Capítulo 1

Introdução à Análise Exploratória de Dados

1.1 O que é Estatística?

Neste capítulo, pretendemos formalizar alguns conceitos que constituem a base de técnicas desenvolvidas com a finalidade de auxiliar a responder, de forma objetiva e segura, situações que envolvem uma grande quantidade de informações. A utilização dessas técnicas, destinadas à análise de situações complexas ou não, tem aumentado e faz parte de nosso cotidiano. Tome-se, por exemplo, as transmissões esportivas. Em jogos de futebol, o número de escanteios, o número de faltas cometidas e o tempo de posse de bola são dados geralmente fornecidos ao telespectador e fazem com que as conclusões sobre qual time foi o melhor em campo, se tornem objetivas (não que isso implique que tenha sido o vencedor...). O que tem levado a essa *quantificação* de nossas vidas no dia a dia? Um fator importante é a popularização dos computadores. No passado, tratar uma grande massa de números era uma tarefa custosa e cansativa, que exigia horas de trabalho tedioso. Recentemente, no entanto, grande quantidade de informações pode ser analisada rapidamente com um computador pessoal e programas adequados. Desta forma, o computador contribui, positivamente, na difusão e uso de métodos estatísticos. Por outro lado, o computador possibilita uma automação que pode levar um indivíduo sem preparo específico a utilizar técnicas inadequadas para resolver um dado problema. Assim, é necessário a compreensão dos conceitos básicos da Estatística, bem como as suposições necessárias para o seu uso de forma criteriosa. Entendemos a Estatística como um conjunto de técnicas que permite, de forma sistemática, organizar, descrever, analisar e interpretar *dados* oriundos de estudos ou experimentos, realizados em qualquer área do conhecimento. Estamos denominando por dados um (ou mais) conjunto de valores, numéricos ou não. A aplicabilidade das técnicas a serem discutidas se dá nas mais variadas áreas da atividade humana.

A grosso modo podemos dividir a Estatística em três áreas:

- Estatística Descritiva
- Probabilidade
- Inferência Estatística

Estatística Descritiva é, em geral, utilizada na etapa inicial da análise, quando tomamos contato com os dados pela primeira vez. Objetivando tirar conclusões de modo informal e direto, a maneira mais simples seria a observação dos valores colhidos. Entretanto, ao depararmos com uma grande massa de dados, percebemos, imediatamente, que a tarefa pode não ser simples. Para tentar depreender dos dados informações a respeito do fenômeno sob estudo, é preciso aplicar alguma técnica que nos permita resumir a informação *daquele particular* conjunto de valores. Em outras palavras, a estatística descritiva pode ser definida como um conjunto de técnicas destinadas a descrever e resumir os dados, a fim de que possamos tirar conclusões a respeito de características de interesse.

Probabilidade pode ser pensada como a teoria matemática utilizada para se estudar a *incerteza* oriunda de fenômenos de caráter *aleatório*. Apesar de ser uma área extremamente atraente e estudada do ponto de vista matemático, abordaremos, aqui, apenas os aspectos necessários para as técnicas estatísticas apresentadas neste livro.

Inferência Estatística é o estudo de técnicas que possibilitam a extrapolação, a um grande conjunto de dados, das informações e conclusões obtidas a partir de subconjuntos de valores, usualmente de dimensão muito menor. Deve ser notado que, se tivermos acesso a todos os elementos que desejamos estudar, não é necessário o uso das técnicas de inferência estatística. Entretanto, elas são indispensáveis quando existe a impossibilidade de acesso a todo o conjunto de dados, por razões de natureza econômica, ética ou física.

Estudos complexos que envolvem o tratamento estatístico dos dados, usualmente, incluem as três áreas mencionadas acima. Na terminologia estatística, o grande conjunto de dados que contém a característica que temos interesse recebe o nome de *população*. Esse termo refere-se não somente a uma coleção de indivíduos, mas também ao alvo sobre o qual reside nosso interesse. Assim, nossa população pode ser tanto todos os habitantes de Sorocaba, como todas as lâmpadas produzidas por uma fábrica em um certo período de tempo, ou todo o sangue no corpo de uma pessoa. Algumas vezes podemos acessar toda a população para estudarmos características de interesse, mas, em muitas situações, tal procedimento não pode ser realizado. Em geral, razões econômicas são as mais determinantes dessas situações. Por exemplo, uma empresa, usualmente, não dispõe de verba suficiente para saber o que pensam todos os consumidores de seus produtos. Há ainda razões éticas, quando, por exemplo, os experimentos de laboratório envolvem o uso de seres vivos. Além disso, existem casos em que a impossibilidade de se acessar toda a população de interesse é incontornável. Na análise do sangue de uma pessoa ou em um experimento para determinar o tempo

de funcionamento das lâmpadas produzidas por uma indústria, não podemos observar toda população de interesse.

Tendo em vista as dificuldades de várias naturezas para se observar todos os elementos da população, tomaremos alguns deles para formar um grupo a ser estudado. Este subconjunto da população, em geral com dimensão sensivelmente menor, é denominado *amostra*. A Figura 1.1 ilustra as etapas da análise estatística.

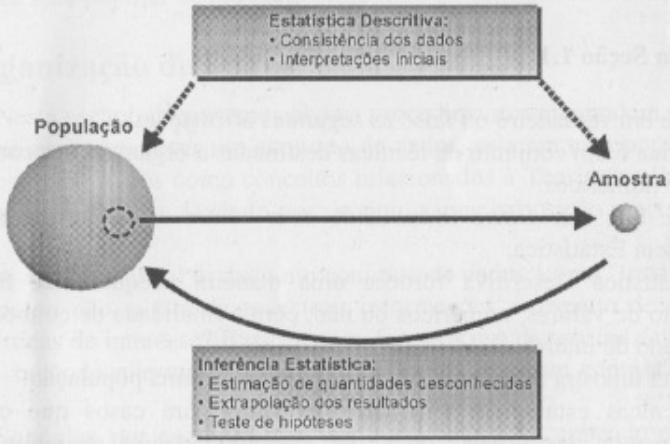


Figura 1.1: População e amostra.

A seleção da amostra pode ser feita de várias maneiras, dependendo, entre outros fatores, do grau de conhecimento que temos da população, da quantidade de recursos disponíveis e assim por diante. Devemos ressaltar que, em princípio, a seleção da amostra tenta fornecer um subconjunto de valores o mais parecido possível com a população que lhe dá origem. A amostragem mais usada é a *amostra casual simples*, em que selecionamos ao acaso, *com* ou *sem reposição*, os itens da população que farão parte da amostra.

Eventualmente, se tivermos informações adicionais a respeito da população de interesse, podemos utilizar outros esquemas de amostragem mais sofisticados. Por exemplo, se numa cidade, tivermos mais mulheres do que homens, podemos selecionar um certo número de indivíduos entre as mulheres e outro número entre os homens. Esse procedimento é conhecido como *amostragem estratificada*. Outras vezes, pode existir uma relação numerada dos

itens da população (uma lista de referência) que nos permitiria utilizar a chamada *amostragem sistemática* em que selecionamos os indivíduos de forma pré-determinada, por exemplo de 8 em 8 ou de 10 em 10. Outros esquemas de amostragem poderiam ser citados e todos fazem parte da chamada *Teoria da Amostragem*, cujos detalhes não serão aprofundados neste livro. Assim sendo, terminamos esta seção mencionando que quanto mais complexa for a amostragem, maiores cuidados deverão ser tomados nas análises estatísticas utilizadas; em contrapartida, o uso de esquemas de amostragem mais elaborados pode levar a uma diminuição no tamanho de amostra necessário para uma dada precisão.

Exercícios da Seção 1.1:

1. Classifique em verdadeiro ou falso as seguintes afirmações:
- **a.** Estatística é um conjunto de técnicas destinadas a organizar um conjunto de valores numéricos.
- **b.** Sempre que estivermos trabalhando com números, deveremos utilizar a Inferência Estatística.
- **c.** A Estatística Descritiva fornece uma maneira adequada de tratar um conjunto de valores, numéricos ou não, com a finalidade de conhecermos o fenômeno de interesse.
- **d.** Qualquer amostra representa, de forma adequada, uma população.
- **e.** As técnicas estatísticas não são adequadas para casos que envolvam experimentos destrutivos como, por exemplo, queima de equipamentos, destruição de corpos de provas, etc.

2. Para as situações descritas a seguir, identifique a população e a amostra correspondente. Discuta a validade do processo de inferência estatística para cada um dos casos.
- **a.** Para avaliar a eficácia de uma campanha de vacinação no Estado de São Paulo, 200 mães de recém-nascidos, durante o primeiro semestre de um dado ano e em uma dada maternidade em São Paulo, foram entrevistadas a respeito da última vez em que vacinaram seus filhos.
- **b.** Uma amostra de sangue foi retirada de um paciente com suspeita de anemia.
- **c.** Para verificar a audiência de um programa de TV, 563 indivíduos foram entrevistados por telefone com relação ao canal em que estavam sintonizados.
- **d.** A fim de avaliar a intenção de voto para presidente dos brasileiros, 122 pessoas foram entrevistadas em Brasília.

3. Discuta, para cada um dos casos abaixo, os cuidados que precisam ser tomados para garantir uma boa conclusão a partir da amostra.
- **a.** Um grupo de crianças será escolhido para receber uma nova vacina contra meningite.
- **b.** Sorteamos um certo número de donas de casa, para testar um novo sabão em pó.
- **c.** Uma fábrica deseja saber se sua produção de biscoitos está com o sabor previsto.
- **d.** Aceitação popular de um certo projeto do governo.

1.2 Organização de Dados

Nesta seção, discutiremos alguns procedimentos que podem ser utilizados para organizar e descrever um conjunto de dados, seja em uma população ou em uma amostra. Veremos como conceitos relacionados à Teoria das Probabilidades aparecem naturalmente, levando-nos, assim, a uma exposição mais criteriosa do assunto.

A questão inicial é: dado um conjunto de dados, como "tratar" os valores, numéricos ou não, a fim de se extrair informações a respeito de uma ou mais características de interesse? Basicamente, faremos uso de *tabelas de freqüências* e *gráficos*, notando que tais procedimentos devem levar em conta a natureza dos dados.

Suponha, por exemplo, que um questionário foi aplicado aos alunos do primeiro ano de uma escola fornecendo as seguintes informações:

Id:	identificação do aluno
Turma:	turma a que o aluno foi alocado (A ou B)
Sexo:	F se feminino, M se masculino
Idade:	idade em anos
Alt:	altura em metros
Peso:	peso em quilogramas
Filhos:	número de filhos na família
Fuma:	hábito de fumar, sim ou não
Toler:	tolerância ao cigarro: (I) indiferente, (P) incomoda pouco e (M) incomoda muito

Exerc:	horas de atividade física, por semana
Cine:	número de vezes em que vai ao cinema por semana
OpCine:	opinião a respeito das salas de cinema na cidade: (B) regular a boa e (M) muito boa
TV:	horas gastas assistindo TV, por semana
OpTV:	opinião a respeito da qualidade da programação na TV: (R) ruim, (M) média, (B) boa e (N) não sabe

O conjunto de informações disponíveis, após a tabulação do questionário ou pesquisa de campo, é denominado de *tabela de dados brutos* e contém os dados da maneira que foram coletados inicialmente. Os valores obtidos para cada uma dessas informações estão apresentados na Tabela 1.1. Cada uma das características perguntadas aos alunos, tais como o peso, a idade e a altura, entre outras, é denominada de *variável*. Assim, a variável Altura assume os valores (em metros) 1,60; 1,58;... e a variável Turma assume os valores *A* ou *B*. Claramente tais variáveis têm naturezas diferentes no que tange aos possíveis valores que podem assumir. Tal fato deve ser levado em conta nas análises e, para fixar idéias, vamos considerar dois grandes tipos de variáveis: numéricas e não numéricas. As numéricas serão denominadas *quantitativas*, ao passo que as não numéricas, *qualitativas*.

A variável é qualitativa quando os possíveis valores que assume representam atributos e/ou qualidades. Se tais variáveis têm uma ordenação natural, indicando intensidades crescentes de realização, então elas serão classificadas como *qualitativas ordinais*. Caso contrário, quando não é possível estabelecer uma ordem natural entre seus valores, elas são classificadas como *qualitativas nominais*. Variáveis tais como Turma (A ou B), Sexo (feminino ou masculino) e Fuma (sim ou não) são variáveis qualitativas nominais. Por outro lado, variáveis como Tamanho (pequeno, médio ou grande), Classe Social (baixa, média ou alta) são variáveis qualitativas ordinais.

Variáveis quantitativas, isto é, variáveis de natureza numérica, podem ser subdivididas em *discretas* e *contínuas*. A grosso modo, variáveis *quantitativas discretas* podem ser vistas como resultantes de contagens, assumindo assim, em geral, valores inteiros. De uma maneira mais formal, o conjunto dos valores assumidos é finito ou enumerável. Já as variáveis *quantitativas contínuas* assumem valores em intervalos dos números reais e, geralmente, são provenientes de uma mensuração. Por exemplo, Número de Irmãos (0, 1, 2, ...) e Número de Defeitos (0, 1, 2, ...) são discretas, enquanto que Peso e Altura são quantitativas contínuas.

Tabela 1.1: Informações de questionário estudantil - dados brutos.

Id	Turma	Sexo	Idade	Alt	Peso	Filh	Fuma	Toler	Exer	Cine	OpCine	TV	OpTV
1	A	F	17	1,60	60,5	2	NAO	P	0	1	B	16	R
2	A	F	18	1,69	55,0	1	NAO	M	0	1	B	7	R
3	A	M	18	1,85	72,8	2	NAO	P	5	2	M	15	R
4	A	M	25	1,85	80,9	2	NAO	P	5	2	B	20	R
5	A	F	19	1,58	55,0	1	NAO	M	2	2	B	5	R
6	A	M	19	1,76	60,0	3	NAO	M	2	1	B	2	R
7	A	F	20	1,60	58,0	1	NAO	P	3	1	B	7	R
8	A	F	18	1,64	47,0	1	SIM	I	2	2	M	10	R
9	A	F	18	1,62	57,8	3	NAO	M	3	3	M	12	R
10	A	F	17	1,64	58,0	2	NAO	M	2	2	M	10	R
11	A	F	18	1,72	70,0	1	SIM	I	10	2	B	8	N
12	A	F	18	1,66	54,0	3	NAO	M	0	2	B	0	R
13	A	F	21	1,70	58,0	2	NAO	M	6	1	M	30	R
14	A	M	19	1,78	68,5	1	SIM	I	5	1	M	2	N
15	A	F	18	1,65	63,5	1	NAO	I	4	1	B	10	R
16	A	F	19	1,63	47,4	3	NAO	P	0	1	B	18	R
17	A	F	17	1,82	66,0	1	NAO	P	3	1	B	10	N
18	A	M	18	1,80	85,2	2	NAO	P	3	4	B	10	R
19	A	F	20	1,60	54,5	1	NAO	P	3	2	B	5	R
20	A	F	18	1,68	52,5	3	NAO	M	7	2	B	14	M
21	A	F	21	1,70	60,0	2	NAO	P	8	2	B	5	R
22	A	F	18	1,65	58,5	1	NAO	M	0	3	B	5	R
23	A	F	18	1,57	49,2	1	SIM	I	5	4	B	10	R
24	A	F	20	1,55	48,0	1	SIM	I	0	1	M	28	R
25	A	F	20	1,69	51,6	2	NAO	P	8	5	M	4	N
26	A	F	19	1,54	57,0	2	NAO	I	6	2	B	5	R
27	B	F	23	1,62	63,0	2	NAO	M	8	2	M	5	R
28	B	F	18	1,62	52,0	1	NAO	P	1	1	M	10	R
29	B	F	18	1,57	49,0	2	NAO	P	3	1	B	12	R
30	B	F	25	1,65	59,0	4	NAO	M	1	2	M	2	R
31	B	F	18	1,61	52,0	1	NAO	P	2	2	M	6	N
32	B	M	17	1,71	73,0	1	NAO	P	1	1	B	20	R
33	B	F	17	1,65	56,0	3	NAO	M	2	1	B	14	R
34	B	F	17	1,67	58,0	1	NAO	M	4	2	B	10	R
35	B	M	18	1,73	87,0	1	NAO	M	7	1	B	25	B
36	B	F	18	1,60	47,0	1	NAO	P	5	1	M	14	R
37	B	M	17	1,70	95,0	1	NAO	P	10	2	M	12	N
38	B	M	21	1,85	84,0	1	SIM	I	6	4	B	10	R
39	B	F	18	1,70	60,0	1	NAO	P	5	2	B	12	R
40	B	M	18	1,73	73,0	1	NAO	M	4	1	B	2	R
41	B	F	17	1,70	55,0	1	NAO	I	5	4	B	10	B
42	B	F	23	1,45	44,0	2	NAO	M	2	2	B	25	R
43	B	M	24	1,76	75,0	2	NAO	I	7	0	M	14	N
44	B	F	18	1,68	55,0	1	NAO	P	5	1	B	8	R
45	B	F	18	1,55	49,0	1	NAO	M	0	1	M	10	R
46	B	F	19	1,70	50,0	7	NAO	M	0	1	B	8	R
47	B	F	19	1,55	54,5	2	NAO	M	4	3	B	3	R
48	B	F	18	1,60	50,0	1	NAO	P	2	1	B	5	R
49	B	M	17	1,80	71,0	1	NAO	P	7	0	M	14	R
50	B	M	18	1,83	86,0	1	NAO	P	7	0	M	20	B

Resumimos a classificação das variáveis no esquema apresentado na Figura 1.2 (a título de exercício, tente classificar todas as variáveis da Tabela 1.1).

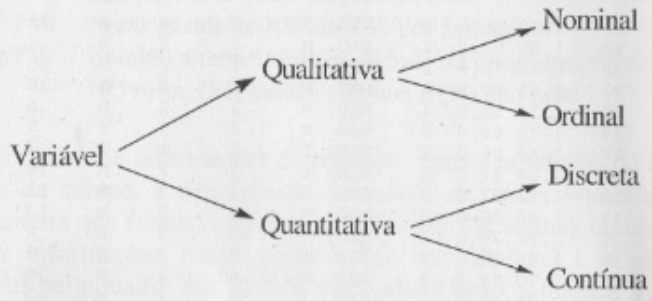


Figura 1.2: Classificação de variáveis.

Vale ressaltar que, em muitas situações práticas, a classificação depende de certas particularidades. Por exemplo, a variável Idade, medida em número de anos, pode ser vista como discreta, entretanto, se levarmos em conta os dias, não é absurdo falar que a idade é 2,5 ou 2,85 anos, dando assim respaldo para classificá-la como contínua. Por outro lado, dependendo da precisão do instrumento utilizado para se obter medidas em um objeto, podemos ter limitações no número de casas decimais e uma variável de mensuração pode se "tornar" discreta. É importante salientar que a classificação apresentada acima se refere à *natureza* da variável e, em geral, devemos utilizar o bom senso na hora de decidir qual procedimento adotar para caracterizar uma variável. Para salientar tal fato, mencionamos que podemos, inclusive, *discretizar* uma variável contínua para obter uma melhor representação da ocorrência de seus valores no conjunto de dados.

Outro ponto que pode trazer confusão é que, muitas vezes, na utilização de programas computacionais, associamos códigos numéricos a uma variável qualitativa. Por exemplo na Tabela 1.1, pode-se associar ao sexo feminino o valor 1 e ao masculino 2. Apesar da variável ser representada por valores numéricos, isso não a torna uma variável quantitativa. Novamente, vemos que a natureza da variável deve sempre ser levada em conta na hora de se interpretar resultados obtidos na análise descritiva.

Apesar de conter muita informação, a tabela de dados brutos pode não ser prática para respondermos às questões de interesse. Por exemplo, da Tabela 1.1 não é imediato dizer se os alunos se incomodam muito ou pouco com os fumantes. Portanto, a partir da tabela de dados brutos, vamos construir uma nova tabela com as informações resumidas, para cada variável. Essa tabela será denominada de *tabela de freqüência* e, como o nome indica, conterá os valores da variável e suas respectivas contagens, as quais são denominadas *freqüências absolutas* ou simplesmente, *freqüências*. No caso de variáveis qualitativas ou quantitativas discretas, a tabela de freqüência consiste em listar os valores possíveis da variável, numéricos ou não e fazer a contagem na tabela de dados brutos do número de suas ocorrências. Representaremos por n_i a freqüência do valor i e por n a freqüência total. Para efeito de comparação com outros grupos ou conjuntos de dados, será conveniente acrescentarmos uma coluna na tabela de freqüência contendo o cálculo da *freqüência relativa,* definida por $f_i = n_i/n$. Convém notar que, quando estivermos comparando dois grupos com relação às freqüências de ocorrência dos valores de uma dada variável, grupos com um número total de dados maior tendem a ter maiores freqüências de ocorrência dos valores da variável. Desta forma, o uso da freqüência relativa vem resolver este problema.

A Tabela 1.2 apresenta as freqüências para a variável Sexo, obtida a partir da Tabela 1.1.

Tabela 1.2: Tabela de freqüência para a variável Sexo.

Sexo	n_i	f_i
F	37	0,74
M	13	0,26
total	$n = 50$	1

Note que, para variáveis cujos valores possuem ordenação natural (qualitativas ordinais e quantitativas em geral), faz sentido incluirmos também uma coluna contendo as *freqüências acumuladas* f_{ac}. A freqüência acumulada até um certo valor é obtida pela soma das freqüências de todos os valores da variável, menores ou iguais ao valor considerado. Sua utilidade principal é ajudar a estabelecer pontos de corte com uma determinada freqüência nos valores da variável. Por exemplo, na Tabela 1.3, observamos que 90% dos alunos têm idades até 21 anos, de fato até 22, uma vez que este valor tem freqüência zero.

Tabela 1.3: Tabela de freqüência para a variável Idade.

Idade	n_i	f_i	f_{ac}
17	9	0,18	0,18
18	22	0,44	0,62
19	7	0,14	0,76
20	4	0,08	0,84
21	3	0,06	0,90
22	0	0	0,90
23	2	0,04	0,94
24	1	0,02	0,96
25	2	0,04	1,00
total	$n = 50$	1	

Com relação à variável Peso, lembremos que foi classificada como quantitativa contínua e assim, teoricamente, seus valores podem ser qualquer número real num certo intervalo. Aqui os valores variam entre 44,0 e 95,0 kg e foram medidos com apenas uma casa decimal. Ainda assim, existe um grande número de valores diferentes de modo que, se a tabela de freqüência fosse feita nos mesmos moldes dos casos anteriores, obteríamos praticamente os valores originais da tabela de dados brutos. A alternativa que vamos adotar consiste em construir *classes* ou *faixas de valores* e contar o número de ocorrências em cada faixa. Para a variável Peso, usamos faixas de amplitude 10, iniciando em 40 kg. Na Tabela 1.4, escolhemos incluir o extremo inferior e excluir o superior. Dessa forma, a freqüência da faixa 40,0 $\vdash$ 50,0 não incluiu os alunos 46 e 48 que tinham peso igual a 50,0 kg. A opção de qual extremo incluir pode ser arbitrária, mas o importante é indicar claramente quais são os valores que estão sendo contados em cada faixa.

Apesar de não adotarmos nenhuma regra formal quanto ao total de faixas, utilizamos, em geral, de 5 a 8 faixas com mesma amplitude. Entretanto, ressaltamos que faixas de tamanho desigual podem ser convenientes para representar valores nas extremidades da tabela.

Tabela 1.4: Tabela de freqüência para a variável Peso.

Peso	n_i	f_i	f_{ac}
40,0 $\vdash$ 50,0	8	0,16	0,16
50,0 $\vdash$ 60,0	22	0,44	0,60
60,0 $\vdash$ 70,0	8	0,16	0,76
70,0 $\vdash$ 80,0	6	0,12	0,88
80,0 $\vdash$ 90,0	5	0,10	0,98
90,0 $\vdash$ 100,0	1	0,02	1,00
total	50	1	

Vamos estudar, agora, a situação em que a variável é por natureza discreta, mas o conjunto de possíveis valores é muito grande. Por exemplo, a variável TV, definida como o número de horas assistindo televisão, tem valores inteiros entre 0 e 30 e uma tabela representando seus valores e respectivas freqüências seria muito extensa e pouco prática. O caminho adequado, nesse caso, é tratar a variável como se fosse contínua e criar faixas para representar seus valores. Assim, passamos a tratar como contínua uma variável que seria, originalmente, classificada como discreta.

Tabela 1.5: Tabela de freqüência para a variável TV.

TV	n_i	f_i	f_{ac}
0 $\vdash$ 6	14	0,28	0,28
6 $\vdash$ 12	17	0,34	0,62
12 $\vdash$ 18	11	0,22	0,84
18 $\vdash$ 24	4	0,08	0,92
24 $\vdash\dashv$ 36	4	0,08	1,00
total	50	1	

A organização dos dados em tabelas de freqüência proporciona um meio eficaz de estudo do comportamento de características de interesse. Muitas vezes, a informação contida nas tabelas pode ser mais facilmente visualizada através de gráficos. Meios de comunicação apresentam, diariamente, gráficos das mais variadas formas para auxiliar na apresentação das informações. Órgãos públicos e empresas se municiam de gráficos e tabelas em documentos internos e relatórios

de atividades e desempenho. Graças à proliferação de recursos gráficos, cuja construção tem sido cada vez mais simplificada em programas computacionais, existe hoje uma infinidade de tipos de gráficos que podem ser utilizados. Como ilustração deste ponto, apresentamos na Figura 1.3 alguns gráficos publicados em órgãos de imprensa.

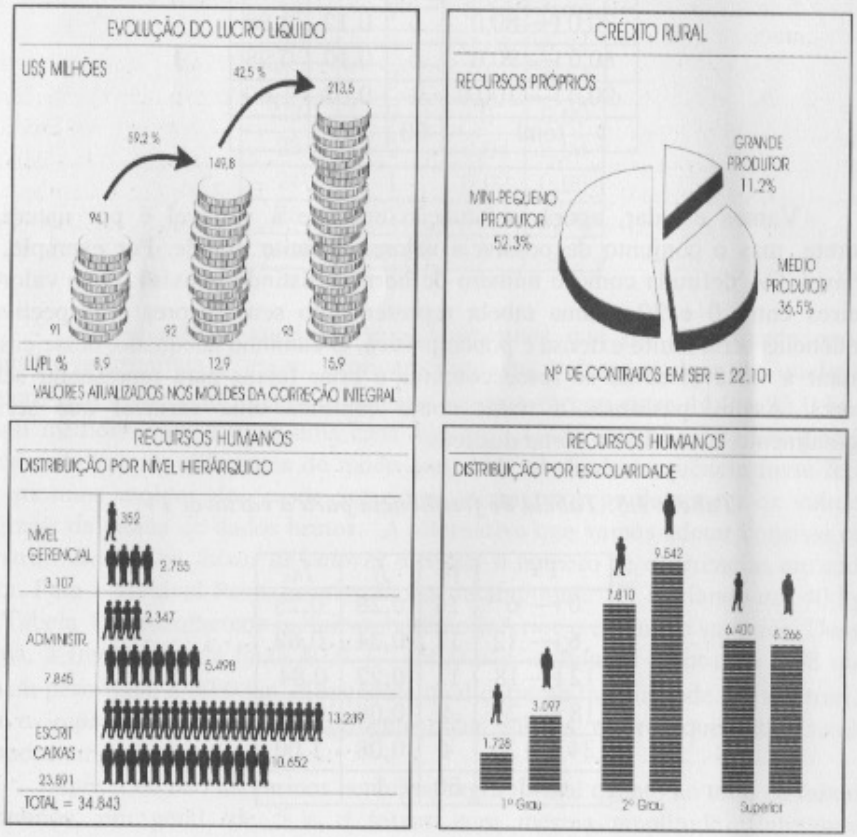


Figura 1.3: Exemplos de gráficos publicados na imprensa.

Deve ser notado, entretanto, que a utilização de recursos visuais na criação de gráficos deve ser feita cuidadosamente; um gráfico desproporcional em suas medidas pode dar falsa impressão de desempenho e conduzir a conclusões

equivocadas. Obviamente, questões de manipulação incorreta da informação podem ocorrer em qualquer área e não cabe culpar a Estatística. O uso e a divulgação ética e criteriosa de dados devem ser pré-requisitos indispensáveis e inegociáveis.

Vamos definir três tipos básicos de gráficos: *disco* ou *pizza*, *barras* e *histograma*. Como dissemos, a criatividade na apresentação gráfica pode ser imensa e os gráficos que discutiremos sintetizam três caminhos, entre vários, de representação.

O gráfico de *disco*, ou *pizza*, ou ainda *diagrama circular*, se adapta muito bem às variáveis qualitativas nominais. Consiste em repartir um disco em setores circulares correspondentes às porcentagens de cada valor, calculadas multiplicando-se por 100 a freqüência relativa f_i. Por exemplo, 0,20 de freqüência relativa corresponde a 20% uma vez que $100 \times 0{,}20 = 20$. A Figura 1.4 apresenta o diagrama de disco para a variável Toler, obtida a partir da Tabela 1.1. Note que a fatia correspondente à categoria "indiferente" foi destacada.

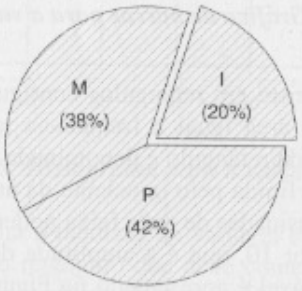


Figura 1.4: Diagrama circular para a variável Toler.

O *gráfico de barras* utiliza o plano cartesiano com os valores da variável no eixo das abcissas e as freqüências ou porcentagens no eixo das ordenadas. Note que para cada valor da variável desenha-se uma barra com altura correspondendo à sua freqüência ou porcentagem. Esse tipo de gráfico se adapta melhor às variáveis discretas ou qualitativas ordinais.

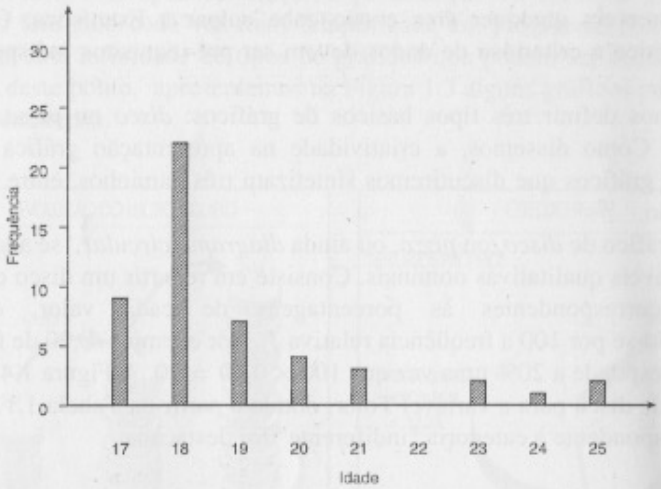


Figura 1.5: Gráfico de barras para a variável Idade.

O *histograma* consiste em retângulos contíguos com base nas faixas de valores da variável e com área igual à freqüência relativa da respectiva faixa. Dessa forma, a altura de cada retângulo é denominada *densidade de freqüência* ou simplesmente *densidade* definida pelo quociente da área pela amplitude da faixa. Para a variável peso, as densidades de cada faixa podem ser obtidas dividindo-se a coluna f_i da Tabela 1.4 por 10, que é a amplitude de cada faixa. O histograma correspondente a essa variável é apresentado na Figura 1.6. Note que incluímos, no topo de cada retângulo, a porcentagem de observações correspondente, para facilitar a interpretação.

É importante ressaltar que alguns autores utilizam a freqüência absoluta ou porcentagem na construção do histograma. Preferimos o uso da densidade de freqüência, pois ela faz com que o histograma não fique distorcido, quando amplitudes diferentes são utilizadas nas faixas. Uma outra vantagem diz respeito à relação entre histograma e gráfico da função densidade de probabilidade, que será visto mais adiante.

O histograma também pode ser utilizado no cálculo da *mediana* (md_{obs}), que é o valor da variável que divide o conjunto de dados ordenados em dois subgrupos de mesmo tamanho. Isto é, das observações ordenadas, 50% estão abaixo e 50% estão acima da mediana. Assumindo que as observações da variável

em cada faixa são homogeneamente distribuídas, para um mesmo retângulo, fatias de mesmo tamanho contém uma mesma porcentagem de observações. Apesar da suposição de homogeneidade não ser sempre verificada, ela é bastante razoável em muitas situações e pode ser uma boa aproximação da realidade.

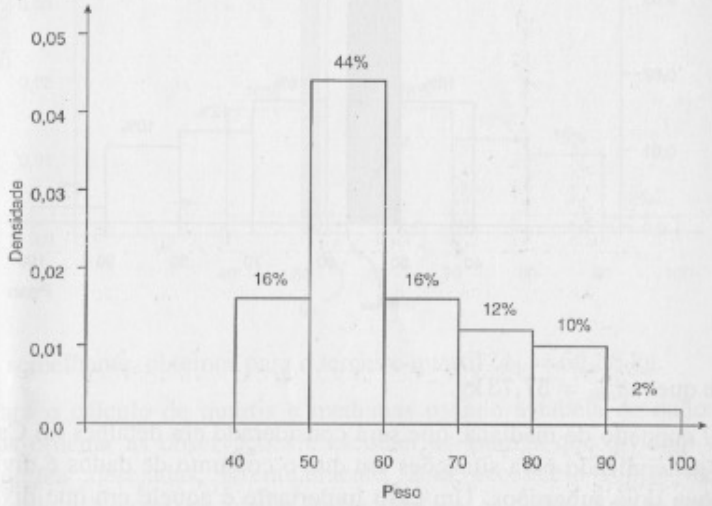
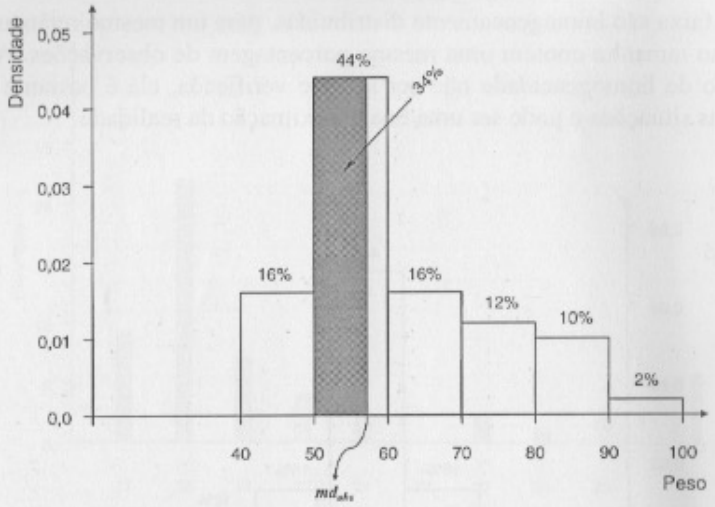


Figura 1.6: Histograma para a variável Peso.

Exemplo 1.1: Vamos calcular a mediana da variável Peso através do histograma. Inicialmente identificamos o retângulo que deve conter a mediana. Uma simples soma das áreas resulta que a mediana pertence ao intervalo [50,0; 60,0), uma vez que até o valor 60,0 temos acumuladas 60% das observações. Dentro dessa faixa, precisamos determinar um retângulo com área igual a 34%, que é o que falta para atingir o valor 50%. A situação é ilustrada na figura a seguir, cujo retângulo procurado está marcado com área mais escura.

Com uso de proporções, estabelecemos a seguinte igualdade:

$$\frac{md_{obs} - 50}{0{,}34} = \frac{60 - 50}{0{,}44}$$

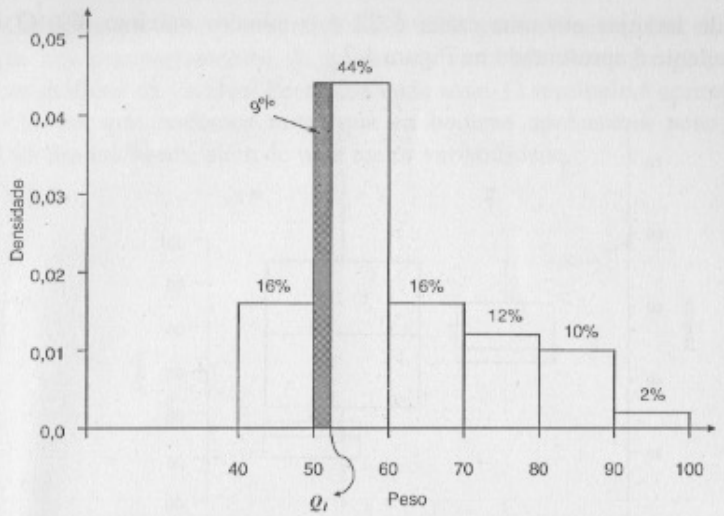


Daí segue que $md_{obs} = 57{,}73$ kg. □

O conceito de mediana, que será considerado em detalhes no Capítulo 4, pode ser generalizado para situações em que o conjunto de dados é dividido em mais do que dois subgrupos. Um caso importante é aquele em que dividimos o conjunto em quatro subgrupos. Para tanto, deveremos determinar, além da mediana, dois valores tais que 25% das observações ordenadas estarão abaixo de um deles e 75% estarão abaixo do outro. Tais valores são denominados, respectivamente, *primeiro quartil* e *terceiro quartil*, usualmente representados por Q_1 e Q_3. Note que a mediana, discutida anteriormente, representa o *segundo quartil*. O cálculo dos valores dos quartis também pode ser feito através do histograma, conforme mostrado no exemplo a seguir.

Exemplo 1.2: No Exemplo 1.1, o valor da mediana (segundo quartil) calculado através do histograma, é 57,73 kg. De forma semelhante, vemos que o valor do primeiro quartil também se encontra no intervalo [50,0; 60,0), isto é, corresponderá ao valor Q_1 que determinará uma área de 9% no retângulo correspondente. Assim, temos (ver figura a seguir)

$$\frac{Q_1 - 50}{0{,}09} = \frac{60 - 50}{0{,}44} \Rightarrow Q_1 = 52{,}05 \text{ kg}.$$



De forma semelhante, obtemos para o terceiro quartil $Q_3 = 69{,}38$ kg. □

Para o cálculo de quartis e medianas usando a tabela de dados brutos, precisamos ordenar as observações e escolher os valores que dividem os dados nas proporções desejadas. Eventualmente, será necessário tomar médias de valores vizinhos. No caso de tabelas de freqüências, os dados já estão ordenados e o procedimento é similar.

Uma representação gráfica envolvendo os quartis é o *box-plot*. Definimos uma "caixa" com o nível superior dado pelo terceiro quartil e o nível inferior pelo primeiro quartil. A mediana é representada por um traço no interior da caixa e segmentos de reta são colocados da caixa até os valores máximo e mínimo, que não sejam observações discrepantes (o critério para decidir se uma observação é discrepante não será discutido aqui, mas, em geral, envolve a diferença entre o terceiro e o primeiro quartis). O próximo exemplo ilustra a construção do *box-plot* para uma variável quantitativa discreta utilizando-se os dados brutos.

Exemplo 1.3: Suponha que um produtor de laranjas costuma guardar as frutas em caixas e está interessado em estudar o número de laranjas por caixa. Após um dia de colheita, 20 caixas foram contadas. Os resultados brutos, após a ordenação, são: 22, 29, 33, 35, 35, 37, 38, 43, 43, 44, 48, 48, 52, 53, 55, 57, 61, 62, 67 e 69. Para esses dados, temos que $md_{obs} = (10º + 11º)/2 = (44 + 48)/2 = 46$. Analogamente, obtemos $Q_1 = 36$ e $Q_3 = 56$. Também observamos que o número

mínimo de laranjas em uma caixa é 22 e o número máximo, 69. O *box-plot* correspondente é apresentado na Figura 1.7.

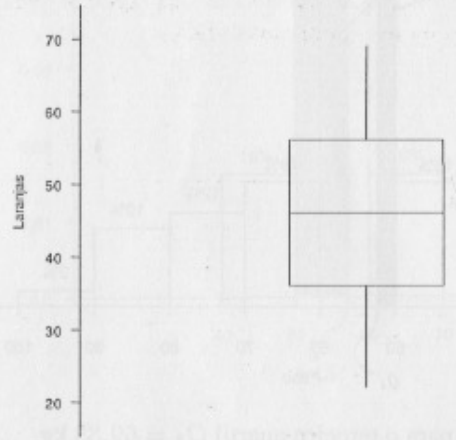


Figura 1.7. Box-plot para o número de laranjas por caixa.

A representação gráfica através do *box-plot* é bastante rica no sentido de informar, entre outras coisas, a variabilidade e simetria dos dados. Note que na Figura 1.7 os dados apresentam simetria acentuada (a distância da mediana para os quartis é a mesma), o mesmo podendo ser observado a respeito da distância dos pontos de mínimo e máximo em relação à mediana. Em contraste, temos na Figura 1.8 o *box-plot* para a variável Peso, que apresenta uma pequena assimetria.

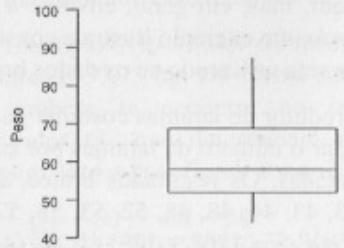


Figura 1.8: Box-plot para a variável Peso.

Gráficos tipo *box-plot* também são úteis para detectar, descritivamente, diferenças nos comportamentos de grupos de variáveis. Por exemplo, podemos considerar gráficos da variável Peso para cada sexo. O resultado é apresentado na Figura 1.9, em que podemos notar que os homens apresentam peso mediano superior ao das mulheres, além de uma maior variabilidade.

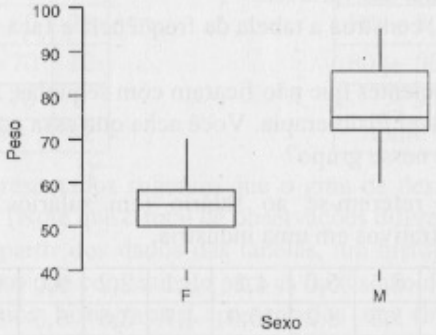


Figura 1.9: Box-plot da variável Peso para cada sexo.

Exercícios da Seção 1.2:

1. Classifique cada uma das variáveis abaixo em qualitativa (nominal / ordinal) ou quantitativa (discreta / contínua):

- **a.** Ocorrência de hipertensão pré-natal em grávidas com mais de 35 anos (*sim* ou *não* são possíveis respostas para esta variável).
- **b.** Intenção de voto para presidente (possíveis respostas são os nomes dos candidatos, além de *não sei*).
- **c.** Perda de peso de maratonistas na Corrida de São Silvestre, em quilos.
- **d.** Intensidade da perda de peso de maratonistas na Corrida de São Silvestre (leve, moderada, forte).
- **e.** Grau de satisfação da população brasileira com relação ao trabalho de seu presidente (valores de 0 a 5, com 0 indicando totalmente insatisfeito e 5 totalmente satisfeito).

2. Quinze pacientes de uma clínica de ortopedia foram entrevistados quanto ao número de meses previstos de fisioterapia, se haverá (S) ou não (N) seqüelas

após o tratamento e o grau de complexidade da cirurgia realizada: alto (A), médio (M) ou baixo (B). Os dados são apresentados na tabela abaixo:

Pacientes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Fisioterapia	7	8	5	6	4	5	7	7	6	8	6	5	5	4	5
Seqüelas	S	S	N	N	N	S	S	N	N	S	S	N	S	N	N
Cirurgia	A	M	A	M	M	B	A	M	B	M	B	B	M	M	A

a. Classifique cada uma das variáveis.

b. Para cada variável, construa a tabela de freqüência e faça uma representação gráfica.

c. Para o grupo de pacientes que não ficaram com seqüelas, faça um gráfico de barras para a variável Fisioterapia. Você acha que essa variável se comporta de modo diferente nesse grupo?

3. Os dados abaixo referem-se ao salário (em salários mínimos) de 20 funcionários administrativos em uma indústria.

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

a. Construa uma tabela de freqüência agrupando os dados em intervalos de amplitude 2 a partir de 1.

b. Construa o histograma e calcule o 1º e o 3º quartil.

4. Um grupo de estudantes do ensino médio foi submetido a um teste de matemática resultando em:

Nota	freqüência
$0 \vdash 2$	14
$2 \vdash 4$	28
$4 \vdash 6$	27
$6 \vdash 8$	11
$8 \vdash 10$	4

a. Construa o histograma.

b. Se a nota mínima para aprovação é 5, qual será a porcentagem de aprovação?

c. Obtenha o *box-plot*.

5. Um estudo pretende verificar se o problema da desnutrição em adultos medida pelo peso, em quilos, em uma região agrícola (denotada por Região A), é maior do que em uma região industrial (Região B). Para tanto, uma amostra foi tomada em cada região, fornecendo a tabela de freqüências a seguir:

Região A

Peso	n_i
< 40	8
$40 \vdash 50$	25
$50 \vdash 60$	28
$60 \vdash 70$	12
≥ 70	9
total	82

Região B

Peso	n_i
< 60	10
$60 \vdash 70$	34
$70 \vdash 80$	109
$80 \vdash 90$	111
≥ 90	55
total	319

a. Os dados apresentados sugerem que o grau de desnutrição é diferente nas duas regiões? (Note que o total de observações difere em cada região).

b. Construa, a partir dos dados das tabelas, um histograma para cada região. Faça uma suposição conveniente para as faixas não delimitadas.

c. Com base nos histogramas apresentados em (b), obtenha as medidas necessárias e construa o *box-plot*, um para cada região. Com base nessa representação gráfica, rediscuta o item (a).

1.3 O Uso de Computadores em Estatística

Foi mencionado anteriormente que o desenvolvimento da indústria de computadores deu grande impulso ao uso da Estatística. Vários programas computacionais de uso comum contém rotinas estatísticas incorporadas às suas funções básicas. É o caso das *planilhas eletrônicas*, usualmente pré-instaladas em computadores novos. Programas especificamente desenvolvidos para efetuar análises estatísticas são conhecidos como *pacotes estatísticos*. Existe um número considerável desses pacotes, alguns voltados para análises mais comuns na área de humanidades, outros para a área de biomédicas; alguns são extremamente simples de se utilizar através de menus, outros pressupõem conhecimento de uma linguagem de programação específica. Qualquer que seja o programa a ser utilizado, três são as etapas que envolvem seu uso:

1. Entrada de Dados
2. Execução da Análise Estatística
3. Interpretação de Resultados

A Entrada de Dados deve assumir certas convenções. Apesar de certos programas terem rotinas desenvolvidas de forma a simplificar a criação do banco

de dados, intrinsecamente o que se tem é a criação de uma *matriz*, em que cada linha corresponde a uma *unidade experimental* e cada coluna a uma variável.

Por unidade experimental, entende-se o elemento da população ou amostra no qual observaremos as variáveis. Por exemplo, na Tabela 1.1, observamos 50 unidades experimentais, os estudantes, nos quais foram observadas 14 variáveis. Assim, os dados podem ser representados por uma matriz com dimensão 50 por 14. Leitores familiares com planilhas eletrônicas não terão problema em visualizar esta situação. Assim, quando estudamos uma única variável, consideramos a coluna correspondente. Se estamos interessados em saber o comportamento desta variável em dois grupos diferentes (como na Figura 1.9), precisamos estudar os valores da coluna em que ela se encontra, conjuntamente com a coluna que contém a informação dos grupos.

A fase da execução da análise estatística pressupõe o conhecimento de como o programa que está sendo utilizado trabalha as informações. Torna-se, assim, importante se ter acesso ao manual do programa.

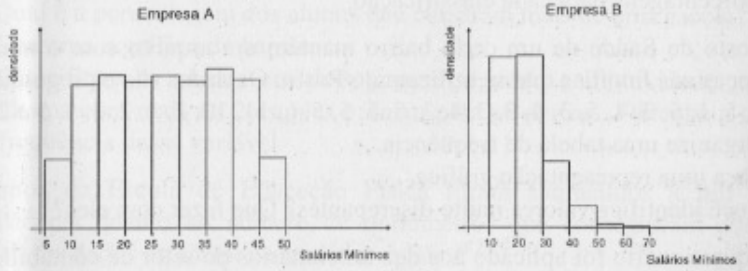
Após as informações terem sido trabalhadas, vem a fase da interpretação dos resultados obtidos. Nesta hora, é aconselhável consultar o manual sempre que houver dúvida, se o que foi calculado relaciona-se, de fato, à análise estatística desejada. Ao interpretar as características observadas, é importante verificar se resultados absurdos não estão ocorrendo. Em caso positivo, releia o manual e certifique-se de ter executado a análise correta para os dados em questão. Em muitos casos, a fase de interpretação é a mais difícil e interessante, pois envolve o equacionamento das características apresentadas na análise com vistas a responder as questões inicialmente colocadas.

Exercícios da Seção 1.3:

1. Utilizando alguma planilha eletrônica ou pacote estatístico disponível e com as informações da Tabela 1.1, construa um banco de dados para os 20 indivíduos iniciais e as 4 primeiras colunas. Imprima e confira os valores digitados.
2. Considerando o banco de dados criado no Exercício 1 desta seção, construa histogramas para as quatro variáveis e, baseado no gráfico, descreva os seus comportamentos.
3. Considerando o banco de dados criado no Exercício 1 desta seção, divida a idade em três categorias (menores de 18 anos, idade entre 18 e 21 inclusive, e maiores de 21 anos). Construa gráficos de barra para essa variável, incluindo todos os indivíduos e um para cada sexo. Interprete os resultados obtidos.

1.4 Exercícios

1. Responda certo ou errado, justificando:
   - **a.** Suponha duas amostras colhidas de uma mesma população, sendo uma de tamanho 100 e outra de tamanho 200. Então, a amostra de tamanho maior é mais representativa da população.
   - **b.** Duas variáveis diferentes podem apresentar histogramas idênticos.
   - **c.** Duas variáveis com *box-plot* iguais não podem ter valores diferentes.
2. Suponha que duas empresas desejam empregá-lo e após considerar as vantagens de cada uma, você vai escolher aquela que lhe pagar melhor. Após certa pesquisa, você consegue a distribuição de salário das empresas, dadas segundo os gráficos abaixo.



   Com base nas informações de cada gráfico, qual seria sua decisão?
3. Uma pesquisa com usuários de transporte coletivo na cidade de São Paulo indagou sobre os diferentes tipos usados nas suas locomoções diárias. Dentre ônibus, metro e trem, o número de diferentes meios de transporte utilizados foi o seguinte: 2, 3, 2, 1, 2, 1, 2, 1, 2, 3, 1, 1, 1, 2, 2, 3, 1, 1, 1, 1, 2, 1, 1, 2, 2, 1, 2, 1, 2 e 3.
   - **a.** Organize uma tabela de freqüência.
   - **b.** Faça uma representação gráfica.
   - **c.** Admitindo que essa amostra represente bem o comportamento do usuário paulistano, você acha que a porcentagem dos usuários que utilizam mais de um tipo de transporte é grande?
4. A idade dos 20 ingressantes num certo ano no curso de pós-graduação em jornalismo de uma universidade foi o seguinte: 22, 22, 22, 22, 23, 23, 24, 24, 24, 24, 25, 25, 26, 26, 26, 26, 27, 28, 35 e 40.

a. Apresente os dados em uma tabela de freqüência, incluindo a freqüência relativa.

b. Idades atípicas parecem ter ocorrido nesse ano. Após sua retirada do conjunto de dados, refaça o item (a). Comente as diferenças encontradas.

5. Um novo medicamento para cicatrização está sendo testado e um experimento é feito para estudar o tempo (em dias) de completo fechamento em cortes provenientes de cirurgia. Uma amostra em trinta cobaias forneceu os valores: 15, 17, 16, 15, 17, 14, 17, 16, 16, 17, 15, 18, 14, 17, 15, 14, 15, 16, 17, 18, 18, 17, 15, 16, 14, 18, 18, 16, 15 e 14.

a. Organize uma tabela de freqüência.

b. Que porcentagem das observações estão abaixo de 16 dias?

c. Classifique como *rápida* as cicatrizações iguais ou inferiores a 15 dias e como *lenta* as demais. Faça um diagrama circular indicando as porcentagens para cada classificação.

6. O Posto de Saúde de um certo bairro mantém um arquivo com o número de crianças nas famílias que se utilizam do Posto. Os dados são os seguintes: 3, 4, 3, 4, 5, 1, 6, 3, 4, 5, 3, 4, 3, 3, 4, 3, 5, 5, 5, 5, 6, 11, 10, 2, 1, 2, 3, 1, 5 e 2.

a. Organize uma tabela de freqüência.

b. Faça uma representação gráfica.

c. Você identifica valores muito discrepantes? Que fazer com eles?

7. Um questionário foi aplicado aos dez funcionários do setor de contabilidade de uma empresa fornecendo os dados apresentados na tabela.

Funcionário	Curso (completo)	Idade	Salário (R$)	Anos de Empresa
1	superior	34	1100,00	5
2	superior	43	1450,00	8
3	médio	31	960,00	6
4	médio	37	960,00	8
5	médio	24	600,00	3
6	médio	25	600,00	2
7	médio	27	600,00	5
8	médio	22	450,00	2
9	fundamental	21	450,00	3
10	fundamental	26	450,00	3

a. Classifique cada uma das variáveis.

b. Faça uma representação gráfica para a variável Curso.

c. Discuta a melhor forma de construir a tabela de freqüência para a variável Idade. Construa uma representação gráfica.

d. Repita o item (c) para a variável Salário.

e. Considerando apenas os funcionários com mais de três anos de casa, descreva o comportamento da variável Salário.

8. Um grupo de pedagogos estuda a influência da troca de escolas no desempenho de alunos do ensino fundamental. Como parte do levantamento realizado, foi anotado o número de escolas cursadas pelos alunos participantes do estudo.

Escolas Cursadas	freqüência
1	46
2	57
3	21
4	15
5	4

a. Qual é a porcentagem dos alunos que cursaram mais de uma escola?

b. Construa o gráfico de barras.

c. Classifique os alunos em dois grupos segundo a rotatividade: *alta* para alunos com mais de 2 escolas e *baixa* para os demais. Obtenha a tabela de freqüência dessa variável.

9. Alunos da Escola de Educação Física foram submetidos a um teste de resistência quanto ao número de quilômetros que conseguiram correr sem parar. Os dados estão apresentados a seguir.

a. Qual é a variável em estudo?

b. Construa o histograma.

c. Obtenha o *box-plot*.

Faixas	freqüência
$0 \vdash 4$	438
$4 \vdash 8$	206
$8 \vdash 12$	125
$12 \vdash 16$	22
$16 \vdash 20$	9

10. O tempo de utilização de caixas eletrônicos depende de cada usuário e das operações efetuadas. Foram coletadas 26 medidas desse tempo (em minutos):

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