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

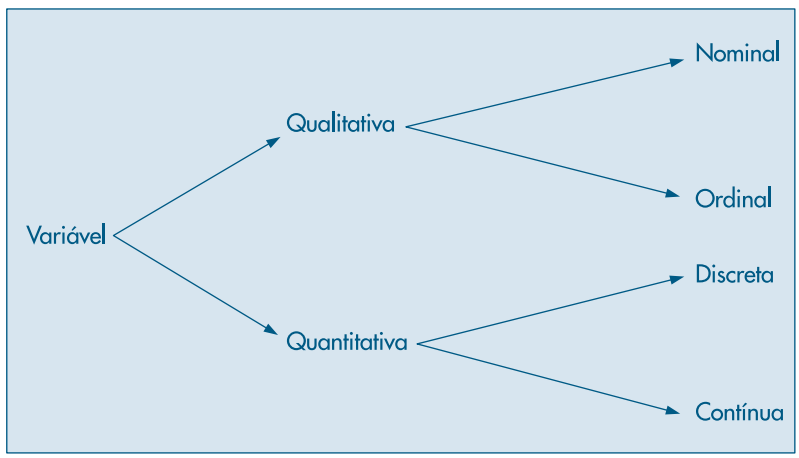
Algumas variáveis, como sexo, educação, estado civil, apresentam como possíveis realizações uma qualidade (ou atributo) do indivíduo pesquisado, ao passo que outras, como número de filhos, salário, idade, apresentam como possíveis realizações números resultantes de uma contagem ou mensuração. As variáveis 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

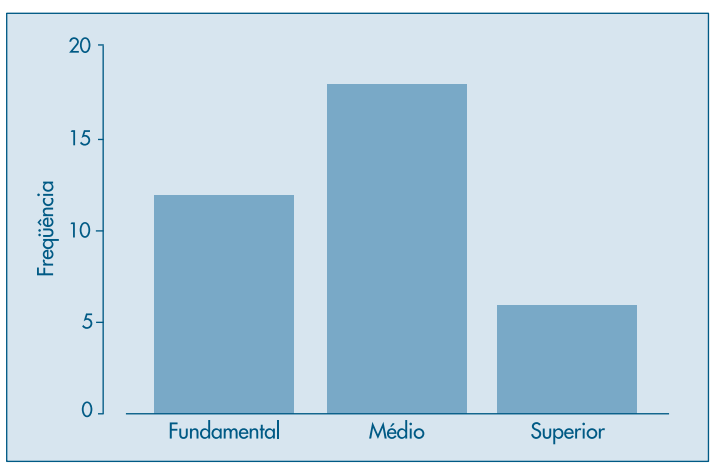
Como já salientamos no Capítulo 1, a representação gráfica da distribuição de uma variável tem a vantagem de, rápida e concisamente, informar sobre sua variabilidade. Existem vários gráficos que podem ser utilizados e abordaremos aqui os mais simples para variáveis quantitativas. No Capítulo 3, voltaremos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

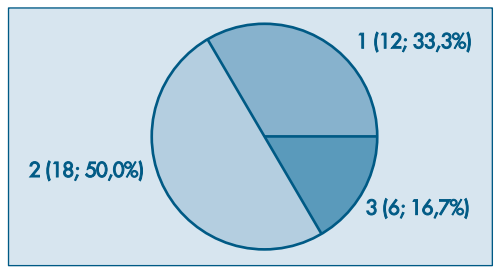
Exemplo 2.4. Tomemos como ilustração a variável Y: grau de instrução, exemplificada nas Tabelas 2.2 e 2.3. O gráfico em barras consiste em construir retângulos ou barras, em que uma das dimensões é proporcional à magnitude a ser representada (n_i ou f_i), sendo a outra arbitrária, porém igual para todas 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.

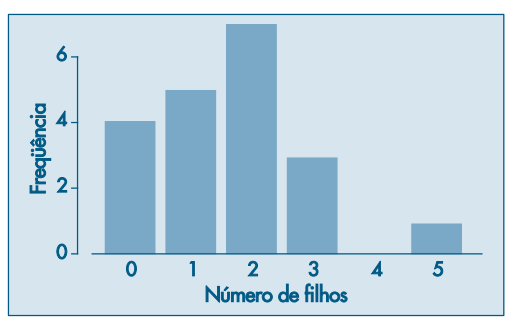
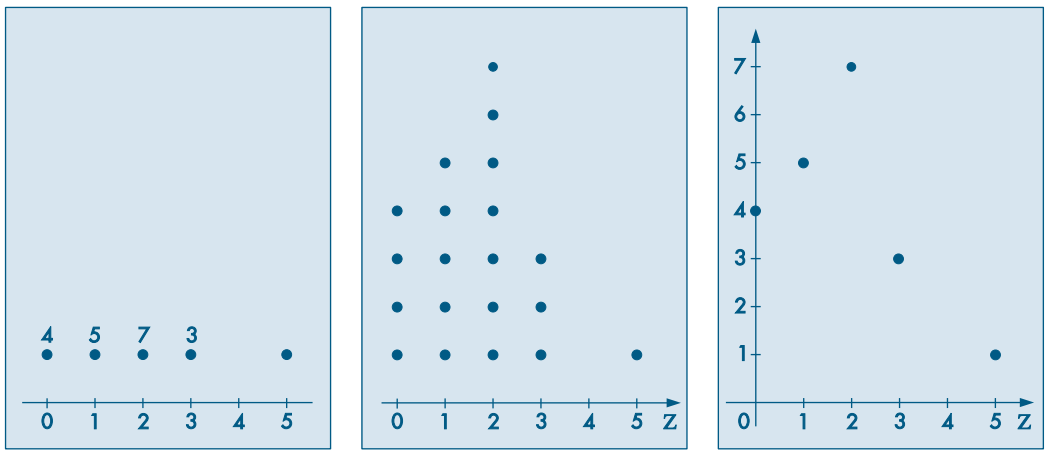


Figura 2.5: Gráficos de dispersão unidimensionais para a variável Z: número de filhos.



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.

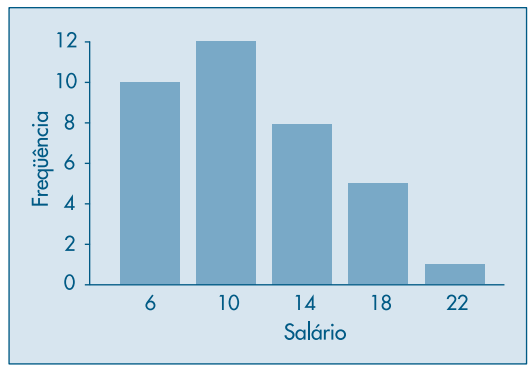
Nº de filhos z_i	Freqüência n_i	Porcentagem $100\, f_i$
0	4	20
1	5	25
2	7	35
3	3	15
5	1	5
Total	20	100

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

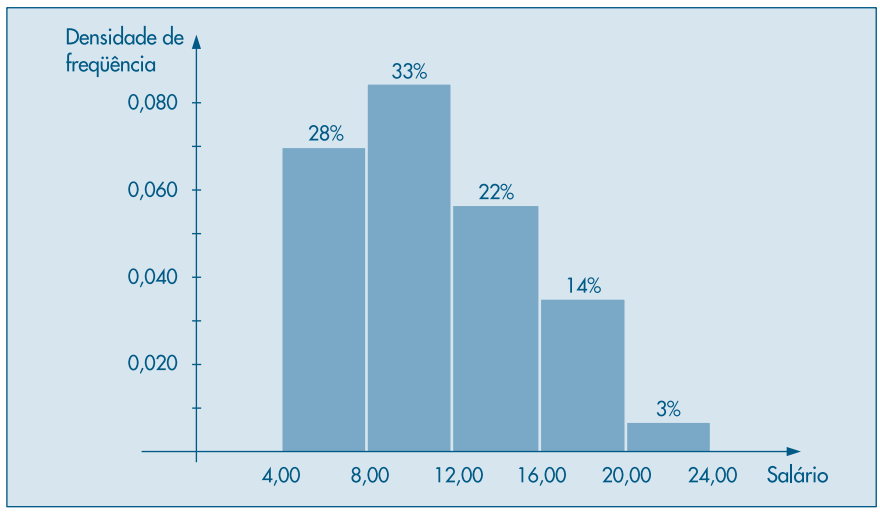
Fonte: Tabela 2.4.

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.

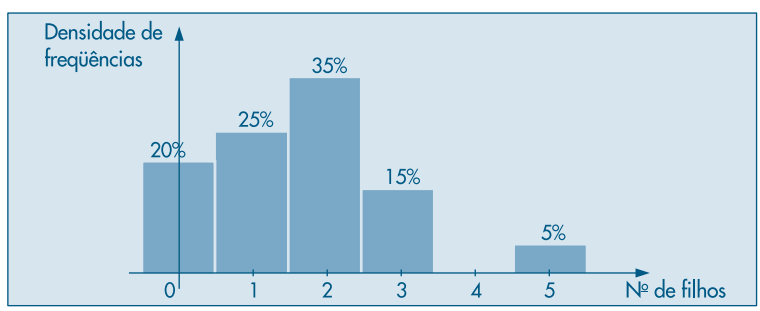
Figura 2.7: Histograma da variável *S*: salários.



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.

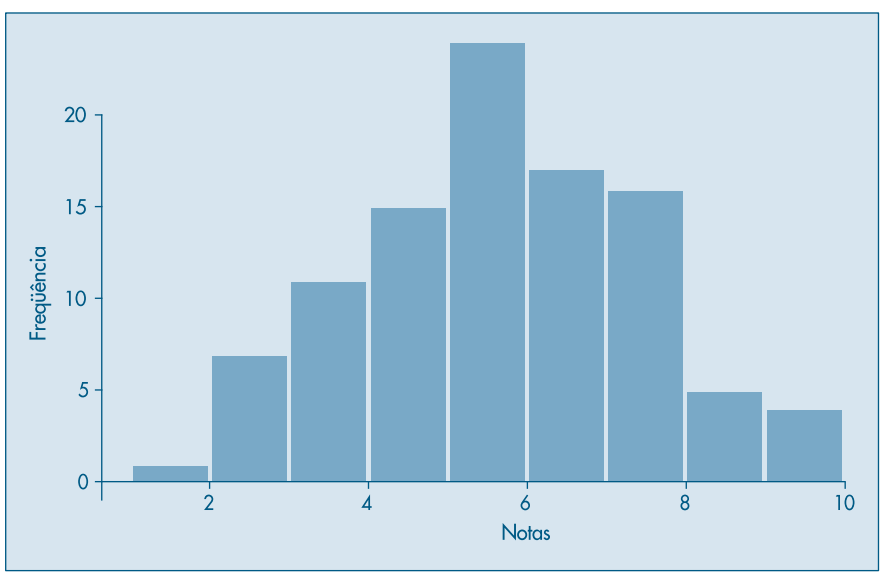


Figura 2.13: Gráfico de dispersão unidimensional para o CD-Notas. Minitab.

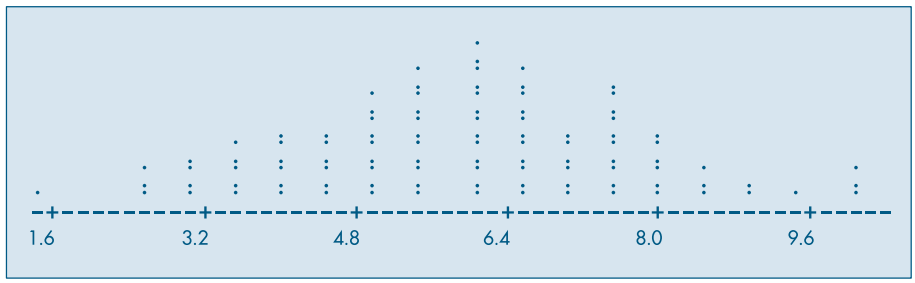
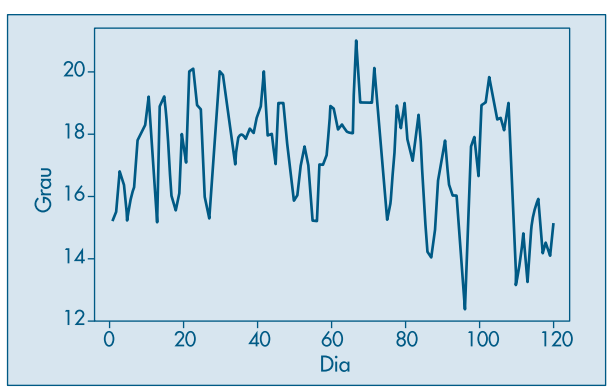


Figura 2.14: Ramo-e-folhas para o CD-Notas. Minitab.

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.

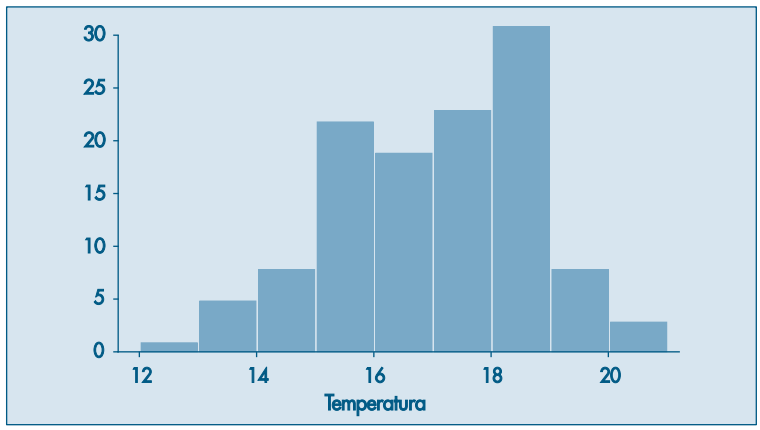


Figura 2.17: Gráfico de dispersão unidimensional para os dados de temperatura de São Paulo. Minitab.

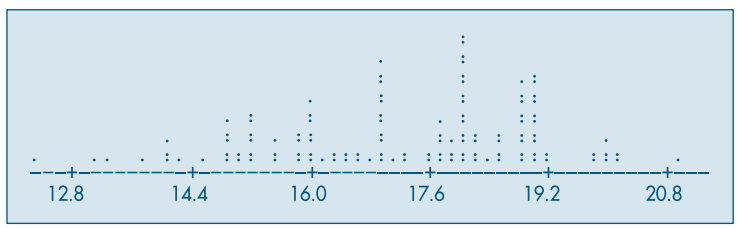


Figura 2.18: Ramo-e-folhas para os dados de temperatura de São Paulo. Minitab.

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.

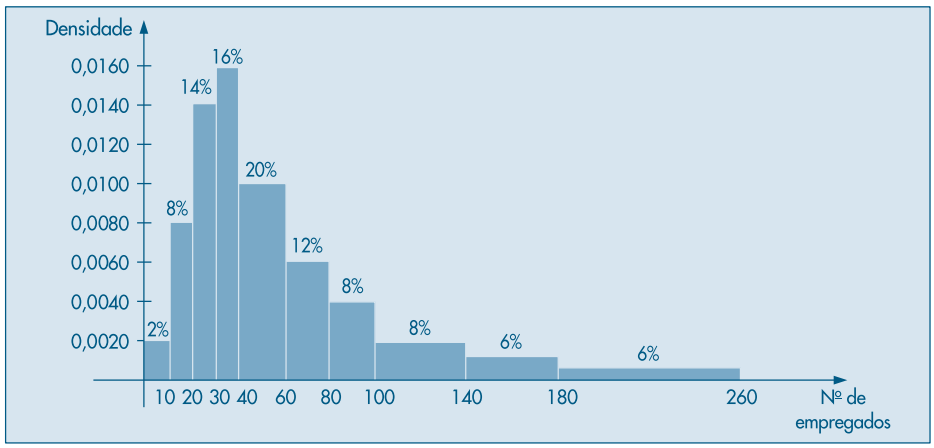
Número de empregados	Freqüência n_i	Amplitude Δ_i	Densidade n_i/Δ_i	Proporção f_i	Densidade f_i/Δ_i
$0 \vdash 10$	5	10	0,50	0,02	0,0020
$10 \vdash 20$	20	10	2,00	0,08	0,0080
$20 \vdash 30$	35	10	3,50	0,14	0,0140
$30 \vdash 40$	40	10	4,00	0,16	0,0160
$40 \vdash 60$	50	20	2,50	0,20	0,0100
$60 \vdash 80$	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.



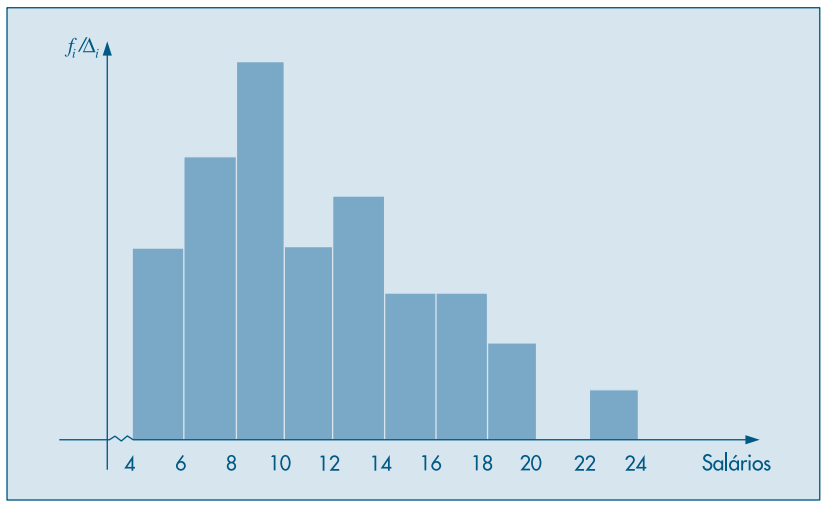
11. Dispomos de uma relação de 200 aluguéis de imóveis urbanos e uma relação de 100 aluguéis rurais.
   (a) Construa os histogramas das duas distribuições.
   (b) Com base nos histogramas, discuta e compare as duas distribuições.

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

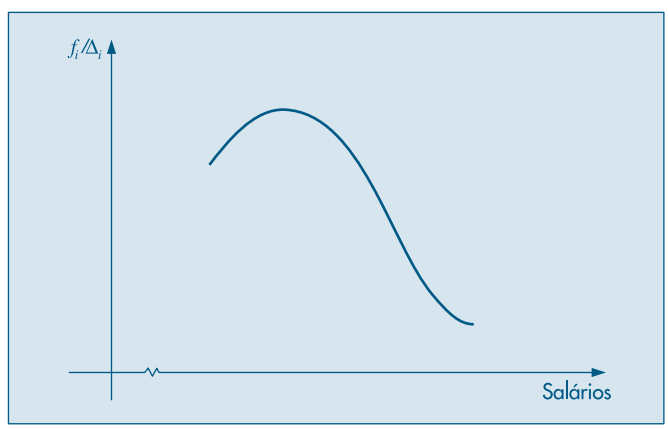
Figura 2.20 (a): Histograma para a variável S: salário, $\Delta = 2$.



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.

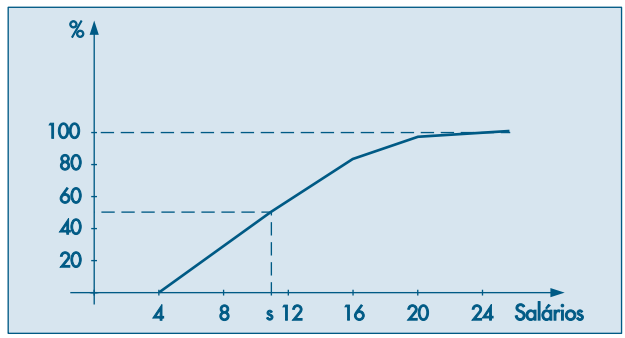
14. Faça no mesmo gráfico um esboço das três distribuições descritas abaixo:
    (a) Distribuição das alturas dos brasileiros adultos.
    (b) Distribuição das alturas dos suecos adultos.
    (c) Distribuição das alturas dos japoneses adultos.

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 $F_e(x)$

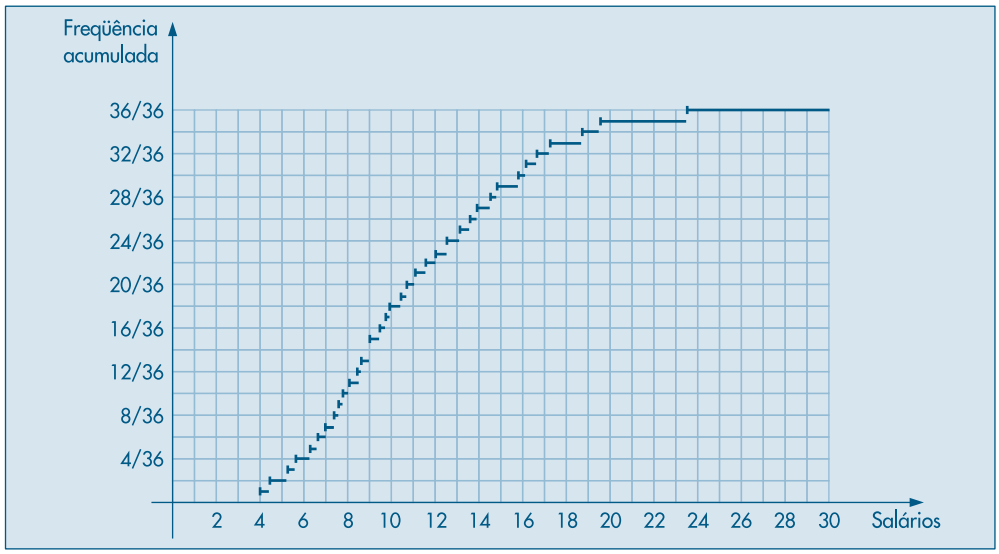
$$F_e(x) = F_n(x) = \frac{N(x)}{n}.$$

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.

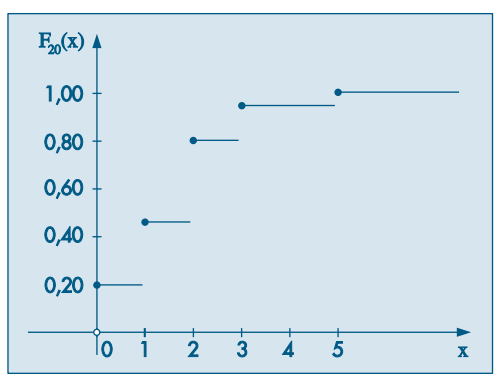


Exemplo 2.13. Esta definição também vale para variáveis quantitativas discretas. Assim, para a variável número de filhos resumida na Tabela 2.5, tem-se a seguinte f.d.e.:

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

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.

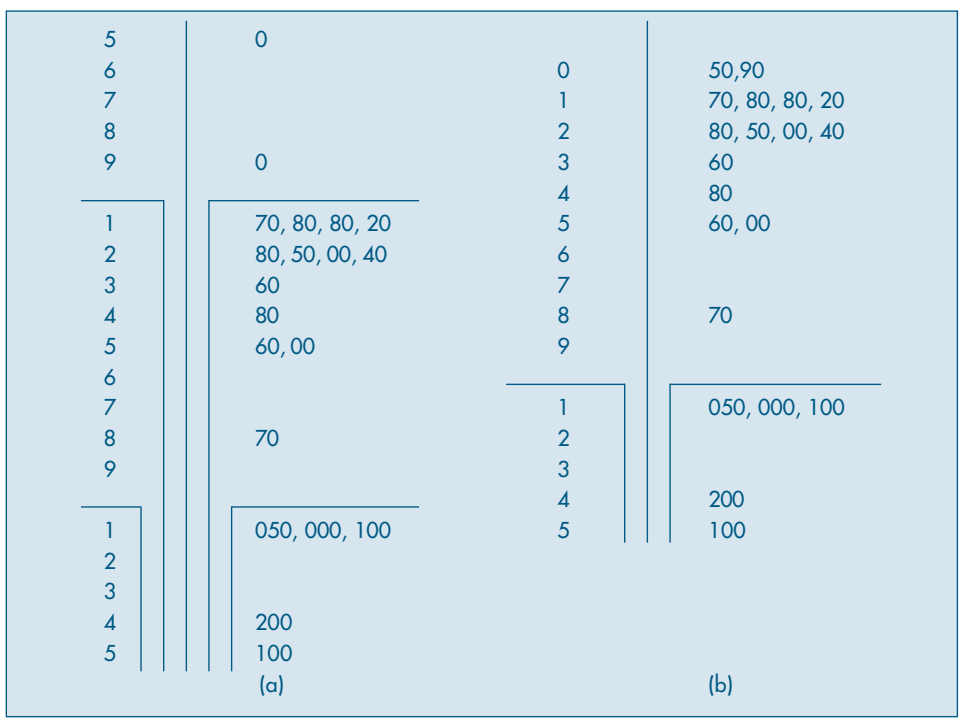


Figura 2.25: Ramo-e-folhas com: (a) freqüências em cada ramo, (b) profundidade.

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).