos retomar os exemplos estudados no Capítulo 2 e complementar as análises feitas com as técnicas aprendidas neste capítulo.

Exemplo 2.10. (continuação) Aqui temos as notas em Estatística de 100 alunos de Economia. Temos no Quadro 3.3 as principais medidas-resumo desse conjunto de dados, fornecidas pelo comando *describe* do Minitab.

Quadro 3.3: Medidas descritivas para o CD-Notas. Minitab.

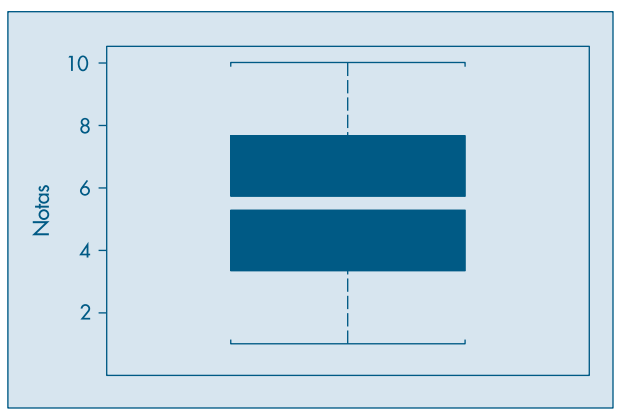
Descriptive Statistics

Variable	N	Mean	Median	Tr mean	StDev	SE Mean
C1	100	5.925	6.000	5.911	1.812	0.181

Variable	Min	Max	Q1	Q3
C1	1.500	10.000	4.625	7.375

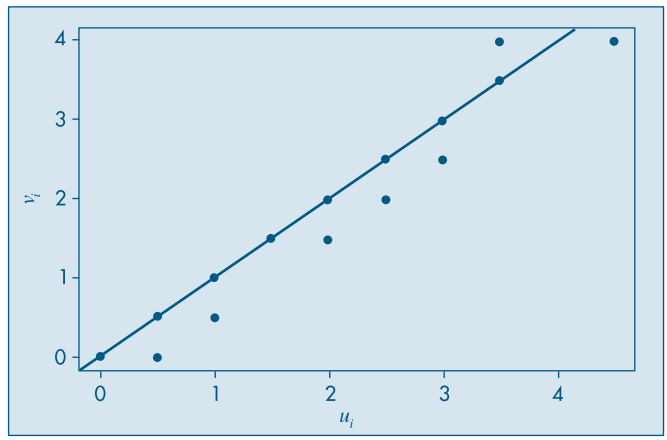
Vemos, por exemplo, que q_1 = 4,625, q_2 = 6,000 e q_3 = 7,375 e, portanto, $d_q = q_3 - q_1$ = 2,75. O desvio padrão é dp = 1,812. Vimos que a distribuição das notas é razoavelmente simétrica, não havendo valores atípicos, o que é confirmado pelo *box plot* da Figura 3.14.

Figura 3.14: *Box plot* para o CD-Notas. SPlus.



O gráfico de simetria está na Figura 3.15, mostrando também a reta $u = v$. Note que os pontos dispõem-se ao redor da reta, estando vários deles sobre ela, indicando a quase-simetria dos dados. Deveríamos ter 50 pontos no gráfico, mas há vários pares (u_i, v_i) repetidos.

Figura 3.15: Gráfico de simetria para o CD-Notas.



Exemplo 2.11. (continuação) Os dados de temperatura (diários) na cidade de São Paulo, no período considerado, são ligeiramente assimétricos à esquerda. O comando *summary* do SPlus fornece as medidas descritivas do Quadro 3.4. Note que o Minitab fornece mais informações que o SPlus por meio desses comandos.

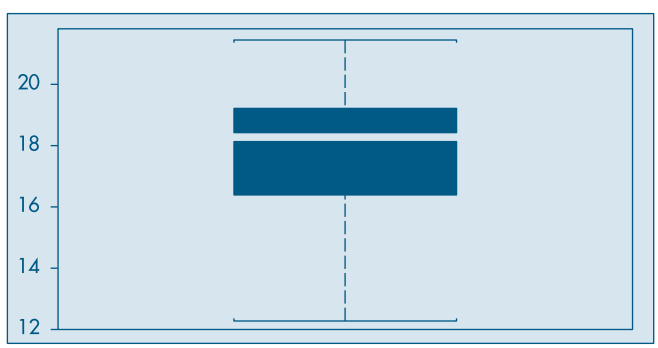
Quadro 3.4. Medidas descritivas para temperaturas. SPlus.

> summary (temp)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
12.3	16	17.7	17.22	18.6	21

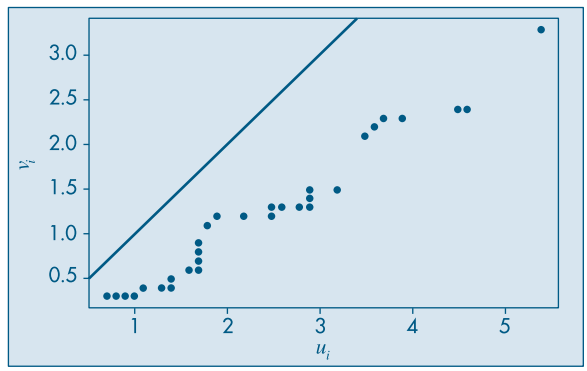
Temos, por exemplo, $q_1 = 16$, $q_2 = 17{,}7$ e $q_3 = 18{,}6$. A amplitude amostral é $x_{(n)} - x_{(1)} = 8{,}7$ e a distância interquartil é $d_q = 2{,}6$. O *box plot* está na Figura 3.16, que mostra a assimetria. Não há valores atípicos.

Figura 3.16: *Box plot* para as temperaturas de São Paulo. CD-Poluição. SPlus.



No gráfico de simetria na Figura 3.17, todos os pontos estão abaixo da reta $u = v$, mostrando que $u_i > v_i$, para todo $i = 1, 2, ..., 60$, ou seja, as distâncias da mediana aos quantis inferiores são maiores do que as distâncias dos quantis superiores à mediana, indicando que a distribuição das observações é assimétrica à esquerda.

Figura 3.17: Gráfico de simetria para as temperaturas de São Paulo. CD-Poluição.



3.8 Problemas e Complementos

14. Mostre que:

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

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

(c) $\sum_{i=1}^{k}n_i(x_i - \bar{x})^2 = \sum_{i=1}^{k}n_i x_i^2 - n\bar{x}^2$

(d) $\sum_{i=1}^{k}f_i(x_i - \bar{x})^2 = \sum_{i=1}^{k}f_i x_i^2 - \bar{x}^2$

15. Usando os resultados da questão anterior, calcule as variâncias dos Problemas 1 e 2 deste capítulo.

16. Os dados abaixo representam as vendas semanais, em classes de salários mínimos, de vendedores de gêneros alimentícios:

Vendas semanais	Nº de vendedores
30 ⊢ 35	2
35 ⊢ 40	10
40 ⊢ 45	18
45 ⊢ 50	50
50 ⊢ 55	70
55 ⊢ 60	30
60 ⊢ 65	18
65 ⊢ 70	2

(a) Faça o histograma das observações.
(b) Calcule a média da amostra, $\bar{x}$.
(c) Calcule o desvio padrão da amostra, s.
(d) Qual a porcentagem das observações compreendidas entre $\bar{x}-2s$ e $\bar{x}+2s$?
(e) Calcule a mediana.

17. Quantis. Para calcular os quantis de uma sequência de valores de uma variável X poderíamos usar a função de distribuição acumulada ou empírica, definida no Problema 17 do Capítulo 2. Essa função fornece, para cada número real x, a proporção das observações menores ou iguais a x. No Exemplo 3.5, temos

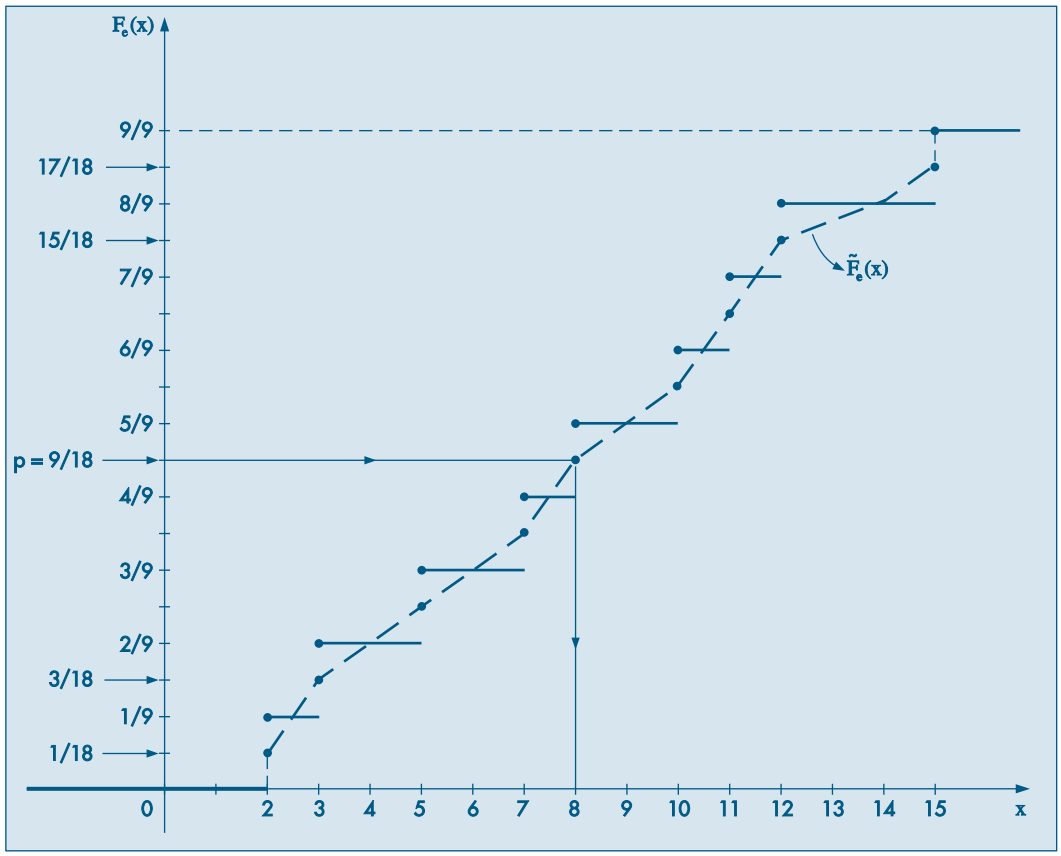
$$F_e(x) = \begin{cases} 0, & \text{se } x < 2 \\ 1/9, & \text{se } 2 \leqslant x < 3 \\ 2/9, & \text{se } 3 \leqslant x < 5 \\ 3/9, & \text{se } 5 \leqslant x < 7 \\ 4/9, & \text{se } 7 \leqslant x < 8 \\ 5/9, & \text{se } 8 \leqslant x < 10 \\ 6/9, & \text{se } 10 \leqslant x < 11 \\ 7/9, & \text{se } 11 \leqslant x < 12 \\ 8/9, & \text{se } 12 \leqslant x < 15 \\ 1, & \text{se } x \geqslant 15. \end{cases} \tag{3.16}$$

O gráfico de $F_e(x)$ está na Figura 3.18. Note que não há nenhum valor de x tal que $F_e(x) = 0{,}5$ e $F_e(2) = 1/9$, $F_e(3) = 2/9$, ..., $F_e(15) = 1$, ou seja, podemos escrever de modo geral

$$F_e(x_{(i)}) = \frac{i}{9}, \; i = 1, 2, \ldots, 9. \tag{3.17}$$

Em particular, $F_e(\text{md}) = F(x_{(5)}) = F_e(8) = 5/9 = 0{,}556$. Portanto, ou mudamos nossa definição de mediana, ou $F_e(.)$ não pode ser usada para definir precisamente mediana ou, em geral, um quantil $q(p)$.

Figura 3.18: Funções de distribuição empírica (F_e) e f.d.e. alisada ($\tilde{F}_e$) para o Exemplo 3.5.



Mas vejamos que $F_e(\cdot)$ pode ser a base para tal definição. Considere "alisar" ou "suavizar" $F_e(\cdot)$, como feito na Figura 3.18, de modo a obter uma curva contínua $\tilde{F}_e(x)$, que passa pelos pontos $(x_{(i)}, p_i)$, onde

$$p_i = \frac{i - 0{,}5}{9}, \; i = 1, 2, ..., 9. \tag{3.18}$$

Observe que $0 < p_1 < 1/9$, $1/9 < p_2 < 2/9$ etc. Com esse procedimento, notamos que

$$\tilde{F}_e(x_{(1)}) = 1/18, ..., \tilde{F}_e(x_{(5)}) = 9/18 = 0{,}5, ..., \tilde{F}_e(x_{(9)}) = 17/18,$$

ou seja, podemos escrever

$$\tilde{F}_e(x_{(i)}) = \frac{i - 0{,}5}{n}, \; i = 1, 2, ..., n, \tag{3.19}$$

sendo que no nosso caso $n = 9$. Com essa modificação, obtemos que $\tilde{F}_e(\text{md}) = \tilde{F}_e(8) = 0{,}5$, e para cada p, $0 \leq p < 1$, podemos obter de modo unívoco o quantil $q(p)$, tomando-se a função inversa $\tilde{F}_e^{-1}(p)$. Ou seja, considere uma reta horizontal passando por p no eixo das ordenadas, até encontrar a curva contínua e depois baixe uma reta vertical até encontrar $q(p)$ no eixo das abscissas.

Uma maneira equivalente de proceder nos leva à seguinte definição para calcular $q(p)$, para qualquer p, $0 < p < 1$.

Definição. O p-quantil é definido por

$$q(p) = \begin{cases} x_{(i)}, & \text{se } p = p_i = \dfrac{i - 0{,}5}{n},\ i = 1, 2, \ldots, n \\ (1 - f_i)x_{(i)} + f_i x_{(i+1)}, & \text{se } p_i < p < p_{i+1} \\ x_{(1)}, & \text{se } p < p_1 \\ x_{(n)}, & \text{se } p > p_n, \end{cases}$$

onde $f_i = \dfrac{(p - p_i)}{(p_{i+1} - p_i)}$.

Notamos, então, que se p coincidir com a proporção p_i, o quantil será a i-ésima observação, $x_{(i)}$. Se $p_i < p < p_{i+1}$, o quantil estará no segmento de reta que une $(p_i, x_{(i)})$ e $(p_{i+1}, x_{(i+1)})$. De fato, a reta passando por $(p_i, x_{(i)})$ e $(p, q(p))$ é

$$q(p) - x_{(i)} = \frac{x_{(i+1)} - x_{(i)}}{p_{i+1} - p_i}\ (p - p_i).$$

Exemplo 3.5. (continuação) Usando a definição obtemos:

$$q(0{,}1) = (0{,}6)x_{(1)} + (0{,}4)x_{(2)} = (0{,}6)(2) + (0{,}4)(3) = 2{,}4;$$

$$q(0{,}2) = (0{,}7)x_{(2)} + (0{,}3)x_{(3)} = (0{,}7)(3) + (0{,}3)(5) = 3{,}6;$$

$$q(0{,}25) = (0{,}25)x_{(2)} + 0{,}75x_{(3)} = 4{,}5;$$

$$q(0{,}5) = x_{(5)} = 8;$$

$$q(0{,}75) = (0{,}75)x_{(7)} + (0{,}25)x_{(8)} = (0{,}75)(11) + (0{,}25)(12) = 11{,}25.$$

18. Considere o CD-Municípios e tome somente os 15 maiores, relativamente à sua população. Calcule $q(0, 1)$, $q(0, 2)$, q_1, q_2, q_3.

19. O número de divórcios na cidade, de acordo com a duração do casamento, está representado na tabela abaixo.
    (a) Qual a duração média dos casamentos? E a mediana?
    (b) Encontre a variância e o desvio padrão da duração dos casamentos.
    (c) Construa o histograma da distribuição.
    (d) Encontre o 1º e o 9º decis.
    (e) Qual o intervalo interquantil?

Anos de casamento	Nº de divórcios
0 ⊢ 6	2.800
6 ⊢ 12	1.400
12 ⊢ 18	600
18 ⊢ 24	150
24 ⊢ 30	50

20. O Departamento Pessoal de uma certa firma fez um levantamento dos salários dos 120 funcionários do setor administrativo, obtendo os resultados (em salários mínimos) da tabela abaixo.
    (a) Esboce o histograma correspondente.
    (b) Calcule a média, a variância e o desvio padrão.
    (c) Calcule o 1º quartil e a mediana.

Faixa salarial	Freqüência relativa
0 ⊢ 2	0,25
2 ⊢ 4	0,40
4 ⊢ 6	0,20
6 ⊢ 10	0,15

    (d) Se for concedido um aumento de 100% para todos os 120 funcionários, haverá alteração na média? E na variância? Justifique sua resposta.
    (e) Se for concedido um abono de dois salários mínimos para todos os 120 funcionários, haverá alteração na média? E na variância? E na mediana? Justifique sua resposta.

21. O que acontece com a mediana, a média e o desvio padrão de uma série de dados quando:
    (a) cada observação é multiplicada por 2?
    (b) soma-se 10 a cada observação?
    (c) subtrai-se a média geral $\bar{x}$ de cada observação?
    (d) de cada observação subtrai-se $\bar{x}$ e divide-se pelo desvio padrão $\mathrm{dp}(x)$?

22. Na companhia A, a média dos salários é 10.000 unidades e o 3º quartil é 5.000.
    (a) Se você se apresentasse como candidato a funcionário nessa firma e se o seu salário fosse escolhido ao acaso entre todos os possíveis salários, o que seria mais provável: ganhar mais ou menos que 5.000 unidades?
    (b) Suponha que na companhia B a média dos salários seja 7.000 unidades, a variância praticamente zero e o salário também seja escolhido ao acaso. Em qual companhia você se apresentaria para procurar emprego?

23. Estamos interessados em estudar a idade dos 12.325 funcionários da Cia. Distribuidora de Leite Teco, e isso será feito por meio de uma amostra. Para determinar que tamanho deverá ter essa amostra, foi colhida uma amostra-piloto. As idades observadas foram: 42, 35, 27, 21, 55, 18, 27, 30, 21, 24.
    (a) Determine as medidas descritivas dos dados que você conhece.
    (b) Qual dessas medidas você acredita que será a mais importante para julgar o tamanho final da amostra? Por quê?

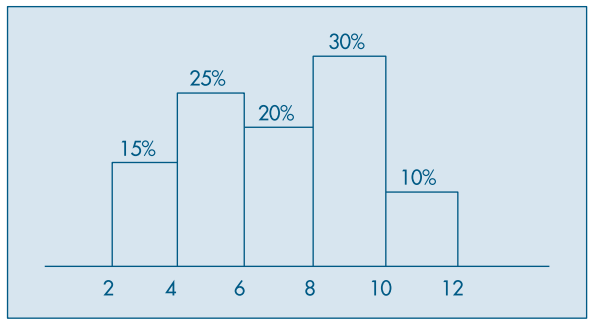
24. Estudando-se o consumo diário de leite, verificou-se que, em certa região, 20% das famílias consomem até um litro, 50% consomem entre um e dois litros, 20% consomem entre dois e três litros e o restante consome entre três e cinco litros. Para a variável em estudo:
    (a) Escreva as informações acima na forma de uma tabela de freqüências.
    (b) Construa o histograma.
    (c) Calcule a média e a mediana.
    (d) Calcule a variância e o desvio padrão.
    (e) Qual o valor do 1º quartil?

25. A distribuição de freqüências do salário anual dos moradores do bairro A que têm alguma forma de rendimento é apresentada na tabela abaixo:

Faixa salarial (× 10 salários mínimos)	Freqüência
0 ⊢ 2	10.000
2 ⊢ 4	3.900
4 ⊢ 6	2.000
6 ⊢ 8	1.100
8 ⊢ 10	800
10 ⊢ 12	700
12 ⊢ 14	2.000
Total	20.500

    (a) Construa um histograma da distribuição.
    (b) Qual a média e o desvio padrão da variável salário?
    (c) O bairro B apresenta, para a mesma variável, uma média de 7,2 e um desvio padrão de 15,1. Em qual dos bairros a população é mais homogênea quanto à renda?
    (d) Construa a função de distribuição acumulada e determine qual a faixa salarial dos 10% mais ricos da população do bairro.
    (e) Qual a "riqueza total" dos moradores do bairro?

26. Dado o histograma abaixo, calcular a média, a variância, a moda, a mediana e o 1º quartil.



27. Em uma granja foi observada a distribuição dos frangos em relação ao peso, que era a seguinte:

Peso (gramas)	n_i
960 ⊢ 980	60
980 ⊢ 1.000	160
1.000 ⊢ 1.020	280
1.020 ⊢ 1.040	260
1.040 ⊢ 1.060	160
1.060 ⊢ 1.080	80

(a) Qual a média da distribuição?
(b) Qual a variância da distribuição?
(c) Construa o histograma.
(d) Queremos dividir os frangos em quatro categorias, em relação ao peso, de modo que:
— os 20% mais leves sejam da categoria D;
— os 30% seguintes sejam da categoria C;
— os 30% seguintes sejam da categoria B;
— os 20% seguintes (ou seja, os 20% mais pesados) sejam da categoria A.

Quais os limites de peso entre as categorias A, B, C e D?

(e) O granjeiro decide separar deste lote os animais com peso inferior a dois desvios padrões abaixo da média para receberem ração reforçada, e também separar os animais com peso superior a um e meio desvio padrão acima da média para usá-los como reprodutores.

Qual a porcentagem de animais que serão separados em cada caso?

28. A idade média dos candidatos a um determinado curso de aperfeiçoamento sempre foi baixa, da ordem de 22 anos. Como esse curso foi planejado para atender a todas as idades, decidiu-se fazer uma campanha de divulgação. Para se verificar se a campanha foi ou não eficiente, fez-se um levantamento da idade dos candidatos à última promoção, e os resultados estão na tabela a seguir.

Idade	Freqüência	Porcentagem
18 ⊢ 20	18	36
20 ⊢ 22	12	24
22 ⊢ 26	10	20
26 ⊢ 30	8	16
30 ⊢ 36	2	4
Total	50	100

(a) Baseando-se nesses resultados, você diria que a campanha produziu algum efeito (isto é, aumentou a idade média)?
(b) Um outro pesquisador decidiu usar a seguinte regra: se a diferença $\bar{x} - 22$ fosse maior que o valor $2\text{dp}(X)/\sqrt{n}$, então a campanha teria surtido efeito. Qual a conclusão dele, baseada nos dados?
(c) Faça o histograma da distribuição.

29. Para se estudar o desempenho de duas corretoras de ações, selecionou-se de cada uma delas amostras aleatórias das ações negociadas. Para cada ação selecionada, computou-se a porcentagem de lucro apresentada durante um período fixado de tempo. Os dados estão a seguir.

Corretora A		
45	60	54
62	55	70
38	48	64
55	56	55
54	59	48
65	55	60

Corretora B		
57	55	58
50	52	59
59	55	56
61	52	53
57	57	50
55	58	54
59	51	56

Que tipo de informação revelam esses dados? (Sugestão: use a análise proposta nas Seções 3.3 e 3.4.)

30. Para verificar a homogeneidade das duas populações do problema anterior, um estatístico sugeriu que se usasse o quociente $F = \dfrac{\text{var}\,(X/A)}{\text{var}\,(X/B)}$, mas não disse qual decisão tomar baseado nesse valor. Que regra de decisão você adotaria para dizer se são homogêneas ou não ($\text{var}(X/A)$ = variância de X, para a corretora A; X = % de lucro)?

31. Faça um desenho esquemático (*box plot*) para os dados da corretora A e um para os dados da corretora B. Compare os dois conjuntos de dados por meio desses desenhos.

32. Para decidir se o desempenho das duas corretoras do exercício 29 são semelhantes ou não, adotou-se o seguinte teste: sejam

$$t = \frac{\bar{x}_A - \bar{x}_B}{S_*^2\sqrt{1/n_A + 1/n_B}}, \quad S_*^2 = \frac{(n_A - 1)\,\text{var}(X/A) + (n_B - 1)\text{var}(X/B)}{n_A + n_B - 2}.$$

Caso $|t| < 2$, os desempenhos são semelhantes, caso contrário, são diferentes. Qual seria a sua conclusão? Aqui, n_A é o número de ações selecionadas da corretora A e nomenclatura análoga para n_B.

33. Um órgão do governo do estado está interessado em determinar padrões sobre o investimento em educação, por habitante, realizado pelas prefeituras. De um levantamento de dez cidades, foram obtidos os valores (codificados) da tabela abaixo:

Cidade	A	B	C	D	E	F	G	H	I	J
Investimento	20	16	14	8	19	15	14	16	19	18

Nesse caso, será considerado como *investimento básico* a *média final* das observações, calculada da seguinte maneira:

1. Obtém-se uma média inicial.
2. Eliminam-se do conjunto aquelas observações que forem superiores à média inicial mais duas vezes o desvio padrão, ou inferiores à média inicial menos duas vezes o desvio padrão.
3. Calcula-se a média final com o novo conjunto de observações.

Qual o investimento básico que você daria como resposta?

Observação: O procedimento do item 2 tem a finalidade de eliminar do conjunto a cidade cujo investimento é muito diferente dos demais.

34. Estudando-se a distribuição das idades dos funcionários de duas repartições públicas, obtiveram-se algumas medidas que estão no quadro abaixo. Esboce o histograma alisado das duas distribuições, indicando nele as medidas descritas no quadro. Comente as principais diferenças entre os dois histogramas.

Repartição	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo	dp
A	18	27	33	33	39	48	5
B	18	23	32	33	42	48	10

35. Decidiu-se investigar a distribuição dos profissionais com nível universitário em duas regiões, A e B. As informações pertinentes foram obtidas e encontram-se no quadro abaixo, expressas em salários mínimos. Esboce a distribuição (histograma alisado) dos salários de cada região, indicando no gráfico as medidas apresentadas no quadro. Faça também uma descrição rápida das principais diferenças observadas nos gráficos.

Região	Média	dp	Mediana	Moda	q_1	q_3	$x_{(1)}$	$x_{(n)}$
A	20,00	4,00	20,32	20,15	17,32	22,68	8,00	32,00
B	20,00	6,00	18,00	17,00	16,00	24,00	14,00	42,00

36. Construa o desenho esquemático para os dados do Problema 6, do Capítulo 2. Obtenha conclusões a respeito da distribuição, a partir desse desenho.

37. Usando os dados da variável qualitativa *região de procedência*, da Tabela 2.1, transforme-a na variável quantitativa X, definida do seguinte modo:

$$X = \begin{cases} 1, \text{ se a região de procedência for capital;} \\ 0, \text{ se a região de procedência for interior ou outra.} \end{cases}$$

(a) Calcule $\bar{x}$ e $\text{var}(X)$.
(b) Qual a interpretação de $\bar{x}$?
(c) Construa um histograma para X.

38. No Problema 9, do Capítulo 2, temos os resultados de 25 funcionários em vários exames a que se submeteram. Sabe-se agora que os critérios adotados em cada exame não são comparáveis, por isso decidiu-se usar o *desempenho relativo* em cada exame. Essa medida será obtida do seguinte modo:

(I) Para cada exame serão calculados a média $\bar{x}$ e o desvio padrão $\text{dp}(X)$.
(II) A nota X de cada aluno será padronizada do seguinte modo:

$$Z = \frac{X - \bar{x}}{\text{dp}(X)}.$$

(a) Interprete o significado de Z.
(b) Calcule as notas padronizadas dos funcionários para o exame de Estatística.
(c) Com os resultados obtidos em *(b)*, calcule $\bar{z}$ e $\text{dp}(Z)$.
(d) Se alguma das notas padronizadas estiver acima de $2\text{dp}(Z)$ ou abaixo de $-2\text{dp}(Z)$, esse funcionário deve ser considerado um caso atípico. Existe algum nessa situação?
(e) O funcionário 1 obteve 9,0 em Direito, em Estatística e em Política. Em que disciplina o seu desempenho relativo foi melhor?

39. Média aparada. Se $0 < \alpha < 1$, uma média aparada a $100\alpha\%$ é obtida eliminando $100\alpha\%$ das menores observações e $100\alpha\%$ das maiores observações e calculando-se a média aritmética das restantes. Por exemplo, se tivermos 10 observações ordenadas $x_{(1)} < x_{(2)} < \ldots < x_{(10)}$, a média aparada a 10% é

$$\bar{x}(0{,}10) = \frac{x_{(2)} + x_{(3)} + \ldots + x_{(9)}}{8}$$

Se $\alpha = 0{,}25$, $\bar{x}(0{,}25)$ é chamada *meia-média*.
Calcule a média aparada a 10% e 25% para os dados de salários da Tabela 2.1.

40. Coeficiente de variação. Como vimos na seção 3.3, o desvio padrão é bastante afetado pela magnitude dos dados, ou seja, ele não é uma medida resistente. Se quisermos comparar a variabilidade de dois conjuntos de dados podemos usar o coeficiente de variação, que é definido como a razão entre o desvio padrão, S, e a média amostral e usualmente expresso em porcentagem:

$$cv = \frac{S}{\bar{x}} 100\%.$$

Calcule o coeficiente de variação para as regiões A e B e do Problema 35 e comente o resultado.

41. Desvio absoluto mediano. Esta é uma medida de dispersão dos dados $x_1, \ldots, x_n$, definida por:

$$\text{dam} = \text{med}_{1 \le j \le n} \left| x_j - \text{med}_{1 \le i \le n}(x_i) \right|.$$

Ou seja, calculamos a mediana dos dados, depois os desvios absolutos dos dados em relação à mediana e, finalmente, a mediana desses desvios absolutos. Vamos considerar os dados abaixo, extraídos de Graedel e Kleiner (1985) e que representam velocidades do vento no aeroporto de Philadelphia (EUA) para os primeiros 15 dias de dezembro de 1974. Vemos que há uma observação muito diferente das demais (61,1), mas que representa um dado real: no dia 2 de dezembro houve uma tempestade forte com chuva e vento.

22,2	61,1	13,0	27,8	22,2
7,4	7,4	7,4	20,4	20,4
20,4	11,1	13,0	7,4	14,8

Calculando-se as medidas de posição e dispersão estudadas, obtemos:

$$\bar{x} = 18{,}4,\ \bar{x}(0{,}20) = 15{,}8;$$
$$\text{md} = 14{,}8,\ q_1 = 8{,}3,\ q_3 = 21{,}8;$$
$$d_q = 14{,}8,\ \text{dam} = 7{,}4,\ \text{dp}(X) = 13{,}5.$$

Observemos que, retirando-se o valor atípico 61,1, a média passa a ser 15,3 e o desvio padrão 6,8, valor este mais próximo do dam.

42. Calcule o desvio absoluto mediano para as populações do CD-Brasil.

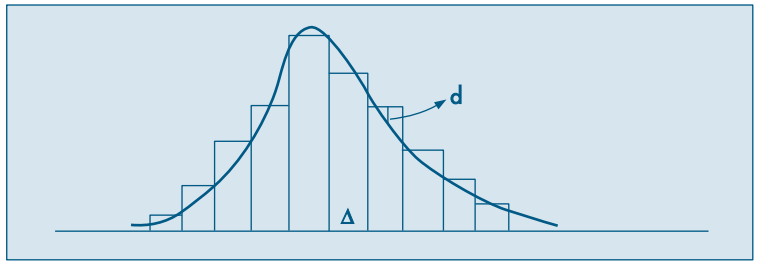
43. Calcule as principais medidas de posição e dispersão (incluindo a média aparada e o dam) para:
    (a) variável CO no CD-Poluição;
    (b) salários de mecânicos, CD-Salários; e
    (c) variável preço, CD-Veículos.

44. Construa os histogramas, ramo-e-folhas e desenhos esquemáticos para as variáveis do problema anterior.

45. Faça um gráfico de quantis e um de simetria para os dados do Problema 3. Os dados são simétricos? Comente.

46. Para o CD-Temperaturas e para a variável *temperatura de Ubatuba*, obtenha um gráfico de quantis e um gráfico de simetria. Os dados são simétricos? Comente.

47. O histograma dá uma idéia de como é a verdadeira *densidade de freqüências* da população da qual os dados foram selecionados. Suponha que tenhamos o histograma da figura abaixo e que a curva suave seja a verdadeira densidade populacional desconhecida.



Considere as distâncias entre o histograma e a densidade. Suponha que queiramos determinar a amplitude de classe Δ do histograma de modo a minimizar a maior distância (em valor absoluto). Freedman e Diaconis (1981) mostraram que o valor de Δ é dado aproximadamente por

$$\Delta = 1{,}349\tilde{S}\left(\frac{\log n}{n}\right)^{1/3},$$

em que $\tilde{S}$ é um estimador robusto do desvio padrão populacional. Por exemplo, podemos tomar

$$\tilde{S} = \frac{d_q}{1{,}349},$$

em que $d_q = q_3 - q_1$ é a distância interquartil, devido ao fato de, numa distribuição normal, $d_q = 1{,}349\sigma$, sendo o σ o desvio padrão. Segue-se que Δ é dado por

$$\Delta = d_q\left(\frac{\log n}{n}\right)^{1/3}.$$

Usando esse resultado, o número de classes a considerar num histograma é obtido por meio de $\dfrac{(x_{(n)} - x_{(1)})}{\Delta}$.

48. Use o problema anterior para construir histogramas para:
   (a) variável umid (umidade) do CD-Poluição;
   (b) variável salário dos professores do CD-Salários; e
   (c) a temperatura de Cananéia, do CD-Temperaturas.