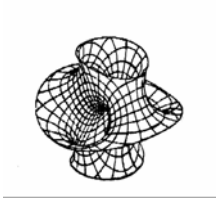
UNIVERSIDADE FEDERAL FLUMINENSE
CENTRO DE ESTUDOS GERAIS
INSTITUTO DE MATEMÁTICA



DEPARTAMENTO DE ESTATÍSTICA

ESTATÍSTICA DESCRITIVA

Ana Maria Lima de Farias
Luiz da Costa Laurencel

Agosto de 2008

.

Conteúdo

Capítulo 1

Introdução

1.1 O que é uma pesquisa estatística?

Freqüentemente nos deparamos com informações estatísticas nos jornais, televisão, empresas públicas ou privadas, etc. Por exemplo, quando a direção do Metrô do Rio de Janeiro informa que transporta 500.000 passageiros por dia, estamos lidando com uma estatística do número de passageiros do metrô. Tal estatística foi obtida com base na análise do movimento diário ao longo de um determinado período de tempo e dessas análises resultou um número que pretende dar uma idéia do movimento diário de passageiros. É claro que isso não significa que todo dia circulam exatamente 500.000 passageiros, mas tal número representa uma estimativa do número de passageiros.

Um outro exemplo que presenciamos periodicamente no Brasil são os Censos Demográficos, que são levantamentos realizados pelos governos com o objetivo de conhecer as características de sua população, suas condições sócio-econômicas, suas características culturais e religiosas, etc. Temos também os Censos Econômicos, com os quais se pretende conhecer as características da população formada pelos estabelecimentos econômicos do país; assim podemos ter o Censo Industrial, o Censo Agropecuário, etc.

Nas *pesquisas censitárias*, o objetivo é que *todos* os elementos da população tenham os seus dados levantados. Nos censos demográficos, isso significa que todas as pessoas e domicílios têm que ser visitados; já no censo industrial, todas as empresas que desenvolvam atividades industriais têm que ser pesquisadas. Com esses exemplos, vê-se que o conceito de *população de uma pesquisa estatística* é mais amplo, não se restringindo a seres humanos; ela é definida exatamente a partir dos objetivos da pesquisa. Mais precisamente, população é o conjunto de elementos para os quais se deseja estudar determinada(s) característica(s).

Um outro exemplo que faz parte do nosso dia-a-dia e que resulta de um levantamento estatístico é o índice de inflação, por exemplo, o Índice Nacional de Preços ao Consumidor (INPC) produzido pelo IBGE[1]. O índice de inflação é um número resultante de um levantamento de preços que resume a variação dos preços durante um determinado período de tempo. Sendo esse levantamente realizado mensalmente, não é possível levantar os preços de todos os produtos em todos os estabelecimentos. Então, é feita uma seleção de produtos e estabelecimentos a serem pesquisados. Temos, assim, um exemplo de *pesquisa por amostragem*. Nessas pesquisas, são selecionados alguns elementos da população, que compõem a *amostra*, e métodos estatísticos de inferência nos permitem generalizar os resultados obtidos com a amostra para toda a população de interesse. Na pesquisa do INPC, temos amostragem dos produtos e serviços, bem como dos locais onde é feito o levantamento dos preços.

[1] Fundação Instituto Brasileiro de Geografia e Estatística

Outro exemplo de pesquisa por amostragem são as pesquisas de intenção de voto: alguns eleitores são entrevistados e daí tiram-se estimativas dos percentuais de votos de cada candidato.

Esses exemplos ilustram, então, o conceito de *pesquisa estatística*, que consiste num trabalho de identificação, reunião, tratamento, análise e apresentação de informações (dados) para satisfazer certa necessidade. Em qualquer levantamento ou pesquisa estatística é fundamental um planejamento cuidadoso de todo o processo, resultando na necessidade da elaboração da *metodologia da pesquisa*, que consiste em um conjunto de definições, procedimentos, rotinas, métodos e técnicas utilizados para a obtenção e apresentação das informações desejadas.

Nas pesquisas por amostragem, em particular, o método de seleção da amostra é uma peça fundamental, pois os elementos da amostra têm que ser *representativos* da população à qual os resultados da pesquisa serão estendidos. Por exemplo, numa pesquisa de intenção de voto para prefeito do município do Rio de Janeiro, a amostra tem que ser representativa de todas as regiões do município; não podemos concentrar a pesquisa em Copacabana, por exemplo, pois o comportamento do eleitorado desse bairro pode ser diferente do comportamento dos eleitores da Rocinha, em São Conrado. Na pesquisa de preços para elaboração do INPC, temos que ter um levantamento nas principais regiões do país para que o índice resultante possa ser representativo do movimento de preços em todo o país.

De posse dos dados levantados, temos que decidir como os resultados serão organizados e apresentados. Do Censo Demográfico, por exemplo, saem diversas tabelas que nos informam a população do Brasil por município, o nível de escolaridade da população, etc. No levantamento de preços para medir a inflação, um dos resultados é um número em forma percentual, que indica a variação dos preços de um mês para outro.

Nas pesquisas por amostragem, temos uma etapa importante, que é a etapa de estimação, onde se decide como os resultados obtidos para a amostra serão estendidos para toda a população e qual o erro máximo que teremos nessa estimativa.

Assim, temos identificadas em diferentes pesquisas as três grandes áreas da Estatística, que, no entanto, não formam ramos isolados:

- Amostragem e Planejamento de Experimentos - processo de obtenção dos dados;
- Estatística Descritiva - organização, apresentação e sintetização dos dados;
- Estatística Inferencial - conjunto de métodos para a tomada de decisão nas situações onde existam incertezas e variações.

Neste curso introdutório, estaremos lidando com a parte da Estatística Descritiva, quando veremos técnicas de análise exploratória de dados. O objetivo é capacitar o aluno a organizar conjuntos de dados, desenvolvendo uma postura crítica na análise dos fenômenos em estudo. Sempre que possível, estaremos utilizando conjuntos de dados reais, referentes à realidade sócio-econômica brasileira.

1.2 Organização das notas de aula

Estas notas de aula estão divididas em 3 capítulos. No Capítulo 2 são apresentados métodos de análise exploratória de dados, tanto tabulares quanto gráficos. No Capítulo 3 apresentam-se as principais medidas estatísticas de posição, dispersão, assimetria e associação entre variáveis.

Os capítulos são divididos em seções e subseções. Ao final de cada seção é dado um conjunto de exercícios resolvidos para auxiliar o aluno na compreensão dos conceitos dados (você deve tentar fazer os exercícios, antes de ler a solução) e em seguida um conjunto de exercícios propostos. Ao final de cada capítulo há um conjunto de exercícios complementares, abrangendo toda a

matéria do capítulo. Os gabaritos completos dos exercícios está disponibilizado no site do curso, www.uff.br/ieeanamariafarias.

Os exemplos apresentados ao longo do texto, sempre que possível, contemplarão dados verídicos, obtidos de diversas fontes pertinentes à realidade brasileira. Os alunos interessados poderão obter cópia do disquete com os dados utilizados no texto com os autores. Vários conjuntos de dados se referem a pesquisas realizadas pela Fundação Instituto Brasileiro de Geografia e Estatística - IBGE - e podem ser encontrados na página www.ibge.gov.br.

Capítulo 2

Apresentação de dados

2.1 Introdução

De posse dos dados obtidos de um levantamento estatístico (censitário ou por amostragem), é importante escolher a forma como esses dados serão apresentados, de modo a facilitar a visualização dos resultados desejados. Neste capítulo serão vistas algumas técnicas de apresentação de dados, tanto tabulares quanto gráficas.

2.2 Níveis de mensuração

Um problema básico que se coloca nos levantamentos estatísticos é o nível de mensuração das informações a serem levantadas. Isto porque a aplicabilidade ou não de modelos e métodos estatísticos a serem utilizados posteriormente na análise do material vai depender em grande parte desse aspecto.

O nível mais elementar de mensuração consiste na classificação dos indivíduos ou objetos de uma população de acordo com uma certa característica, isto é, tenta-se separar os elementos em grupos, conforme possuam essa ou aquela característica em questão. É o que sucede, por exemplo, quando a característica estudada é sexo, religião, estado civil, etc. Nesses casos, as categorias se expressam nominalmente e para a aplicação de técnicas estatísticas adequadas, é necessário que as categorias sejam *exaustivas* (isto é, cubram todos os elementos da população) e *mutuamente exclusivas* (isto é, um elemento não pode pertencer simultaneamente a duas categorias distintas). Nesses casos, diz-se que a característica em estudo é expressa segundo uma *escala nominal*. Assim, as operações usuais de aritmética não podem ser realizadas sobre esse tipo de escala, mesmo que as categorias estejam expressas em números. No processamento de dados, é bastante comum representar as categorias de sexo Feminino e Masculino por números, como 1 e 2. Naturalmente, não faz sentido dizer que o Masculino é duas vezes o Feminino; o 1 e o 2 são apenas substitutos dos nomes das categorias.

Num nível de mensuração seguinte, podemos ordenar as categorias de uma determinada característica. É o que ocorre com o nível de escolaridade, quando uma população pode ser classificada em 4 categorias: analfabeto, 1º grau, 2º grau, 3º grau, por exemplo. Aqui podemos dizer que o nível de escolaridade de um indivíduo da categoria 2º grau é maior que o de um indivíduo da categoria 1º grau, mas não podemos dizer que é duas vezes maior. Nesta escala, chamada *escala ordinal*, valem apenas as operações de ordenação, maior do que ou menor do que.

Passa-se deste tipo de escala para um nível de mensuração propriamente dito quando, além da ordenação das categorias, pode-se dizer quanto valem exatamente as diferenças entre essas categorias. Um exemplo típico dessa situação é a medição de temperatura: a diferença entre 90°C e 70°C é 20°C e é igual à diferença entre 30°C e 10°C. No entanto, como o zero (0°C) nesta escala é definido

arbitrariamente (não existe naturalmente), não podemos dizer que 90°C é três vezes mais quente que 30°C. Dizemos, então, que a temperatura está medida em uma *escala intervalar.*

Quando o zero na escala puder ser estabelecido de forma não arbitrária, todas as operações aritméticas poderão ser realizadas sobre os valores tomados pela característica em estudo. Nesse caso, dizemos que a característica está medida em uma *escala de razão* ou *proporcional.* É o caso da idade, que é contada a partir da data de nascimento do indivíduo.

É comum denominar de *variável qualitativa* as características medidas em escala nominal ou ordinal. Já as variáveis medidas em escala intervalar ou proporcional são chamadas *variáveis quantitativas.* As variáveis quantitativas, por sua vez, podem ser discretas ou contínuas. Quando a variável puder assumir qualquer valor numérico em um determinado intervalo de variação, ela será uma variável *contínua.* Essas variáveis resultam normalmente de medições: peso, altura, dosagem de hemoglobina, renda, etc. A interpretação desse tipo de variável leva à noção de valor aproximado, pois não existe instrumento de medição capaz de fornecer precisão absoluta na informação. Assim, quando uma balança mostra o peso de uma pessoa como 65,5 kg, esse valor, na verdade, é uma aproximação para qualquer valor entre, digamos, 65,495 e 65,505 kg. Por outro lado, a variável quantitativa *discreta* só poderá assumir valores pertencentes a um conjunto enumerável; os valores normalmente são obtidos através de algum processo de contagem. Alguns exemplos são: número de filhos de um casal, número de empregados de uma firma de contabilidade, etc.

Exemplo 2.1 *A Pesquisa Mensal de Emprego*

"A Pesquisa Mensal de Emprego[1] - PME - é uma das principais fontes das estatísticas do trabalho, no âmbito do IBGE. Mensalmente são produzidas e divulgadas distintas estatísticas sobre a estrutura e a distribuição da *população economicamente ativa*, sobre os *níveis de ocupação e de desocupação*, sobre os *rendimentos médios* da população ocupada, entre outras.

Essas estatísticas, sob diferentes cruzamentos, como a *idade*, o *sexo*, a *ocupação*, a *atividade*, entre outros, são essenciais a uma ampla análise do desempenho da economia de um país. Pela compreensão do estado de sua força de trabalho, um país poderá implementar políticas econômicas e sociais que o levem a um desenvolvimento mais racional."

Vamos identificar as variáveis envolvidas na PME, segundo o texto acima.

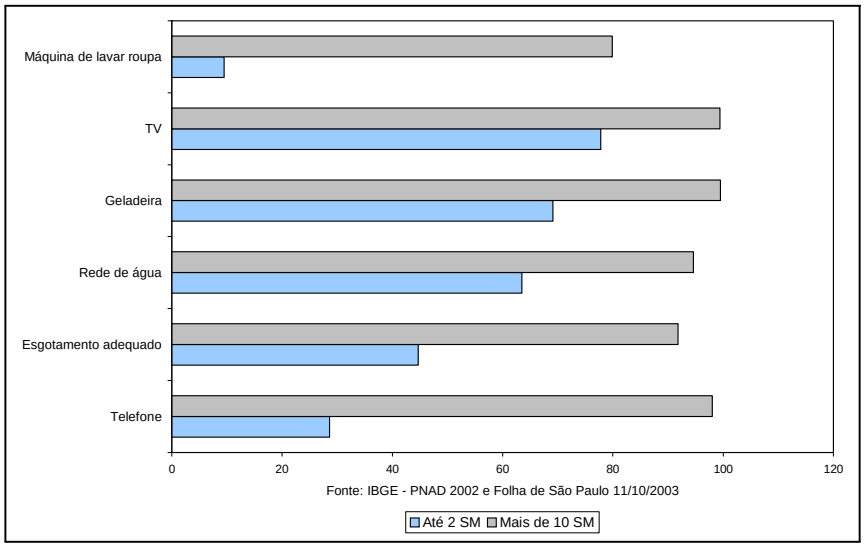
- População economicamente ativa: é uma variável quantitativa discreta, que mede o número de pessoas (potencial de mão de obra) com que o setor produtivo pode contar;
- Nível de ocupação e de desocupação: são variáveis quantitativas contínuas, que medem a taxa de emprego e desemprego;
- Rendimento médio: é uma variável quantitativa contínua;
- Idade: variável quantitativa discreta;
- Sexo: variável qualitativa nominal;
- Ocupação: variável qualitativa nominal;
- Atividade econômica: variável qualitativa nominal.

[1] *Para Compreender a PME - Um Texto Simplificado*, IBGE, 1991.

2.2.1 Exercícios propostos da Seção 2.2

2.1 *Na Figura 2.1 temos um gráfico que ilustra a presença de bens e serviços nos domicílios das duas classes de renda extremas, segundo a Pesquisa Nacional porAmostra de Domicílios realizada pelo IBGE. Defina e classifique todas as variáveis envolvidas; tente imaginar como esses dados foram coletados na pesquisa.*

Figura 2.1: Bens e serviços nos domicílios por classe de renda



2.2 *Na Tabela 2.1 apresentam-se dados referentes aos estabelecimentos de ensino brasileiros. Defina e classifique as variáveis envolvidas na tabela.*

Tabela 2.1: Dados gerais dos estabelecimentos de ensino(1994) para o Exercício 2.2

Especificação	Pré-escolar	1º grau	2º grau	Superior
Estabelecimentos	**115 318**	**195 545**	**13 178**	**851**
Público	99 529	181 586	9 013	218
Privado	15 789	13 959	4 165	633
Matrículas	**5 339 288**	**31 091 662**	**4 426 543**	**1 661 034**
Público	4 121 188	27 508 600	3 383 822	690 450
Privado	1 218 100	3 583 062	1 042 721	970 584

Fonte: Brasil em números, vol. 4, 1995-1996 - IBGE

2.3 Distribuição univariada de freqüências: Representação tabular

Considere os dados da Tabela 2.2, onde temos informações sobre a turma, o sexo, a matéria predileta (**P**ortuguês, **M**atemática, **H**istória, **G**eografia ou **C**iências) no 2º grau e a nota (número de questões certas) em um teste de múltipla escolha com 10 questões de matemática, ministrado no primeiro

dia de aula dos calouros de Economia. As três primeiras variáveis são qualitativas, enquanto nota é uma variável quantitativa discreta.

Como podemos resumir essas informações de uma forma mais clara e objetiva? Afinal, o que nos interessa é saber quantas mulheres e quantos homens há em cada turma, quantas pessoas tiraram 10, e assim por diante. Para isso, vamos construir *tabelas ou distribuições de freqüência*.

Tabela 2.2: Dados sobre sexo, matéria predileta e nota de alunos de 2 turmas

Turma	Sexo	Predileta	Nota	Turma	Sexo	Predileta	Nota	Turma	Sexo	Predileta	Nota
A	F	H	5	A	M	M	2	B	F	G	6
A	M	M	8	A	M	G	4	B	F	M	4
A	F	P	8	A	M	G	9	B	M	M	6
A	F	H	6	A	M	M	7	B	F	P	5
A	M	C	5	A	M	M	1	B	M	G	3
A	M	H	6	A	F	P	8	B	F	M	5
A	F	M	8	A	F	G	5	B	M	P	3
A	F	P	4	A	M	G	9	B	M	M	4
A	F	H	2	A	M	P	5	B	F	C	8
A	M	C	6	A	F	M	8	B	F	H	3
A	F	P	8	A	F	G	6	B	M	G	4
A	M	H	3	A	F	P	9	B	M	P	5
A	M	M	5	A	M	M	8	B	M	P	4
A	F	P	5	B	F	H	6	B	M	H	6
A	F	G	5	B	M	M	3	B	M	M	6
A	M	C	7	B	F	P	4	B	M	G	6
A	M	H	4	B	M	H	8	B	M	H	6
A	F	M	7	B	M	G	10	B	M	H	6
A	F	P	7	B	F	M	5	B	F	M	8
A	F	M	6	B	F	P	7	B	F	M	8
A	M	G	6	B	F	P	5	B	F	G	5
A	M	H	9	B	M	M	6	B	M	C	5
A	F	M	8	B	F	M	5				
A	M	P	5	B	M	G	5				
A	M	G	6	B	F	H	8				
A	F	M	7	B	F	G	5				
A	M	P	5	B	M	G	6				
A	F	M	5	B	F	M	5				
A	F	M	5	B	M	G	2				

2.3.1 Variáveis qualitativas

Vamos começar com a variável qualitativa sexo. Analisando as duas turmas conjuntamente, vemos que há um total de 41 alunos e 39 alunas. Essas contagens são chamadas *freqüências absolutas*.

Poderíamos resumir essa informação em forma de tabela:

Sexo	Número de alunos
Masculino	41
Feminino	39
Total	80

Note a linha referente ao total!

Caso quiséssemos a informação por turma, a tabela seria a seguinte:

Sexo	Número de alunos		
	Turma A	Turma B	Total
Masculino	21	20	41
Feminino	21	18	39
Total	42	38	80

Note a coluna referente ao total!

Uma dificuldade que surge na comparação das duas turmas é o fato de o total de alunos ser diferente. Assim, é comum acrescentar, à tabela de freqüências, uma nova coluna com as *freqüências relativas*, que nada mais são que as freqüências em forma percentual, representando a participação da freqüência de cada uma das categorias da variável sexo no total de alunos. Na Tabela 2.3 temos a versão completa; note que foi acrescentado um título e a fonte dos dados, informações imprescindíveis na apresentação de dados.

Tabela 2.3: Distribuição da variável Sexo por turma

Sexo	Freqüência na Turma A		Freqüência na Turma B		Freqüência Total	
	Absoluta	Relativa (%)	Absoluta	Relativa (%)	Absoluta	Relativa (%)
Masculino	21	50,00	20	52,63	41	51,25
Feminino	21	50,00	18	47,37	39	48,75
Total	42	100,00	38	100,00	80	100,00

Fonte: Dados hipotéticos

Vamos ver com detalhes como foi feito o cálculo das freqüências relativas. Lembre-se que a idéia é analisar o percentual de cada sexo no respectivo grupo:

	Turma A	Turma B
Masculino	$\frac{21}{42} \times 100 = 50,000000$	$\frac{20}{38} \times 100 = 52,631579$
Feminino	$\frac{21}{42} \times 100 = 50,000000$	$\frac{18}{38} \times 100 = 47,368421$
Total	$\frac{41}{80} \times 100 = 51,250000$	$\frac{39}{80} \times 100 = 48,750000$

Vale a pena salientar, neste momento, a questão do arredondamento de resultados. Nos cálculos acima, as freqüências estão apresentadas com 6 casas decimais, enquanto que, na Tabela 2.3, os resultados estão com 2 casas decimais, que é a forma usual. Existe a seguinte **regra de arredondamento**:

Regra 2.1 *Regra de Arredondamento*

Quando o primeiro algarismo a ser suprimido é menor ou igual a 4 (isto é, é igual a 0, 1, 2, 3 ou 4), o algarismo final (depois do arredondamento) permanece inalterado. Quando o primeiro algarismo a ser suprimido é igual a 5, 6, 7, 8 ou 9, o algarismo final (depois do arredondamento) é acrescido de 1.

Vamos arredondar as freqüências para a Turma B mantendo 2 casas decimais no resultado final. Para o sexo masculino, o primeiro algarismo a ser suprimido (terceira casa decimal) é 1 e, assim, o algarismo final permanece inlaterado; esse algarismo (segunda casa decimal) é 3 que, depois do arredondamento, continua sendo 3, o que resulta na freqüência relativa arredondada de 52,63. Para o sexo feminino, o primeiro algarismo a ser suprimido (terceira casa decimal) é 8 e, assim, o algarismo final é acrescido de 1; esse algarismo (segunda casa decimal) é 6, que depois do arredondamento passa a ser 7, o que resulta na freqüência relativa arredondada de 47,37.

A título de ilustração, apresenta-se na Tabela 2.4 a distribuição para a variável qualitativa "matéria predileta no segundo grau".

Tabela 2.4: Distribuição da variável Matéria Predileta no Segundo Grau por turma

Matéria Predileta	Freqüência na Turma A		Freqüência na Turma B		Freqüência Total	
no Segundo Ggrau	Absoluta	Relativa (%)	Absoluta	Relativa (%)	Absoluta	Relativa (%)
Português	10	23,81	7	18,42	17	21,25
Matemática	14	33,33	12	31,58	26	32,50
História	7	16,67	7	18,42	14	17,50
Geografia	8	19,05	10	26,32	18	22,50
Ciências	3	7,14	2	5,26	5	6,25
Total	42	100,00	38	100,00	80	100,00

2.3.2 Variáveis quantitativas

Vamos, agora, analisar a variável Nota, que é uma variável quantitativa discreta. Na Tabela 2.5 temos as notas ordenadas. A listagem dos dados, mesmo ordenados, é de pouca utilidade nas situações práticas, uma vez que, em geral, o número de observações é muito grande. Além disso, ao se analisarem dados estatísticos, muitas vezes o interesse não está na observação individual, mas, sim, no comportamento de grupos. Mais difícil ainda é a comparação entre os resultados das duas turmas, uma vez que as turmas têm números de alunos diferentes.

Tabela 2.5: Notas ordenadas por turma

Turma A											Turma B										
1	2	2	3	3	3	3	5	5	5	5	2	3	3	3	3	4	4	4	4	4	5
5	5	5	5	5	5	5	6	6	6	6	5	5	5	5	5	5	5	5	5	5	6
6	6	6	7	7	7	7	7	8	8	8	6	6	6	6	6	6	6	6	6	7	8
8	8	8	8	8	9	9	9	9			8	8	8	8	10						

A partir dos dados ordenados, podemos saber rapidamente os valores mínimo e máximo: na Turma A as notas variam de 2 a 10 e na Turma B, de 1 a 9. Esse é o conceito de *amplitude* de um conjunto de dados.

Definição 2.1 *A **amplitude** de um conjunto de dados, representada por Δ_{total}, é definida como a diferença entre os valores máximo e mínimo:*

$$\Delta_{total} = V_{Máx} - V_{Mín} \tag{2.1}$$

A amplitude das notas da turma A é $10 - 2 = 8$ e da turma B é $9 - 1 = 8$, ou seja, ambas as turmas têm a mesma amplitude, embora os valores extremos sejam diferentes.

Considere novamente os dados da Tabela 2.5. Um primeiro fato que chama a atenção é a existência de vários alunos com notas iguais. Então, uma forma mais simplificada de apresentar os dados, sem nenhuma perda de informação, é construir uma tabela ou distribuição de freqüências, da mesma forma que fizemos para as variáveis qualitativas. Em uma coluna colocamos as diferentes notas existentes e nas colunas adjacentes, as freqüências absolutas e relativas. Na Tabela 2.6 temos uma apresentação inicial para as notas das turmas A e B.

Tabela 2.6: Freqüências absolutas e relativas das notas de um teste de múltipla escolha

Turma A		
Nota	Freqüência	
	Absoluta	Relativa
1	1	2,38
2	2	4,76
3	1	2,38
4	3	7,14
5	11	26,19
6	7	16,67
7	5	11,91
8	8	19,05
9	4	9,52
Total	42	100,00

Turma B		
Nota	Freqüência	
	Absoluta	Relativa
2	1	2,63
3	4	10,53
4	5	13,16
5	11	28,95
6	10	26,32
7	1	2,63
8	5	13,16
9	0	0,00
10	1	2,63
Total	38	100,00

No cálculo das freqüências relativas, o arredondamento se fez segundo a regra dada anteriormente. No total das freqüências relativas, o resultado é 100,00 para ambas as turmas, uma vez que esse é o objetivo das freqüências relativas em forma percentual: os totais passam a ser 100. No entanto, ao somarmos as freqüências relativas da turma B aí apresentadas, o resultado não é exatamente 100; mais precisamente,

$$2,63 + 10,53 + 13,16 + \cdots + 2,63 = 100,01$$

Isso se deve aos arredondamentos efetuados. No entanto, é comum apresentar o total como 100, ficando subentendido que qualquer diferença é devida a arredondamentos. Em geral, essas diferenças são pequenas, desde que se mantenha um procedimento coerente de arredondamento. Voltaremos a apresentar mais exemplos sobre essa questão para ilustrar alguns procedimentos comuns e aconselháveis no processo de arredondamento.

Na apresentação de tabelas de freqüências para variáveis quantitativas, é comum acrescentar mais duas colunas com as freqüências acumuladas. Por exemplo, se, para aprovação, o aluno precisa tirar no mínimo 6, quantos alunos foram aprovados em cada turma? Para facilitar a resposta de perguntas desse tipo, é costume acrescentar uma coluna com as *freqüências acumuladas*, que dão, para cada nota (linha da tabela), o total de notas *menores ou iguais* à nota em questão. Na turma A, a menor nota é 2; assim, não há notas menores que 2 e a freqüência acumulada (nota ≤ 2) para essa nota é igual à freqüência simples. Para a nota 3, há $1 + 4 = 5$ notas menores ou iguais a 3; assim, a freqüência acumulada para a nota 3 é 5. Há $1 + 4 + 5 = 5 + 5 = 10$ notas menores ou

iguais a 4; assim, a freqüência acumulada para a nota 4 é 10. Continuando com esse procedimento, obtemos as Tabelas 2.7 e 2.8 para as turmas A e B, respectivamente. Note que, agora, mudamos os nomes para freqüências simples e freqüências acumuladas (absolutas ou relativas) para diferenciar os dois tipos de freqüência.

Tabela 2.7: Distribuição de freqüências das notas de um teste de múltipla escolha - Turma A

Nota	Freqüência Simples		Freqüência Acumulada	
	Absoluta	Relativa	Absoluta	Relativa
1	1	2,38	1	2,38
2	2	4,76	3	7,14
3	1	2,38	4	9,52
4	3	7,14	7	16,66
5	11	26,19	18	42,85
6	7	16,67	25	59,52
7	5	11,91	30	71,43
8	8	19,05	38	90,48
9	4	9,52	42	100,00
Total	42	100,00		

Fonte: Dados fictícios

Tabela 2.8: Distribuição de freqüências das notas de um teste de múltipla escolha - Turma B

Nota	Freqüência Simples		Freqüência Acumulada	
	Absoluta	Relativa	Absoluta	Relativa
2	1	2,63	1	2,63
3	4	10,53	5	13,16
4	5	13,16	10	26,32
5	11	28,95	21	55,26
6	10	26,32	31	81,58
7	1	2,63	32	84,21
8	5	13,16	37	97,37
9	0	0,00	37	97,37
10	1	2,63	38	100,00
Total	38	100,00		

Fonte: Dados fictícios

Novamente, vamos fazer uma observação sobre os cálculos efetuados, concentrando nossa atenção na turma B, ou seja, na Tabela 2.8. Há duas maneiras possíveis de se calcularem as freqüências acumuladas relativas. Da mesma forma como feito para as freqüências absolutas acumuladas, podemos acumular as freqüências simples relativas:

$$\begin{aligned} 2,63 + 10,53 &= 13,16 \\ 2,63 + 10,53 + 13,16 &= 26,32 \\ 2,63 + 10,53 + 13,16 + 28,95 &= 55,27 \\ 2,63 + 10,53 + 13,16 + 28,95 + 26,32 &= 81,59 \end{aligned}$$

e assim por diante. Note que com esse procedimento obteremos a freqüência 100,01 na última classe. Outra possibilidade, que, em geral, fornece resultados mais precisos, consiste em calcular as

freqüências acumuladas relativas a partir das freqüências acumuladas simples, dividindo pelo total de observações. Isto é,

$$\begin{aligned} 100 \times 5/38 &= 13,16 \\ 100 \times 10/38 &= 26,32 \\ 100 \times 21/38 &= 55,26 \\ 100 \times 31/38 &= 81,58 \end{aligned}$$

e assim por diante. Note que, a partir da quarta freqüência acumulada, já desaparece a diferença de 0,01 nos resultados (55,26 em vez de 55,27), o que faz com que o total neste caso seja 100,00 e não 100,01.

É importante observar que, para variáveis qualitativas, como sexo, não faz sentido trabalharmos com as freqüências acumuladas, uma vez que não existe relação de grandeza entre as categorias de uma variável qualitativa. Por exemplo, não podemos falar "menor ou igual a Masculino".

O procedimento apresentado acima pode ser usado para dados quantitativos discretos em geral, desde que não haja muitos valores distintos. No exemplo das notas, o número de notas diferentes era 9 em ambas as turmas e, assim, a tabela resultante tinha um tamanho razoável. Consideremos, agora, os dados da Tabela 2.9, onde temos o número de empregados das Unidades Locais[2] industriais de empresas[3] industriais no estado do Rio de Janeiro, que é também uma variável quantitativa discreta.

Tabela 2.9: Número de empregados das UL's industriais - RJ

6	11	21	28	14	21	110	14	6	7	503	120	5	5	6	17	11
13	20	6	10	73	8	9	72	17	22	27	80	7	12	24	13	15
6	6	33	40	16	13	51	47	12	11	40	73	56	26	9	8	461
38	19	23	5	49	29	7	33	21	55	11	9	13	19	15	10	8
35	30	14	16	7	26	56	36	40	6	837	6	9	9	19	5	10
6	21	30	14	55	11	5	8	6	5	12	8	6	5	20	24	6
19	6	15	6	15	8	7	7	9	9	18	17	54	6	13	12	17
10	8	11	12	7	28	8	18	9	25	8	16	274	5	37	45	7
53	7	5	8	26	5	11	6	7	7	12	705	6	23	10	13	351
204	22	5	9	38	5	11	10	98	216	10	6	18	20	14	32	20
7																

Fonte: Pesquisa Industrial Mensal de Emprego e Salário - PIMES -IBGE

Nesta tabela, além do número total de observações ser bem maior (171), há também muitos valores distintos: 55. Por exemplo, temos 12 ULs com 5 empregados, 18 com 6 empregados e assim por diante. Uma tabela com 55 linhas é difícil de analisar; além disso, não há necessidade de sermos tão detalhistas. Por exemplo, em se tratando de número de empregados em ULs industriais, não há diferença significativa entre uma UL com 5 e outra com 6 empregados ou uma com 100 e outra com 101. Nesses casos, é comum *agrupar os dados em classes*. A idéia, então, é definir limites de classes de tal modo que, se o número de empregados de uma UL estiver entre determinados limites, ela será classificada como micro indústria, por exemplo. A construção da distribuição de freqüências se faz de maneira idêntica à vista anteriormente; a diferença é que as freqüências agora se referem

[2] Unidade Local é o endereço de atuação de uma empresa, ocupando geralmente uma área contínua na qual são desenvolvidas uma ou mais atividades econômicas.

[3] Empresa é a unidade jurídica que responde por uma firma ou Razão Social, englobando o conjunto de atividades econômicas exercidas em uma ou mais unidades locais.

às freqüências de classes de valores, em vez de se referirem a um único valor. Por essa razão, tais distribuições são chamadas às vezes de *distribuição de freqüências agrupadas*.

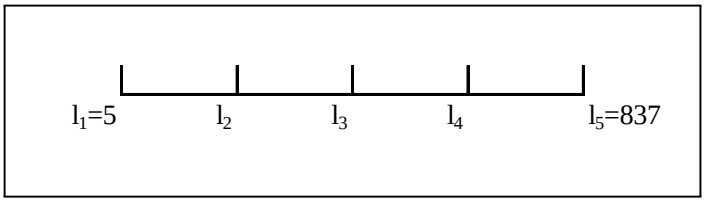
Há duas regras fundamentais que têm que ser seguidas quando da definição das classes de distribuições de freqüências agrupadas.

Regra 2.2 *Definição das classes em uma distribuição de freqüências agrupadas*

1. *As classes têm que ser exaustivas, isto é, todos os elementos devem pertencer a alguma classe.*
2. *As classes têm que ser mutuamente exclusivas, isto é, cada elemento tem que pertencer a uma única classe.*

Para simplificar a questão, suponhamos inicialmente que queiramos trabalhar com 4 classes e que todas as classes devam ter comprimentos iguais. Como determinar os limites das classes? O procedimento está ilustrado na Figura 2.2 para os dados da Tabela 2.9, onde o valor mínimo é 5 e o valor máximo é 837.

Figura 2.2: Definição dos limites de classe



Como cada classe tem que ter comprimento igual e o comprimento total de variação, isto é, a amplitude é $837 - 5 = 832$, cada intervalo deve ter comprimento

$$\delta = \frac{832}{4} = 208;$$

logo, os limites das classes são:

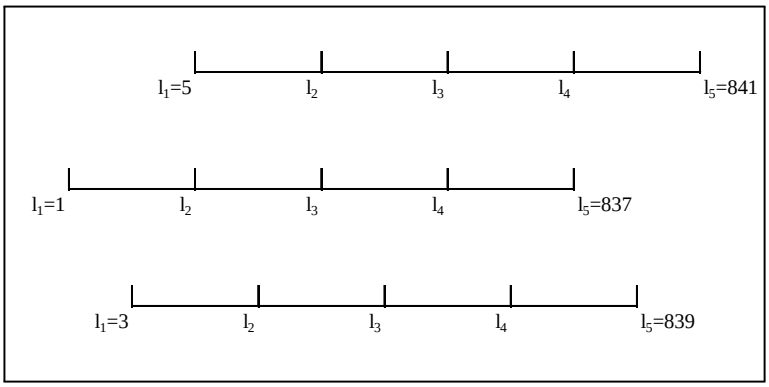
$$\begin{aligned} l_1 &= 5 \\ l_2 &= l_1 + \delta = 5 + 208 = 213 \\ l_3 &= l_2 + \delta = 213 + 208 = 421 \\ l_4 &= l_3 + \delta = 421 + 208 = 629 \\ l_5 &= 503 + 166 = 629 + 208 = 837 \end{aligned}$$

Dessa forma, as ULs com número de empregados entre 5 e 213 seriam classificadas como micro, entre 213 e 421 como pequenas, entre 421 e 629 como médias e entre 629 e 837 como grandes. O problema agora é definir o tratamento a ser dado às ULs com número de empregados exatamente igual a um dos limites. Obviamente, as ULs com 5 empregados têm que ser incluídas na primeira classe. Se incluirmos o 213 na primeira classe, isto é, trabalharmos com o intervalo fechado $[5, 213]$, a próxima classe teria que ser do tipo $(213, 421]$[4], pois as classes têm que ser mutuamente exclusivas. Mas é totalmente inadequado trabalhar com classes de tipos diferentes. A solução, então, é definir

[4] O parênteses indica que o valor não está incluído no intervalo e o colchete indica que o valor está incluído no intervalo. Essa notação é equivalente a $213 < x \leq 421$.

a primeira classe como $[5, 213)$ e a segunda como $[213, 421)$. Continuando com esse procedimento, as outras classes seriam $[421, 629)$ e $[629, 837)$. Note a última classe! Ela não inclui o valor máximo 837! Esse problema surgiu porque utilizamos a amplitude exata dos dados. Uma solução é aumentar um pouco a amplitude e repetir o procedimento. Só que o mais conveniente é aumentar a amplitude para o próximo múltiplo do número de classes, para não termos limites de classes fracionários, uma vez que a variável em estudo (número de empregados) só assume valores inteiros. A amplitude exata é 832; o próximo múltiplo de 4 é 836, implicando num aumento de 4 unidades na amplitude. Na Figura 2.3 temos a ilustração de diferentes maneiras de redefinir as classes.

Figura 2.3: Método de correção da definição dos limites de classe



Na primeira opção, toda a diferença de 4 unidades foi alocada na cauda superior da distribuição, enquanto que, na segunda, essas 4 unidades foram alocadas na cauda inferior. Na terceira opção, as 4 unidades foram igualmente distribuídas, 2 unidades em cada cauda da distribuição. Em geral, esse último procedimento é o mais recomendado. Utilizando-o, a amplitude de classe passa a ser

$$\delta = \frac{836}{4} = 209$$

e as classes passam a ser $[3, 212)$, $[212, 421)$, $[421, 630)$, $[630, 839)$.

A construção da tabela se faz de maneira análoga à descrita nas Tabelas 2.7 e 2.8, só que agora contamos o número de ocorrências em cada classe, isto é, cada freqüência simples aboluta se refere ao número de valores em cada classe. Para agilizar o processo de contagem manual (em geral, essas tabelas são construídas com o auxílio de algum programa de computador), podemos fazer um esquema de marcação, de modo que só precisamos "varrer" o conjunto de dados uma única vez. Por exemplo, varrendo o conjunto de dados por linha (linha 1, depois linha 2, etc), obtemos as seguintes marcações e respectivas contagens referentes às 3 primeiras linhas:

$[3, 212)$	||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||	49
$[212, 421)$		
$[421, 630)$	||	2
$[630, 839)$		

Continuando com a contagem, obtemos a Tabela 2.10.

Uma observação interessante sobre essa distribuição é a alta concentração de ULs na primeira classe. Esse fato caracteriza a *assimetria* da distribuição, como veremos adiante, e é bastante comum

Tabela 2.10: Distribuição de freqüência do número de empregados das ULs industriais - RJ

Classe	Freqüência Simples		Freqüência Acumulada	
de PO	Absoluta	Relativa	Absoluta	Relativa
[3,212)	164	95,91	164	95,91
[212,421)	3	1,75	167	97,66
[421,630)	2	1,17	169	98,83
[630,835)	2	1,17	171	100,00
Total	171	100,00		

Fonte: PIMES - IBGE

para esse tipo de variável, ou seja, na maioria dos setores industriais, existem muitas indústrias com poucos empregados e poucas indústrias com muitos empregados. O mais razoável para esse tipo de distribuição é trabalhar com *classes de tamanhos diferentes*. Por exemplo, o IBGE, na elaboração da amostra da PIMES - Pesquisa Industrial Mensal de Emprego e Salário - definiu as seguintes classes de pessoal ocupado (PO): $[5, 30), [30, 100), [100, 500)$ e $PO \geq 500$. Note que a última classe não tem limite superior; na verdade, em cada unidade da federação, o máximo do PO é um número diferente mas só estamos interessados nas ULs com 500 ou mais empregados. Usando essas classes, a distribuição de freqüências passa a ser como a da Tabela 2.11.

Tabela 2.11: Distribuição de freqüência do número de empregados das ULs industriais - RJ

Número de	Freqüência Simples		Freqüência Acumulada	
empregados	Absoluta	Relativa	Absoluta	Relativa
$5 \vdash 30$	133	77,778	133	77,778
$30 \vdash 100$	28	16,374	161	94,152
$100 \vdash 500$	7	4,094	168	98,246
≥ 500	3	1,754	171	100,000
Total	171	100,000		

Fonte: Tabela 2.9

O procedimento de construção de distribuição de freqüências agrupadas foi ilustrado usando-se uma variável quantitativa discreta mas pode também ser aplicado a variáveis quantitativas contínuas, conforme veremos a seguir, onde vamos trabalhar com o preço da dúzia de ovos (em centavos) nos estados americanos em 1990, apresentados na Tabela 2.12 [cf. Gujarati(1995), *Basic Econometrics*, McGraw-Hill, 3ª ed, Tabela 1.1].

O valor mínimo é 48,0 centavos e o valor máximo é 151,0; sendo assim, a amplitude é $151,0 - 48,0 = 103,0$. Trabalhando com 5 classes de mesmo tamanho, devemos arredondar a amplitude para o próximo múltiplo de 5, que é 105, e definir a amplitude de cada classe como $\delta = 105/5 = 21$. Distribuindo as 2 unidades a mais $(105-103 = 2)$ nas duas caudas da distribuição, as classes passam a ser $[47, 68)$; $[68, 89)$; $[89, 110)$; $[110; 131)$; $[131, 152)$. Na Tabela 2.13 temos a distribuição final.

É importante notar que as regras apresentadas para definição das classes de uma distribuição de freqüências agrupadas não são rígidas; o importante é ter bom senso. As únicas exigências são que as classes sejam exaustivas e mutuamente exclusivas. É usual trabalhar com limites inteiros (muitas vezes, múltiplos de 10), para facilitar a leitura da tabela. Além disso, o número de classes, em geral, não deve ser inferior a 5 nem superior a 25.

Uma forma de se determinar um número razoável, k, de classes consiste em aplicar a fórmula de

Tabela 2.12: Produção de ovos nos Estados Unidos em 1990

Estado	Preço/dz (cents)	Estado	Preço/dz (cents)	Estado	Preço/dz (cents)	Estado	Preço/dz (cents)	Estado	Preço/dz (cents)
AK	151,0	HI	85,0	ME	101,0	NJ	85,0	SD	48,0
AL	92,7	IA	56,5	MI	58,0	NM	74,0	TN	71,0
AR	86,3	ID	79,1	MN	57,7	NV	53,9	TX	76,7
AZ	61,0	IL	65,0	MO	55,4	NY	68,1	UT	64,0
CA	63,4	IN	62,7	MS	87,8	OH	59,1	VA	86,3
CO	77,8	KS	54,5	MT	68,0	OK	101,0	VT	106,0
CT	106,0	KY	67,7	NC	82,8	OR	77,0	WA	74,1
DE	117,0	LA	115,0	ND	55,2	PA	61,0	WI	60,1
FL	62,0	MA	105,0	NE	50,3	RI	102,0	WV	104,0
GA	80,6	MD	76,6	NH	109,0	SC	70,1	WY	83,0

Tabela 2.13: Distribuição de freqüências dos preços de ovos - EUA - 1990

Preço dos ovos	Freqüência Simples		Freqüência Acumulada	
(cents/dúzia)	Absoluta	Relativa	Absoluta	Relativa
$[47, 68)$	19	$38,0$	19	$38,0$
$[68, 89)$	19	$38,0$	38	$76,0$
$[89, 110)$	9	$18,0$	47	$94,0$
$[110, 131)$	2	$4,0$	49	$98,0$
$[131, 152)$	1	$2,0$	50	$100,0$
Total	50	$100,0$		

Fonte: Gujarati(1995)

Sturges, que sugere o cálculo de k mediante a expressão:

$$k = 1 + \log_2 N = 1 + \frac{\log n}{\log 2} \tag{2.2}$$

onde n é o número de observações. No entanto, dadas as características da função logaritmo, um dos problemas na utilização dessa fórmula é que ela fornece um número grande de classes para valores pequenos de n e um número pequeno de classes para valores grandes de n, como pode ser observado na Tabela 2.14, onde os resultados foram arredondados para o próximo inteiro.

Tabela 2.14: Número de classes pela fórmula de Sturges

n	k
30	6
35	6
40	6
50	7
100	8
200	9
500	10
1000	11

Assim, a decisão final sobre o número de classes deve se basear na natureza dos dados e da

unidade de medida, com essa ou outra fórmula servindo apenas de referência.

2.3.3 Notação para distribuições univariadas de freqüências

Para generalizarmos o procedimento de construção de uma tabela de freqüências, vamos adotar a seguinte notação, descrita na Tabela 2.15 a seguir.

Tabela 2.15: Construção de uma tabela de freqüências

Nome da variável	Freqüência Simples		Freqüência Acumulada	
	Absoluta (n_i)	Relativa (%) (f_i)	Absoluta (N_i)	Relativa (%) (F_i)
Valor 1	n_1	f_1	N_1	F_1
Valor 2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Valor k	n_k	f_k	N_k	F_k
Total	n	$100,00$		

Cada valor n_i é obtido através da contagem do número de ocorrências do i-ésimo valor ou categoria. Essa é a informação primária, específica do conjunto de dados em análise. A partir dos valores dos n_i, $i = 1, \ldots k$ (k é o número de valores distintos ou classes), obtém-se o número total de observações como

$$n = \sum_{i=1}^{k} n_i \,; \tag{2.3}$$

cada freqüência simples relativa é obtida como

$$f_i = 100 \times \frac{n_i}{n} \,. \tag{2.4}$$

As freqüências absolutas acumuladas são obtidas como

$$N_i = n_1 + n_2 + \cdots + n_{i-1} + n_i \,; \tag{2.5}$$

ou de forma recursiva como

$$N_i = N_{i-1} + n_i \,; \tag{2.6}$$

com

$$N_1 = n_1. \tag{2.7}$$

Com relação às freqüências acumuladas relativas, devemos notar o seguinte:

$$\begin{aligned} F_i &= f_1 + f_2 + \cdots + f_{i-1} + f_i = \\ &= 100 \times \left(\frac{n_1}{n} + \frac{n_2}{n} + \cdots + \frac{n_{i-1}}{n} + \frac{n_i}{n} \right) = \\ &= 100 \times \frac{n_1 + n_2 + \cdots + n_{i-1} + n_i}{n} = \\ &= 100 \times \frac{N_i}{n} \end{aligned}$$

Matematicamente, todas essas expressões são equivalentes mas, quando estamos trabalhando com calculadoras e mesmo com computadores, devemos tomar cuidado com a precisão dos resultados, por causa de arredondamentos. A operação de divisão é uma operação que, em geral, resulta em

números fracionários; assim, sempre que possível, devemos fazer o menor número possível de divisões. Seguindo esse raciocínio, a freqüência acumulada relativa deve ser calculada a partir das freqüências absolutas acumuladas, isto é:

$$F_i = 100 \times \frac{N_i}{n} \tag{2.8}$$

2.3.4 Exercícios resolvidos da Seção 2.3

Considere os dados das Tabelas 2.16 a 2.18, referentes a um levantamento feito por professores da Universidade Federal de Santa Catarina (UFSC), onde o principal objetivo era avaliar os efeitos políticos dos programas de alimentação popular. Aqui temos dados referentes a 120 famílias residentes em três locais: Conjunto Residencial Monte Verde, Conjunto Residencial Parque da Figueira e na Encosta do Morro.[5] As variáveis apresentadas são:

- **PAP**: variável indicadora de uso de programas de alimentação (1 = Sim; 0 = Não);
- **GI**: grau de instrução do chefe da casa (1 = nenhum grau oficialmente completo; 2 = primeiro grau completo; 3 = segundo grau completo);
- **RES**: número de pessoas residentes na casa;
- **RENDA**: renda familiar mensal, em salários mínimos.

Tabela 2.16: Conjunto residencial Monte Verde

Ident.	PAP	GI	RES	RENDA	Ident.	PAP	GI	RES	RENDA
1	0	3	4	10,3	21	1	3	5	5,8
2	0	3	4	15,4	22	1	3	5	12,9
3	1	2	4	9,6	23	0	3	5	7,7
4	0	2	5	5,5	24	0	2	4	1,1
5	1	3	4	9,0	25	0	2	8	7,5
6	1	1	1	2,4	26	1	3	4	5,8
7	0	3	2	4,1	27	1	1	5	7,2
8	1	3	3	8,4	28	0	3	3	8,6
9	1	3	6	10,3	29	1	2	4	5,1
10	1	2	4	4,6	30	0	3	5	2,6
11	0	2	6	18,6	31	1	3	5	7,7
12	1	1	4	7,1	32	1	2	2	2,4
13	0	2	4	12,9	33	1	3	5	4,8
14	0	2	6	8,4	34	1	1	2	2,1
15	0	3	3	19,3	35	1	1	6	4,0
16	0	2	5	10,4	36	1	1	8	12,5
17	1	3	3	8,9	37	1	3	3	6,8
18	0	3	4	12,9	38	1	3	5	3,9
19	0	3	4	5,1	39	1	3	5	9,0
20	1	3	4	12,2	40	1	3	3	10,9

[5] Dados extraídos de Barbetta (1994)

Tabela 2.17: Conjunto residencial Parque da Figueira

Ident.	PAP	GI	RES	RENDA	Ident.	PAP	GI	RES	RENDA
41	1	2	5	5,4	63	1	1	3	5,5
42	1	1	3	6,4	64	1	1	7	3,5
43	1	1	6	4,4	65	1	3	3	9,0
44	1	1	5	2,5	66	1	3	6	5,8
45	0	1	6	5,5	67	0	1	6	4,2
46	1	1	8	.	68	1	3	3	6,8
47	1	3	4	14,0	69	1	2	5	4,8
48	1	2	4	8,5	70	1	3	5	6,0
49	1	1	5	7,7	71	1	2	7	9,0
50	0	2	3	5,8	72	1	1	4	5,3
51	1	3	5	5,0	73	1	3	4	3,1
52	0	1	3	4,8	74	0	3	1	6,4
53	1	2	2	2,8	75	1	1	3	3,9
54	1	2	4	4,2	76	1	2	3	6,4
55	1	3	3	10,2	77	1	3	4	2,7
56	1	2	4	7,4	78	0	2	4	2,4
57	1	2	5	5,0	79	0	2	4	3,6
58	0	3	2	6,4	80	0	3	5	6,4
59	0	3	4	5,7	81	0	3	2	11,3
60	1	2	4	10,8	82	1	1	5	3,8
61	0	3	1	2,3	83	1	2	3	4,1
62	1	1	7	6,1					

Tabela 2.18: Encosta do Morro

Ident.	PAP	GI	RES	RENDA	Ident.	PAP	GI	RES	RENDA
84	1	1	5	1,8	103	0	1	6	2,3
85	1	3	5	7,1	104	1	2	5	4,9
86	0	1	3	13,9	105	1	1	5	2,3
87	1	2	6	4,0	106	1	1	3	3,9
88	1	1	6	2,9	107	1	1	4	2,1
89	1	2	9	3,9	108	1	1	4	2,7
90	1	1	4	2,2	109	1	2	5	11,1
91	0	2	3	5,8	110	1	1	6	6,4
92	0	2	5	2,8	111	0	3	7	25,7
93	1	2	5	4,5	112	1	1	4	0,9
94	0	2	4	5,8	113	1	3	5	3,9
95	0	3	8	3,9	114	1	1	5	5,1
96	0	2	7	2,8	115	1	2	6	4,2
97	1	1	3	1,3	116	1	1	6	4,4
98	1	3	5	3,9	117	1	1	7	7,9
99	1	3	5	5,0	118	0	1	4	4,2
100	1	1	5	0,1	119	0	1	4	3,5
101	0	2	3	4,6	120	0	2	6	11,4
102	1	2	4	2,6					

1. Classifique as variáveis da pesquisa de acordo com o seu tipo.

   Solução:

   LOCAL: variável qualitativa

   PAP: variável qualitativa

   GI: variável qualitativa ordinal

   RES: variável quantitativa discreta

   RENDA: variável quantitativa contínua.

2. Para as variáveis qualitativas e quantitativa discreta, construa tabelas de freqüencia sem perda de informação, considerando as três localidades em conjunto.

   Solução:

   Das definições das variáveis dadas no enunciado do exercício, sabemos que LOCAL pode assumir os valores 1, 2, e 3, que representam as localidades do Conj. Res. Monte Verde, Conj. Res. Parque da Figueira e da Encosta do Morro. A variável PAP pode assumir os valores 1 e 0, indicando que a família tem ou não acesso a programas alimentares. Analogamente, a variável GI pode assumir os valores 1, 2, 3. Note que essas são codificações para as variáveis qualitativas. Provavelmente, no questionário a pergunta era feita de modo que o entrevistador assinalava com um X o quadrinho correspondente à resposta dada pelo informante. A codificação é feita para facilitar o processamento das informações pelo computador. Analisando os dados, podemos ver que o valor mínimo para RES é 1 e o valor máximo é 9. Com essas informações, constroem-se as Tabelas 2.19 a 2.22 abaixo. Note que aí as freqüências relativas não estão multiplicadas por 100 e, portanto, somam 1. No caso de se apresentarem essas freqüências em forma percentual, é comum colocar o título da coluna como Relativa (%).

Tabela 2.19: Distribuição das famílias por local de residência

Local	Freqüência simples	
	Absoluta	Relativa
Monte Verde	40	0,3333
Parque da Figueira	43	0,3583
Encosta do Morro	37	0,3083
Total	120	1,0000

Tabela 2.20: Distribuição do número de famílias com relação ao uso de programas de alimentação

Uso de programa de alimentação	Freqüência simples	
	Absoluta	Relativa
Sim	78	0,65
Não	42	0,35
Total	120	1,00

3. Para a variável RENDA, construa uma tabela de freqüências trabalhando com 4 classes de mesmo tamanho.

   Solução:

   Uma primeira observação diz respeito à família identificada pelo número 46: para essa família, não há informação disponível sobre a renda. Vamos, então, trabalhar com as 119 famílias

Tabela 2.21: Distribuição do grau de instrução do chefe de família

Grau de Instrução	Freqüência simples	
	Absoluta	Relativa
Nenhum completo	38	0,3167
1º grau completo	38	0,3167
2º grau completo	44	0,3667
Total	120	1,0000

Tabela 2.22: Distribuição do número de moradores

Número de residentes	Freqüência simples		Freqüência acumulada	
	Absoluta	Relativa	Absoluta	Relativa
1	3	0,0250	3	0,0250
2	6	0,0500	9	0,0750
3	21	0,1750	30	0,2500
4	32	0,2667	62	0,5167
5	32	0,2667	94	0,7833
6	15	0,1250	109	0,9083
7	6	0,0500	115	0,9583
8	4	0,0333	119	0,9917
9	1	0,0083	120	1,0000

restantes. O valor mínimo é 0,1 e o valor máximo é 25,7, o que resulta em uma amplitude exata de 25,6. Como os dados estão em forma decimal, não há necessidade de trabalharmos com limites de classe inteiros; assim, vamos arredondar a amplitude para 26 e trabalhar com comprimento de classe igual a $\frac{26}{4} = 6,5$. Como o menor valor é 0,1, vamos definir como 0 o limite inferior da primeira classe. Para construir a tabela à mão, é interessante fazer um esquema de contagem para as diferentes classes, de modo a não precisarmos ordenar os dados. Uma possibilidade é ir marcando com um tracinho cada ocorrência nas diferentes classes, à medida que vamos varrendo os dados:

$0,0 \vdash 6,5$	||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
	||||| ||||| ||||| ||||| ||||| ||||| ||||| |||
$6,5 \vdash 13,0$	||||| ||||| ||||| ||||| |||||
$13,0 \vdash 19,5$	|||||
$19,5 \vdash 26,0$ |	

Resulta a Tabela 2.23.

Tabela 2.23: Distribuição da renda

Renda (salários mínimos)			Freqüência Simples		Freqüência Acumulada	
			Absoluta	Relativa	Absoluta	Relativa
0,0	$\vdash$	6,5	78	0,6555	78	0,6555
6,5	$\vdash$	13,0	35	0,2941	113	0,9496
13,0	$\vdash$	19,5	5	0,0420	118	0,9916
19,5	$\vdash$	26,0	1	0,0084	119	1,0000
Total			119	1,0000		

Obs.: Esse exercício mostra a dificuldade de se construir tabelas à mão! É importante aprender a usar recursos computacionais.

4. Analisando a tabela da distribuição de renda, pode-se ver que há uma grande concentração nas duas classes iniciais. Trabalhar com classes de mesmo tamanho não é recomendável nesse caso, pois, como sabemos, no Brasil há um grande número de famílias de baixa renda. Vamos definir as seguintes classes: [0,2), [2,3), [3,4), [4,5), [5,6), [6,8), [8,10), [10,15), ≥ 15. Com essas classes obtemos a Tabela 2.24, onde fica mais detalhada a distribuição das classes de renda mais baixas:

Tabela 2.24: Distribuição de renda - Classes desiguais

Renda (sal. mín.	Freqüência Simples		Freqüência Acumulada	
	Absoluta	Relativa	Absoluta	Relativa
0 ⊢ 2	5	0,04202	5	0,04202
2 ⊢ 3	18	0,15126	23	0,19328
3 ⊢ 4	12	0,10084	35	0,29412
4 ⊢ 5	17	0,14286	52	0,43697
5 ⊢ 6	18	0,15126	70	0,58824
6 ⊢ 8	19	0,15966	89	0,74790
8 ⊢ 10	10	0,08403	99	0,83193
10 ⊢ 15	16	0,13445	115	0,96639
15 ⊢ 26	4	0,03361	119	1,00000
Total	119	1,00000		

2.3.5 Exercícios propostos da Seção 2.3

2.3 *Na Tabela 2.25 temos o número de questões acertadas por 50 alunos em um teste de múltipla escolha com 10 questões. Construa uma tabela de freqüências para representar esses dados, de modo que não haja perda de informação.*

Tabela 2.25: Notas de 50 alunos em um teste múltipla escolha para o Exercício 2.3

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

Fonte: Dados hipotéticos

2.4 *Na Tabela 2.26 temos dados sobre a produção de ovos nos 50 estados dos Estados Unidos no ano de 1990. Construa uma tabela de freqüências para a variável Quantidade Produzida de Ovos utilizando 5 classes de mesmo tamanho.*

2.5 *Estudando-se o consumo diário de leite, verificou-se que em certa localidade, 20% das famílias consomem até 1 litro, 50% consomem entre 1 e 2 litros, 20% entre 2 e 3 litros e o restante entre 3 e 5 litros. Para a variável em estudo, escreva as informações dadas em forma de tabela.*

Tabela 2.26: Produção de ovos nos Estados Unidos em 1990 para o Exercício 2.4

Estado	Quant. (milhões)	Estado	Quant. (milhões)	Estado	Quant. (milhões)
AK	0,7	MA	235,0	OR	652,0
AL	2206,0	MD	885,0	PA	4976,0
AR	3620,0	ME	1069,0	RI	53,0
AZ	73,0	MI	1406,0	SC	1422,0
CA	7472,0	MN	2499,0	SD	435,0
CO	788,0	MO	1580,0	TN	277,0
CT	1029,0	MS	1434,0	TX	3317,0
DE	168,0	MT	172,0	UT	456,0
FL	2586,0	NC	3033,0	VA	943,0
GA	4302,0	ND	51,0	VT	31,0
HI	227,5	NE	1202,0	WA	1287,0
IA	2151,0	NH	43,0	WI	910,0
ID	187,0	NJ	442,0	WV	136,0
IL	793,0	NM	283,0	WY	1,7
IN	5445,0	NV	2,2		
KS	404,0	NY	975,0		
KY	412,0	OH	4667,0		
LA	273,0	OK	869,0		

Fonte: Gujarati (1995) - Tabela 1.1

2.6 *Em um levantamento feito pela revista Exame-Maiores e Melhores 1998 para as 100 maiores empresas brasileiras, em termos de vendas, nem todas informaram o número de empregados*[6]*. Na Tabela 2.27 abaixo temos os dados obtidos, ordenados pelo volume de vendas das empresas e na Tabela 2.28 temos os dados ordenados pelo número de empregados. Identifique a variável de estudo e construa uma tabela de frequência, utilizando 5 classes de mesmo tamanho.*

2.7 *Na Tabela 2.29 tem-se as médias dos alunos de 2 turmas de Introdução à Estatística Econômica da Faculdade de Economia da UFF no primeiro semestre de 2000. Segundo o critério de aprovação da UFF, o aluno que obtiver média inferior a 4 estará reprovado. O aluno que obtiver média maior ou igual a 4 mas menor que 6 terá direito à Verificação Suplementar (VS) e os alunos com média maior ou igual a 6 estarão aprovados. A partir desses dados, construa uma tabela de freqüências que ilustre o número de alunos reprovados, com direito à VS e aprovados.*

[6] Dados extraídos de Lopes (1999).

Tabela 2.27: Número de empregados das 100 maiores empresas para o Exercício 2.6 - Dados originais

Ordem	Número de Empregados	Ordem	Número de Empregados	Ordem	Número de Empregados	Ordem	Número de Empregados
1	30775	26	14020	48	4700	71	3616
2	21411	27	987	50	10465	72	3500
3	24045	29	2666	51	2147	73	6084
4	1763	30	5588	52	4500	78	5543
5	7840	31	6700	53	2141	79	3581
7	1932	32	5132	54	7092	80	9564
8	13038	33	7926	55	5254	83	4621
9	5242	34	2788	57	9443	86	3073
10	12097	35	11439	58	3622	88	590
11	9378	36	18093	59	2356	90	6468
12	1303	38	8237	60	1082	91	1754
13	1047	39	950	61	1020	92	6025
15	17812	40	8177	62	746	93	2616
16	10865	41	3996	64	3354	94	2237
17	198	42	11484	65	4973	95	3014
18	11360	43	2415	66	4859	96	154
19	10995	44	4208	67	3326	97	4019
22	11522	45	5817	68	1688	98	5113
24	19896	46	7820	69	5840	99	4087
25	8949	47	11028	70	383	100	1873

Tabela 2.28: Número de empregados das 100 maiores empresas para o Exercício 2.6 - Dados ordenados

154	198	383	590	746	950	987	1020	1047	1082	1303	1688
1754	1763	1873	1932	2141	2147	2237	2356	2415	2616	2666	2788
3014	3073	3326	3354	3500	3581	3616	3622	3996	4019	4087	4208
4500	4621	4700	4859	4973	5113	5132	5242	5254	5543	5588	5817
5840	6025	6084	6468	6700	7092	7820	7840	7926	8177	8237	8949
9378	9443	9564	10465	10865	10995	11028	11360	11439	11484	11522	12097
13038	14020	17812	18093	19896	21411	24045	30775				

Tabela 2.29: Médias dos alunos de Int.Est.Econômica (1/2000-UFF) para o Exercício 2.7

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

2.4 Distribuição univariada de freqüências: Representação gráfica

2.4.1 Gráfico de setores

Este gráfico é usado quando cada valor representa uma parte de um todo. É, então, usado um círculo de raio qualquer, com a área ou ângulo total sendo proporcional ao total (100%) da série de dados a representar e a área ou ângulo de cada setor circular sendo proporcional a cada dado da série.

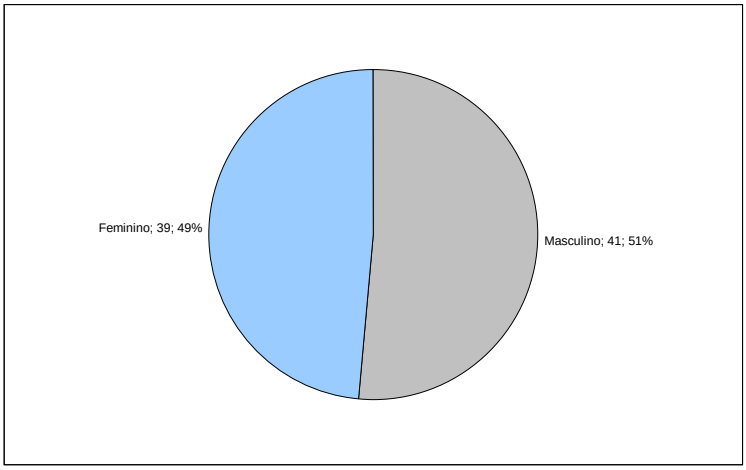
Vamos ilustrar a construção deste tipo de gráfico com os dados da Tabela 2.2 referentes à variável sexo. De 80 alunos, 41 são do sexo masculino e 39 do sexo feminino. Como os ângulos dos setores são diretamente proporcionais às respectivas freqüências, temos a seguinte regra de três:

$$\frac{80}{360^o} = \frac{41}{x^o} \Rightarrow x = 184,5^o$$

$$\frac{80}{360^o} = \frac{39}{x^o} \Rightarrow x = 175,5^o$$

Na Figura 2.4 temos o gráfico resultante, construído com o programa de planilhas Excel.

Figura 2.4: Distribuição dos alunos por sexo



De forma análoga obtemos o gráfico para a variável matéria predileta no segundo grau, dado na Figura 2.5. Note que esses gráficos podem ser construídos com base nas freqüências absolutas ou relativas.

2.4.2 Gráfico de colunas

No caso de variáveis qualitativas, outra representação gráfica apropriada se faz através do gráfico de colunas; nesse gráfico, as categorias são colocadas sobre um eixo horizontal e as freqüências simples, absolutas ou relativas, são indicadas através de colunas cujas alturas representam essas freqüências. Os mesmos dados sobre sexo e matéria predileta no segundo grau podem ser representados pelos gráficos dados nas Figuras 2.6 e 2.7.

Note que nesse tipo de gráfico não há uma escala no eixo horizontal, uma vez que aí temos representadas as *categorias* da variável em estudo. Se um gráfico de colunas é usado para representar uma variável quantitativa discreta, há que se tomar cuidado pois, nesse caso, existe uma escala, que

Figura 2.5: Distribuição dos alunos por matéria predileta no segundo grau

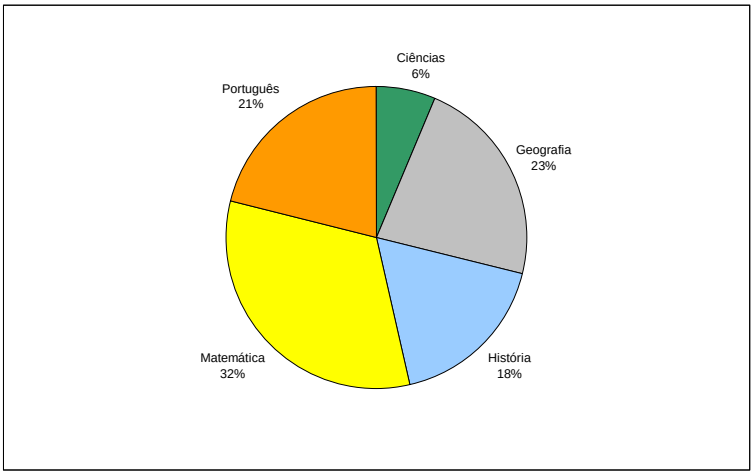


Figura 2.6: Distribuição dos alunos por sexo

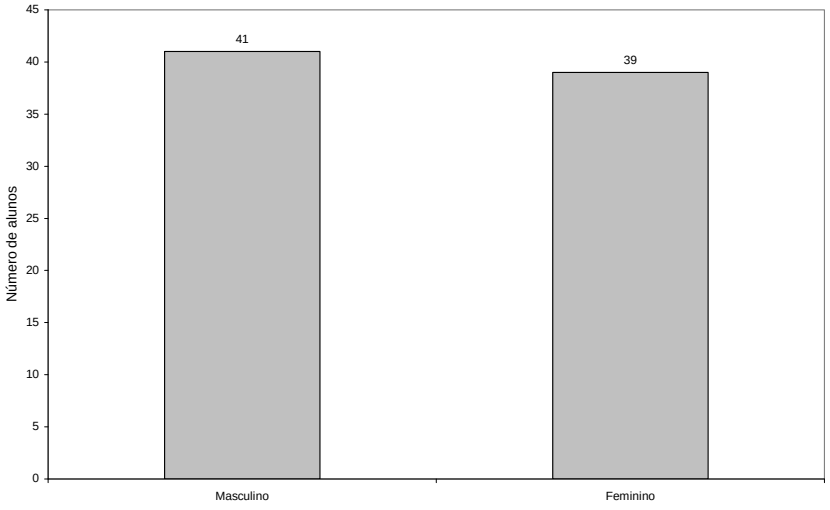
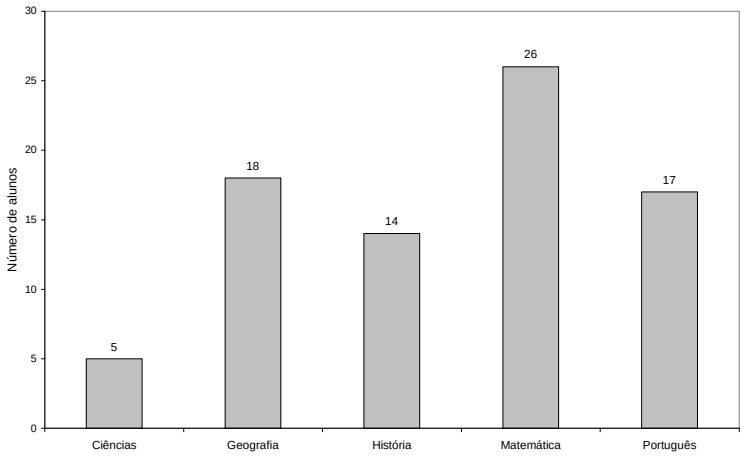
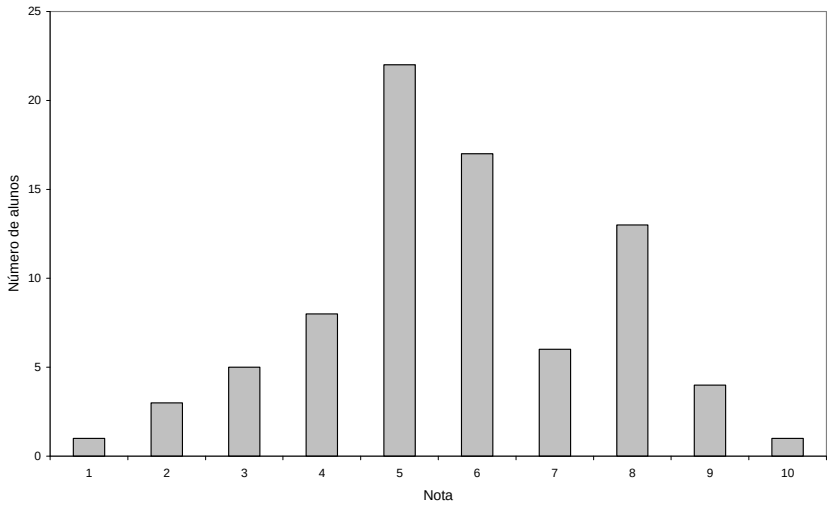


Figura 2.7: Distribuição dos alunos por matéria predileta no segundo grau



deve ser bem representada. No EXCEL, a opção de gráfico de colunas considera a variável como uma variável qualitativa. Na Figura 2.8 temos o gráfico que representa a distribuição das notas dos 80 alunos.

Figura 2.8: Distribuição das notas de 80 alunos



2.4.3 Histograma e polígono de freqüências

A apresentação tabular dos dados através de uma distribuição de freqüências fica complementada com uma representação gráfica desses mesmos dados. O histograma e o polígono de freqüências são tipos de gráficos usados para representar uma distribuição de freqüências simples de uma variável quantitativa contínua.

Um histograma é um conjunto de retângulos com bases sobre um eixo horizontal dividido de acordo com os comprimentos de classes, centros nos pontos médios das classes e *áreas proporcionais ou iguais às freqüências*. Um polígono de freqüências é um gráfico de linha que se obtém unindo por uma poligonal os pontos correspondentes às freqüências das diversas classes, centradas nos respectivos pontos médios. Para obter as interseções da poligonal com o eixo, cria-se em cada extremo uma classe com freqüência nula. Note que esses gráficos podem ser construídos com base nas freqüências absolutas ou relativas. O importante é que a escala nos eixos horizontal e vertical, bem como os retângulos, sejam construídos de forma a que suas áreas espelhem a proporcionalidade dessas freqüências.

Na Figura 2.9 apresentamos o histograma para a distribuição de freqüências dada na Tabela 2.13, referente ao preço da dúzia de ovos nos estados americanos em 1990. Aqui cabe uma observação sobre o histograma, que foi construído com o programa XLSTAT: cada retângulo foi construído de modo que sua *área* fosse exatamente igual à *freqüência relativa*. Por exemplo, todos os retângulos têm base 21, que é a amplitude de classe. A altura dos dois primeiros retângulos é área/base $= 0,38\,/\,21 = 0,0180952$, de modo que a área resultante é 0,38. Para a terceira classe, temos que altura = área/base $= 0,18\,/\,21 = 0,0085714$. Voltaremos a discutir esse assunto quando da apresentação dos histogramas com classes desiguais. O polígono de freqüência está na Figura 2.10.

O ponto fundamental na interpretação de um histograma é compreender que as áreas dos retângulos representam as freqüências de cada classe. Como a variável é contínua e a freqüência dada se refere a uma classe de valores, a suposição que se faz é que essa freqüência se distribui uniformemente pela classe. Na Figura 2.9, a freqüência relativa da classe $[47;68)$ é 0,38 (ou 38%) e ela está uniformemente distribuída pela classe, o que significa que sub-classes de mesmo comprimento teriam a mesma freqüência. Por exemplo, as freqüências das classes $[47;57,5)$ e $[57,5;68)$ seriam ambas iguais 0,19. Já a sub-classe [89;95) teria uma freqüência de $0,0085714 \times (95-89) = 0,0514286$. Mais uma vez, o princípio é que área = freqüência. Com relação ao polígono de freqüências, a idéia é representar o comportamento "típico" de cada classe através do seu ponto médio.

Figura 2.9: Histograma da distribuição de freqüência dos preços dos ovos nos estados americanos - Fonte: Tabela 2.13

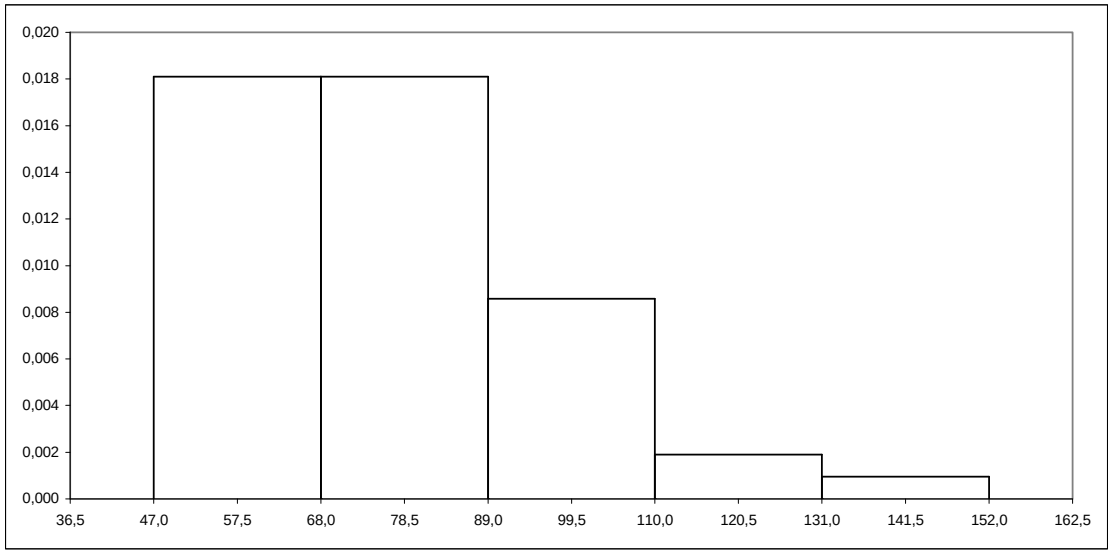
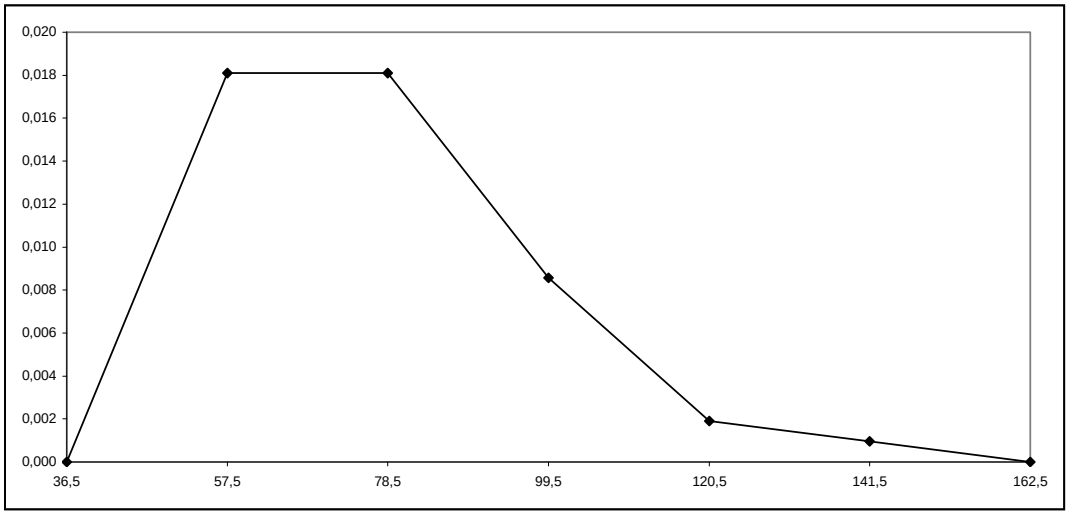


Figura 2.10: Polígono de freqüência dos preços dos ovos nos estados americanos - Fonte: Tabela 2.13



2.4.4 Gráfico das distribuições de freqüências acumuladas

As freqüências acumuladas também podem ser representadas graficamente através do gráfico da função acumulada das freqüências absolutas, $N(x)$, que é definida para todo $x \in (-\infty, +\infty)$ da seguinte forma: para cada valor $x \in \mathbb{R}$, $N(x)$ é definida como o número ou freqüência absoluta das observações para as quais a variável X em estudo é menor ou igual a x.

Se a variável é discreta assumindo os valores $a_1 < a_2 < a_3 < \ldots$ com freqüências absolutas simples iguais a $n_1, n_2, n_3, \ldots$ respectivamente, então podemos calcular $N(x)$ em termos dos a_i's observando o seguinte:

- se $x < a_1$ então $N(x) = 0$ pois não há nenhuma observação com valor menor que a_1;
- se $a_1 \leq x < a_2$, então $N(x) = n_1$, uma vez que as únicas observações menores ou iguais a x são aquelas para as quais a variável é igual a a_1 e sabemos que há n_1 delas;
- se $a_2 \leq x < a_3$, então $N(x) = n_1 + n_2 = N_2$, uma vez que as únicas observações menores ou iguais a x são aquelas para as quais a variável é igual a a_1 ou a a_2 e sabemos que há n_1 delas iguais a a_1 e n_2 iguais a a_2;
- se $a_3 \leq x < a_4$, então $N(x) = n_1 + n_2 + n_3 = N_3$, uma vez que as únicas observações menores ou iguais a x são aquelas para as quais a variável é igual a a_1 ou a a_2 ou a a_3 e sabemos que há n_1 delas iguais a a_1, n_2 iguais a a_2 e n_3 iguais a a_3.
- Em geral, $N(x) = n_1 + n_2 + \cdots + n_{i-1} = N_{i-1}$ para $a_{i-1} \leq x < a_i$.

Note que $N(x)$ é uma função não-decrescente e cada diferença $N(a_i) - N(a_{i-1}) = n_i$. De maneira análoga, pode-se definir a função acumulada das freqüências relativas $F(x)$, trabalhando-se com as freqüências relativas. Mais precisamente, $F(x)$ é definida para todo $x \in (-\infty, +\infty)$ como a freqüência relativa das observações para as quais a variável X em estudo é menor ou igual a x.

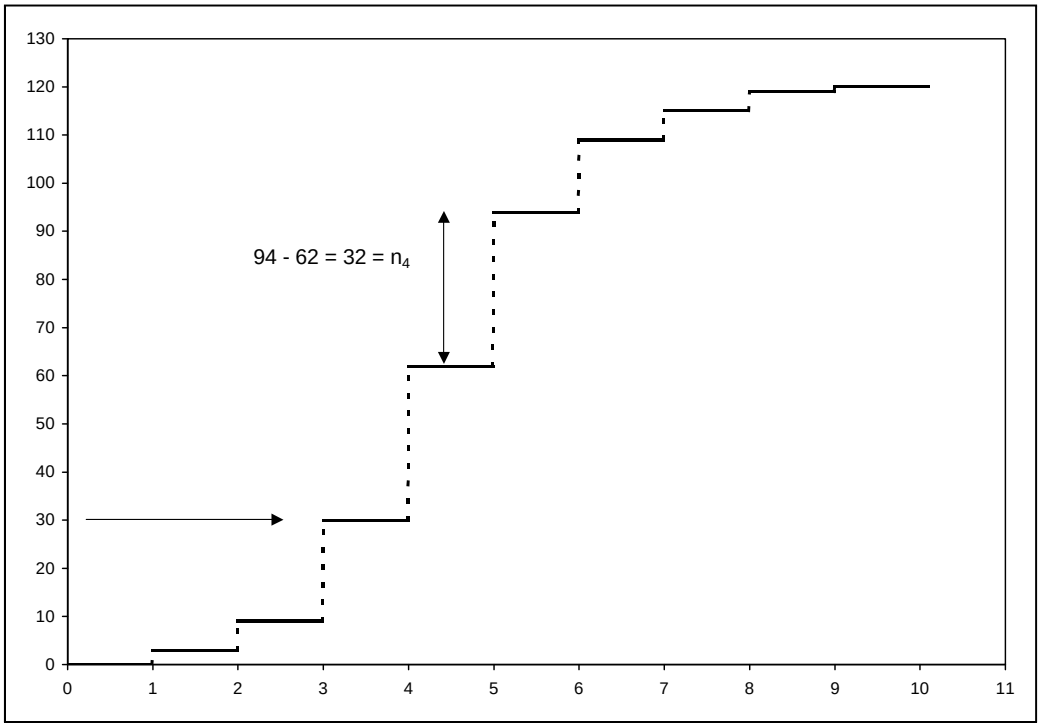
A título de ilustração, consideremos a variável RES, número de residentes por domicílio, da Tabela 2.22, que assume os valores $1, 2, 3, 4, 5, 6, 7, 8, 9$ com freqüências $3, 6, 21, 32, 32, 15, 6, 4, 1$.

Seguindo o raciocínio acima, podemos ver que as funções $N(x)$ e $F(x)$ são definidas como

$$N(x) = \begin{cases} 0 & \text{se} \quad x < 1 \\ 3 & \text{se} \quad 1 \leq x < 2 \\ 9 & \text{se} \quad 2 \leq x < 3 \\ 30 & \text{se} \quad 3 \leq x < 4 \\ 62 & \text{se} \quad 4 \leq x < 5 \\ 94 & \text{se} \quad 5 \leq x < 6 \\ 109 & \text{se} \quad 6 \leq x < 7 \\ 115 & \text{se} \quad 7 \leq x < 8 \\ 119 & \text{se} \quad 8 \leq x < 9 \\ 120 & \text{se} \quad x \geq 9 \end{cases} \qquad F(x) = \begin{cases} 0,0000 & \text{se} \quad x < 1 \\ 0,0250 & \text{se} \quad 1 \leq x < 2 \\ 0,0750 & \text{se} \quad 2 \leq x < 3 \\ 0,2500 & \text{se} \quad 3 \leq x < 4 \\ 0,5167 & \text{se} \quad 4 \leq x < 5 \\ 0,7833 & \text{se} \quad 5 \leq x < 6 \\ 0,9083 & \text{se} \quad 6 \leq x < 7 \\ 0,9583 & \text{se} \quad 7 \leq x < 8 \\ 0,9917 & \text{se} \quad 8 \leq x < 9 \\ 1,0000 & \text{se} \quad x \geq 9 \end{cases}$$

Na Figura 2.11 temos o gráfico da função acumulada das freqüências absolutas. Esse gráfico ilustra a característica discreta da variável. Cada "degrau" ou segmento de reta horizontal tem uma bola fechada na extremidade esquerda para indicar que estamos trabalhando com intervalos do tipo $\leq$. A altura de cada degrau dá a freqüência simples de cada classe, conforme ilustrado.

Figura 2.11: Função de distribuição acumulada para o número de moradores por domicílio



A análise desse gráfico nos leva a estabelecer as seguintes características da função acumulada das freqüências relativas:

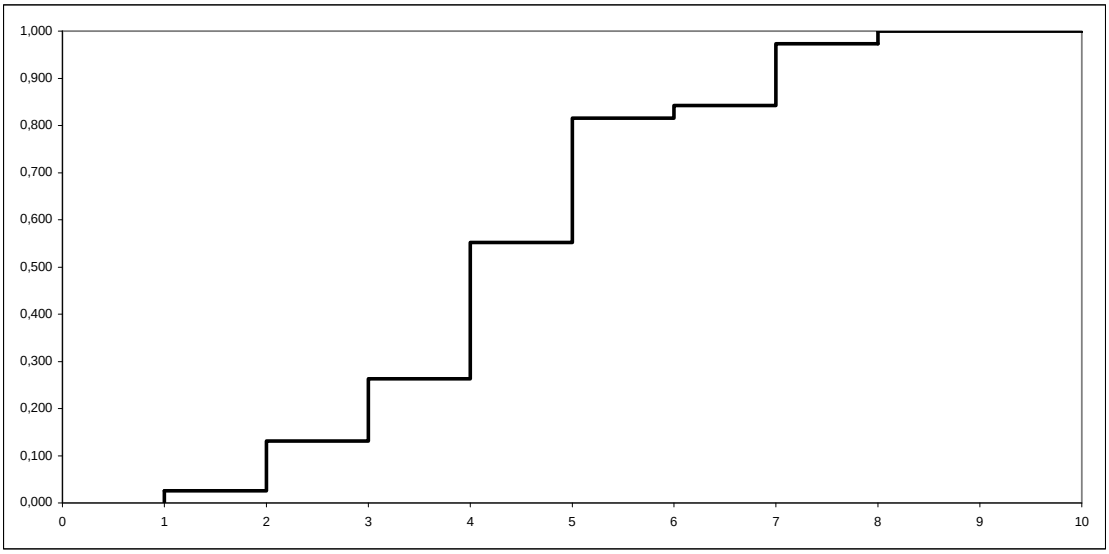
- $\lim_{x \to -\infty} = 0$
- $\lim_{x \to +\infty} = 1$
- $F(x)$ é uma função não-decrescente

- $F(x)$ é uma função contínua à direita

Vale a pena observar que alguns autores definem $N(x)$ ou $F(x)$ como a freqüência absoluta ou relativa das observações menores que x (e não menores ou iguais a x); nesse caso, as funções são contínuas à esquerda (isto é, no gráfico cada segmento teria uma "bola" no extremo superior direito).

Na Figura 2.12 temos o gráfico da função acumulada das freqüências relativas construído pelo programa XLSTAT: note que a "escada" é apresentada em uma linha sólida. Esse procedimento é usual, ficando subentendida a característica discreta (ou "saltos") da função.

Figura 2.12: Gráfico da função acumulada das freqüências relativas do número de moradores por domicílio

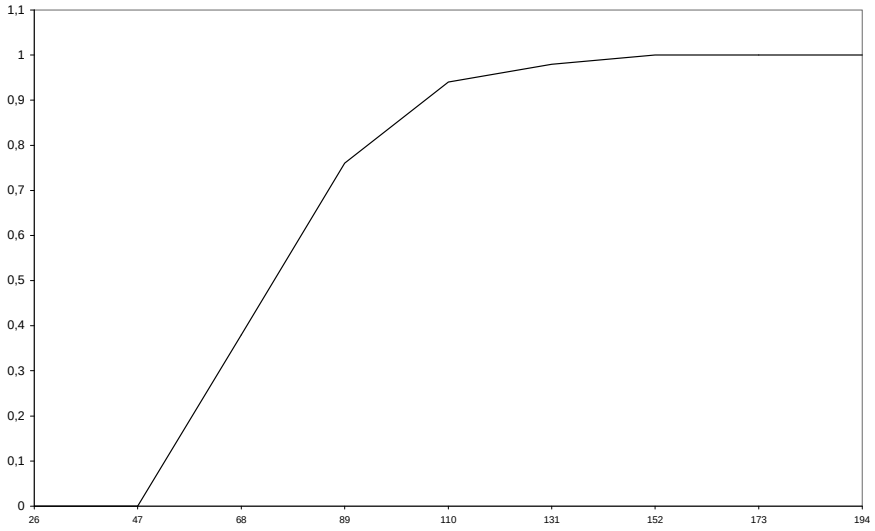


Quando a variável representada na tabela de freqüências é contínua, a diferença fundamental está na interpretação das freqüências: cada freqüência n_i ou f_i se refere a uma classe de valores e supõe-se que essa freqüência se distribua uniformemente ao longo da classe (e não em apenas um ponto, como ocorre com variáveis discretas). A função acumulada das freqüências continua sendo definida como a freqüência das observações menores ou iguais a x. Nesse caso, o gráfico da função acumulada de freqüências é, em geral, chamado *ogiva de freqüências*. Esse gráfico é sempre uma poligonal não-descendente, pela própria definição de freqüência acumulada. Para os extremos das classes, as funções $N(x)$ e $F(x)$ são iguais às freqüências acumuladas absolutas ou relativas das respectivas classes. A questão a resolver é como "ligar" esses pontos. Para responder essa questão, vamos recorrer ao histograma da Figura 2.9, lembrando que área = freqüência. Para qualquer ponto x na primeira classe, $F(x)$ é a área de um retângulo de base $x - 47$ e altura 0,01809524, ou seja, $F(x) = (x - 47) \times 0,01809524$ e essa é a equação de uma reta que passa pelos pontos $(47; 0)$ e $(68; 0,38)$, uma vez que não há observações menores que 47 e 38% das observações são menores que 68. Para qualquer ponto na segunda classe, $F(x)$ é igual à área do retângulo correspondente à primeira classe (ou seja, a freqüência da primeira classe) mais a área de um retângulo com base $(x - 68)$ e altura 0,01809524, ou seja, $F(x) = 0,38 + (x - 68) \times 0,01809524$ e essa é a equação de uma reta que passa pelos pontos $(68; 0,38)$ e $(89; 0,76)$. Analogamente, para qualquer ponto na terceira classe, $F(x)$ é igual à área dos dois primeiros retângulos mais a área de um retângulo com

base $x - 89$ e altura 0,00857143, ou seja, $F(x) = 0,76 + (x - 89) \times 0,00857143$ e essa é a equação de uma reta que passa pelos pontos $(89; 0,76)$ e $(110; 0,94)$.

Generalizando esse raciocínio, vemos que a ogiva de freqüências é formada por segmentos de reta que ligam os pontos no plano cujas abscissas são os extremos superiores das classes e cujas ordenadas são as freqüencias acumuladas das respectivas classes. Assim como no caso discreto, $N(x)$ ou $F(x)$ é igual a 0 para qualquer x menor que o valor mínimo e é igual a n (número total de observações) ou 1 para qualquer valor maior que o valor máximo dos dados. Na Figura 2.13 temos a ogiva das freqüências relativas para o preço dos ovos nos estados americanos.

Figura 2.13: Distribuição de freqüência acumulada dos preços dos ovos nos estados americanos - Fonte: Tabela 2.13



2.4.5 Gráfico de Linhas

O gráfico de linhas é usado principalmente para representar observações feitas ao longo do tempo, isto é, observações de uma série de tempo. No eixo horizontal colocam-se as datas em que foram realizadas as observações e no eixo vertical, os valores observados. Os pontos assim obtidos são unidos por segmentos de reta para facilitar a visualização do comportamento dos dados ao longo do tempo.

Na Tabela 2.30 são apresentados os resultados referentes à taxa de desemprego aberto total (semana), produzidos pela Pesquisa Mensal de Emprego e na Figura 2.14 temos o gráfico desta série temporal.

2.4.6 Histograma com classes desiguais

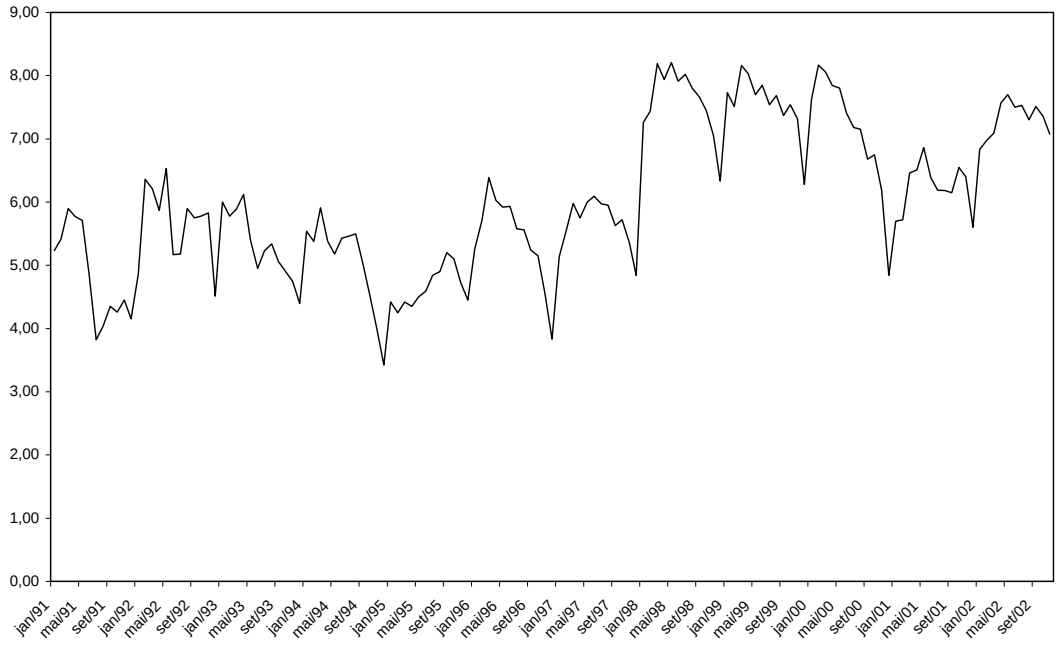
Embora não seja muito usual, é possível construir um histograma quando as classes têm tamanhos diferentes. Mas para que a representação seja correta, as áreas dos retângulos têm que ser proporcionais às freqüências das classes. No caso de classes iguais, como as bases dos retângulos são as mesmas, a diferenciação das áreas se faz simplesmente através das alturas mas esse não é o caso quando as classes são desiguais. Para a construção do histograma, serão acrescentadas

Tabela 2.30: Taxa de de desemprego aberto - semana - Total das áreas - PME

	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Jan	5,23	4,86	6,00	5,54	4,42	5,26	5,14	7,26	7,73	7,63	5,70	6,83
Fev	5,41	6,36	5,78	5,38	4,25	5,71	5,55	7,43	7,51	8,17	5,72	6,98
mar	5,90	6,21	5,89	5,91	4,42	6,39	5,98	8,19	8,16	8,06	6,46	7,09
Abr	5,77	5,87	6,12	5,38	4,35	6,03	5,75	7,94	8,03	7,84	6,51	7,57
Mai	5,71	6,53	5,40	5,18	4,50	5,92	6,00	8,21	7,70	7,80	6,86	7,70
Jun	4,87	5,17	4,95	5,43	4,59	5,93	6,09	7,91	7,85	7,41	6,38	7,50
Jul	3,82	5,18	5,23	5,46	4,84	5,58	5,97	8,02	7,54	7,18	6,19	7,53
Ago	4,04	5,90	5,34	5,50	4,90	5,56	5,95	7,80	7,68	7,15	6,18	7,30
Set	4,35	5,75	5,06	5,05	5,20	5,24	5,63	7,66	7,37	6,68	6,15	7,51
Out	4,27	5,78	4,90	4,53	5,10	5,15	5,72	7,45	7,54	6,75	6,55	7,36
Nov	4,45	5,83	4,75	4,01	4,73	4,56	5,36	7,05	7,32	6,19	6,40	7,07
Dez	4,15	4,51	4,40	3,42	4,45	3,83	4,84	6,33	6,28	4,84	5,60	

Fonte: IBGE - Pesquisa Mensal de Emprego

Figura 2.14: Taxa de desemprego aberto - semana - Total das áreas da PME



à tabela de freqüências duas colunas: a primeira dá o comprimento de cada classe; a segunda, chamada densidade, é obtida dividindo-se as freqüências simples (absoluta ou relativa) das classes pelos respectivos comprimentos. Então, essa coluna nos dá a concentração em cada classe por unidade da variável. É um conceito análogo ao conceito de densidade populacional, que mede a concentração da população por unidade de área. Em termos geométricos, a concentração nada mais é que a altura do retângulo que representa a freqüência de cada classe.

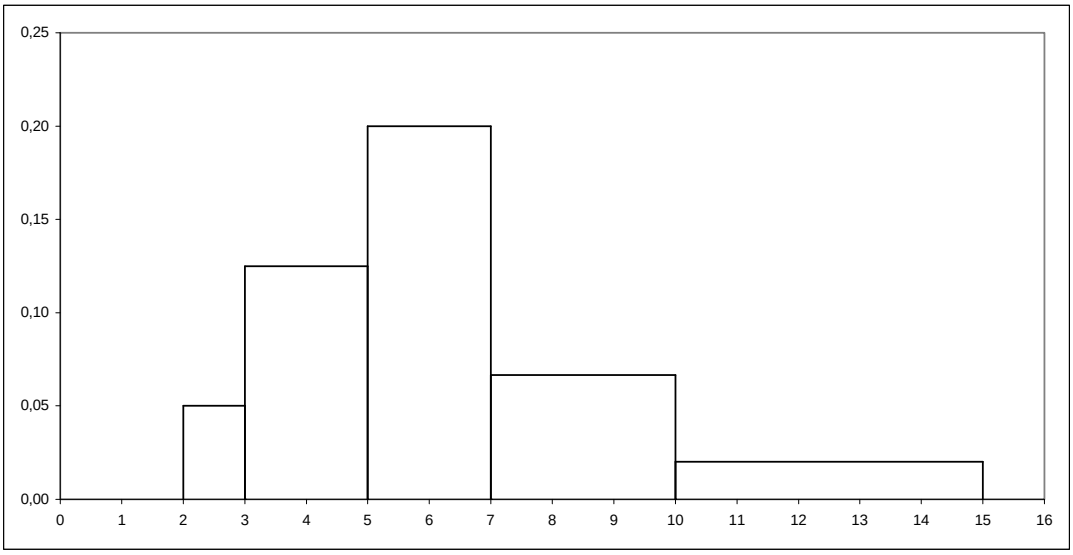
A título de ilustração do procedimento, consideremos os dados sobre aluguéis de imóveis urbanos dados na Tabela 2.31 cujo histograma se encontra na Figura 2.15. Note que a base de cada retângulo, representada na escala horizontal, tem comprimento δ_i e a altura é a densidade, de modo que, como antes, área = freqüência.

Tabela 2.31: Aluguéis de 200 imóveis urbanos

Aluguéis (u.m.)	Comprimento de classe	Freqüência Simples		Freqüência Acumulada		Densidade
		Absoluta	Relativa	Absoluta	Relativa	
	δ_i	n_i	f_i	N_i	F_i	f_i/δ_i
$2 \vdash 3$	1	10	0,05	10	0,05	0,050
$3 \vdash 5$	2	50	0,25	60	0,30	0,125
$5 \vdash 7$	2	80	0,40	140	0,70	0,200
$7 \vdash 10$	3	40	0,20	180	0,90	0,067
$10 \vdash 15$	5	20	0,10	200	1,00	0,020
Total		200	1,00			

Fonte: Dados hipotéticos

Figura 2.15: Distribuição de freqüências dos aluguéis de 200 imóveis urbanos



2.4.7 Observações sobre a construção de gráficos

Os gráficos são apresentados em uma moldura retangular, formada pelos eixos de referência. Tal moldura é construída, em geral, de forma a se manter a proporcionalidade entre a largura e a altura

de 1,414 ($\sqrt{2}$) para 1, que é a mesma razão entre a diagonal e o lado de um quadrado.

Definida a moldura, o próximo ponto é a definição de uma escala adequada para cada eixo. Para isso, deve-se observar a amplitude dos dados a serem representados no eixo e, a partir dela, definir o tamanho do intervalo que definirá a unidade de medida. Esse é o procedimento adotado nos pacotes computacionais, se o usuário não define a escala.

Os gráficos apresentados nessas notas foram todos construídos utilizando o programa de planilhas EXCEL e o programa XLSTAT . A construção dos gráficos de setores, de linhas e de barras é automática no EXCEL, bastando para isso selecionar o tipo adequado. Já os histogramas e o polígono de freqüências foram construídos usando o XLSTAT.

2.4.8 Ramo e folhas

Um outro gráfico usado para mostrar a forma da distribuição de um grupo de dados é o ramo-e-folhas, desenvolvido pelo estatístico americano John Tukey. Este gráfico é constituído de uma linha vertical com a escala indicada à esquerda desta linha. A escala, naturalmente, depende dos valores observados, mas deve ser escolhida de tal forma que cada valor observado possa ser "quebrado" em duas partes: uma primeira parte quantificada pelo valor da escala e a segunda quantificada pelo último algarismo do número correspondente à observação. Os ramos do gráfico correspondem aos números da escala, à esquerda da linha vertical. Já as folhas são os números que aparecem na parte direita. Na Figura 2.16 temos o ramo-e-folhas das notas da Verificação Suplementar de Introdução à Estatística Econômica no primeiro semestre de 2003. Note que a "quebra" dos valores nesse caso é bastante natural: os ramos são formados pelo algarismo inteiro e as folhas pelos algarismos decimais, o que é indicado pela unidade no gráfico.

Figura 2.16: Notas da VS de Introdução à Estatística Econômica - Semestre 1/2003

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

Um outro exemplo utiliza os dados da Tabela 2.32, onde temos dados sobre as quilometragens médias por litro de óleo diesel percorridas por ônibus de 97 empresas de Belo Horizonte. Na Figura 2.17 temos o respectivo ramo-e-folhas gerado pelo programa XLSTAT. Com relação a esse conjunto de dados, as folhas são formadas pela segunda casa decimal; para passar essa informação, é colocado um cabeçalho indicando a unidade dos dados. Uma outra observação importante diz respeito aos valores extremos: se fôssemos representá-los em ramos específicos, a árvore ficaria muito longa, com vários ramos vazios. Uma solução, em geral adotada pelos programas computacionais, é listar os valores com saltos na escala e para chamar a atenção para a quebra de escala, pode-se colocar uma linha divisória, como indicado na figura.

Tabela 2.32: Quilometragem média por litro de óleo diesel de 97 empresas de ônibus de BH

Quilometragem média por litro de óleo diesel									
1,02	1,07	1,16	1,30	1,38	1,43	1,47	1,54	1,56	1,57
1,67	1,67	1,72	1,72	1,72	1,79	1,79	1,82	1,82	1,85
1,85	1,89	1,89	1,92	1,92	1,92	1,96	1,96	2,00	2,00
2,04	2,04	2,04	2,08	2,08	2,08	2,08	2,08	2,08	2,13
2,13	2,13	2,13	2,13	2,13	2,13	2,13	2,17	2,17	2,17
2,17	2,17	2,22	2,22	2,22	2,27	2,27	2,27	2,27	2,27
2,33	2,33	2,33	2,33	2,33	2,33	2,38	2,43	2,44	2,44
2,44	2,44	2,50	2,50	2,56	2,56	2,56	2,56	2,56	2,63
2,63	2,70	2,70	2,78	2,78	2,78	2,78	2,86	2,86	2,94
2,94	3,13	3,23	3,44	3,85	4,08	6,67			

Fonte: Soares, Farias e Cesar (1991)

Figura 2.17: Ramo-e-folhas para os dados da Tabela 2.32

```
Unidade:   0,01
      10 | 2 = 1,02


      10 | 2 7
      11 | 6
      12 |
      13 | 0 8
      14 | 3 7
      15 | 4 6 7
      16 | 7 7
      17 | 2 2 2 9 9
      18 | 2 2 5 5 9 9
      19 | 2 2 2 6 6
      20 | 0 0 4 4 4 8 8 8 8 8 8
      21 | 3 3 3 3 3 3 3 3 7 7 7 7 7
      22 | 2 2 2 7 7 7 7 7
      23 | 3 3 3 3 3 3 8
      24 | 3 4 4 4 4
      25 | 0 0 6 6 6 6 6
      26 | 3 3
      27 | 0 0 8 8 8 8
      28 | 6 6
      29 | 4 4
      30 |
      31 | 3
      32 | 3
      33 |
      34 | 4
      38 | 5
      40 | 8
      66 | 7
```

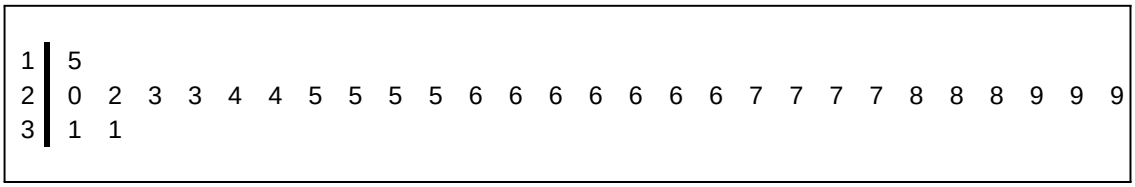
Note que, se olharmos o ramo-e-folhas na posição invertida (isto é,"deitado"), temos o mesmo efeito visual de um gráfico de barras.

Para certos conjuntos de dados, pode acontecer que alguns ramos apresentem muitas folhas, dificultando a sua interpretação. Considere, por exemplo, os dados da Tabela 2.33, onde temos os dados referentes ao consumo de combustível (milhas por galão, MPG) para diferentes modelos de carro. O ramo-e-folhas para esses dados está na Figura 2.18.

Tabela 2.33: Consumo de combustível de 30 modelos de carro

Modelo	MPG	Modelo	MPG
BMW 740i	23	Hyundai Sonata	27
Buick Century	31	Infinity Q45	22
Buick LeSabre	28	Lexus LS400	23
Buick Park Avenue	27	Lincoln Continental	26
Buick Regal	29	Lincoln Mark VIII	25
Buick Roadmaster	25	Mazda 626	31
Cadillac DeVille	25	Mazda 929	24
Chevrolet Caprice	26	Mercedes-Benz S320	24
Chevrolet Lumina	29	Mercedes-Bens S420	20
Chrysler Concorde	28	Nissan Maxima	26
Chrysler New Yorker	26	Rolls-Royce Silver Stone	15
Dodge Spirit	27	Saab 900	26
Fort LTD	25	Saab 9000	27
Ford Taurus	29	Toyota Camry	28
Ford Thunderbird	26	Volvo 850	26

Figura 2.18: Ramo-e-folhas para os dados da Tabela 2.33



Uma forma alternativa de construir esse gráfico é "quebrando" cada ramo em duas partes: a primeira referente aos números terminando com algarismos menores que 5 e a segunda aos números terminando com algarismos maiores ou iguais a 5. Na Figura 2.19 temos essa nova versão.

O ramo-e-folhas comparativo pode ser usado para comparar os resultados referentes a dois grupos. Na Figura 2.20 temos um exemplo baseado nas médias finais (antes da VS) dos alunos de Introdução à Estatística Econômica no primeiro semestre de 2003. Note que na parte esquerda do gráfico, as folhas são anotadas crescentemente da direita para a esquerda, enquanto que na parte direita do gráfico, as folhas são anotadas crescentemente da esquerda para a direita. A análise desses gráficos nos permite ver que a turma da noite teve um comportamento mais homogêneo, com notas, em média, mais altas que a turma da tarde.

Para maiores detalhes sobre esse gráfico e outras técnicas de análise de dados, o leitor pode consultar Tukey(1977), Velleman e Hoaglin(1981) e Murteira(1983).

Figura 2.19: Ramo-e-folhas alternativo para dados da Tabela 2.33

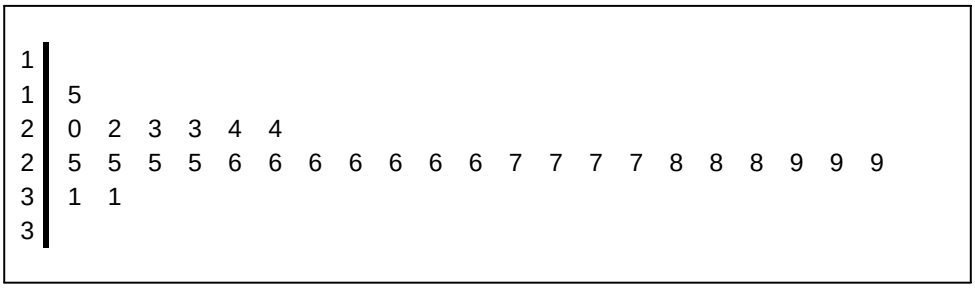
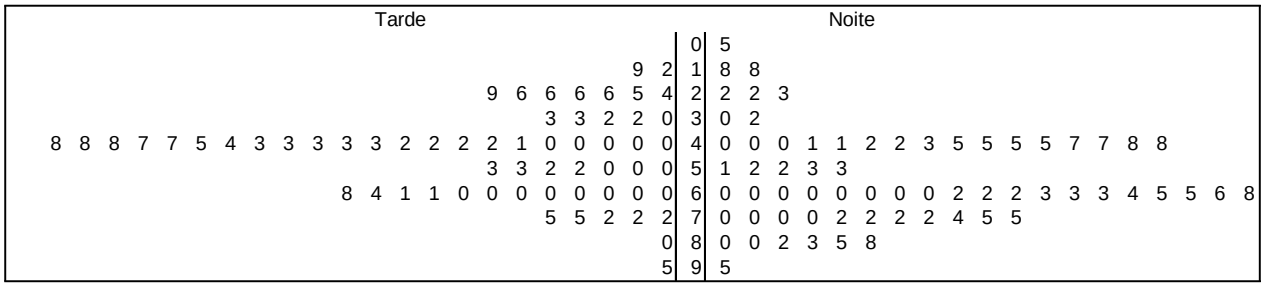


Figura 2.20: Ramo-e-folhas comparativo das notas de alunos



2.4.9 Exercícios resolvidos da Seção 2.4

1. Considere a população total de cada região geográfica do Brasil, conforme exibido na Tabela 2.34. Construa gráficos de setores e de colunas para representar a população total por região e um gráfico de colunas para comparar as populações masculina e feminina por região.

Tabela 2.34: População por região geográfica do Brasil

Região	População		
	Masculina	Feminina	Total
Norte	6.533.555	6.367.149	12.900.704
Nordeste	23.413.914	24.327.797	47.741.711
Sudeste	35.426.091	36.986.320	72.412.411
Sul	12.401.450	12.706.166	25.107.616
Centro-Oeste	5.801.005	5.835.723	11.636.728
Total	83.576.015	86.223.155	169.799.170

Solução:

Para determinar a área ou ângulo de cada setor, usam-se as seguintes regras de três:

Região Norte: $\dfrac{x}{12900704} = \dfrac{360}{169799170} \Rightarrow x = 27,351^{\text{o}}$

Região Nordeste: $\dfrac{x}{47741711} = \dfrac{360}{169799170} \Rightarrow x = 101,220^{\text{o}}$

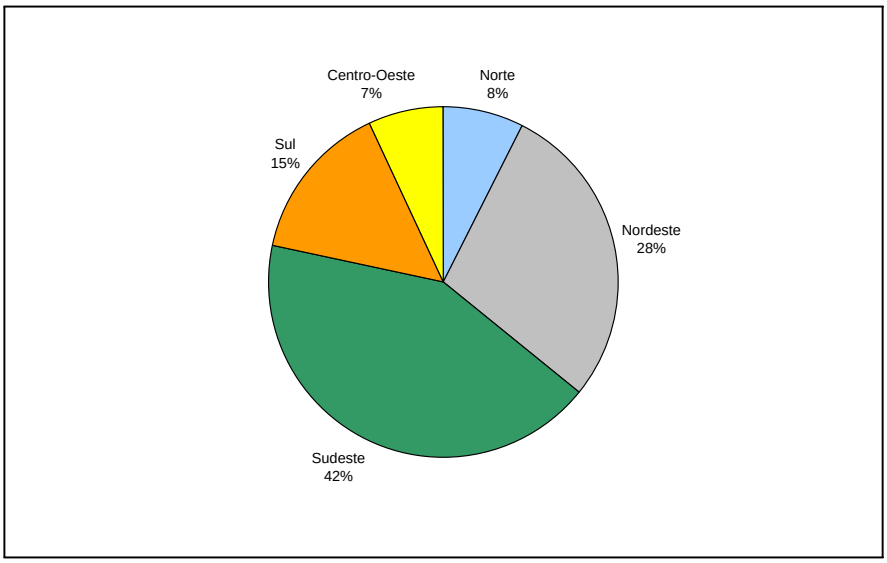
Região Sudeste: $\dfrac{x}{72412411} = \dfrac{360}{169799170} \Rightarrow x = 153,525^{\text{o}}$

Região Sul: $\frac{x}{25107616} = \frac{360}{169799170} \Rightarrow x = 53,232^{o}$

Região Centro-Oeste: $\frac{x}{11636728} = \frac{360}{169799170} \Rightarrow x = 24,672^{o}$

Os gráfico de setores e de colunas são apresentados na Figuras 2.21 e 2.22.

Figura 2.21: População por região geográfica do Brasil



2. Na seção anterior, construímos as tabelas para representar as variáveis LOCAL, PAP, RES e RENDA dos dados das Tabelas 2.16 a 2.18. Vamos agora construir os gráficos apropriados para cada uma delas. Esses gráficos estão apresentados nas Figuras 2.23 a 2.27.

   Solução:

   O interesse na variável LOCAL está em ver a distribuição dos domicílios pelos três locais pesquisados; assim, pode-se usar um gráfico de barras ou um gráfico de setores (ver Figura 2.23).

   A variável PAP indica se a família participa ou não de programas de alimentação; essa informação pode ser representada por um gráfico de barras ou de setores (ver Figura 2.24).

   A variável RES fica bem ilustrada com um gráfico de barras (ver Figura 2.25).

   Para a variável RENDA, vamos usar a distribuição com classes desiguais, dada na Tabela 2.24; a representação gráfica é, então, feita através de um histograma, construído com base nas densidades de cada classe (ver Figura 2.26).

   Outra possibilidade é representar a renda através de um ramo-e-folhas (ver Figura 2.27).

2.4.10 Exercícios propostos da Seção 2.4

2.8 *Construa os gráficos apropriados para representar as tabelas construídas nos Exercícios 2.3 a 2.7.*

Figura 2.22: População (em milhões de habitantes) por região geográfica do Brasil

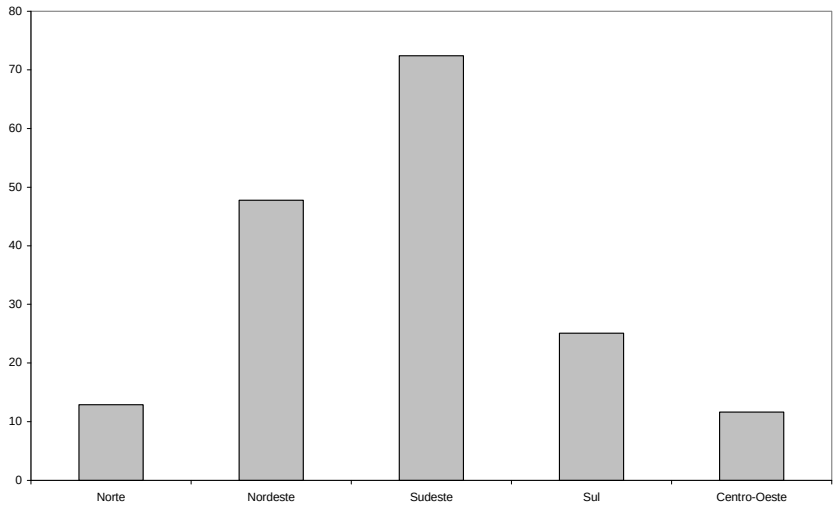


Figura 2.23: Distribuição dos domicílios por localização (LOCAL)

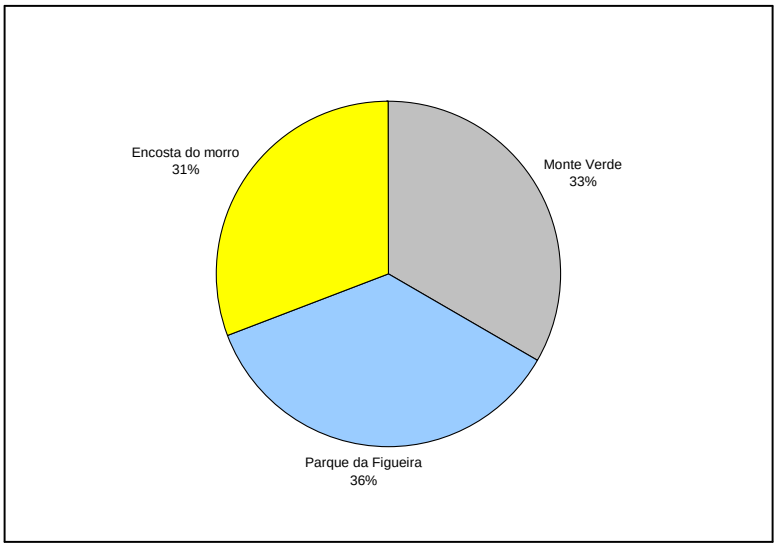


Figura 2.24: Participação em programas de alimentação (PAP)

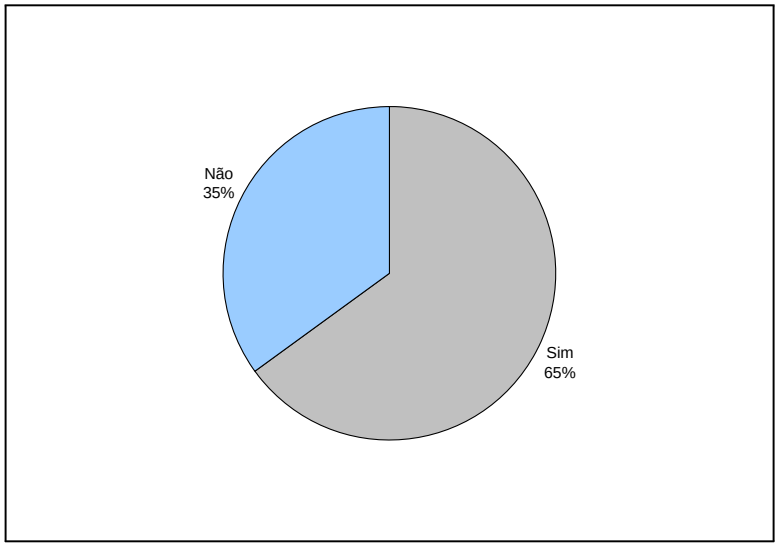


Figura 2.25: Número de residentes por domicílio (RES)

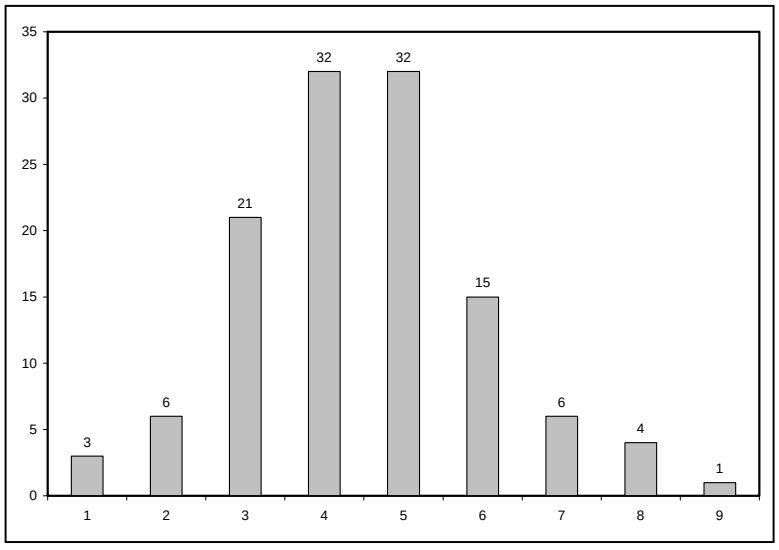


Figura 2.26: Distribuição da renda dos 119 domicílios

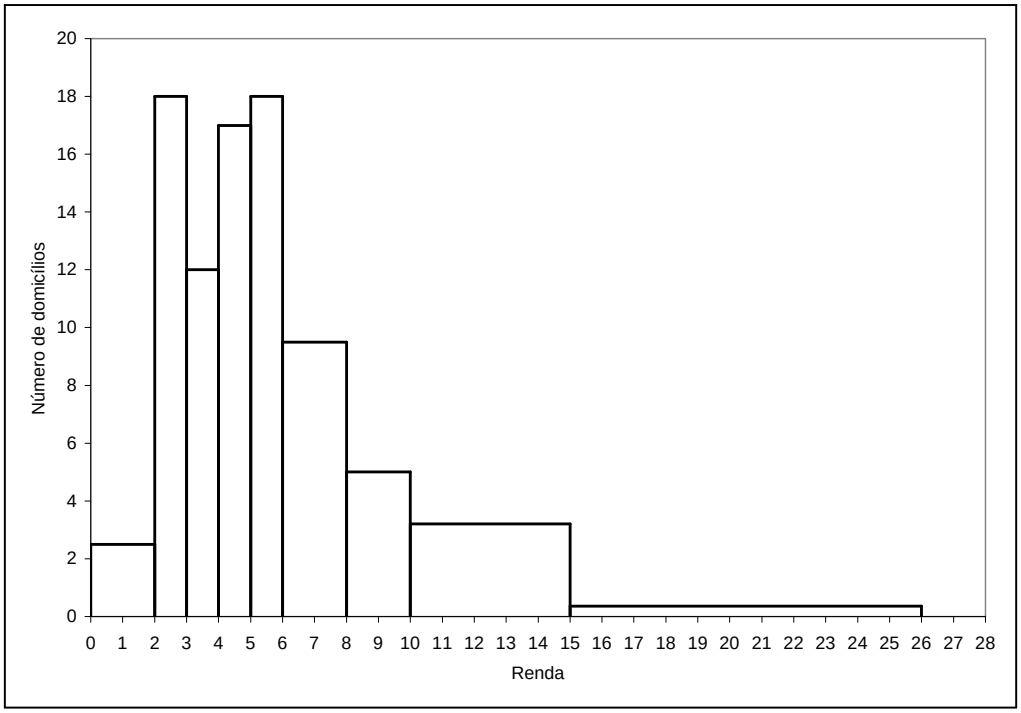


Figura 2.27: Ramo-e-folhas da renda das 119 famílias

0	1 9
1	0 0 0
2	1 1 2 3 3 3 4 4 4 5 6 6 7 7 8 8 8 9
3	1 5 5 6 8 9 9 9 9 9 9 9
4	0 0 1 1 2 2 2 2 4 4 5 6 6 8 8 8 9
5	0 0 0 1 1 1 3 4 5 5 5 7 8 8 8 8 8 8
6	0 1 4 4 4 4 4 4 8 8
7	1 1 2 4 5 7 7 7 9
8	4 4 5 6 9
9	0 0 0 0 6
10	2 3 3 4 8 9
11	1 3 4
12	2 5 9 9 9
13	9
14	0
15	4
16	
17	
18	6
19	3
20	
21	
22	
23	
24	
25	7

2.5 Representação tabular: Distribuição bivariada de freqüências

2.5.1 Variáveis qualitativas

Até o momento, vimos como organizar e resumir informações referentes a uma única variável. No entanto, é bastante freqüente depararmos com situações onde há interesse em estudar conjuntamente duas ou mais variáveis. Para os dados da Tabela 2.2, por exemplo, podemos estudar se há alguma relação entre sexo e a matéria predileta no segundo grau. Num estudo sobre mortalidade infantil, é importante acompanhar também o tratamento pré-natal da mãe; espera-se, neste caso, que haja uma diminuição da taxa de mortalidade infantil com o aumento dos cuidados durante a gravidez.

Nesta seção nos deteremos no estudo de distribuições bidimensionais, dando ênfase à forma de representação tabular. Seguindo uma convenção usual, denotaremos por uma letra maiúscula a variável em estudo e pela letra minúscula correspondente o valor observado da variável.

Consideremos inicialmente o caso de duas variáveis *qualitativas*. Como exemplo, vamos trabalhar com os dados apresentados na Tabela 2.2, onde temos a matéria predileta no segundo grau e o sexo de 80 alunos.

Uma forma de representar conjuntamente as informações referentes a essas duas variávies é através de uma distribuição ou tabela conjunta de freqüências. Como temos duas variáveis de interesse, precisamos de duas dimensões, linha e coluna, para representar as informações disponíveis, que serão apresentadas em forma de contagem ou freqüência. A escolha da variável linha e da variável coluna depende do objetivo do estudo. Se existe entre as variáveis uma relação do tipo dependente/explanatória, isto é, se queremos usar uma das variáveis para "explicar" a outra, então é costume colocar a variável explanatória na coluna e denotá-la por X. A variável dependente, que é explicada pela variável explanatória, é colocada na linha e indicada pela letra Y. Caso contrário, qualquer uma das duas pode ser a variável coluna. No exemplo, poderíamos estar interessados em analisar o efeito do sexo sobre a matéria predileta (obviamente, não podemos explicar o sexo...); sendo assim, o sexo é a variável explanatória X e a matéria predileta no segundo grau é a variável explicada ou dependente Y. Cada aluno dá origem a um par de valores (x_i, y_i), por exemplo, (masculino, história).

Na Tabela 2.35 apresentamos a distribuição conjunta dessas variáveis. Em cada cela temos o número de alunos que pertencem simultaneamente às respectivas categorias. Assim, podemos ver que há 12 homens que preferiam geografia no segundo grau, enquanto que, entre as mulheres, apenas 6 preferiam essa matéria. Como já visto no caso univariado, essa forma de apresentação é mais interessante, uma vez que não estamos interessados na observação individual e, sim, no comportamento dos grupos.

Tabela 2.35: Distribuição conjunta das variáveis sexo e matéria predileta no segundo grau

Matéria predileta	Sexo		Total
no segundo grau	Masculino	Feminino	
Ciências	4	1	5
Geografia	12	6	18
História	8	6	14
Matemática	11	15	26
Português	6	11	17
Total	41	39	80

Além das contagens em cada cela, acrescentamos também a linha e a coluna com os respectivos totais. Os totais das linhas, então, nos dizem que há 5 alunos que preferiam Ciências, 18 que

preferiam Geografia, e assim por diante. Já os totais das colunas nos dizem que há 41 alunos do sexo masculino e 39 do sexo feminino. O total de alunos (80) pode ser obtido somando-se os totais das linhas (matéria predileta): $5 + 88 + 14 + 26 + 17 = 80$ ou das colunas (sexo): $41 + 39 = 80$.

Na construção de tabelas de freqüências univariadas, foi acrescentada à tabela a coluna de freqüências relativas, que davam a proporção de elementos em cada classe com relação ao número total de elementos. Um procedimento análogo pode ser feito para as tabelas bidimensionais; a diferença é que, neste caso, existem três possibilidades para expressarmos as proporções de cada cela: (i) com relação ao total geral; (ii) com relação ao total de cada linha e (iii) com relação ao total de cada coluna. A escolha entre essas três possibilidades deverá ser feita de acordo com o objetivo da análise. Nas Tabelas 2.36 a 2.38 temos as três versões para os dados da Tabela 2.35 usando freqüências relativas.

Tabela 2.36: Distribuição conjunta relativa das variáveis sexo e matéria predileta no segundo grau

Matéria predileta	Sexo		Total
no segundo grau	Masculino	Feminino	
Ciências	5,00	1,25	6,25
Geografia	15,00	7,50	22,50
História	10,00	7,50	17,50
Matemática	13,75	18,75	32,50
Português	7,50	13,75	21,25
Total	51,25	48,75	100,00

Tabela 2.37: Distribuição condicional do sexo dada a matéria predileta no segundo grau

Matéria predileta	Sexo		Total
no segundo grau	Masculino	Feminino	
Ciências	80,00	20,00	100,00
Geografia	66,67	33,33	100,00
História	57,14	42,86	100,00
Matemática	42,31	57,69	100,00
Português	35,29	64,71	100,00
Total	51,25	48,75	100,00

Tabela 2.38: Distribuição condicional da matéria predileta no segundo grau dado o sexo do aluno

Matéria predileta	Sexo		Total
no segundo grau	Masculino	Feminino	
Ciências	9,76	2,56	6,25
Geografia	29,27	15,38	22,50
História	19,51	15,38	17,50
Matemática	26,83	38,46	32,50
Português	14,63	28,21	21,25
Total	100,00	100,00	100,00

Da Tabela 2.36 podemos concluir que 5% dos alunos são do sexo Masculino e preferiam Ciências no segundo grau, enquanto 18,75% eram do sexo feminino e preferiam Matemática. Essa é a *tabela da distribuição conjunta relativa*; em cada cela temos a freqüência dos indivíduos que pertencem

simultaneamente às duas categorias em questão relativa ao total geral. A título de ilustração dos cálculos, temos:

$$\text{Masculino e Ciências:} \quad \frac{4}{80} \times 100 = 5,00\%$$

$$\text{Feminino e Matemática:} \quad \frac{15}{80} \times 100 = 18,75\%$$

Da Tabela 2.37 conclui-se, por exemplo, que, dos alunos que preferiam Ciências no segundo grau, 80% são homens e 20% são mulheres, enquanto que, dos alunos que preferiam Matemática, 42,31% são homens e 57,69% são mulheres. Essa é a *distribuição condicional do sexo (variável coluna) dada a matéria predileta no segundo grau (variável linha).* Na linha Total temos a distribuição por sexo na população completa, que coincide com os totais das linhas da Tabela 2.35: 51,25% dos alunos são do sexo masculino e 48,75% são do sexo feminino. Os detalhes dos cálculos são o seguintes:

	Masculino	Feminino
Ciências no segundo grau	$\frac{4}{5} \times 100 = 80,0$	$\frac{1}{5} \times 100 = 20,0$
Matemática no segundo grau	$\frac{11}{26} \times 100 = 42,31$	$\frac{15}{26} \times 100 = 57,69$

Da Tabela 2.38 podemos ver que 9,76% dos homens preferiam Ciências no segundo grau, enquanto 15,38% das mulheres preferiam Geografia. Essa tabela nos dá a *distribuição condicional da matéria predileta no segundo grau (variável linha), dado o sexo (variável coluna).* Na coluna Total temos a distribuição da variável matéria predileta no segundo grau (variável linha) na população completa. Esse total, obviamente, coincide com os totais das colunas na Tabela 2.35. Essa é a tabela apropriada para a análise desejada, de comparar os sexos segundo a matéria predileta. Os detalhes dos cálculos são o seguintes:

$$\text{Ciências no segundo grau, dado que é homem} \quad \frac{4}{41} \times 100 = 9,76$$

$$\text{Geografia no segundo grau, dado que é mulher} \quad \frac{6}{39} \times 100 = 15,38$$

Mais uma vez, é importante salientar que, na construção de tabelas com freqüências relativas, um cuidado especial deve ser tomado com relação ao arredondamento dos números. Arredondamentos excessivos podem fazer com que os totais de linhas e/ou colunas não somem 100%!

É possível também usar o gráfico de barras para representar distribuições conjuntas de variáveis. Consideremos novamente o exemplo de sexo e matéria predileta no segundo grau, conforme dados na Tabela 2.35. O gráfico apresentado na Figura 2.28 representa essas variáveis, levando em conta o fato de que sexo é a variável explicativa.

2.5.2 Variáveis quantitativas

No caso de variáveis quantitativas discretas com poucos valores, a construção de tabelas bivariadas é feita de maneira análoga às variáveis qualitativas. Para variáveis quantitativas contínuas ou discretas com muitos valores, a construção é possível, mas não muito usual, uma vez que há muita perda de informação pois, assim como no caso univariado, é preciso agrupar os dados em classes.

Figura 2.28: Distribuição da matéria predileta no segundo grau por sexo dos alunos

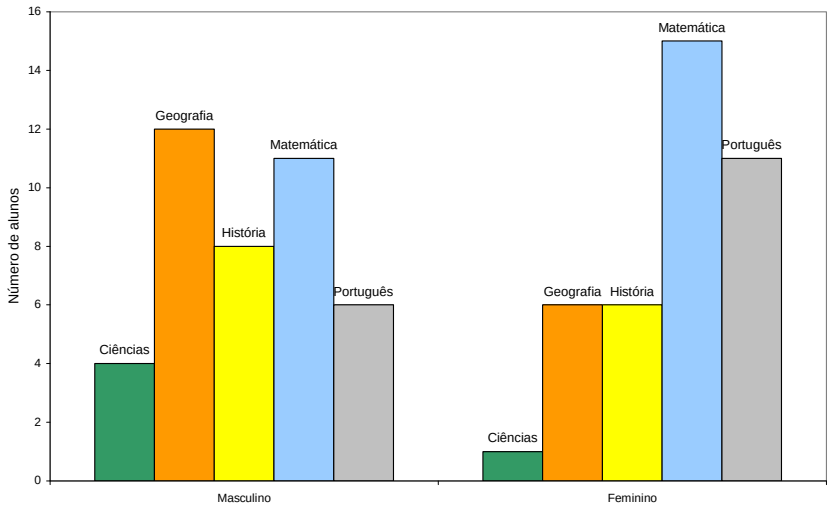


Diagrama de dispersão

O diagrama de dispersão é um gráfico utilizado para representar conjuntamente os valores de duas variáveis quantitativas, com o objetivo de se estudar uma possível relação entre as duas. Como exemplo, consideremos os dados da Tabela 2.39 sobre despesas com alimentação (Y) e renda (X). Nesse caso, espera-se que, ao aumentar a renda, aumentem também as despesas com alimentação. Como verificar isso graficamente? Para cada domicílio há um par de valores (x_i, y_i). O que vamos fazer é simplesmente representar esses pontos em um sistema de eixos cartesianos. Na Figura 2.29 temos o diagrama de dispersão para esses dados.

Na Figura 2.30 temos alguns diagramas de dispersão que ilustram possíveis padrões de relação entre duas variáveis. Na linha superior da figura, no gráfico à esquerda há uma relação quase linear crescente, enquanto que no gráfico à direita há uma relação decrescente, também quase linear. Na linha inferior, no gráfico à esquerda não podemos identificar qualquer relação entre as variáveis, enquanto que no gráfico à direita, a relação não é linear, aproximando-se bastante de uma relação quadrática. No próximo capítulo voltaremos a abordar situações como essas.

Tabela 2.39: Despesas com alimentação e renda

	Despesas com Alimentação (u.m.)	Renda Mensal (u.m.)		Despesas com Alimentação (u.m.)	Renda Mensal (u.m.)
1	52,25	258,3	21	98,14	719,80
2	58,32	343,1	22	123,94	720,00
3	81,79	425,00	23	126,31	722,30
4	119,9	467,50	24	146,47	722,30
5	125,8	482,90	25	115,98	734,40
6	100,46	487,70	26	207,23	742,50
7	121,51	496,50	27	119,80	747,70
8	100,08	519,40	28	151,33	763,30
9	127,75	543,30	29	169,51	810,20
10	104,94	548,70	30	108,03	818,50
11	107,48	564,60	31	168,90	825,60
12	98,48	588,30	32	227,11	833,30
13	181,21	591,30	33	84,94	834,00
14	122,23	607,30	34	98,70	918,10
15	129,57	611,20	35	141,06	918,10
16	92,84	631,00	36	215,40	929,60
17	117,92	659,60	37	112,89	951,70
18	82,13	664,00	38	166,25	1014,00
19	182,28	704,20	39	115,43	1141,30
20	139,13	704,80	40	269,03	1154,60

Figura 2.29: Diagrama de dispersão para renda e despesas com alimentação

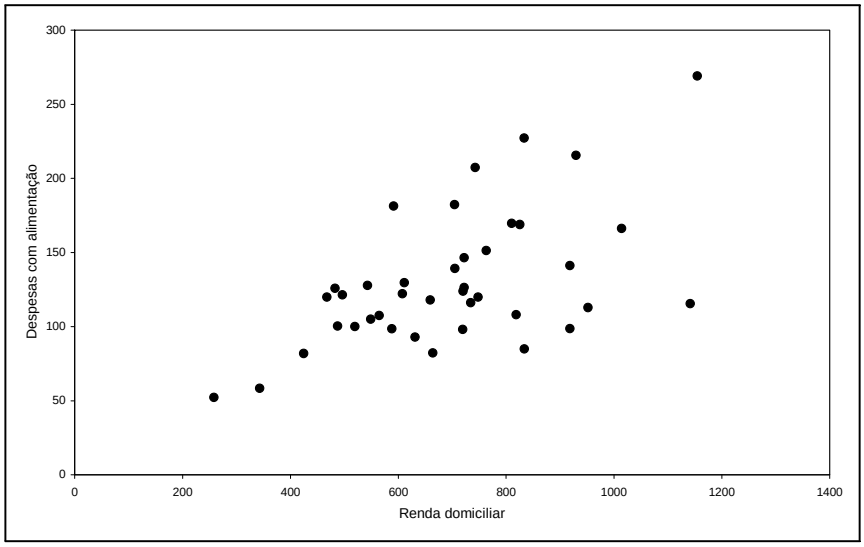
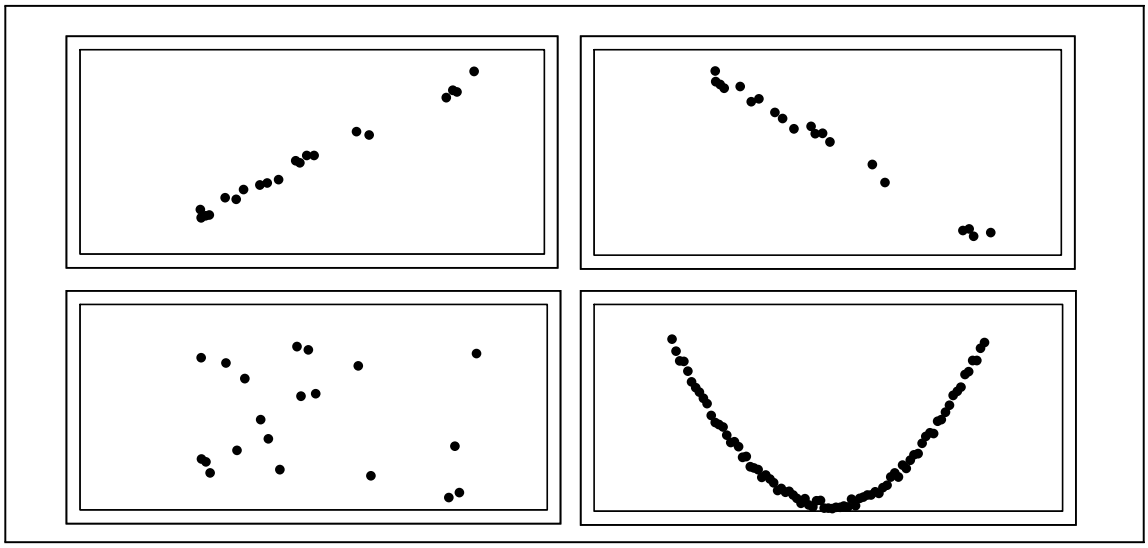


Figura 2.30: Exemplos de diagramas de dispersão que ilustram diferentes relações entre as variáveis



2.5.3 Exercícios resolvidos da Seção 2.5

1. Considere a população, por sexo, de cada região geográfica do Brasil, conforme exibido na Tabela 2.40. Construa um gráfico de colunas para comparar as populações masculina e feminina por região.

Tabela 2.40: População por região geográfica do Brasil para o Exercício Resolvido 1

Região	População	
	Masculina	Feminina
Norte	6.533.555	6.367.149
Nordeste	23.413.914	24.327.797
Sudeste	35.426.091	36.986.320
Sul	12.401.450	12.706.166
Centro-Oeste	5.801.005	5.835.723
Total	83.576.015	86.223.155

Solução:

O gráfico que compara as populações masculina e feminina por região está na Figura 2.31.

2. Na tabela abaixo temos dados sobre hábitos de fumo de uma amostra de moradores de uma pequena cidade (dados fictícios).

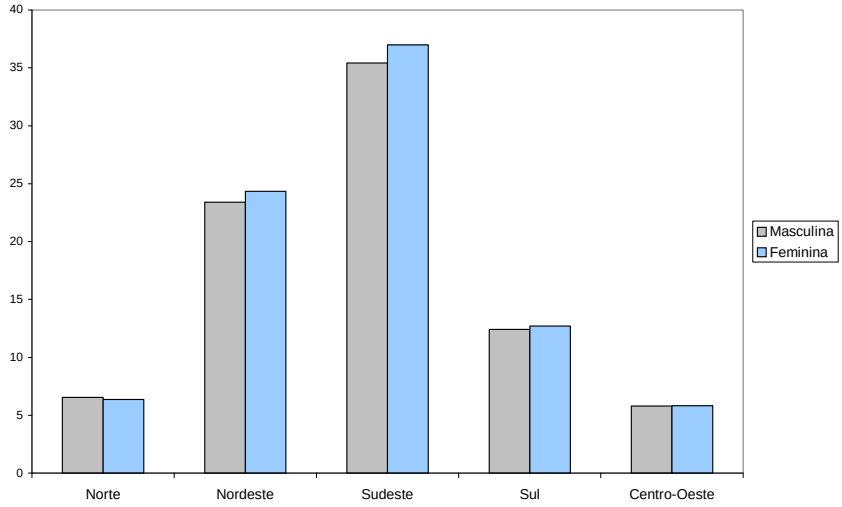
   (a) Defina claramente as variáveis envolvidas, estabelecendo o tipo de cada uma.

   (b) É possível estabelecer uma relação dependente/explanatória entre elas? Em caso afirmativo, qual é a variável explanatória e qual é a variável dependente?

   (c) Complete a tabela, acrescentando os totais.

   (d) Construa as três tabelas possíveis de freqüências relativas.

Figura 2.31: População (em milhões de habitantes) por sexo nas regiões geográficas do Brasil



(e) Construa o gráfico apropriado para representar esses dados.

Hábitos de fumo	Idade		
	< 20	$[20, 30)$	≥ 30
Fumante	143	171	40
Ex-fumante	11	152	140
Nunca fumou	66	57	20

Solução:

As variáveis envolvidas são Hábito de Fumo e Idade. Ambas são qualitativas, uma vez que a idade foi dada em classes.

A única possibilidade é explicar o hábito de fumo pela idade, ou seja, idade é a variável explicativa ou independente e Hábito de Fumo é a variável dependente.

A seguir temos a tabela com os totais de linha e de coluna

Hábitos de fumo	Idade			
	< 20	$[20, 30)$	≥ 30	Total
Fumante	143	171	40	354
Ex-fumante	11	152	140	303
Nunca fumou	66	57	20	143
Total	220	380	200	800

A distribuição de freqüência relativa conjunta é a seguinte:

Hábitos	Idade			
de fumo	< 20	$[20, 30)$	≥ 30	Total
Fumante	17,875	21,375	5,000	44,250
Ex-fumante	1,375	19,000	17,500	37,875
Nunca fumou	8,250	7,125	2,500	17,875
Total	27,500	47,500	25,000	100,000

Em termos da distribuição condicional do hábito de fumo por faixa etária (total por coluna) temos a seguinte tabela:

Hábitos	Idade			
de fumo	< 20	$[20, 30)$	≥ 30	Total
Fumante	65,00	45,00	20,00	44,250
Ex-fumante	5,00	40,00	70,00	37,875
Nunca fumou	30,00	15,00	10,00	17,875
Total	100,00	100,00	100,00	100,000

E para a distibuição condicional da idade pelo hábito de fumo (total por linha) a tabela é:

Hábitos	Idade			
de fumo	< 20	$[20, 30)$	≥ 30	Total
Fumante	40,3955	48,3051	11,2994	100,0000
Ex-fumante	3,6304	50,1650	46,2046	100,0000
Nunca fumou	46,1538	39,8601	13,9860	100,0000
Total	27,5000	47,5000	25,0000	100,0000

Como idade é a variável explicativa, o gráfico apropriado é o gráfico de colunas apresentado na Figura 2.32.

3. Construa um gráfico para comparar as três localidades com relação à variável Grau de Instrução para os dados das Tabelas 2.16 a 2.18.

   Solução:

   Uma possibilidade é o gráfico de colunas (Figura 2.33) e outra é o gráfico de colunas empilhadas (Figura 2.34).

4. Na Tabela 2.41 temos o consumo de cigarros per capita (X) em 1930 e as mortes (Y) por 1.000.000 habitantes am 1950, causadas por câncer de pulmão em 11 países. Para os dados em questão, construa o diagrama de dispersão.

   Solução:

   O diagrama de dispersão para esses dados está na Figura 2.35 abaixo.

Figura 2.32: Distribuição do hábito de fumar por faixa etária

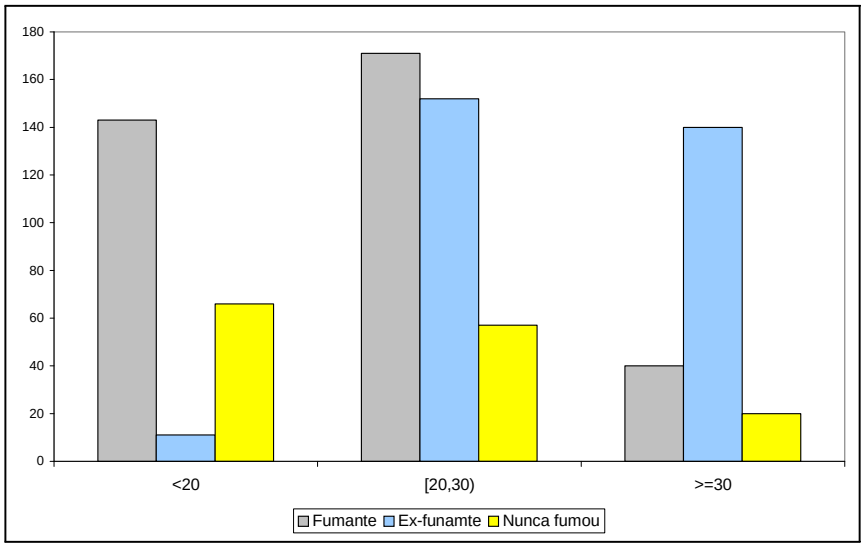


Figura 2.33: Grau de instrução dos chefes de família

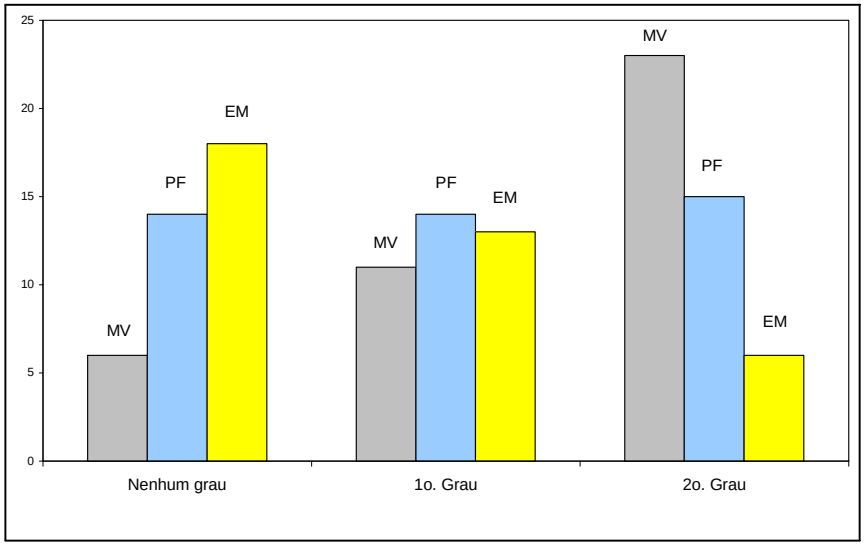


Tabela 2.41: Consumo de cigarros (X) e morte por câncer de pulmão (Y) para o Exercício Resolvido 4 da Seção 2.4

País	X	Y	País	X	Y
Islândia	240	63	Holanda	490	250
Noruega	255	100	Suiça	180	180
Suécia	340	140	Finlândia	1125	360
Dinamarca	375	175	Grã-Bretanha	1150	470
Canadá	510	160	Estados Unidos	1275	200
Austrália	490	180			

Figura 2.34: Grau de instrução dos chefes de família

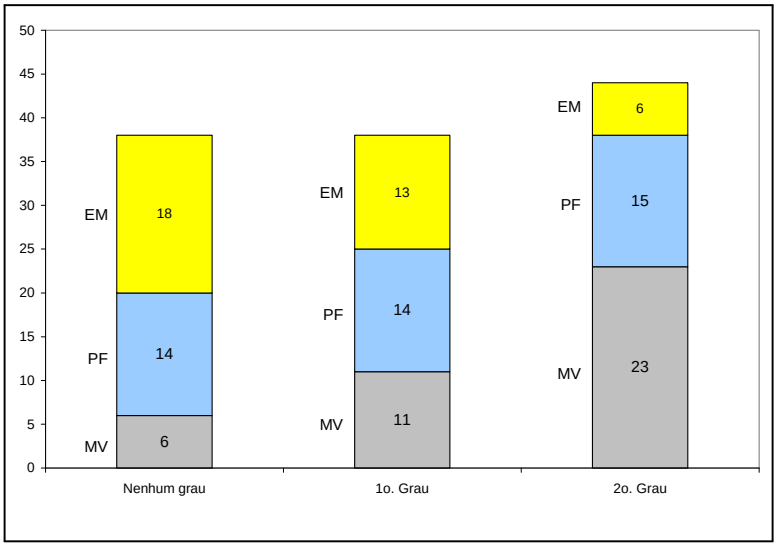
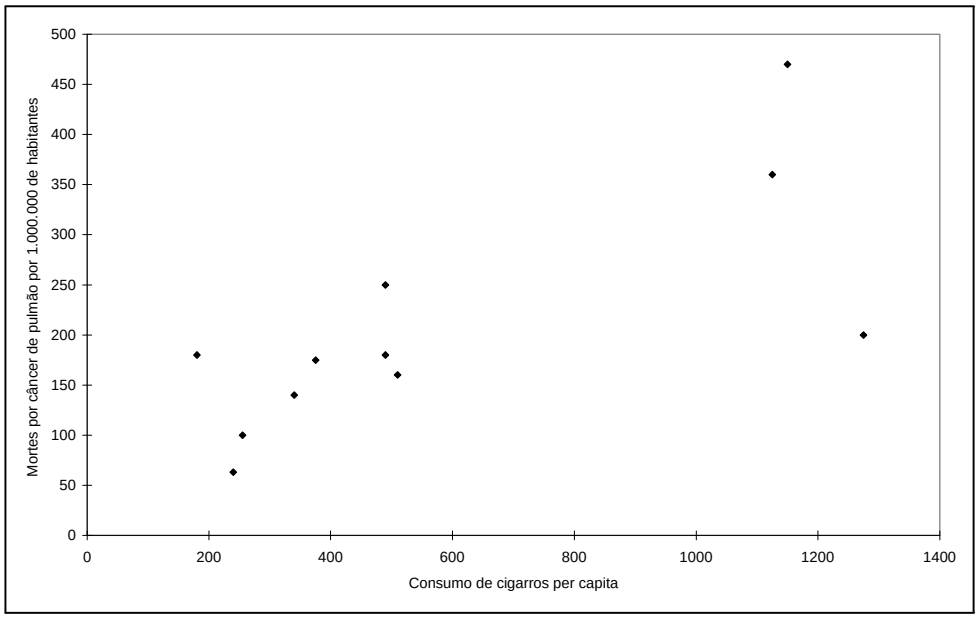


Figura 2.35: Consumo de cigarro e mortes por câncer de pulmão para o Exercício Resolvido 4 da Seção 2.4



2.6 Exercícios Complementares

2.9 *Os pesos dos jogadores de um time de futebol variam de 75 a 95 quilos. Quais seriam os extremos se quiséssemos grupá-los em 5 classes de mesmo tamanho?*

2.10 *Em certa época, os salários mensais dos operários de uma indústria eletrônica variavam de 1.500 a 3.150 unidades monetárias. Quais seriam os limites se quiséssemos grupá-los em 6 classes de mesmo tamanho?*

2.11 *Na Tabela 2.42 abaixo temos as notas de 50 alunos em um teste. Construa uma tabela de freqüências, usando as classes* $20 \vdash 30, 30 \vdash 40, 40 \vdash 50, \cdots, 90 \vdash 100$. *Construa o histograma, o polígono de freqüências e a ogiva de freqüências.*

Tabela 2.42: Notas de 50 alunos para o Exercício 2.11

29	37	38	47	49	52	56	58	60	62
63	63	63	65	65	66	68	68	69	69
70	70	71	73	73	74	74	75	75	76
76	77	77	79	81	81	82	82	83	83
84	85	87	87	88	89	90	91	94	97

2.12 *Num estudo sobre a jornada de trabalho das empresas de Produtos Alimentares foram levantados os dados da Tabela 2.43 relativos ao total de horas trabalhadas pelos funcionários no mês de agosto (dados hipotéticos). Construa uma tabela de freqüências usando 5 classes de mesmo tamanho; construa também o histograma e a ogiva de freqüências. Para facilitar a solução, os valores mínimo e máximo são: 1.815 e 118.800.*

Tabela 2.43: Jornada de trabalho de empresas alimentares para o Exercício 2.12

3.960	5.016	13.015	8.008	6.930	5.544	4.224	6.138
118.800	57.904	72.600	100.100	55.935	7.223	3.775	4.224
3.216	7.392	2.530	6.930	1.815	4.338	8.065	10.910
8.408	8.624	6.864	5.742	5.749	8.514	2.631	5.236
8.527	3.010	5.914	11.748	8.501	6.512	11.458	10.094
6.721	2.631	7.082	10.318	8.008	3.590	7.128	7.929
10.450	6.780	5.060	5.544	6.178	13.763	9.623	14.883
17.864	34.848	25.300	52.800	17.732	63.923	30.360	18.876
30.800	19.562	49.240	49.434	26.950	22.308	21.146	14.212
25.520	49.251	30.976	23.338	43.648	26.796	44.880	30.008
30.769	16.907	33.911	27.034	16.500	14.445	28.160	42.442
16.507	36.960	67.760	84.084	89.888	65.340	82.280	86.152
91.080	99.792	77.836	76.032				

2.13 *Na Tabela 2.44 temos a densidade populacional (hab/km*^{2}*) das unidades da federação brasileira. Construa um gráfico ramo-e-folhas para esses dados.*

2.14 *Na Tabela 2.45 temos a população dos municípios de MG com mais de 50.000 habitantes, com base nos dados do Censo Demográfico 2000. Excluindo a capital Belo Horizonte, construa uma tabela de freqüências e o respectivo histograma, trabalhando com as seguintes classes (em 1.000 hab.): [50,60), [60,70), [70,80), [80,100), [100,200), [200, 500) e 500 ou mais.*

Tabela 2.44: Densidade populacional dos estados brasileiros, para o Exercício 2.13

UF	Densidade Populacional (hab/km^2)	UF	Densidade Populacional (hab/km^2)
RO	6	SE	81
AC	4	BA	24
AM	2	MG	31
RR	2	ES	68
PA	5	RJ	328
AP	4	SP	149
TO	5	PR	48
MA	17	SC	57
PI	12	RS	37
CE	51	MS	6
RN	53	MT	3
PB	61	GO	15
PE	81	DF	353
AL	102		

Fonte: IBGE - Censo Demográfico 2000

Tabela 2.45: População dos municípios de MG com mais de 50.000 habitantes, para o Exercício 2.14

Município	População	Município	População	Município	População
Leopoldina	50.097	Timóteo	71.478	Varginha	108.998
Pirapora	50.300	Pará de Minas	73.007	Barbacena	114.126
três Pontas	51.024	Patrocínio	73.130	Sabará	115.352
São Francisco	51.497	Paracatu	75.216	Patos de Minas	123.881
Pedro Leopoldo	53.957	Vespasiano	76.422	Teófilo Otoni	129.424
Ponte Nova	55.303	Itaúna	76.862	Ibirité	133.044
S.Seb.do Paraíso	58.335	Caratinga	77.789	Poços de Caldas	135.627
Janaúba	61.651	S.João del Rei	78.616	Divinópolis	183.962
Formiga	62.907	Lavras	78.772	Sete Lagoas	184.871
Januária	63.605	Araxá	78.997	Santa Luzia	184.903
Cataguases	63.980	Itajubá	84.135	Ipatinga	212.496
Nova Lima	64.387	Ubá	85.065	Ribeirão das Neves	246.846
Viçosa	64.854	Ituiutaba	89.091	Gov.Valadares	247.131
Três Corações	65.291	Muriaé	92.101	Uberaba	252.051
Ouro Preto	66.277	Passos	97.211	Betim	306.675
João Monlevade	66.690	Cor. Fabriciano	97.451	Montes Claros	306.947
Alfenas	66.957	Itabira	98.322	Juiz de Fora	456.796
Manhuaçu	67.123	Araguari	101.974	Uberlândia	501.214
Curvelo	67.512	Cons.Lafaiete	102.836	Contagem	538.017
Unaí	70.033	Pouso Alegre	106.776	Belo Horizonte	2.238.526

Fonte: IBGE - Censo Demográfico 2000

2.15 *Na Tabela 2.46 temos os dados que ilustram a seguinte manchete do jornal Folha de São Paulo:*

VAREJO

Preços sobem 1,37% em SP, em média, na semana;
setor não vê anormalidade e diz que é só acomodação.
Hipermercados têm a maior alta do ano.

Construa o gráfico apropriado para ilustrar o fato descrito na manchete.

Tabela 2.46: Preços no varejo, para o Exercício 2.15

Variação % semanal do s preços					
Semana	%	Semana	%	Semana	%
17/11	2,05	28/12	1,23	09/02	-0,13
24/11	0,18	05/01	-0,39	16/02	0,43
01/12	-0,26	12/01	0,57	23/02	0,71
08/12	0,68	19/01	0,58	01/03	0,53
15/12	0,84	26/01	0,30	08/03	0,64
21/12	1,12	02/02	-0,40	15/03	1,37

Fonte: Folha de São Paulo

2.16 *Para a seguinte notícia, extraída do jornal Folha de São Paulo, construa um gráfico para ilustrar o texto da notícia.*

"Dentro de dez anos, 90% do mercado automobilístico mundial estará nas mãos de meia dúzia de conglomerados. A previsão consta de estudo produzido pela consultoria especializada britânica Autopolis, que dá assessoria técnica a montadoras que estão instaladas no Reino Unido.

... Dados levantados pela Autopolis mostram que, hoje, a concentração de mercado já é grande. Cerca de 75% do setor é dominado por somente seis conglomerados, liderados por General Motors (22,8%), Ford (16,8%), Volkswagen (9,4%), Toyota (9,2%, incluindo Daihatsu), Reanult-Nissan (8,7%) e Daimler-Chrysler (8,3%). Os outros 24,8% do mercado são dominados por uma infinidade de empresas pequenas e médias, como Fiat, BMW, Peugeot e Honda, entre outras.".

2.17 *Com base na Tabela 2.47, construa um gráfico para mostrar a distribuição da população por sexo nas 27 unidades da federação (UF) brasileiras.*

2.18 *Na Tabela 2.48 temos dados referentes ao número de pulsos excedentes na conta de telefone de uma residência para os meses de janeiro de 98 a junho de 99. Construa o gráfico adequado para representar esses dados.*

2.19 *Na Tabela 2.49, temos dados sobre casas vendidas na região de Boulder, Colorado (EUA)*[7]*, no primeiro semestre de 1995. Vamos denotar por X a variável área (em m^2) e por Y o preço de venda (em 1000 US$).*

*(a) Construa uma tabela de freqüências completa para a variável Y (preço de venda) usando 5 classes de mesmo comprimento. **Atenção**: na definição das classes, tome como limite inferior da primeira classe o valor 110 e trabalhe com amplitude de classe inteira!*

[7] Dados extraídos de Moore e McCabe (1999)

Tabela 2.47: População brasileira por UF e sexo, para o Exercício 2.17

	População			População	
UF	Homens	Mulheres	UF	Homens	Mulheres
RO	708.140	671.647	SE	874.906	909.569
AC	280.983	276.543	BA	6.462.033	6.608.217
AM	1.414.367	1.398.190	MG	8.851.587	9.039.907
RR	166.037	158.360	ES	1.534.806	1.562.426
PA	3.132.768	3.059.539	RJ	6.900.335	7.490.947
AP	239.453	237.579	SP	18.139.363	18.893.040
TO	591.807	565.291	PR	4.737.420	4.826.038
MA	2.812.681	2.838.794	SC	2.669.311	2.687.049
PI	1.398.290	1.444.988	RS	4.994.719	5.193.079
CE	3.628.474	3.802.187	MS	1.040.024	1.037.977
RN	1.359.953	1.416.829	MT	1.287.187	1.217.166
PB	1.671.978	1.771.847	GO	2.492.438	5.510.790
PE	3.826.657	4.091.687	DF	981.356	1.069.790
AL	1.378.942	1.443.679			

Fonte: IBGE - Censo Demográfico 2000

Tabela 2.48: Número de pulsos excedentes, para o Exercício 2.18

Jan/98	110	Jul/98	340	Jan/99	290
Fev/98	0	Ago/98	198	Fev/99	48
Mar/98	212	Set/98	141	Mar/99	303
Abr/98	239	Out/98	195	Abr/99	223
Mai/98	120	Nov/98	398	Mai/99	296
Jun/98	174	Dez/98	377	Jun/99	383

(b) Construa um ramo-e-folhas para a variável Área.

(c) Construa um diagrama de dispersão para as variáveis Área e Preço.

Tabela 2.49: Vendas de casas em Boulder, Colorado (1995) para o Exercício 2.19

Preço (Y) (1000 US$)	Área (X) (m^2)	Preço (Y) (1000 US$)	Área (X) (m^2)	Preço (Y) (1000 US$)	Área (X) (m^2)
113	126	163	227	186	228
114	158	168	228	187	219
120	126	168	249	187	222
120	126	169	244	188	279
122	158	169	263	188	249
123	126	170	234	190	317
129	229	171	283	192	304
137	196	172	286	193	195
140	262	173	268	195	217
142	272	175	223	195	232
143	189	175	270	200	234
146	158	175	231	200	322
146	218	176	249	200	304
148	276	177	285	207	300
149	218	178	243	270	252
152	302	178	251	290	322
153	168	180	279	300	353
157	302	180	189	320	349
157	289	181	153	328	388
160	277	185	316		

2.20 *Represente graficamente os dados da Tabela 2.50 sobre o consumo diário médio de energia elétrica em uma residência.*

Tabela 2.50: Consumo diário médio de energia para o Exercício 2.20

Mês	Consumo (kWh)	Mês	Consumo (kWh)
Jan/00	6,41	Ago/00	8,00
Fev/00	14,00	Set/00	8,21
Mar/00	15,64	Out/00	8,90
Abr/00	11,63	Nov/00	10,50
Mai/00	9,43	Dez/00	10,34
Jun/00	8,45	Jan/01	8,93
Jul/00	8,10		

2.21 *Na Tabela 2.51 temos as freqüências acumuladas do número de sinistros por apólice de seguro do ramo Automóveis. Complete a tabela, calculando as freqüencias simples absolutas e relativas e também as freqüências acumuladas relativas.*

Tabela 2.51: Número de sinistros por apólice, para o Exercício 2.21

Número de sinistros	Número de apólices
0	2913
1	4500
2	4826
3	4928
4	5000

2.22 *Em uma pesquisa realizada em uma cidade, entrevistou-se uma amostra de moradores. Dentre as variáveis pesquisadas estava a classe de renda e o jornal preferido, dentre os três maiores da cidade. Os dados constam da Tabela 2.52. Construa a tabela de freqüências relativas apropriada e utilize um gráfico para ilustrá-la.*

Tabela 2.52: Jornais preferidos

Jornal	Classe social			
	Pobre	Média inferior	Média	Alta
A	15	27	44	22
B	20	27	26	11
C	13	18	14	3

2.23 *Considere os dados da tabela a seguir, onde temos a opinião de 228 indivíduos norte-americanos sobre o aborto, segundo a afiliação partidária. Os dados constam da Tabela 2.53. Construa a tabela de freqüências relativas apropriada e utilize um gráfico para ilustrá-la.*

Tabela 2.53: Opinião sobre aborto

Opinião sobre aborto	Partido	
	Democrata	Republicano
A favor	78	34
Neutro	8	5
Contra	37	66

Capítulo 3

Medidas Estatísticas

3.1 Introdução

A redução dos dados através de tabelas de freqüências ou gráficos é um dos meios disponíveis para ilustrar o comportamento de um conjunto de dados. No entanto, muitas vezes queremos resumir ainda mais esses dados, apresentando um único valor que seja "representativo" do conjunto original. Como, ao fazermos isso, perdemos informação sobre a variabilidade dos dados, é importante que se tenha também um valor que "represente" a dispersão dos dados.

Neste capítulo estudaremos algumas medidas de posição, que são medidas que sintetizam, em um único valor, o conjunto original, e também algumas medidas de dispersão. Para completar a caracterização da distribuição univariada dos dados, serão dadas algumas medidas de assimetria e curtose. A covariância e o coeficiente de correlação serão também apresentados como medidas de associação linear entre variáveis quantitativas.

3.2 Medidas de posição

3.2.1 Média aritmética simples

No nosso dia-a-dia, o conceito de média é bastante comum, quando nos referimos, por exemplo, à altura média dos brasileiros, à temperatura média dos últimos anos, etc.

Definição 3.1 *Dado um conjunto de n observações $x_1, x_2, \ldots, x_n$, a* ***média aritmética simples*** *é definida como*

$$\boxed{\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i}. \tag{3.1}$$

Como exemplo, considere os dados da Tabela 2.2, referentes às notas de duas turmas; a nota média para a turma A é

$$\overline{x}_A = \frac{5 + 8 + 8 + \cdots + 9 + 8}{42} = \frac{252}{42} = 6,0$$

e para a turma B é

$$\overline{x}_B = \frac{6 + 3 + 4 + \cdots + 5 + 5}{38} = \frac{206}{38} = 5,4211$$

Como os dados originais representam número de questões corretas em um teste de múltipla escolha, a média representa o número médio de questões corretas. Em geral, *a média de um conjunto de dados tem a mesma unidade dos dados originais.*

Nas Figuras 3.1 e 3.2 temos os gráficos ou diagramas de pontos[1] representando as notas de ambas as turmas. Nessas figuras, a setinha indica a média do conjunto de dados. A interpretação física da média aritmética é que ela representa o centro de gravidade da distribuição; nas figuras, ela é o ponto de equilíbrio, indicado pela seta.

Figura 3.1: Gráfico de pontos das notas da Turma A

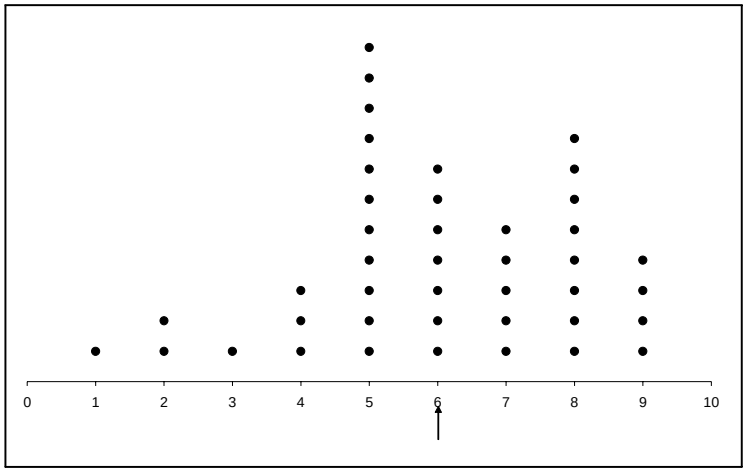
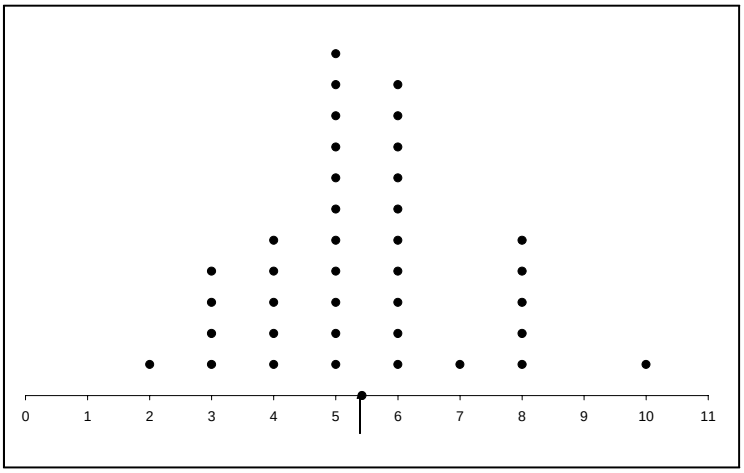


Figura 3.2: Gráfico de pontos das notas da Turma B



Considerando os dados sobre número de empregados das ULs industriais do Rio de Janeiro apresentados na Tabela 2.9, a tabela de freqüências sem perda de informação, dada na Tabela 3.1, nos auxilia no cálculo de várias medidas descritivas.

Como há vários valores repetidos, podemos calcular a média como

$$\overline{x} = \frac{12 \times 5 + 18 \times 6 + 12 \times 7 + \cdots + 1 \times 503 + 1 \times 705 + 1 \times 837}{12 + 18 + 12 + \cdots + 1 + 1 + 1} = \frac{6774}{171} = 39,614$$

[1] Esses gráficos são construídos usando-se uma "pilha de pontos" para representar as freqüências de cada valor. Note que os pontos têm que estar equi-espaçados.

Tabela 3.1: Número de empregados - RJ

Num. Emp.	Freq.	Num. Emp	Freq.	Num. Emp	Freq.	Num. Emp	Freq.
5	12	19	4	35	1	73	2
6	18	20	4	36	1	80	1
7	12	21	4	37	1	98	1
8	10	22	2	38	2	110	1
9	9	23	2	40	3	120	1
10	7	24	2	45	1	204	1
11	8	25	1	47	1	216	1
12	6	26	3	49	1	274	1
13	6	27	1	51	1	351	1
14	5	28	2	53	1	461	1
15	4	29	1	54	1	503	1
16	3	30	2	55	2	705	1
17	4	32	1	56	2	837	1
18	3	33	2	72	1		

Note que o valor da média aritmética é um valor tal que, se substituíssemos todos os dados por ela, isto é, se todas as observações fossem iguais à média aritmética, a soma total seria igual à soma dos dados originais. Então, a média aritmética é uma forma de se distribuir o total observado pelos n elementos, de modo que todos tenham o mesmo valor. Considere os seguintes dados fictícios referentes aos salários de 5 funcionários de uma firma: $136, 210, 350, 360, 2500$. O total da folha de pagamentos é 3236, havendo um salário bastante alto, discrepante dos demais. A média para esses dados é 647,20. Se todos os 5 funcionários ganhassem esse salário, a folha de pagamentos seria a mesma e todos teriam o mesmo salário.

3.2.2 Moda

Analisando os gráficos de pontos das notas das turmas A e B, podemos ver que, em ambas as turmas, a nota que mais se repete é a nota 5. Esse é o conceito de *moda*.

Definição 3.2 *A **moda** de uma distribuição ou conjunto de dados, que representaremos por x^*, é o valor que mais se repete, ou seja, o valor mais freqüente.*

Podemos ter distribuições amodais (todos os valores ocorrem o mesmo número de vezes), unimodais (uma moda), bimodais (duas modas), etc. Para as notas das turmas A e B, os diagramas de pontos das Figuras 3.1 e 3.2 nos permitem ver rapidamente que

$$x^*_A = x^*_B = 5;$$

para a Tabela 3.1, temos uma única moda $x^* = 6$.

3.2.3 Mediana

Considere os seguintes conjuntos de dados (hipotéticos) referentes aos salários de empregados de duas firmas, medidos em alguma unidade monetária (u.m.):

- Firma 1: 300 350 600 700 800

- Firma 2: 300 350 600 700 3000

Para a firma 1, o salário médio é $\overline{x} = 550$ e para a firma 2, $\overline{x} = 990$ u.m.. A diferença entre os 2 conjuntos é o salário mais alto: na firma 1, os salários são mais homogêneos, enquanto na firma 2 o maior salário é muito mais alto que os restantes. A conseqüência disso é que o salário médio para a firma 2 fica muito influenciado por esse valor alto, fazendo com que a média não seja um bom "representante" dos salários. Esse exemplo ilustra um fato geral sobre a média aritmética: ela é muito influenciada por *valores discrepantes* (em inglês, *outliers*), isto é, valores muito grandes (ou muito pequenos) que sejam distintos da maior parte dos dados. Nesses casos é necessário utilizar uma outra medida de posição para representar o conjunto; uma medida possível é a *mediana*.

Definição 3.3 *Seja $x_1, x_2, \ldots, x_n$ um conjunto de n observações e seja $x_{(i)}, i = 1, \ldots, n$ o conjunto das observações ordenadas, de modo que $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Então, a **mediana** Q_2 é definida como o valor tal que 50% das observações são menores que ela e 50% são maiores que ela.*

Para efeito de cálculo, valem as seguintes regras:

$$\begin{aligned} n \text{ ímpar}: &\qquad Q_2 = x_{\left(\frac{n+1}{2}\right)} \\ n \text{ par}: &\qquad Q_2 = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} \end{aligned} \tag{3.2}$$

Dessa definição, podemos ver que a mediana é o valor central dos dados.

Voltando às notas das turmas A e B, na turma A temos 42 notas e na turma B, 38 notas.Assim, a mediana da turma A é a média da 21ª e da 22ª notas; para a turma B, é a média da 19ª e da 20ª notas. Os diagramas de pontos facilitam a identificação da mediana:

$$\begin{aligned} Q_{2,A} &= \frac{x_{(19)} + x_{(20)}}{2} = \frac{5+5}{2} = 5 \\ Q_{2,A} &= \frac{x_{(21)} + x_{(22)}}{2} = \frac{6+6}{2} = 6 \end{aligned}$$

Para os dados da Tabela 3.1, como o número de observações é ímpar, $n = 171$, temos que (note que $\frac{171+1}{2} = 86$) :

$$Q_{2,RJ} = x_{(86)} = 13.$$

Compare esse valor com a respectiva média $\overline{x}_{RJ} = 39,58$: os valores altos "puxam" a média para cima.

3.2.4 Separatrizes

A mediana é um caso particular de um conjunto mais amplo de medidas estatísticas, chamadas *separatrizes*.

Definição 3.4 *A **separatriz de ordem** p é um valor tal que pelo menos p% dos dados são menores do que ele e pelo menos $(1-p)$% são maiores.*

As separatrizes mais comuns são os *quartis*, *decis* e *percentis*, cujos fatores de divisão são 4, 10 e 100. Mais precisamente, existem 3 quartis, 9 decis e 99 percentis. Os quartis serão representados pela letra Q e são eles:

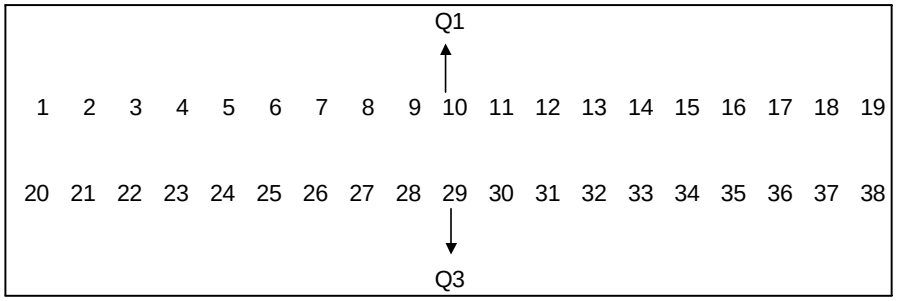
- primeiro quartil Q_1 : deixa pelo menos 25% das observações abaixo dele e pelo menos 75% acima;
- segundo quartil Q_2 : deixa pelo menos 50% das observações abaixo dele e pelo menos 50% acima; é a mediana;
- terceiro quartil Q_3 : deixa pelo menos 75% das observações abaixo dele e pelo menos 25% acima.

Os decis serão representados pela letra D e os percentis pela letra P; assim, por exemplo:

- o terceiro decil D_3 deixa pelo menos 30% das observações abaixo e pelo menos 70% acima;
- o quinto decil e o 50º percentil são a mediana;
- o octagésimo percentil deixa pelo menos 80% das observações abaixo e pelo menos 20% acima.

No cálculo das separatrizes quase sempre será necessário algum procedimento de arredondamento e aproximação. Para os quartis, podemos adotar o seguinte procedimento: depois de calculada a mediana, considere as duas partes dos dados, a parte abaixo da mediana e a parte acima da mediana, em ambos os casos excluindo a mediana. O primeiro quartil é calculado como a mediana da parte abaixo da mediana original e o terceiro quartil é calculado como a mediana da parte acima da mediana original. Consideremos as notas da turma B: temos 38 observações e a mediana é a média dos valores centrais (19ª e 20ª observações). Então, as duas partes consistem nas 19 observações inferiores e nas 19 observações superiores, respectivamente (ver Figura 3.3). Como 19 é um número ímpar, a mediana é o valor central, ou seja, a 10ª observação; então, o primeiro quartil é a 10ª observação e o terceiro quartil é a 10ª observação contada a partir da posição 19, ou seja, é calculado como a observação de posição ordenada $19 + 10 = 29$. Resulta $Q_{1,B} = 4$ e $Q_{3,B} = 6$.

Figura 3.3: Cálculo dos quartis - $n = 38$



Analogamente, para a turma A, que tem 42 notas, o primeiro e terceiro quartis são calculados como

$$\begin{aligned} Q_{1,A} &= x_{(11)} = 5 \\ Q_{3,A} &= x_{(21+11)} = x_{(32)} = 8 \end{aligned}$$

Para os dados do Rio de Janeiro, o número de observações é ímpar (171) e a mediana é a observação de posição ordenada 86; excluindo essa observação, restam 85 observações abaixo e 85 acima. Com 85 observações, a mediana é a observação de posição 43. Logo, o primeiro quartil é a

observação original de posição ordenada 43 e o terceiro quartil é a observação original de posição ordenada $86 + 43 = 129$. Resulta $Q_1 = 8$ e $Q_3 = 26$.

O primeiro decil para as notas da turma B pode ser calculado como (note que $\frac{38}{10} = 3,8$) :

$$D_{1,B} = x_{(4)} = 4$$

e o quarto decil como (note que $4 \times \frac{38}{10} = 15,2 \simeq 16$) :

$$D_{4,B} = x_{(16)} = 5$$

Todos esses arredondamentos são necessários mas um pouco arbitrários; não existe uma regra definida para tratar as diversas situações, por isso trabalha-se com a definição "pelo menos $p\%$ abaixo e $(1-p)\%$ acima". Uma boa prática é manter a simetria das separatrizes; por exemplo, o primeiro e o terceiro quartis são simétricos com relação à mediana, assim como o primeiro e o nono decis. Então, se o primeiro decil deixa 5 observações abaixo, por exemplo, o nono decil deve deixar 5 observações acima. Para as notas da turma B, o nono decil deve ser calculado como

$$D_9 = x_{(38-3)} = x_{(35)} = 8$$

e o sexto decil, simétrico ao quarto, como

$$D_6 = x_{(38-15)} = x_{(23)} = 6$$

3.2.5 Média aritmética ponderada

Em algumas situações, os números que queremos sintetizar têm graus de importância diferentes. Por exemplo, o Índice Nacional de Preços ao Consumidor (INPC) é calculado com base nos Índices de Preço ao Consumidor (IPC) de diversas regiões metropolitanas do Brasil mas a importância dessas regiões é diferente. Uma das variáveis que as diferencia é a população residente.

Nesse tipo de situação, em vez de se usar a média aritmética simples, usa-se a *média aritmética ponderada*, que será representada por $\overline{x}_p$.

Definição 3.5 *A **média aritmética ponderada** de números $x_1, x_2, \ldots, x_n$ com pesos $\rho_1, \rho_2, \ldots, \rho_n$ é definida como*

$$\overline{x}_p = \frac{\rho_1 x_1 + \rho_2 x_2 + \cdots + \rho_n x_n}{\rho_1 + \rho_2 + \ldots + \rho_n} = \frac{\sum_{i=1}^{n} \rho_i x_i}{\sum_{i=1}^{n} \rho_i}.$$

Se definimos

$$\omega_i = \frac{\rho_i}{\sum_{j=1}^{n} \rho_j}$$

então a média aritmética ponderada pode ser reescrita como

$$\overline{x}_p = \sum_{i=1}^{n} \omega_i x_i \tag{3.3}$$

onde $\sum_{i=1}^{n} \omega_i = 1$.

Note que a média aritmética simples é um caso particular da média aritmética ponderada, onde todas as observações têm o mesmo peso e, portanto, peso igual a $\frac{1}{n}$.

Para a construção do Índice Nacional de Preços ao Consumidor - INPC, o peso de cada índice regional é definido pela população residente urbana, conforme dados da Tabela 3.2. Os pesos em porcentagem aí apresentados representam a participação da população residente urbana da região metropolitana no total da população residente urbana das 11 regiões metropolitanas pesquisadas. O índice geral é dado pela média ponderada:

$$\begin{aligned} \text{INPC}_{09/03} &= 0,0572 \times 0,98 + 0,0620 \times 0,36 + 0,0721 \times 0,85 + 0,1030 \times 1,82 + \\ &\quad 0,1102 \times 0,69 + 0,1080 \times 0,39 + 0,2679 \times 0,94 + 0,0709 \times 0,51 + \\ &\quad 0,0766 \times 0,36 + 0,0502 \times 0,67 + 0,0219 \times 1,34 \\ &= 0,82382 \end{aligned}$$

Tabela 3.2: Estrutura básica de ponderação regional para cálculo do INPC

Área Geográfica	Peso (%)	IPC - Set/03
Belém	5,72	0,98
Fortaleza	6,20	0,36
Recife	7,21	0,85
Salvador	10,30	1,82
Belo Horizonte	11,02	0,69
Rio de Janeiro	10,80	0,39
São Paulo	26,79	0,94
Curitiba	7,09	0,51
Porto Alegre	7,66	0,36
Goiânia	5,02	0,67
Distrito Federal	2,19	1,34
INPC - Geral		0,82

Fonte: IBGE

3.2.6 Média geométrica

Definição 3.6 *A **média geométrica** de n valores positivos $x_1, x_2, \ldots, x_n$ é definida como*

$$\overline{x}_g = \sqrt[n]{x_1 \times x_2 \times \cdots \times x_n}\,. \tag{3.4}$$

Em Demografia, a média geométrica pode ser usada para se estimar a população de uma determinada localidade num ano t_x. É usual que os países realizem Censos Demográficos a cada 10 anos, quando, então, é obtido o número de residentes no país. Para estimar a população em algum ano entre dois censos, podemos usar a média geométrica, desde que se suponha que a taxa de crescimento entre os 2 censos seja constante. Sejam P_0 a população no 1º censo, realizado na data t_0, P_N a população do 2º censo realizado na data t_N e P_x a população que se quer estimar na data t_x ($t_0 < t_x < t_N$). O crescimento da população entre os dois censos é igual a $\frac{P_N}{P_0}$; se a taxa de crescimento é constante igual a r, isso significa que ao fim do primeiro período a população é igual a

$$P_1 = P_0 + P_0 \times r = P_0 \times (1 + r)$$

Ao final do segundo período,

$$P_2 = P_1 + P_1 \times r = P_1 \times (1+r) = P_0 \times (1+r) \times (1+r) = P_0 \times (1+r)^2$$

Ao final do último período,

$$P_N = P_0 \times (1+r)^N$$

Logo,

$$\frac{P_N}{P_0} = (1+r)^N \qquad \Rightarrow \qquad r = \sqrt[N]{\frac{P_N}{P_0}} - 1$$

A população em qualquer período x entre os censos, então, é dada por

$$P_x = P_0 \times (1+r)^x = P_0 \times \left(\sqrt[N]{\frac{P_N}{P_0}} \right)^x$$

Lembrando que

$$\sqrt[n]{x} = x^{\frac{1}{n}}$$

podemos escrever

$$P_x = P_0 \times \left(\frac{P_N}{P_0} \right)^{\frac{x}{N}} = (P_0)^{1-\frac{x}{N}} \times (P_N)^{\frac{x}{N}} = \sqrt[N]{(P_0)^{N-x} (P_N)^x}$$

Vê-se, então, que P_x é uma média geométrica de $N - x$ valores iguais a P_0 e de x valores iguais a P_N. Em particular, se o instante de tempo x é o período central, isto é, $x = \frac{N}{2}$, então

$$P_x = \sqrt[N]{(P_o)^{\frac{N}{2}} (P_N)^{\frac{N}{2}}} = \left[(P_o)^{\frac{N}{2}} (P_N)^{\frac{N}{2}} \right]^{\frac{1}{N}} = \sqrt{P_0 \times P_N}$$

a média geométrica de P_0 e P_N. De acordo com os Censos Demográficos realizados pelo IBGE, a população (recenseada) do estado do Rio de Janeiro em 1/9/1980 era de 11.489.797 habitantes e em 1/9/1991 de 12.783.761. Admitindo um crescimento geométrico constante, uma estimativa para a população desse estado em 1985 pode ser calculada como

$$\begin{aligned} P_{85} &= \sqrt[11]{(11.489.797)^6 (12.783.761)^5} = 11.489.797 \times \left(\sqrt[11]{\frac{12.783.761}{11.489.797}} \right)^5 = \\ &= 11.489.797 \times (1,009748691)^5 = 12.060.876 \end{aligned}$$

3.2.7 Média harmônica

Considere o seguinte exemplo: uma pessoa viaja num fim de semana do Rio de Janeiro para São Paulo, dirigindo seu próprio carro. Na ida, ela desenvolve uma velocidade média de 70km/h mas, na volta, por estar o tráfego na via Dutra mais tranqüilo, ela desenvolve uma velocidade média de 90km/h. Qual a velocidade média para a viagem completa? Para responder esta pergunta, temos que lembrar que a velocidade média é dada pela razão entre a distância percorrida e o tempo gasto para percorrê-la. Para simplificar, suponhamos que a distância entre as duas cidades seja de 450 km. Então, a distância total percorrida é de $2 \times 450 = 900$km. Por outro lado, o tempo gasto na ida foi de $\frac{450}{70}$h e na volta, $\frac{450}{90}$h. Logo, a velocidade média para a viagem completa é de

$$\overline{x}_h = \frac{2 \times 450}{\frac{450}{70} + \frac{450}{90}} = \frac{2}{\frac{1}{70} + \frac{1}{90}} = \frac{1}{\frac{\frac{1}{70} + \frac{1}{90}}{2}}.$$

Essa última expressão nos leva à definição de *média harmônica*.

Definição 3.7 *A **média harmônica** de um conjunto de valores $x_1, x_2, \ldots, x_n$ é o inverso da média aritmética dos inversos dos valores, isto é:*

$$\boxed{\overline{x}_h = \frac{1}{\frac{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}{n}} = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}}. \tag{3.5}$$

Analisando essa expressão, conclui-se que a velocidade média para a viagem completa é a média harmônica das velocidades médias desenvolvidas na ida e na volta.

3.2.8 Algumas propriedades das medidas de posição

Média

1. A média aritmética de um conjunto de valores $x_1, x_2, \ldots, x_n$ é maior ou igual ao menor dos números e menor ou igual ao maior dos números. Em outras palavras, a média aritmética está compreendida entre o menor e o maior valor dos dados. Para demonstrar esse fato, sejam $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ as observações ordenadas, isto é, $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Temos que:

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \geq \frac{x_{(1)} + x_{(1)} + \cdots + x_{(1)}}{n} = x_{(1)}$$

e

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \leq \frac{x_{(n)} + x_{(n)} + \cdots + x_{(n)}}{n} = x_{(n)}$$

Logo,

$$x_{\min} \leq \overline{x} \leq x_{\max} \tag{3.6}$$

Como já visto, o conceito de média aritmética simples corresponde ao conceito de centro de gravidade estudado em Física. Baseado nesse fato, é fácil verificar as seguintes propriedades da média.

2. Somando-se um mesmo valor a cada um dos elementos de um conjunto de observações, a média aritmética simples fica somada desse valor. Note que essa operação equivale a um deslocamento constante e rígido dos dados (uma translação), o que desloca igualmente o centro de gravidade. Para demonstrar formalmente esse resultado, seja $x_1, x_2, \ldots, x_n$ um conjunto de observações, às quais somamos uma constante k, isto é, criamos uma nova série de observações $y_1, y_2, \ldots, y_n$ definida por $y_i = x_i + k, \forall k = 1, \ldots, n$. Então

$$\begin{aligned} \overline{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(x_i + k) = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} k = \\ &= \overline{x} + \frac{1}{n}(k + k + \cdots + k) = \overline{x} + \frac{1}{n}(nk) = \overline{x} + k. \end{aligned}$$

Resumindo:

$$y_i = x_i + k \Rightarrow \overline{y} = \overline{x} + k \tag{3.7}$$

3. Multiplicando cada observação por uma mesma constante não nula k, a média aritmética simples fica multiplicada por essa constante. Definindo a nova série de observações por $y_i = kx_i$, temos que

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} kx_i = k \times \frac{1}{n}\sum_{i=1}^{n} x_i = k\overline{x}.$$

Resumindo:

$$y_i = kx_i \Rightarrow \overline{y} = k\overline{x}$$

Mediana e moda

Para a mediana e a moda, valem as mesmas propriedades acima. Embora mais trabalhosas para demonstrar formalmente, elas são intuitivas: ao se somar a mesma constante, a relação de ordenação entre os dados não se altera; logo, a mediana fica somada da mesma constante. O valor mais freqüente dos novos dados, isto é, a moda, passa a ser a moda original mais a constante. Se multiplicamos por uma constante positiva, a ordenação não se altera; logo a nova mediana é a mediana original multiplicada pela constante. Se a constante é negativa, há uma inversão na ordenação mas os valores centrais se mantêm.

Relação entre as médias aritmética, geométrica e harmônica

Para um conjunto de observações não-negativas, valem as seguintes relações:

$$\overline{x}_h \leq \overline{x}_g \leq \overline{x}. \tag{3.8}$$

Vamos provar esse resultado para o caso em que temos apenas 2 observações não negativas, isto é, $x_1 \geq 0$ e $x_2 \geq 0$. As médias aritmética, geométrica e harmônica, neste caso, são:

$$\overline{x} = \frac{x_1 + x_2}{2} \qquad \overline{x}_g = \sqrt{x_1 x_2} \qquad \overline{x}_h = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}$$

Quaisquer que sejam x_1, x_2 temos que

$$\begin{aligned}
(x_1 - x_2)^2 &\geq 0 \Leftrightarrow \\
x_1^2 + x_2^2 - 2x_1x_2 &\geq 0 \Leftrightarrow \\
x_1^2 + x_2^2 &\geq 2x_1x_2 \Leftrightarrow \\
x_1^2 + x_2^2 + 2x_1x_2 &\geq 4x_1x_2 \Leftrightarrow \\
(x_1 + x_2)^2 &\geq 4x_1x_2 \Leftrightarrow \\
\frac{(x_1 + x_2)^2}{4} &\geq x_1x_2 \Leftrightarrow \\
\frac{(x_1 + x_2)}{2} &\geq \sqrt{x_1x_2} \Leftrightarrow \\
\overline{x} &\geq \overline{x}_g
\end{aligned} \tag{3.9}$$

A penúltima desigualdade foi obtida extraindo-se a raiz quadrada de ambos os lados; essa operação é possível pois os números envolvidos são todos não-negativos.

O resultado provado acima é válido para quaisquer dois números positivos; em particular, vale para $y_1 = \dfrac{1}{x_1}$ e $y_2 = \dfrac{1}{x_2}$. Para esses números temos que

$$
\begin{aligned}
\overline{y} &\geq \overline{y}_g \Leftrightarrow \\
\frac{y_1 + y_2}{2} &\geq \sqrt{y_1 y_2} \Leftrightarrow \\
\frac{\frac{1}{x_1} + \frac{1}{x_2}}{2} &\geq \sqrt{\frac{1}{x_1} \times \frac{1}{x_2}} \Leftrightarrow \\
\frac{\frac{1}{x_1} + \frac{1}{x_2}}{2} &\geq \frac{1}{\sqrt{x_1 x_2}} \Leftrightarrow \\
\frac{1}{\frac{\frac{1}{x_1} + \frac{1}{x_2}}{2}} &\leq \frac{1}{\frac{1}{\sqrt{x_1 x_2}}} \Leftrightarrow \\
\frac{2}{\frac{1}{x_1} + \frac{1}{x_2}} &\leq \sqrt{x_1 x_2} \Leftrightarrow \\
\overline{x}_h &\leq \overline{x}_g
\end{aligned} \tag{3.10}
$$

A demonstração desse resultado para o caso geral (n qualquer) é dada no Anexo 1 deste capítulo.

3.2.9 Exercícios resolvidos da Seção 3.2

1. Considere os dados da Tabela 3.3 abaixo, onde temos as notas dos 50 alunos, já analisadas no Exercício 2.3 do capítulo anterior. Calcule a nota média, a nota modal, a nota mediana, o primeiro e terceiro quartis e o oitavo decil.

Tabela 3.3: Notas de 50 alunos em um teste múltipla escolha para o Exercício Resolvido 1 da Seção 3.2

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

Fonte: Dados hipotéticos

Solução:

Para facilitar a solução do exercício, consideremos a distribuição de freqüências dada na Tabela 3.4 abaixo.

A nota média é

$$
\begin{aligned}
\overline{x} &= \frac{3 \times 1 + 4 \times 2 + 4 \times 3 + 5 \times 4 + 8 \times 5 + 8 \times 6 + 5 \times 7 + 5 \times 8 + 5 \times 9 + 3 \times 10}{50} = \\
&= \frac{281}{50} = 5,62
\end{aligned}
$$

A distribuição é bimodal, com as modas sendo as notas 5 e 6. Como temos um número par de observações, a mediana é a média dos valores centrais, que ocupam as posições 25 e 26.

Tabela 3.4: Notas de 50 alunos para a solução do Exercício Resolvido 1 da Seção 3.2

Nota x_i	Freqüência simples n_i	Freqüência acumulada N_i	$n_i \times x_i$
1	3	3	3
2	4	7	8
3	4	11	12
4	5	16	20
5	8	24	40
6	8	32	48
7	5	37	35
8	5	42	40
9	5	47	45
10	3	50	30
Total	50		281

Das freqüências acumuladas, podemos ver que esses valores são ambos iguais a 6 (note que as observações $x_{(25)}$ a $x_{(32)}$ são todas iguais a 6) , ou seja::

$$Q_2 = \frac{x_{(25)} + x_{(26)}}{2} = \frac{6+6}{2} = 6$$

O cálculo dos outros quartis se faz notando que a mediana é a média dos valores centrais e, portanto, as duas partes dos dados são formadas por 25 observações. Para $n = 25$, a mediana é a observação de posição ordenada 13 e, assim, o primeiro quartil é a observação original de posição ordenada 13 e o terceiro quartil é a observação original de posição ordenada $25 + 13 = 38$, ou seja:

$$Q_1 = x_{(13)} = 4 \qquad\qquad Q_3 = x_{(38)} = 8$$

Para o cálculo do oitavo decil, note que $\frac{50}{10} = 5$ e, portanto, o oitavo decil deve deixar pelo menos $8 \times 5 = 40$ observações abaixo dele; assim, podemos calcular o oitavo decil como $D_8 = x_{(41)} = 8$.

2. Considere os dados referentes à população dos municípios mineiros com mais de 50.000 habitantes da Tabela 3.5. Calcule os três quartis, o primeiro e o nono decis.

   Solução:

   Temos 60 municípios; logo, a mediana é a média das observações de posições ordenadas 30 (Araxá) e 31 (Itajubá), ou seja:

$$Q_2 = \frac{78997 + 84315}{2} = 81656$$

   Excluída a mediana, que não é uma observação real, restam 30 observações acima e abaixo. Logo, o primeiro quartil é a média das observações de posições ordenadas 15 (Ouro Preto) e 16 (João Monlevade) e o terceiro quartil é a média das observações de posições ordenadas $30 + 15 = 45$ (Teófilo Otoni) e $30 + 16 = 46$ (Ibirité), ou seja:

$$\begin{aligned} Q_1 &= \frac{66277 + 66690}{2} = 66.483,5 \\ Q_3 &= \frac{129.424 + 133.044}{2} = 131.234 \end{aligned}$$

Tabela 3.5: População dos municípios de MG com mais de 50.000 habitantes, para o Exercício 2

Município	População	Município	População	Município	População
Leopoldina	50.097	Timóteo	71.478	Varginha	108.998
Pirapora	50.300	Pará de Minas	73.007	Barbacena	114.126
três Pontas	51.024	Patrocínio	73.130	Sabará	115.352
São Francisco	51.497	Paracatu	75.216	Patos de Minas	123.881
Pedro Leopoldo	53.957	Vespasiano	76.422	Teófilo Otoni	129.424
Ponte Nova	55.303	Itaúna	76.862	Ibirité	133.044
S.Seb.do Paraíso	58.335	Caratinga	77.789	Poços de Caldas	135.627
Janaúba	61.651	S.João del Rei	78.616	Divinópolis	183.962
Formiga	62.907	Lavras	78.772	Sete Lagoas	184.871
Januária	63.605	Araxá	78.997	Santa Luzia	184.903
Cataguases	63.980	Itajubá	84.135	Ipatinga	212.496
Nova Lima	64.387	Ubá	85.065	Ribeirão das Neves	246.846
Viçosa	64.854	Ituiutaba	89.091	Gov.Valadares	247.131
Três Corações	65.291	Muriaé	92.101	Uberaba	252.051
Ouro Preto	66.277	Passos	97.211	Betim	306.675
João Monlevade	66.690	Cor. Fabriciano	97.451	Montes Claros	306.947
Alfenas	66.957	Itabira	98.322	Juiz de Fora	456.796
Manhuaçu	67.123	Araguari	101.974	Uberlândia	501.214
Curvelo	67.512	Cons.Lafaiete	102.836	Contagem	538.017
Unaí	70.033	Pouso Alegre	106.776	Belo Horizonte	2.238.526

Fonte: IBGE - Censo Demográfico 2000

Para o cálculo dos decis, note que $\frac{60}{10} = 6$. O primeiro decil deve deixar pelo menos 6 observações abaixo e, assim, $D_1 = x_{(7)} = 58.335$ e o nono decil, por simetria, é $D_9 = x_{(54)} = 252.051$.

3. Vamos fazer uma comparação entre as médias aritmética e geométrica através de um exemplo de matemática financeira elementar.

   No regime de capitalização simples (juros simples), apenas o capital inicial rende juros. Já no regime de capitalização composta (juros compostos), os rendimentos incorporados ao capital inicial, em cada período, também rendem juros no período seguinte. Vamos analisar os resultados da aplicação de umcapital inicial C_0 durante um perído de n meses, com taxas de juros $i_1, i_2, i_3, \ldots, i_n$ que, para simplificar, vamos supor que não estejam em forma percentual.

   Capitalização Simples:

   Como os juros só incidem sobre o capital inicial, em cada mês o valor dos juros J_t (em u.m.) é calculado como

   $$J_t = C_0 \times i_t$$

   e ao final do período o montante é

   $$C_t = C_{t-1} + J_t$$

   Então, para o primeiro mês temos

   $$\begin{aligned} J_1 &= C_0 \times i_1 \\ C_1 &= C_0 + J_1 = C_0 + C_0 \times i_1 \end{aligned}$$

Para o segundo mês,

$$\begin{aligned} J_2 &= C_0 \times i_2 \\ C_2 &= C_1 + J_2 = C_1 + C_0 \times i_2 = C_0 + C_0 \times i_1 + C_0 \times i_2 = C_0 + C_0(i_1 + i_2) \end{aligned}$$

Para o terceiro mês,

$$\begin{aligned} J_3 &= C_0 \times i_3 \\ C_3 &= C_2 + J_2 = C_2 + C_0 \times i_3 = C_0 + C_0(i_1 + i_2) + C_0 \times i_3 = C_0 + C_0(i_1 + i_2 + i_3) \end{aligned}$$

Continuando com esses cálculos, obtemos para o $n^{\underline{o}}$ mês

$$\begin{aligned} J_n &= C_0 \times i_n \\ C_n &= C_{n-1} + J_n = C_0 + C_0(i_1 + i_2 + i_3 + \cdots + i_{n-1}) + C_0 \times i_n = \\ &= C_0 + C_0(i_1 + i_2 + i_3 + \cdots + i_{n-1} + i_n) \\ &= C_0 + C_0 \sum_{t=1}^{n} i_t \end{aligned} \tag{3.11}$$

Vamos considerar agora o conceito de taxa média de juros. A taxa média de juros é uma taxa constante que leva ao mesmo capital final, isto é, obtemos o mesmo rendimento mas, a cada mês, a taxa de juros é a mesma. Da Eq. 3.11 vemos que, para obter o mesmo capital final a uma taxa constante i, temos que ter

$$\begin{aligned} C_0 + C_0 \sum_{t=1}^{n} i_t &= C_0 + C_0 \sum_{t=1}^{n} i \Rightarrow \\ \sum_{t=1}^{n} i_t &= \sum_{t=1}^{n} i \Rightarrow \\ \sum_{t=1}^{n} i_t &= ni \Rightarrow \\ i &= \frac{1}{n} \sum_{t=1}^{n} i_t \end{aligned} \tag{3.12}$$

ou seja, a taxa de juros média tem que ser igual à média aritmética das taxas mensais.

A título de ilustração, considere as seguintes taxas de juros mensais: $i_1 = 2,5\%$; $i_2 = 3,8\%$; $i_3 = 4,5\%$; $i_4 = 4,9\%$; $i_5 = 6,2\%$ e $i_6 = 7,8\%$; suponha também que uma pessoa tenha um capital inicial de $C_0 = 150$ u.m. (unidades monetárias). Na Tabela 3.6 resumimos os resultados da aplicação com as taxas mensais variáveis e com a taxa mensal média. Note que a taxa média é dada por

$$i = \frac{2,5 + 3,8 + 4,5 + 4,9 + 6,2 + 7,8}{6} = 4,95\%$$

Capitalização Composta:

No regime de capitalização composta, os juros incidem também sobre os rendimentos mensais; assim, o valor dos juros para cada mês é dado por

$$J_t = C_{t-1} \times i_t$$

Tabela 3.6: Cálculo dos juros em regime de capitalização simples

Mês	Taxa de juros variável			Taxa de juros constante		
	Juros		Montante	Juros		Montante
	Taxa (%)	Valor (u.m.)	(u.m.)	Taxa (%)	Valor (u.m.)	(u.m.)
1	2,5	3,75	153,75	4,95	7,425	157,425
2	3,8	5,70	159,45	4,95	7,425	164,850
3	4,5	6,75	166,20	4,95	7,425	172,275
4	4,9	7,35	173,55	4,95	7,425	179,700
5	6,2	9,30	182,85	4,95	7,425	187,125
6	7,8	11,70	194,55	4,95	7,425	194,550

e o montante é

$$C_t = C_{t-1} + J_t$$

Então, para o primeiro mês temos

$$\begin{aligned} J_1 &= C_0 \times i_1 \\ C_1 &= C_0 + J_1 = C_0 + C_0 \times i_1 = C_0(1+i_1) \end{aligned}$$

Para o segundo mês,

$$\begin{aligned} J_2 &= C_1 \times i_2 \\ C_2 &= C_1 + J_2 = C_1 + C_1 \times i_2 = C_1(1+i_2) = C_0(1+i_1)(1+i_2) \end{aligned}$$

Para o terceiro mês,

$$\begin{aligned} J_3 &= C_2 \times i_3 \\ C_3 &= C_2 + J_3 = C_2 + C_2 \times i_3 = C_2(1+i_3) = C_0(1+i_1)(1+i_2)(1+i_3) \end{aligned}$$

Continuando com esses cálculos, obtemos para o $n^{\underline{o}}$ mês:

$$\begin{aligned} J_n &= C_{n-1} \times i_n \\ C_n &= C_{n-1} + J_n = C_{n-1} + C_{n-1} \times i_n = C_{n-1}(1+i_n) = C_0(1+i_1)(1+i_2)(1+i_3)\cdots(1+i_n) \end{aligned} \quad (3.13)$$

Para obter o mesmo capital final a uma taxa constante i,a taxa média constante i tem que ser tal que

$$\begin{aligned} (1+i_1)(1+i_2)(1+i_3)\cdots(1+i_n) &= (1+i)(1+i)(1+i)\cdots(1+i) \Rightarrow \\ (1+i_1)(1+i_2)(1+i_3)\cdots(1+i_n) &= (1+i)^n \Rightarrow \\ 1+i &= \sqrt[n]{(1+i_1)(1+i_2)(1+i_3)\cdots(1+i_n)} \end{aligned} \quad (3.14a)$$

Então, a taxa comum é calculada como uma média geométrica, não das taxas mensais, mas dos valores $1+i$, chamados relativos. O "1" aparece exatamente por que os juros incidem sobre o capital do mês anterior. Logo, a taxa comum, em forma percentual, é

$$i = \sqrt[n]{(1+i_1)(1+i_2)(1+i_3)\cdots(1+i_n)} - 1$$

No nosso exemplo, essa taxa comum é

$$\begin{aligned} i &= 100 \times \left(\sqrt[6]{1,025 \times 1,038 \times 1,045 \times 1,049 \times 1,062 \times 1,078} - 1 \right) = \\ &= 100 \times \left(\sqrt[6]{1.33523059526495} - 1 \right) \approx 4,936372179 \end{aligned}$$

Na Tabela 3.7 temos os cálculos para as taxas variáveis e constantes. Os valores estão com um número excessivo de casas decimais para ilustrar a exatidão dos resultados.

Tabela 3.7: Cálculo dos juros em regime de capitalização composta

Mês	Taxa de juros variável			Taxa de juros constante		
	Juros		Montante	Juros		Montante
	Taxa (%)	Valor (u.m.)	(u.m.)	Taxa (%)	Valor (u.m.)	(u.m.)
1	2,5	3,750000000	153,750000000	4,936372179	7,404558269	157,404558269
2	3,8	5,842500000	159,592500000	4,936372179	7,770074823	165,174633092
3	4,5	7,181662500	166,774162500	4,936372179	8,153634635	173,328267728
4	4,9	8,171933963	174,946096463	4,936372179	8,556128387	181,884396115
5	6,2	10,846657981	185,792754443	4,936372179	8,978490728	181,884396115
6	7,8	14,491834847	200,284589290	4,936372179	9,421702447	200,284589290

4. Um capital inicial de 1200 u.m. foi aplicado em um regime de capitalização composta, rendendo ao final de um trimestre (3 meses) juros de 126,52. Qual foi a taxa média mensal?

   Solução:

   Note que da equação (3.13) obtemos

$$\frac{C_n}{C_0} = \left(1 + \frac{i_1}{100}\right) \times \cdots \times \left(1 + \frac{i_n}{100}\right)$$

   Em termos da taxa média comum,

$$\frac{C_n}{C_0} = \left(1 + \frac{i}{100}\right)^n \Rightarrow \sqrt[n]{\frac{C_n}{C_0}} = \left(1 + \frac{i}{100}\right) \Rightarrow i = 100 \times \left(\sqrt[n]{\frac{C_n}{C_0}} - 1 \right)$$

   No exercício, o capital final é 1326,52 e, portanto, a variação nos três meses é

$$\frac{C_3}{C_0} = \frac{1326,52}{1200,00} = 1,105433333$$

   Logo, a taxa média mensal é

$$i = 100 \times \left(\sqrt[3]{1,105433333} - 1 \right) = 3,3976937\%$$

5. Resolva o exercício anterior para um regime de capitalização simples.

   Solução:

   Da equação (3.11), obtemos que

$$\frac{C_n - C_0}{C_0} = \frac{i_1 + \cdots + i_n}{100}$$

Em termos da média comum,

$$\frac{C_n - C_0}{C_0} = \frac{i + \cdots + i}{100}$$

ou seja,

$$i = 100 \times \frac{\left(\frac{C_n - C_0}{C_0}\right)}{n} = 100 \times \frac{\frac{C_n - C_0}{C_0}}{n}$$

Note que $\frac{C_t - C_0}{C_0}$ é a variação relativa; dividindo pelo número de períodos, obtemos a variação média. No nosso exercício,

$$i = 100 \times \frac{\left(\frac{1326,52}{1200,00} - 1\right)}{3} = 100 \times \frac{0,105433333}{3} = 3,5144444$$

6. No ano de 2004, uma escola estadual recebeu, em cada trimestre, uma verba de R$500,00 para comprar folhas de cartolina. A compra é sempre feita na primeira semana do trimestre e os preços de cada folha de cartolina estão na Tabela 3.8 abaixo.

Tabela 3.8: Preço da cartolina para o Exercício Resolvido 6 da Seção 3.2

Trimestre	Preço (R$)
jan-mar	0,35
abr-jun	0,45
jul-set	0,50
out-dez	0,52

Qual o preço médio da folha de cartolina pago pela escola no ano de 2004?

Solução:

O preço médio é calculado como a razão entre o valor total gasto e o número total de folhas de cartolina compradas. Assim, o preço médio não é a média dos preços unitários 0,35, 0,45, 0,50 e 0,52, porque as quantidades compradas variaram a cada trimestre. O valor total gasto foi de 4×500. O número de folhas de cartolina compradas em cada trimestre foi:

$$q_I = \frac{500}{0,35} \qquad q_{II} = \frac{500}{0,45} \qquad q_{III} = \frac{500}{0,50} \qquad q_{IV} = \frac{500}{0,52}$$

Logo, o preço médio é

$$p_m = \frac{4 \times 500}{\frac{500}{0,35} + \frac{500}{0,45} + \frac{500}{0,50} + \frac{500}{0,52}} = \frac{4}{\frac{1}{0,35} + \frac{1}{0,45} + \frac{1}{0,50} + \frac{1}{0,52}} = 0,4443$$

e, portanto, é a média harmônica dos preços unitários em cada ano.

7. Em 1973, em certa localidade, o custo da alimentação aumentou 58%, os aluguéis subiram 47% e o transporte subiu 49%. Se um assalariado gasta 35% do seu salário com alimentação, 25% com aluguel e 12% com transporte, qual o aumento percentual dos gastos dessa pessoa com esses três itens?

Solução:

Para cada unidade do seu salário, a pessoa gastava, antes do aumento, 0,35 com alimentação, 0,25 com aluguel e 0,12 com transporte. Depois do aumento, ela passa a ter uma despesa adicional de $0,58 \times 0,35 = 0,203$ com alimentação, $0,47 \times 0,25 = 0,118$ com aluguel e $0,49 \times 0,12 = 0,059$ com transporte, o que totaliza $0,203 + 0,118 + 0,059 = 0,380$. Então, para cada unidade do seu salário, ela tem um aumento de 0,38 nos gastos com esses três itens, ou seja,um aumento de 38%. Note que ela já gastava $0,35+,025 + 0,12 = 0,72 = 72\%$ do salário com esses três itens. Agora, ela passa a gastar, só com esses itens, 110%, ou seja, mais do que ganha! Esse cálculo corresponde a uma média ponderada das taxas de aumento, onde os fatores de ponderação correspondem às parcelas do salário gastas com os diferentes itens.

8. No mês do dissídio de uma determinada categoria, uma firma deu um aumento de 20% a todos os seus funcionários. Se, antes do aumento, o salário médio dos funcionários era de R$780,00, qual será o novo salário médio? No Natal seguinte, a firma dá um abono de R$50,00 para todos os funcionários. Se a firma tem 22 funcionários, qual o valor da folha de pagamentos neste mês de dezembro?

   Solução:

   Quando todos os funcionários têm aumento de 20%, isso significa que cada salário fica multiplicado por 1,2, ou seja, o salário de cada funcionário é o salário antigo mais 20%. Ao multiplicar todos os números por uma mesma constante, a média fica multiplicada por essa constante. Então, o salário médio fica multiplicado por 1,2, ou seja, passa a ser $1,2 \times 780,00 = R\$936,00$. Como a firma tem 22 funcionários, a folha de pagamentos passa a ser $22 \times 936 = R\$20.592,00$. No Natal, os salários de todos os funcionários ficam somados de R$50,00; logo o salário médio também fica somado de 50,00 e a folha de pagamentos será de $20.592 + 22 \times 50 = R\$21.692,00$.

3.2.10 Exercícios propostos da Seção 3.2

3.1 *O peso médio dos jogadores de um time de futebol é de 81 kg. Se nenhum pesa menos do que 72 kg, quantos podem pesar 95 kg?*

3.2 *Os dados a seguir representam o número de apólices de seguro que um corretor conseguiu vender em cada um de seus 20 primeiros dias em um emprego novo: 2, 4, 6, 3, 2, 1, 4, 3, 5, 2, 1, 1, 4, 0, 2, 2, 5, 2, 2, 1. Calcule a média, a mediana e a moda desses dados, interpretando os resultados obtidos.*

3.3 *O NASDAQ Composite Index dá o preço médio de ações comuns negociadas no balcão, isto é, fora das bolsas de valores. Em 1991, a capitalização média das companhias no índice NASDAQ foi de US$ 80 milhões, e a capitalização mediana foi de US$ 20 milhões. (A capitalização de uma companhia é o valor total de mercado de suas ações). Explique por que a capitalização média é muito superior à capitalização mediana.*

3.4 *Considere os dados da Tabela 2.28 do Exercício 2.6 do Capítulo 2. Sabendo que o total de empregados das 80 empresas é de 517.462, calcule o número médio e o número mediano de empregados das empresas. Interprete a diferença obtida entre a média e a mediana.*

3.5 *Na Tabela 3.9 temos as variações mensais do IPCA (Índice de Preços ao Consumidor Amplo) calculadas pelo IBGE para o ano de 1999. Segundo previsões feitas pelo então secretário-adjunto de Política Econômica (Folha de São Paulo, 11/12/1999), o IPCA no ano de 1999 deveria ficar abaixo de 9%. Para que as previsões do secretário se confirmassem, qual deveria ter sido a taxa máxima do IPCA em dezembro?*

Tabela 3.9: IPCA 1999 para o Exercício 3.6 do Capítulo 3

Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov
0,70	1,05	1,10	0,56	0,30	0,19	1,09	0,56	0,31	1,19	0,95

Fonte: IBGE

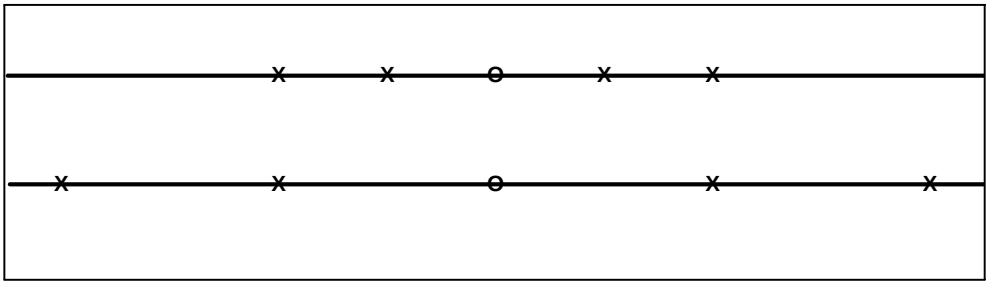
3.6 *A contagem de bactérias em uma cultura aumentou de 2500 para 9200 em três dias. Qual o acréscimo percentual diário médio?*

3.3 Medidas de dispersão

3.3.1 Amplitude

Considere os seguintes conjuntos de dados, representados esquematicamente na Figura 3.4 abaixo.

Figura 3.4: Conjuntos de dados com mesma média e mediana



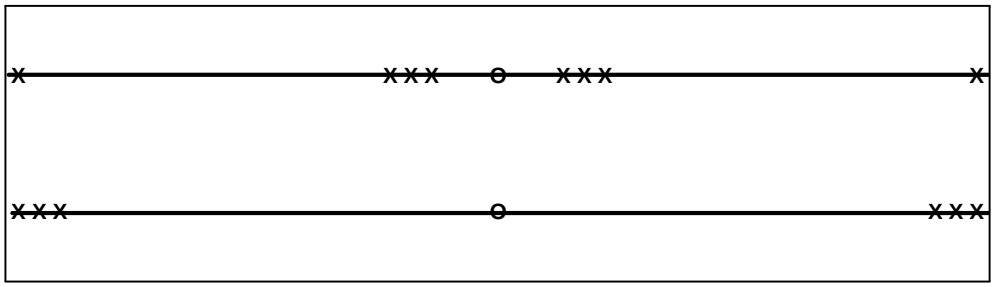
Da interpretação física da média aritmética como centro de gravidade e da definição de mediana, segue que ambos os conjuntos têm a mesma média e a mesma mediana, representadas pelo símbolo **O**. No entanto, esses conjuntos têm características diferentes e ao sintetizá-los por uma dessas medidas, essa característica se perderá. Tal característica é a *dispersão* dos dados: no primeiro conjunto, os dados estão mais concentrados em torno da média do que no segundo conjunto. Como poderíamos "medir" essa dispersão? Uma primeira idéia é considerar a *amplitude* dos dados, que é, como já visto, a diferença entre o maior e o menor valor.

Definição 3.8 *A **amplitude** de um conjunto de dados é a distância entre o maior valor e o menor valor.*

$$\Delta_{total} = V_{\max} - V_{\min}. \tag{3.15}$$

A amplitude tem a mesma unidade dos dados, mas ela tem algumas limitações, conforme ilustrado na Figura 3.5. Aí os dois conjuntos têm a mesma média, a mesma mediana e a mesma amplitude mas essas medidas não conseguem caracterizar o fato de a distribuição dos valores entre o mínimo e o máximo ser diferente nos dois conjuntos. A limitação da amplitude também fica patente pelo fato de ela se basear em apenas duas observações, independentemente do número total de observações.

Figura 3.5: Conjuntos de dados com mesma amplitude



3.3.2 Desvio médio absoluto

Uma maneira de medir a dispersão dos dados seria considerar os tamanhos dos desvios $x_i - \overline{x}$ de cada observação em relação à média. Note nas figuras acima que quanto mais disperso o conjunto de dados, maiores esses desvios tendem a ser. Para obter uma medida-resumo, isto é, um único número, poderíamos somar esses desvios, ou seja, considerar a seguinte medida:

$$D = \sum_{i=1}^{n}(x_i - \overline{x}). \tag{3.16}$$

Vamos desenvolver tal fórmula, usando as propriedades de somatório e a definição da média amostral.

$$\begin{aligned} D &= \sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\overline{x} = \sum_{i=1}^{n} x_i - n\overline{x} = \\ &= \sum_{i=1}^{n} x_i - n \times \frac{1}{n}\sum_{i=1}^{n} x_i = 0. \end{aligned}$$

Ou seja: essa medida, que representa a soma dos desvios em relação à média, é sempre nula, não importa o conjunto de dados! Logo, ela não serve para diferenciar quaisquer conjuntos!

Vamos dar uma explicação intuitiva para esse fato, que nos permitirá obter correções para tal fórmula. Ao considerarmos as diferenças entre cada valor e o valor médio, obtemos valores negativos e positivos, pois, pela definição de média, sempre existem valores menores e maiores que a média; esses valores positivos e negativos, ao serem somados, se anulam.

Bom, se o problema está no fato de termos valores positivos e negativos, por que não trabalhar com o valor absoluto das diferenças? De fato, esse procedimento nos leva à definição de *desvio médio absoluto.*

Definição 3.9 *O desvio médio absoluto de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definido por*

$$DMA = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}| \tag{3.17}$$

onde as barras verticais representam o valor absoluto ou módulo.

Note que nesta definição estamos trabalhando com o desvio médio, isto é, tomamos a média dos desvios absolutos. Isso evita interpretações equivocadas, pois, se trabalhássemos apenas com a soma

dos desvios absolutos, um conjunto com um número maior de observações tenderia a apresentar um resultado maior para a soma devido apenas ao fato de ter mais observações. Esta situação é ilustrada com os seguintes conjuntos de dados:

- Conjunto 1: $\{1, 3, 5\}$
- Conjunto 2: $\left\{1, \frac{5}{3}, 3, \frac{13}{3}, 5\right\}$

Para os dois conjuntos, $\overline{x} = 3$ e para o conjunto 1

$$\sum_{i=1}^{3} |x_i - \overline{x}| = |1-3| + |3-3| + |5-3| = 4$$

e para o conjunto 2

$$\sum_{i=1}^{5} |x_i - \overline{x}| = |1-3| + \left|\frac{5}{3} - 3\right| + |3-3| + \left|\frac{13}{3} - 3\right| + |5-3| = \frac{20}{3} = 6,667.$$

Então, o somatório para o segundo conjunto é maior mas o desvio absoluto médio é o mesmo para ambos; de fato, para o primeiro conjunto temos

$$DMA = \frac{4}{3}$$

e para o segundo conjunto

$$DMA = \frac{\frac{20}{3}}{5} = \frac{4}{3};$$

ao dividirmos o somatório pelo número de observações, compensamos o fato de o segundo conjunto ter mais observações que o primeiro.

O desvio médio absoluto tem a mesma unidade dos dados.

3.3.3 Variância e desvio padrão

Considerar o valor absoluto das diferenças $(x_i - \overline{x})$ é uma das maneiras de se contornar o fato de que $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$. No entanto, a função módulo tem a desvantagem de ser não diferenciável no ponto zero. Outra possibilidade de correção, com propriedades matemáticas mais adequadas, é considerar o quadrado das diferenças. Isso nos leva à definição de *variância*.

Definição 3.10 *A variância*[2] *de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definida por*

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 . \tag{3.18}$$

Suponhamos que os valores x_i representem os pesos, em quilogramas, de um conjunto de pessoas. Então, o valor médio $\overline{x}$ representa o peso médio dessas pessoas e sua unidade também é quilogramas, o mesmo acontecendo com as diferenças $(x_i - \overline{x})$. Ao elevarmos essas diferenças ao quadrado, passamos a ter a variância medida em quilogramas ao quadrado, uma unidade que não tem interpretação física. Uma solução é tomar a raiz quadrada da variância.

[2] É possível definir a variância usando o divisor $n-1$ no lugar de n; essa é a diferença entre os conceitos de variância populacional e variância amostral, que será mais relevante num segundo curso de inferência estatística.

Definição 3.11 *O desvio padrão de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definido por*

$$\sigma = \sqrt{Variância} = \sqrt{\sigma^2} \tag{3.19}$$

Consideremos a expressão 3.18 que define a variância; desenvolvendo o quadrado obtemos:

$$\begin{aligned}\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\overline{x} + \overline{x}^2\right) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\sum_{i=1}^{n}2\overline{x}x_i + \frac{1}{n}\sum_{i=1}^{n}\overline{x}^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\overline{x}\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) + \frac{1}{n}n\overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\overline{x}^2 + \overline{x}^2\end{aligned}$$

ou seja

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2 \tag{3.20}$$

Essa forma de reescrever a variância facilita quando os cálculos têm que ser feitos à mão ou em calculadoras menos sofisticadas, pois o número de cálculos envolvidos é menor. Note que ela nos diz que a variância é a "*média dos quadrados menos o quadrado da média*".

A título de ilustração, vamos calcular a variância das notas das turmas A e B. Como visto na Seção 3.2.1, a nota média da turma A é dados é $\overline{x}_A = 6,0$ e da turma B é $\overline{x}_B = 5,4211$. Usando a fórmula 3.20 para calcular a variância, tem-se que

$$\begin{aligned}\sigma_A^2 &= \left[\frac{1}{42}(5^2 + 8^2 + 8^2 + \cdots + 9^2 + 8^2)\right] - (6,0)^2 \\ &= \frac{1674}{42} - 36 = 3,8571\end{aligned}$$

e o desvio padrão é

$$\sigma_A = \sqrt{3,8571} = 1,964$$

Para a turma B temos que

$$\begin{aligned}\sigma_B^2 &= \left[\frac{1}{38}(6^2 + 3^2 + 4^2 + \cdots + 5^2 + 5^2)\right] - (5,4211)^2 \\ &= \frac{1224}{38} - 29,38781163 = 2,8227\end{aligned}$$

e o desvio padrão é

$$\sigma_B = \sqrt{2.8227} = 1,6801.$$

O desvio médio absoluto para a turma A é:

$$DMA_B = \frac{1}{42}(|5-6| + |8-6| + |8-6| + \cdots + |9-6| + |8-6|) = \frac{66}{42} = 1,5714$$

Na definição da variância, tomam-se os desvios com relação à média $x_i - \overline{x}$. A escolha da média como o ponto de referência, além de resultar em propriedades estatísticas interessantes, tem a seguinte característica:

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \min_a \sum_{i=1}^{n}(x_i - a)^2$$

Isto é, qualquer que seja o ponto de referência a, a variância σ^2 resulta no menor valor da função $f(a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2$. A demonstração se faz usando os métodos clássicos de cálculo.

$$\begin{aligned} f'(a) &= 0 \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n}(-1)2(x_i - a) = 0 \Leftrightarrow \sum_{i=1}^{n}(x_i - a) = 0 \Leftrightarrow \\ \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} a &= 0 \Leftrightarrow \sum_{i=1}^{n} x_i - na = 0 \Leftrightarrow a = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} \end{aligned}$$

O ponto $a = \overline{x}$ corresponde a um mínimo pois $f''(a) = -\frac{2}{n}\sum_{i=1}^{n}(0-1) = -\frac{2}{n}(-n) = 2 > 0$.

3.3.4 Propriedades das medidas de dispersão

1. Somando-se uma mesma constante a todas as observações, as medidas de dispersão não se alteram. Antes de demonstrar formalmente o resultado, note que ele tem que ser verdadeiro pois, sendo medidas de dispersão, ao se somar uma constante aos dados não estamos alterando a dispersão dos mesmos. A demonstração formal é a seguinte: seja $y_i = x_i + k$.

   (a) Amplitude

   $$\begin{aligned} y_{\max} &= x_{\max} + k \\ y_{\min} &= x_{\min} + k \\ \Delta_y &= y_{\max} - y_{\min} = (x_{\max} + k) - (x_{\min} + k) = x_{\max} - x_{\min} = \Delta_x \end{aligned}$$

   (b) Desvio médio absoluto
   Vimos que $\overline{y} = \overline{x} + k$. Logo

   $$DMA_y = \frac{1}{n}\sum_{i=1}^{n}|y_i - \overline{y}| = \frac{1}{n}\sum_{i=1}^{n}|(x_i + k) - (\overline{x} + k)| = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}| = DMA_x$$

   (c) Variância e desvio padrão

   $$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2;$$

   Como $\overline{y} = \overline{x} + k$, resulta que

   $$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}[(x_i + k) - (\overline{x} + k)]^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sigma_x^2.$$

   Naturalmente, se a variância não se altera, o mesmo ocorre com o desvio padrão.

   Resumindo:

   $$y_i = x_i + k \Rightarrow \begin{cases} \Delta_y = \Delta_x \\ DMA_y = DMA_x \\ \sigma_y^2 = \sigma_x^2 \\ \sigma_y = \sigma_x \end{cases} \tag{3.21}$$

2. Vamos ver o que acontece quando multiplicamos os dados por uma constante não nula. Seja $y_i = kx_i$; nesse caso, $\overline{y} = k\overline{x}$.

(a) Amplitude

Vamos considerar os casos em que $k > 0$ e $k < 0$ separadamente. Se $k > 0$

$$\begin{aligned} x_{(1)} &= x_{\min} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)} = x_{\max} \Rightarrow \\ kx_{(1)} &= kx_{\min} \leq kx_{(2)} \leq \cdots \leq kx_{(n-1)} \leq kx_{(n)} = kx_{\max} \Rightarrow \\ & \left\{ \begin{array}{l} y_{\min} = kx_{\min} \\ y_{\max} = kx_{\max} \end{array} \right. \end{aligned}$$

e, portanto,

$$\Delta_y = y_{\max} - y_{\min} = kx_{\max} - kx_{\min} = k\Delta_x = |k|\,\Delta_x$$

Se $k < 0$

$$\begin{aligned} x_{(1)} &= x_{\min} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)} = x_{\max} \Rightarrow \\ kx_{(1)} &= kx_{\min} \geq kx_{(2)} \geq \cdots \geq kx_{(n-1)} \geq kx_{(n)} = kx_{\max} \Rightarrow \\ & \left\{ \begin{array}{l} y_{\max} = kx_{\min} \\ y_{\min} = kx_{\max} \end{array} \right. \end{aligned}$$

e, portanto,

$$\Delta_y = y_{\max} - y_{\min} = kx_{\min} - kx_{\max} = -k\,(x_{\max} - x_{\min}) = -k\Delta_x = |k|\,\Delta_x$$

(b) Desvio médio absoluto

$$DMA_y = \frac{1}{n}\sum_{i=1}^{n} |y_i - \overline{y}| = \frac{1}{n}\sum_{i=1}^{n} |kx_i - k\overline{x}| = \frac{1}{n}\sum_{i=1}^{n} |k|\,|x_i - \overline{x}| = |k|\,DMA_x$$

(c) Variância

$$\sigma_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(kx_i - k\overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} k^2(x_i - \overline{x})^2 = k^2\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right] = k^2\sigma_X^2.$$

(d) Desvio padrão

$$\sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{k^2\sigma_X^2} = |k|\sqrt{\sigma_X^2} = |k|\,\sigma_X.$$

Resumindo:

$$y_i = kx_i \Rightarrow \left\{ \begin{array}{c} \Delta_y = |k|\ \Delta_x \\ DMA_y = |k|\ DMA_x \\ \sigma_y^2 = k^2\sigma_x^2 \\ \sigma_y = |k|\ \sigma_x \end{array} \right. \tag{3.22}$$

3. Das definições, resulta que todas as medidas de dispersão são não negativas!

$$\begin{aligned} \Delta &\geq 0 \\ DMA &\geq 0 \\ \sigma^2 &\geq 0 \\ \sigma &\geq 0 \end{aligned}$$

3.3.5 Coeficiente de variação

Considere a seguinte situação: uma fábrica de ervilhas comercializa seu produto em embalagens de 300 gramas e em embalagens de um quilo. Para efeitos de controle do processo de enchimento das embalagens, sorteia-se uma amostra de 10 embalagens de cada uma das máquinas de enchimento, obtendo-se os seguintes resultados:

$$300g \qquad \begin{cases} \overline{x} = 295g \\ \sigma = 5g \end{cases}$$

$$1000g \qquad \begin{cases} \overline{x} = 995g \\ \sigma = 5g \end{cases}$$

Em qual das duas situações a variabilidade parece ser maior? Ou seja, em qual das duas máquinas parece haver um problema mais sério? Note que em ambos os casos há uma dispersão de 5g em torno da média mas 5g em 1000g é menos preocupante que 5g em 300g.

Como um exemplo mais extremo, um desvio padrão de 10 unidades em um conjunto cuja observação típica é 100 é muito diferente de um desvio padrão de 10 unidades em um conjunto cuja observação típica é 10000. Surge, assim, a necessidade de uma medida de *dispersão relativa*, que permita comparar, por exemplo, esses dois conjuntos. Uma dessas medidas é o *coeficiente de variação.*

Definição 3.12 *Dado um conjunto de observações $x_1, x_2, \ldots, x_n$, o coeficiente de variação (CV) é definido como a razão entre o desvio padrão dos dados e sua média, ou seja:*

$$CV = \frac{\sigma}{\overline{x}}\,. \tag{3.23}$$

Note que, como o desvio padrão e a média são ambos medidos na mesma unidade dos dados originais, o coeficiente de variação é *adimensional*. Este fato permite comparações entre conjuntos de dados diferentes, medidos em unidades diferentes.

No exemplo das latas de ervilha, os coeficientes de variação para as embalagens oriundas das 2 máquinas são

$$\begin{aligned} 300g \qquad CV &= \frac{5}{300} \times 100 = 1,6667 \\ 1000g \qquad CV &= \frac{5}{1000} \times 100 = 0,5 \end{aligned}$$

o que confirma a nossa observação anterior: a variabilidade na máquina de 300g é relativamente maior.

3.3.6 Intervalo interquartil

Quando introduzimos o conceito de mediana, chamamos a atenção para o fato de que a média é bastante afetada pela presença de valores discrepantes. Como a variância e o desvio padrão dependem da média, eles também ficarão afetados. Torna-se necessário, então, definir uma medida de dispersão que seja "robusta" para outliers. Uma dessas medidas é o intervalo interquartil.

Definição 3.13 *O **intervalo interquartil** é a distância entre o terceiro e o primeiro quartis, isto é:*

$$IQ = Q_3 - Q_1\,. \tag{3.24}$$

Pela definição dos quartis, resulta que, entre os valores Q_1 e Q_3, sempre temos 50% das observações. Assim, quanto maior for o intervalo interquartil, mais dispersos serão os dados.

3.3.7 Exemplo: escores padronizados

Considere os dois conjuntos de dados abaixo, que representam as notas em Estatística e Cálculo dos alunos de uma determinada turma.

Aluno	1	2	3	4	5	6	7	8	9
Estatística	6	4	5	7	8	3	5	5	7
Cálculo	7	8	9	10	6	7	8	9	5

As notas médias nas duas disciplinas são:

- Estatística: $\overline{x}_E = \dfrac{6+4+5+7+8+3+5+5+7}{9} = 5,56$
- Cálculo: $\overline{x}_C = \dfrac{7+8+9+10+6+7+8+9+5}{9} = 7,67$

As variâncias são:

- Estatística: $\sigma_E^2 = \dfrac{6^2+4^2+5^2+7^2+8^2+3^2+5^2+5^2+7^2}{9} - (5,56)^2 = 2,2469$
- Cálculo: $\sigma_C^2 = \dfrac{7^2+8^2+9^2+10^2+6^2+7^2+8^2+9^2+5^2}{9} - (7,67)^2 = 2,2222$

Analisando os dois conjuntos de notas, pode-se ver que o aluno 1 tirou 6 em Estatística e o aluno 5 tirou 6 em Cálculo. No entanto, a nota máxima em Estatística foi 8, enquanto que em Cálculo a nota máxima foi 10. Assim, o 6 em Estatística "vale mais" que o 6 em Cálculo, no sentido de que ele está mais próximo da nota máxima. Uma forma de medir tal fato é considerar a posição relativa de cada aluno no grupo. Para isso, o primeiro passo consiste em comparar a nota do aluno com a média do grupo, considerando o seu afastamento da média. Se x_i é a nota do aluno, passamos a trabalhar com $x_i - \overline{x}$. O segundo passo consiste em padronizar a escala, já que no primeiro conjunto as notas variam de 3 a 8 e no segundo, de 5 a 10. Essa padronização da escala se faz dividindo os desvios pelo desvio padrão do conjunto, o que nos dá o escore padronizado:

$$z_i = \frac{x_i - \overline{x}}{\sigma_X}. \tag{3.25}$$

O desvio padrão das notas de Estatística é $\sigma_E = 1,49897$ e das notas de Cálculo é $\sigma_C = 1,49071$. Na tabela a seguir temos os escores padronizados; podemos ver aí que o escore relativo à nota 6 em Estatística é maior que o escore da nota 6 em Cálculo, indicando que a primeira "vale mais" que a segunda.

Aluno	1	2	3	4	5	6	7	8	9
Estatística	0,297	−1,038	−0,371	0,964	1,631	−1,705	−0,371	−0,371	0,964
Cálculo	−0,447	0,224	0,894	1,565	−1,118	−0,447	0,224	0,894	−1,789

Usando as propriedades da média e do desvio padrão pode-se ver que que os escores padronizados têm média zero e desvio padrão (e, portanto, variância) um.

Os coeficientes de variação das notas de Estatística e Cálculo são

$$CV_E = \frac{1,49897}{5,56} = 0,2696$$

$$CV_C = \frac{1,49071}{7,67} = 0,1944$$

indicando uma maior variabilidade relativa nas notas de Estatística.

3.3.8 Exercícios resolvidos da Seção 3.3

1. Considere novamente as notas dos 50 alunos, reproduzidas na Tabela 3.10. Calcule o desvio padrão, o desvio médio absoluto e o intervalo interquartil das notas.

Tabela 3.10: Notas de 50 alunos em um teste múltipla escolha para o Exercício Resolvido 1 da Seção 3.2

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

Fonte: Dados hipotéticos

Solução:

Como visto na seção anterior, a nota média é $\overline{x} = 5,62$. O desvio médio absoluto e a variância utilizam os desvios de cada observação em torno da média. Como temos vários valores repetidos, podemos usar o mesmo tipo de procedimento para calcular a média, resumido na Tabela 3.11:

Tabela 3.11: Notas de 50 alunos para a solução do Exercício Resolvido 1 da Seção 3.3

Nota x_i	Freqüência simples n_i	Freqüência acumulada N_i	Desvio $x_i - \overline{x}$	Desvio absoluto $n_i \times \lvert x_i - \overline{x}\rvert$	Desvio ao quadrado $n_i \times (x_i - \overline{x})^2$
1	3	3	-4,62	13,86	64,0332
2	4	7	-3,62	14,48	52,4176
3	4	11	-2,62	10,48	27,4576
4	5	16	-1,62	8,10	13,1220
5	8	24	-0,62	4,96	3,0752
6	8	32	0,38	3,04	1,1552
7	5	37	1,38	6,90	9,5220
8	5	42	2,38	11,90	28,3220
9	5	47	3,38	16,90	57,1220
10	3	50	4,38	13,14	57,5532
Total	50			103,76	313,7800

O desvio médio absoluto é calculado como

$$D_m = \frac{3 \times |1 - 5,62| + 4 \times |2 - 5,62| + \cdots + 3 \times |10 - 5,62|}{50} = \frac{103,76}{50} = 2,0752$$

e a variância como

$$\sigma^2 = \frac{3 \times (1 - 5,62)^2 + 4 \times (2 - 5,62)^2 + \cdots + 3 \times (10 - 5,62)^2}{50} = \frac{313,78}{50} = 6,2756$$

e, portanto, o desvio padrão é

$$\sigma = \sqrt{6,2756} = 2,505115$$

Note que a soma dos desvios em torno da média é, de fato, zero, ou seja:

$$\sum (x_i - \overline{x}) = 3 \times (1 - 5,62) + 4 \times (2 - 5,62) + \cdots + 3 \times (10 - 5,62) = -51,88 + 51,88 = 0$$

Na seção anterior, calculamos

$$Q_1 = x_{(13)} = 4 \qquad Q_3 = x_{(38)} = 8$$

Logo, o intervalo interquartil é

$$IQ = 8 - 4 = 4$$

2. Durante 13 dias, uma pessoa anotou o tempo de espera na fila do ônibus, quando se dirigia ao trabalho. Os valores obtidos são (em minutos): 15, 10, 2, 17, 6, 8, 3, 10, 2, 9, 5, 9, 13. Calcule o desvio padrão do tempo de espera. Não esqueça de indicar a unidade!

   Solução:

   A média dos dados é $\overline{x} = \dfrac{15+10+2+17+6+8+3+10+2+9+5+9+13}{13} = \dfrac{109}{13} = 8,3846$ minutos. Usando a fórmula (3.20), a variância é

$$\begin{aligned}\sigma^2 &= \frac{15^2+10^2+2^2+17^2+6^2+8^2+3^2+10^2+2^2+9^2+5^2+9^2+13^2}{13} - \left(\frac{109}{13}\right)^2 = \\ &= \frac{1187}{13} - \frac{109^2}{13^2} = \frac{1187\times 13 - 109^2}{13^2} = \frac{15431-11881}{169} = \frac{3550}{169} = 21,005917\end{aligned}$$

   e o desvio padrão é

$$\sigma = \sqrt{21,005917} = 4,58322 \text{ minutos.}$$

3. Uma pesquisa sobre consumo de gasolina deu os seguintes valores para a quilometragem percorrida por três marcas de carro (de mesma classe), em cinco testes com um tanque de 40 litros.

Carro A	400	397	401	389	403
Carro B	403	401	390	378	395
Carro C	399	389	403	387	401

   Compare o desempenho dos três carros.

   Solução:

   O consumo médio e o desvio padrão do consumo nos 5 testes estão resumidos na tabela abaixo:

Carro	Média	Desvio padrão	Coeficiente de variação
A	398,00	4,89898	0,01231
B	393,40	8,95768	0,02277
C	395,80	6,52380	0,01648

   O carro A tem o melhor desempenho, não só porque a média é maior, mas também porque apresenta a menor variabilidade relativa (CV). O carro B certamente tem o pior desempenho.

3.3.9 Exercícios propostos da Seção 3.3

3.7 *Calcule todas as medidas de dispersão para os dados do Exercício 3.2 do Capítulo 3, referentes ao número de apólices vendidas por um corretor de seguros.*

3.8 *O Departamento de Proteção ao Meio Ambiente dos Estados Unidos exige que os fabricantes de automóveis indiquem, para cada modelo de carro, o consumo de combustível por milha, na cidade e na rodovia. Dá-se, na Tabela 3.12, o consumo de combustível na rodovia (milhas por galão, MPG) para 30 modelos médios e grandes de carros do ano de 1994.*

(a) Construa o gráfico ramo-e-folhas e comente suas principais características.

(b) Calcule a mediana e o intervalo interquartil IQ.

(c) O governo taxa os "bebedores de combustível " (carros com baixa milhagem) de acordo com a seguinte regra: todos os modelos com consumo abaixo da mediana por mais de 1,5 vezes o intervalo interquartil serão taxados. Segundo esses dados, quais os modelos taxados?

Tabela 3.12: Consumo de gasolina para 30 modelos para o Exercício 3.8 do Capítulo 3

Modelo	MPG	Modelo	MPG
BMW 740i	23	Hyundai Sonata	27
Buick Century	31	Infinity Q45	22
Buick LeSabre	28	Lexus LS400	23
Buick Park Avenue	27	Lincoln Continental	26
Buick Regal	29	Lincoln Mark VIII	25
Buick Roadmaster	25	Mazda 626	31
Cadillac DeVille	25	Mazda 929	24
Chevrolet Caprice	26	Mercedes-Benz S320	24
Chevrolet Lumina	29	Mercedes-Bens S420	20
Chrysler Concorde	28	Nissan Maxima	26
Chrysler New Yorker	26	Rolls-Royce Silver Stone	15
Dodge Spirit	27	Saab 900	26
Fort LTD	25	Saab 9000	27
Ford Taurus	29	Toyota Camry	28
Ford Thunderbird	26	Volvo 850	26

3.9 *Para se estudar o desempenho de 2 companhias corretoras de ações, selecionou-se de cada uma delas amostras das ações negociadas. Para cada ação selecionada, computou-se a porcentagem de lucro apresentada durante um período fixado de tempo, obtendo-se os dados abaixo. Com base nos coeficientes de variação, qual companhia teve melhor desempenho?*

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

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

3.10 *Faça uma análise comparativa dos dados apresentados no ramo-e-folhas da Figura 2.20 do Cap. 2, utilizando medidas estatísticas apropriadas.*

3.4 Momentos

Os momentos são quantidades numéricas calculadas a partir de um conjunto de dados, usadas também para descrever resumidamente a distribuição. A definição de momentos é bastante genérica e abrange diversos tipos de medidas.

Definição 3.14 *Seja $x_1, x_2, \ldots, x_n$ um conjunto de n observações. Então, o* ***momento natural*** *de ordem r, representado por m'_r é definido como:*

$$m'_r = \frac{x_1^r + x_2^r + \cdots + x_n^r}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i^r \,. \tag{3.26}$$

Das definições de média e variância dadas em (3.1) e (3.20), seguem as seguintes equivalências:

$$\overline{x} = m'_1$$

$$\sigma^2 = m'_2 - (m'_1)^2 \,.$$

Muitas vezes é interessante considerar os momentos com relação a uma origem que não o zero, sendo a média dos dados uma das origens bastante utilizada.

Definição 3.15 *Seja $x_1, x_2, \ldots, x_n$ um conjunto de n observações. Então, o momento de ordem r centrado na média é definido como:*

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^r \,. \tag{3.27}$$

Da definição de variância, segue que $\sigma^2 = m_2$.

Desenvolvendo a fórmula que define o momento de ordem r centrado na média e usando coeficientes binomiais, é possível estabelecer uma relação entre o momento de ordem r centrado na média e os momentos naturais de ordem menor ou igual a r.

3.5 Medidas de assimetria

Considere os diagramas de pontos dados nas Figuras 3.6 a 3.8, onde a seta indica a média dos dados. Analisando-os, podemos ver que a principal e mais marcante diferença entre eles diz respeito à simetria da distribuição. A segunda distribuição é simétrica, enquanto as outras duas são assimétricas.

Figura 3.6: Assimetria positiva

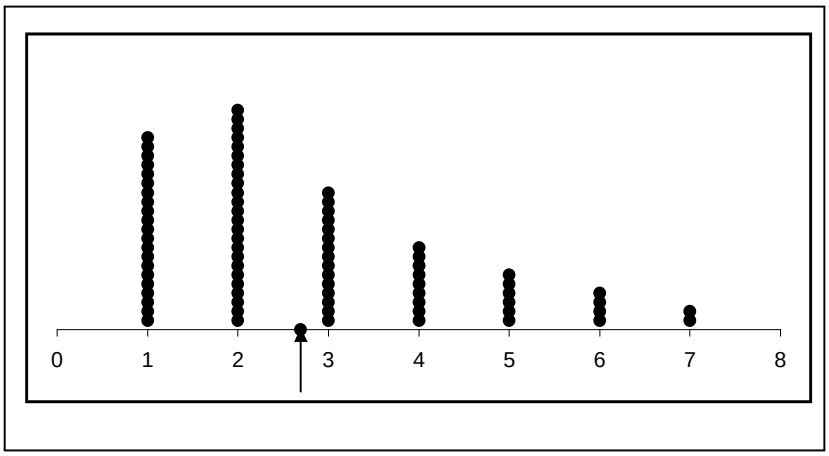


Figura 3.7: Simetria

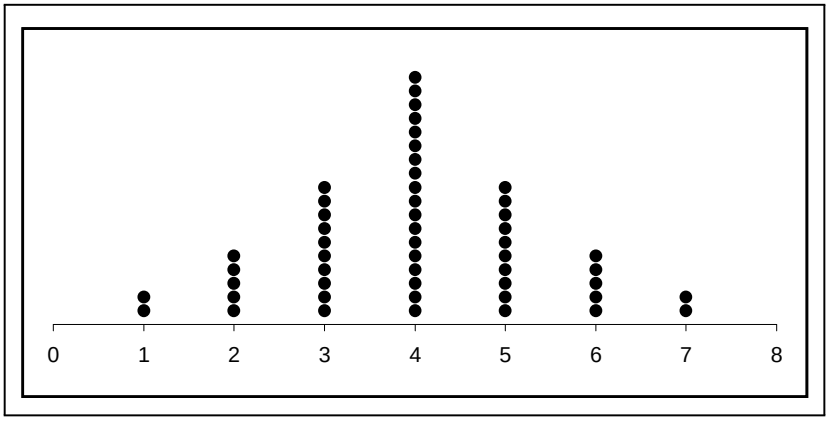
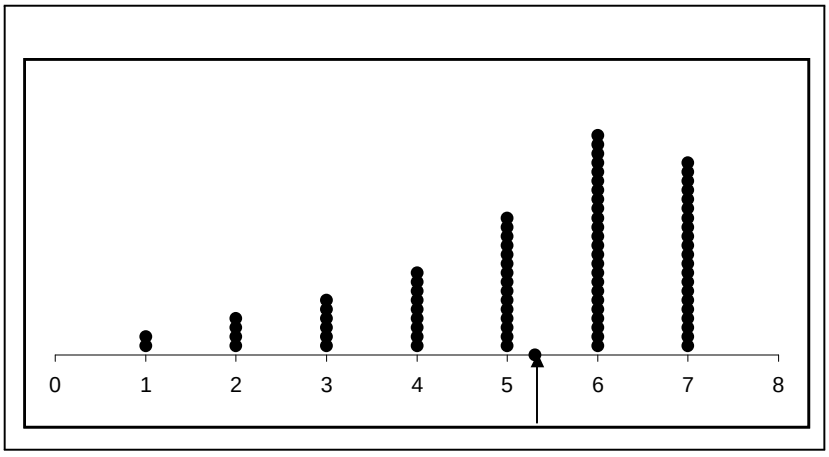


Figura 3.8: Assimetria negativa

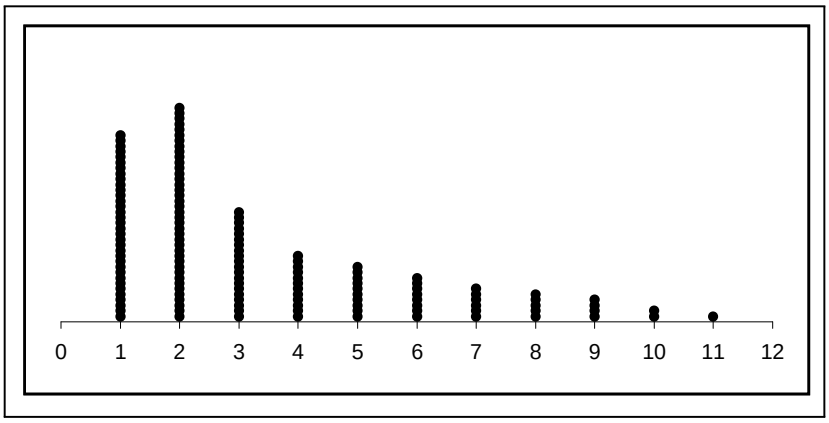


No primeiro diagrama a assimetria é tal que há maior concentração na cauda inferior, enquanto no terceiro, a concentração é maior na cauda superior. Visto de outra maneira, na Figura 3.6, os dados se estendem para o lado positivo da escala, enquanto na Figura 3.8, os dados se estendem para o lado negativo da escala. Esses dois fatos caracterizam o primeiro tipo de assimetria como assimetria positiva e o terceiro como assimetria negativa. Na Figura 3.7 temos uma simetria perfeita. Esses três tipos de assimetria podem ser caracterizados pela posição da moda com relação à média dos dados. No primeiro tipo, a moda tende a estar à esquerda da média, enquanto no terceiro tipo, a moda tende a estar à direita de média (lembre-se que a média é o centro de gravidade ou ponto de equilíbrio da distribuição). Para distribuições simétricas, a moda coincide com a média. Definem-se, assim, os três tipos de assimetria:

- se a média é maior que a moda ($\overline{x} > x^*$), dizemos que a distribuição é assimétrica à direita ou tem assimetria positiva (Figura 3.6);
- se a média é igual à moda ($\overline{x} = x^*$), dizemos que a distribuição é simétrica ou tem assimetria nula (Figura 3.7);
- se a média é menor que a moda ($\overline{x} < x^*$), dizemos que a distribuição é assimétrica à esquerda ou tem assimetria negativa (Figura 3.8).

Essas definições, no entanto, não permitem "medir" diferentes graus de assimetria. Por exemplo, considere os histogramas dados nas Figuras 3.6 e 3.9, ambos assimétricos à direita.

Figura 3.9: Outra distribuição assimétrica positiva



Uma forma de medir essas diferentes assimetrias seria através da distância $\overline{x} - x^*$ entre a média e a moda mas como as distribuições podem ter graus de dispersão diferentes, é importante que consideremos a diferença acima na mesma escala. Assim, define-se o coeficiente de assimetria de Pearson como:

$$\boxed{e = \frac{\overline{x} - x^*}{\sigma}}; \tag{3.28}$$

se o coeficiente é negativo, temos assimetria negativa; se é positivo, tem-se assimetria positiva e se é nulo, tem-se uma distribuição simétrica.

Para os dados da Figura 3.6, temos que $x^* = 2$, $\overline{x} = 2,691358$ e $\sigma = 1,576384$; logo,

$$e = \frac{2,691358 - 2}{1,576384} = 0,43857$$

Para os dados da Figura 3.9, $x^* = 2$, $\overline{x} = 3,312057$ e $\sigma = 2,400162$; logo,

$$e = \frac{3,312057 - 2}{2,400162} = 0,54662$$

o que resulta em uma assimetria mais acentuada.

É possível também definir um coeficiente de assimetria através do momento centrado na média de ordem 3, mas também tomado em sua forma padronizada, ou seja:

$$\boxed{E = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{\sigma^3}}. \tag{3.29}$$

Esse coeficiente tem as mesmas características do coeficiente de Pearson: será negativo, nulo ou positivo, dependendo se a distribuição tem assimetria negativa, nula ou positiva.

Uma medida de assimetria robusta pode ser definida em termos das distâncias de Q_1 e Q_3 à mediana Q_2. Se essas distâncias forem iguais, temos uma distribuição simétrica. Para as distribuições assimétricas, temos o seguinte:

$$\begin{aligned} Q_2 - Q_1 &< Q_3 - Q_2 \Rightarrow \text{assimetria positiva} \\ Q_2 - Q_1 &> Q_3 - Q_2 \Rightarrow \text{assimetria negativa} \end{aligned}$$

Assim, define-se o coeficiente de assimetria de Bowley como

$$e' = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

Note que aqui o denominador é o intervalo interquartil, que também é uma medida de dispersão, o que torna o coeficiente adimensional. Como antes, $e' > 0, e' < 0$ e $e' = 0$ correspondem respectivamente a distribuições assimétricas positivas, assimétricas negaticas e simétricas. Um resultado interessante pode ser obtido notando-se que $Q_3 - Q_1 = (Q_3 - Q_2) + (Q_2 - Q_1)$. Quanto mais acentuada for a assimetria positiva de uma distribuição, menor será $Q_2 - Q_1$ e e' tende a $+1$.Analogamente, quanto mais acentuada for a assimetria negativa, menor será $Q_3 - Q_2$ e e' tende a -1. Resulta que $-1 < e' < 1$.

3.6 Uma estratégia alternativa para análise de dados

Já foi visto que a média aritmética simples é muito afetada pelos valores discrepantes, ou seja, esses valores exercem grande influência na média, "puxando" esse valor em sua direção. Consideremos um exemplo para relembrar o que foi dito: em um levantamemto sobre as rendas familiares dos funcionários de uma empresa, os valores obtidos foram (em u.m.) 7, 9, 10, 15, 25 mas o digitador se equivocou e digitou 250 no lugar de 25. Vamos ver os efeitos desse erro nas médias e desvios padrões dos conjuntos de dados. Para os valores corretos temos:

$$\overline{x} = \frac{7 + 9 + 10 + 15 + 25}{5} = 13,2$$

$$\sigma^2 = \frac{1}{5}\left(7^2 + 9^2 + 10^2 + 15^2 + 25^2\right) - 13,2^2 = 216 - 174,24 = 41,76$$

$$\sigma = 6,4622$$

Para os valores incorretos temos que:

$$\overline{x} = \frac{7+9+10+15+250}{5} = 58,2$$

$$\sigma^2 = \frac{1}{5}\left(7^2+9^2+10^2+15^2+250^2\right) - 13,2^2 = 12591 - 3387,24 = 9203,76$$

$$\sigma = 95,9362$$

Vê-se que há um aumento acentuado nos valores das estatísticas acima. Nesse exemplo, o valor discrepante foi resultado de um erro mas nem sempre é assim. Existem valores discrepantes que refletem algum acontecimento especial. Nessas situações, é importante ter uma estratégia alternativa de análise dos dados, que permita compreender melhor o fenômeno em estudo.

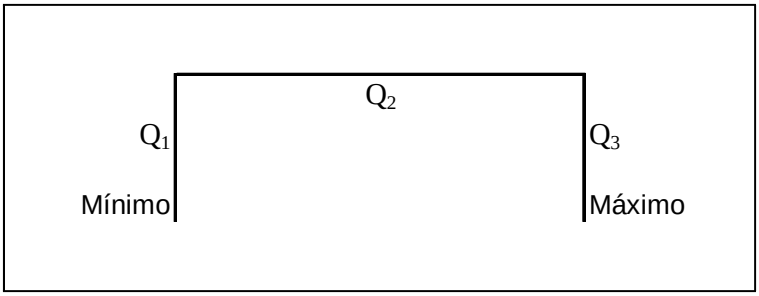
Agora vamos apresentar um conjunto de medidas estatísticas (a mediana é uma delas), que chamamos medidas robustas, por não serem afetadas pela presença de poucos valores discrepantes.

3.6.1 O esquema dos cinco números

Com relação à medida de posição, já vimos que a mediana é uma medida robusta; ela será, então, usada nesse tipo de análise. Com relação à medida de dispersão, será usado o intervalo interquartil que, como já visto, é definido a partir dos quartis. Como é importante também saber os valores extremos, estes também serão usados na análise.

Uma forma de apresentar esses valores é através do esquema dos cinco números, cuja representação genérica está na Figura 3.10 a seguir.

Figura 3.10: Esquema dos 5 números



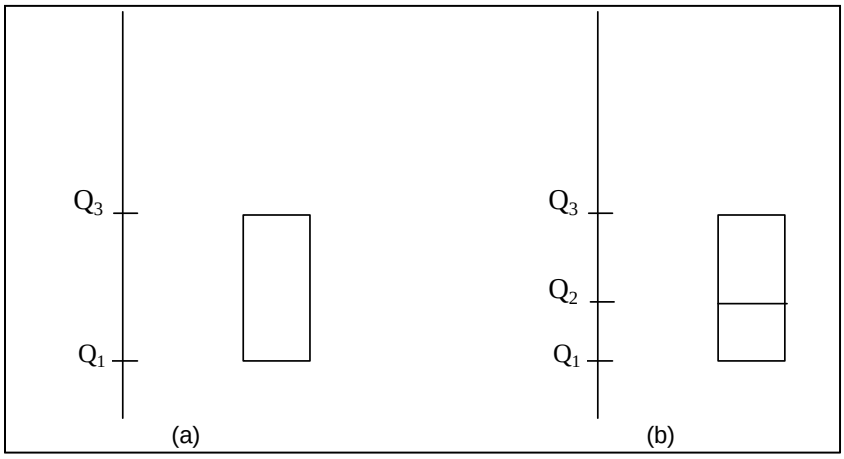
3.6.2 O boxplot

A partir dessas medidas constrói-se também um gráfico chamado *gráfico de caixas* (em inglês, *boxplot*) que ilustra os principais aspectos da distribuição, tomando por base essas medidas robustas. O boxplot é um gráfico muito útil também na comparação de distribuições.

O boxplot é formado basicamente por um retângulo vertical (ou horizontal). O comprimento do lado vertical (ou horizontal) é dado pelo intervalo interquartil (Figura 3.11(a) , onde estamos trabalhando com um retângulo vertical). O tamanho do outro lado é indiferente, sugerindo-se apenas uma escala razoável. Na altura da mediana, traça-se uma linha, dividindo o retângulo em duas partes (Figura 3.11(b)).

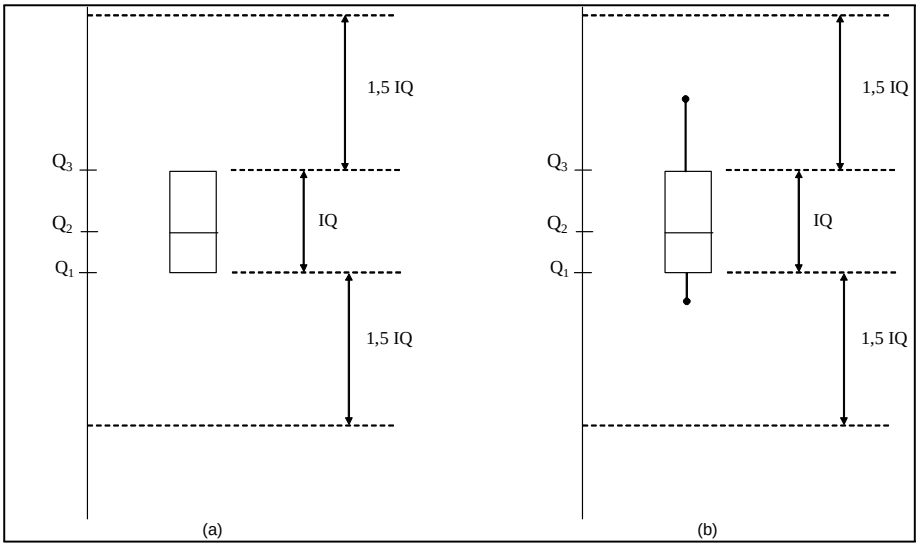
Note que aí já temos representados 50% da distribuição e também já temos idéia da assimetria da mesma. Para representar os 25% restantes em cada cauda da distribuição temos que cuidar primeiro da presença de possíveis *outliers* ou valores discrepantes.

Figura 3.11: Construção do boxplot - Etapa 1



Um dado será considerado *outlier* se ele for menor que $Q_1 - 1,5\,IQ$ ou maior que $Q_3 + 1,5\,IQ$ [Figura 3.12(a)]. Para representar o domínio de variação dos dados que não são *outliers*, traça-se, a partir do retângulo, uma linha para cima e outra para baixo até o ponto mais remoto que não seja *outlier* Figura 3.12(b)]. Esses pontos são chamados *juntas*.

Figura 3.12: Construção do boxplot - Etapa 2



Quanto aos *outliers*, eles são representados individualmente por um X (ou algum outro tipo de carácter), explicitando, de preferência, os seus valores mas com quebra de escala no eixo (Figura 3.13).

Como exemplo, vamos construir o esquema dos 5 números e o boxplot para os dados apresentados na Tabela 3.13, onde temos as populações, em 1000 habitantes, dos estados brasileiros ordenadas crescentemente.

Como temos 27 estados, a mediana é o valor central, correspondente à 14ª observação, ou seja, a

Figura 3.13: Construção do boxplot - Etapa 3

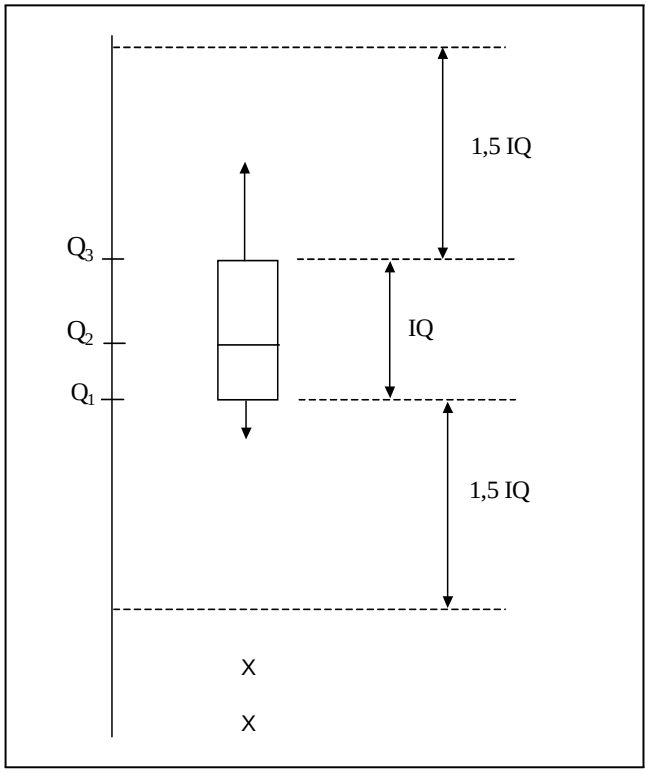


Tabela 3.13: População dos estados brasileiros (em 1000 hab.)

RR	325	MS	2079	PB	3444	PR	9564
AP	478	MT	2505	GO	5004	RS	10188
AC	558	RN	2777	SC	5357	BA	13071
TO	1158	AM	2813	MA	5652	RJ	14392
RO	1380	AL	2823	PA	6193	MG	17892
SE	1785	PI	2844	CE	7431	SP	37033
DF	2052	ES	3098	PE	7919		

observação correspondente ao estado do Espírito Santo. Tirando a mediana, sobram 13 observações em cada metade dos dados; logo, o primeiro quartil é a sétima maior observação (DF) e o terceiro quartil é a 21ª (14+7) maior observação (PE).

$$Q_1 = 2052 \qquad Q_2 = 3098 \qquad Q_3 = 7919$$

O intervalo interquartil é:

$$IQ = 7919 - 2052 = 5867$$

Com relação aos *outliers*, temos que:

$$Q_1 - 1,5\,IQ = 2052 - 1.5 \times 5867 = -6748,5$$

$$Q_3 + 1,5\,IQ = 7919 + 1.5 \times 5867 = 16720$$

Logo, não há *outliers* na cauda inferior mas na cauda superior, os estados de Minas Gerais e São Paulo são *outliers*. Nas Figuras 3.14 e 3.15 temos o esquema dos 5 números e o boxplot para esses dados.

Figura 3.14: População das UFs brasileiras (em 1000 hab) - Esquema dos 5 números

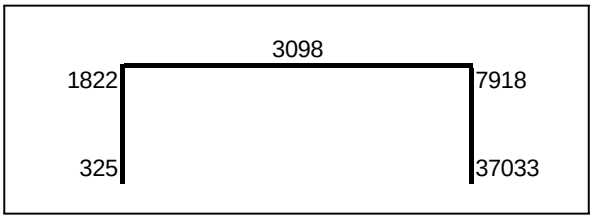
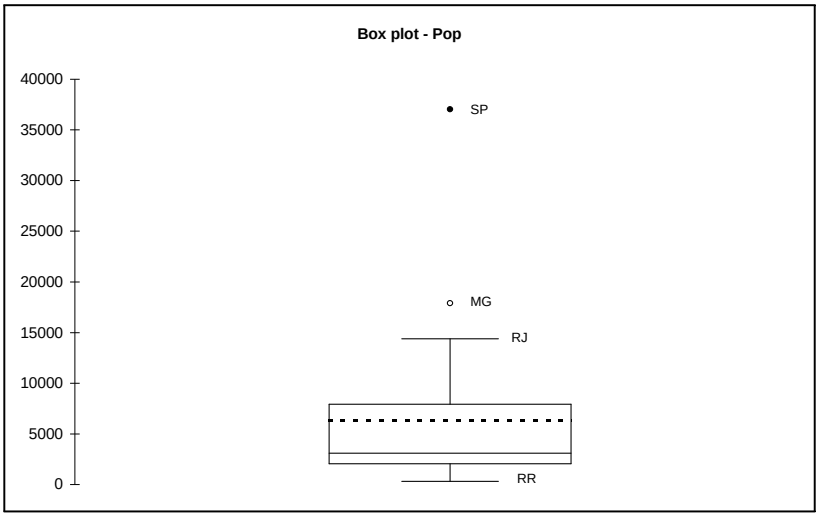


Figura 3.15: População das UFs brasileiras (em 1000 hab) - Boxplot



Note que as linhas ou pontos correspondentes aos limites $Q_1 - 1,5IIQ$ e $Q_3 + 1,5IIQ$ não são represetados no gráfico; eles servem apenas para identificar os outliers.

O boxplot é muito usado também para se fazerem comparações entre conjuntos de dados. Considere, por exemplo, os dados da Tabela 3.14, correspondentes à população urbana e rural das 27 UFs brasileiras, segundo dados do Censo Demográfico 2000. Esses dados encontram-se representados na Fig. 3.16. Podemos ver que a população urbana apresenta maior variabilidade e também uma forte assimetria positiva. Há 3 UFs que são discrepantes: São Paulo, Minas Gerias e Rio de Janeiro.

Tabela 3.14: População urbana e rural das UFs brasileiras (em 1000 hab.)

UF	População		UF	População	
	Urbana	Rural		Urbana	Rural
RO	885	496	MG	14672	3220
AC	371	188	ES	2464	635
AM	2108	706	RJ	13822	570
RR	248	78	SP	34593	2440
PA	4121	2072	PR	7787	1778
AP	425	53	SC	4218	1139
TO	860	298	RS	8318	1870
MA	3365	2288	MS	1748	331
PI	1789	1055	MT	1988	517
CE	5316	2116	GO	4397	607
RN	2037	741	DF	1962	90
PB	2448	997			
PE	6059	1861			
AL	1920	903			
SE	1274	512			
BA	8773	4298			

Fonte: IBGE - Censo Demográfico 2000

3.7 Medidas de posição e dispersão para dados agrupados

Nesta seção serão vistas algumas medidas de posição e dispersão para dados agrupados em classes. Embora seja recomendável calcular tais medidas para um conjunto de dados antes de agrupá-los, às vezes não é possível; por exemplo, os dados originais podem não estar disponíveis.

A idéia básica subjacente aos cálculos a serem feitos é a seguinte: ao agruparmos os dados em classes, estamos perdendo informação, a individualidade dos valores. Informar apenas que existem 5 valores na classe $2 \vdash 5$ nos obriga a escolher um valor típico, representante de tal classe. Esse valor será sempre o ponto médio da classe. Então a informação anterior é interpretada como a existência de 5 valores iguais a 3,5. Essa é a interpretação básica da tabela de freqüências: todos os valores de uma classe são considerados iguais ao ponto médio da classe. A partir dessa interpretação, o cálculo das principais medidas de posição e dispersão se faz usando as definições usuais, apenas aplicadas a um novo conjunto de dados, representado pelos pontos médios das classes.

Vamos ilustrar todos os conceitos com os dados da Tabela 2.31, que reproduzimos na Tabela 3.15 para facilitar a apresentação. Note que nessa nova versão da tabela acrescentamos a coluna do ponto médio da classe, que será denotado por x_i.

Figura 3.16: População urbana e rural das UFs brasileiras (em 1000 hab)

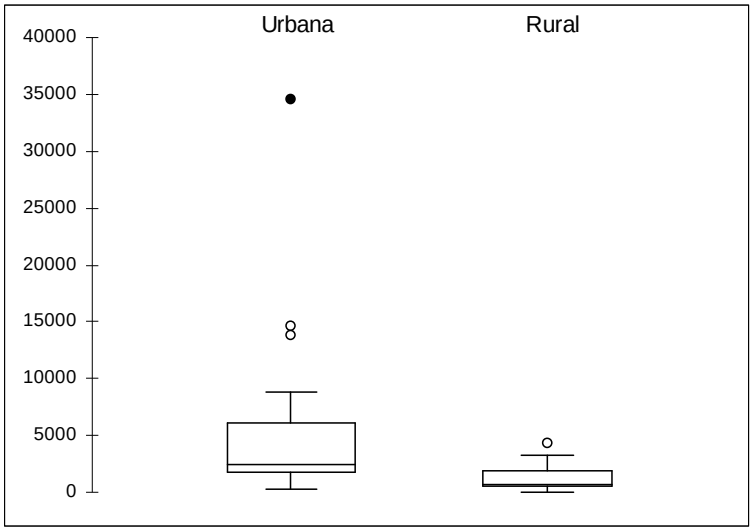


Tabela 3.15: Aluguéis de 200 imóveis urbanos

Classes de aluguéis (u.m.)	Ponto médio $\mathbf{x}_i$	Freqüência Simples		Freqüência Acumulada	
		Absoluta $\mathbf{n}_i$	Relativa $\mathbf{f}_i$	Absoluta $\mathbf{N}_i$	Relativa $\mathbf{F}_i$
$2 \vdash 3$	$2,5$	10	0,05	10	0,05
$3 \vdash 5$	$4,0$	50	0,25	60	0,30
$5 \vdash 7$	$6,0$	80	0,40	140	0,70
$7 \vdash 10$	$8,5$	40	0,20	180	0,90
$10 \vdash 15$	$12,5$	20	0,10	200	1,00

3.7.1 Média simples

A interpretação da tabela de freqüências nos diz que há 10 observações iguais a 2,5; 50 observações iguais a 4,0; 80 iguais a 6,0; 40 iguais a 8,5 e 20 iguais a 12,5. Então esses dados podem ser vistos como o seguinte conjunto de observações:

$$\left.\begin{matrix} 2,5 \\ \vdots \\ 2,5 \end{matrix}\right\} 10 \text{ ocorrências} \tag{3.30}$$

$$\left.\begin{matrix} 4,0 \\ \vdots \\ 4,0 \end{matrix}\right\} 50 \text{ ocorrências}$$

$$\left.\begin{matrix} 6,0 \\ \vdots \\ 6,0 \end{matrix}\right\} 80 \text{ ocorrências}$$

$$\left.\begin{matrix} 8,5 \\ \vdots \\ 8,5 \end{matrix}\right\} 40 \text{ ocorrências}$$

$$\left.\begin{matrix} 12,5 \\ \vdots \\ 12,5 \end{matrix}\right\} 20 \text{ ocorrências}$$

Para calcular a média desse novo conjunto de dados temos que fazer:

$$\begin{aligned} \overline{x} &= \frac{10 \times 2,5 + 50 \times 4,0 + 80 \times 6,0 + 40 \times 8,5 + 20 \times 12,5}{200} = \\ &= \frac{10}{200} \times 2,5 + \frac{50}{200} \times 4,0 + \frac{80}{200} \times 6,0 + \frac{40}{200} \times 8,5 + \frac{20}{200} \times 12,5 = \\ &= 0,05 \times 2,5 + 0,25 \times 4,0 + 0,40 \times 6,0 + 0,20 \times 8,5 + 0,10 \times 12,5 = \\ &= 6,475 \end{aligned}$$

Note, na penúltima linha da equação anterior, que os pontos médios de cada classe são multiplicados pela freqüência relativa da classe. Então, a média dos dados agrupados em classes é uma média ponderada dos pontos médios, onde os pesos são definidos pelas freqüências das classes. Em geral temos:

$$\boxed{\overline{x} = \sum_{i=1}^{k} f_i x_i} \tag{3.31}$$

Os pesos aparecem exatamente para compensar o fato de que as classes têm números diferentes de observações.

3.7.2 Variância

O cálculo da variância é feito de modo análogo, só que agora temos que considerar os desvios dos pontos médios em torno da média, ou seja:

$$\sigma^2 = \frac{1}{200}\left(\begin{array}{c} 10 \times (2,5 - 6,475)^2 + 50 \times (4,0 - 6,475)^2 + 80 \times (6,0 - 6,475)^2 + \\ 40 \times (8,5 - 6,475)^2 + 20 \times (12,5 - 6,475)^2 \end{array}\right) =$$

$$
\begin{aligned}
&= \frac{10}{200} \times (2,5 - 6,475)^2 + \frac{50}{200} \times (4,0 - 6,475)^2 + \frac{80}{200} \times (6,0 - 6,475)^2 + \\
&\quad + \frac{40}{200} \times (8,5 - 6,475)^2 + \frac{20}{200} \times (12,5 - 6,475)^2 \\
&= 0,05 \times (2,5 - 6,475)^2 + 0,25 \times (4,0 - 6,475)^2 + 0,40 \times (6,0 - 6,475)^2 + \\
&\quad + 0,20 \times (8,5 - 6,475)^2 + 0,10 \times (12,5 - 6,475)^2 \\
&= 6,861875
\end{aligned}
$$

Novamente, temos uma média ponderada dos desvios ao quadrado, com os pesos sendo as freqüências relativas. Em geral temos:

$$\boxed{\sigma^2 = \sum_{i=1}^{k} f_i (x_i - \overline{x})^2}. \tag{3.32}$$

Desenvolvendo o quadrado obtemos uma fórmula mais simples de ser utilizada:

$$
\begin{aligned}
\sigma^2 &= \sum_{i=1}^{k} f_i (x_i - \overline{x})^2 = \sum_{i=1}^{k} f_i (x_i^2 - 2x_i\overline{x} + \overline{x}^2) = \\
&= \sum_{i=1}^{k} f_i x_i^2 - \sum_{i=1}^{k} 2 f_i x_i \overline{x} + \sum_{i=1}^{k} f_i \overline{x}^2 = \sum_{i=1}^{k} f_i x_i^2 - 2\overline{x} \sum_{i=1}^{k} f_i x_i + \overline{x}^2 \sum_{i=1}^{k} f_i = \\
&= \sum_{i=1}^{k} f_i x_i^2 - 2\overline{x}\,\overline{x} + \overline{x}^2 \times 1
\end{aligned}
$$

onde usamos a definição da média de dados agrupados dada em (3.31) e o fato de as freqüências relativas somarem 1. Logo, a variância de dados agrupados é dada por:

$$\boxed{\sigma^2 = \sum_{i=1}^{k} f_i (x_i - \overline{x})^2 = \sum_{i=1}^{k} f_i x_i^2 - \overline{x}^2}. \tag{3.33}$$

Note que continua valendo que a variância é "a média dos quadrados menos o quadrado da média", uma vez que o somatório acima calcula a média - ponderada - dos quadrados dos x_i.

3.7.3 Mediana

Como já visto, a mediana é o valor que deixa 50% das observações acima e 50% abaixo dela. Estando os dados agrupados em classes, podemos usar a interpretação da tabela de freqüências para calcular a mediana. Considere novamente os dados da Tabela 3.15, cuja interpretação é dada em (3.30). Como temos 200 observações, a mediana é o valor que deixa 100 observações abaixo dela. A centésima observação ocorre na terceira classe, pois nas duas primeiras temos apenas 60 e nas três primeiras temos 140. Logo, a mediana pode ser definida como o ponto médio da terceira classe. Essa é a definição de *mediana bruta*, que é sempre o ponto médio da classe onde se completam 50% das observações, que, por sua vez, é chamada *classe mediana*.

No entanto, existe um método geométrico que produz uma estimativa da mediana um pouco mais refinada. As idéias subjacentes a esse método são que a mediana divide ao meio o conjunto de dados (ou seja, a definição de mediana) e que, no histograma da distribuição, as áreas dos retângulos são proporcionais às freqüências relativas.

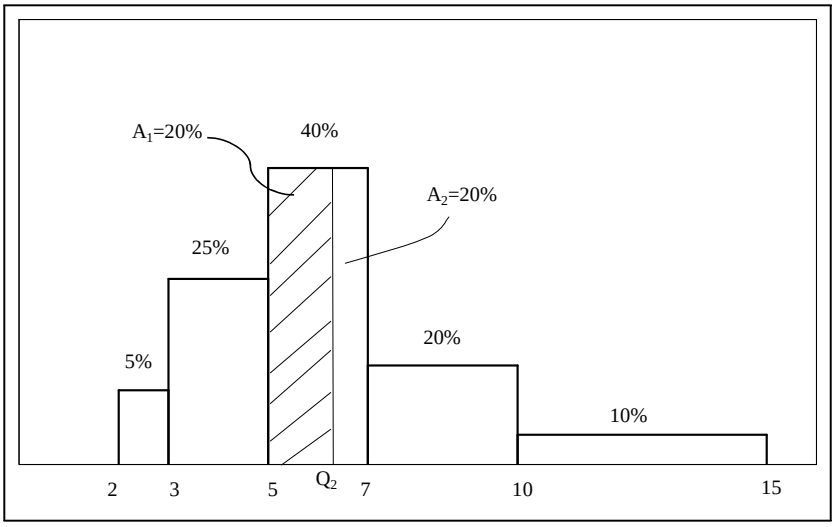
Consideremos o histograma da Figura 3.17, referente aos dados da Tabela 3.15. Nas duas primeiras classes temos 30% das observações e nas três primeiras classes temos 70%; logo, a mediana é algum ponto da classe mediana $5 \vdash 7$. Para identificá-la, devemos notar que na classe mediana ficam faltando $50\% - 30\% = 20\%$ da distribuição para completar 50%. Então a área A_1 do retângulo hachurado deve ser igual a 20%, enquanto o retângulo da classe mediana tem área $A_m = 40\%$. Usando a fórmula que dá a área de um retângulo obtém-se:

$$\begin{aligned} A_1 &= 0,20 = (Q_2 - 5) \times h \\ A_m &= 0,40 = (7 - 5) \times h \end{aligned}$$

onde h é a altura comum dos dois retângulos. Dividindo as duas igualdades termo a a termo obtém-se:

$$\frac{0,20}{0,40} = \frac{Q_2 - 5}{2} \Rightarrow Q_2 = 6\,.$$

Figura 3.17: Cálculo da mediana de dados agrupados



O retângulo correspondente à classe mediana fica dividido em dois subretângulos. Os cálculos acima foram feitos com base no subretângulo inferior mas poderiam também ser feitos com base no subretângulo superior, o que resultaria na seguinte regra de três:

$$\frac{7 - Q_2}{0,70 - 0,50} = \frac{7 - 5}{0,40} \Rightarrow 7 - Q_2 = 0,2 \times \frac{2}{0,40} \Rightarrow Q_2 = 6.$$

Para generalizar este procedimento, vamos estabelecer a seguinte notação:

- ℓ_I limite inferior da classe mediana;
- ℓ_S limite superior da classe mediana;
- δ_m comprimento da classe mediana ($\delta_m = \ell_S - \ell_I$);
- F_{ant} freqüência relativa acumulada da classe anterior à classe mediana;
- F_m freqüência relativa acumulada da classe mediana;

- f_m freqüência relativa simples da classe mediana ($f_m = F_m - F_{\text{ant}}$).

Trabalhando com o subretângulo inferior, temos que as áreas envolvidas são:

$$A_1 = 50 - F_{\text{ant}} = (Q_2 - \ell_I) \times h$$

$$A_m = F_m - F_{\text{ant}} = (\ell_S - \ell_I) \times h$$

Dividindo membro a membro obtém-se:

$$\frac{50 - F_{\text{ant}}}{F_m - F_{\text{ant}}} = \frac{Q_2 - \ell_I}{\ell_S - \ell_I} \Rightarrow \frac{50 - F_{\text{ant}}}{f_m} = \frac{Q_2 - \ell_I}{\delta_m} \Rightarrow$$

$$\boxed{Q_2 = \ell_I + \frac{50 - F_{\text{ant}}}{f_m} \times \delta_m} \tag{3.34}$$

Trabalhando com o subretângulo superior, as áreas envolvidas são:

$$A_2 = F_m - 50 = (\ell_S - Q_2) \times h$$

$$A_m = f_m = (\ell_S - \ell_I) \times h$$

o que resulta em:

$$\frac{F_m - 50}{f_m} = \frac{\ell_S - Q_2}{\ell_S - \ell_I} \Rightarrow \frac{F_m - 50}{f_m} = \frac{\ell_S - Q_2}{\delta_m}$$

ou:

$$\boxed{Q_2 = \ell_S - \frac{F_m - 50}{f_m} \times \delta_m} \tag{3.35}$$

Talvez essa última fórmula seja mais fácil de memorizar, em virtude de só envolver dados relativos à classe mediana.

Para o exemplo tratado, esses valores são:

$$\begin{aligned} \ell_I &= 5 \\ \ell_S &= 7 \\ \delta_m &= 2 \end{aligned}$$

$$\begin{aligned} F_{\text{ant}} &= 30 \\ F_m &= 70 \\ f_m &= 40 \end{aligned}$$

resultando, como antes,

$$Q_2 = 5 + \frac{50 - 30}{40} \times 2 = 6$$

ou

$$Q_2 = 7 - \frac{70 - 50}{40} \times 2 = 6.$$

3.7.4 Outras separatrizes

O cálculo de qualquer separatriz de dados agrupados em classes é feito de maneira análoga ao cálculo da mediana. A diferença é que, em vez de estarmos lidando com a classe mediana, estaremos lidando com a classe onde se completa o percentual da separatriz desejada e com a diferença entre esse percentual e o percentual acumulado até a classe anterior. O procedimento genérico pode ser estabelecido com a seguinte notação: seja p o percentual abaixo da separatriz S desejada (no caso da mediana, $p = 50\%$). Identificada a classe onde se completa essa freqüência acumulada, vamos denotá-la por classe p-separatriz. Sejam:

- ℓ_I limite inferior da classe p-separatriz;
- ℓ_S limite superior da classe p-separatriz;
- δ_p comprimento da classe p-separatriz ($\delta_m = \ell_S - \ell_I$);
- F_p freqüência relativa acumulada da classe p-separatriz;
- f_p freqüência relativa simples da classe p-separatriz.

Trabalhando com o subretângulo superior, como fizemos na mediana, as áreas envolvidas são:

$$A_2 = F_p - p = (\ell_S - S) \times h$$

$$A_p = f_p = (\ell_S - \ell_I) \times h$$

Dividindo membro a membro obtém-se:

$$\frac{F_p - p}{f_p} = \frac{\ell_S - S}{\ell_S - \ell_I} \Rightarrow \frac{F_p - p}{f_p} = \frac{\ell_S - S}{\delta p} \Rightarrow$$

$$\boxed{S = \ell_S - \frac{F_p - p}{f_p} \times \delta_p} \tag{3.36}$$

Vamos aplicar essa fórmula para determinar o terceiro quartil dos dados da Tabela 3.15, analisada no caso da mediana. Nesse caso, $p = 75\%$ e a classe 75-separatriz é a classe $7 \vdash 10$. Logo,

$$\begin{aligned} \ell_I &= 7 \\ \ell_S &= 10 \\ \delta_{75} &= 3 \end{aligned}$$

$$\begin{aligned} F_{75} &= 90\% \\ f_{75} &= 20\% \end{aligned}$$

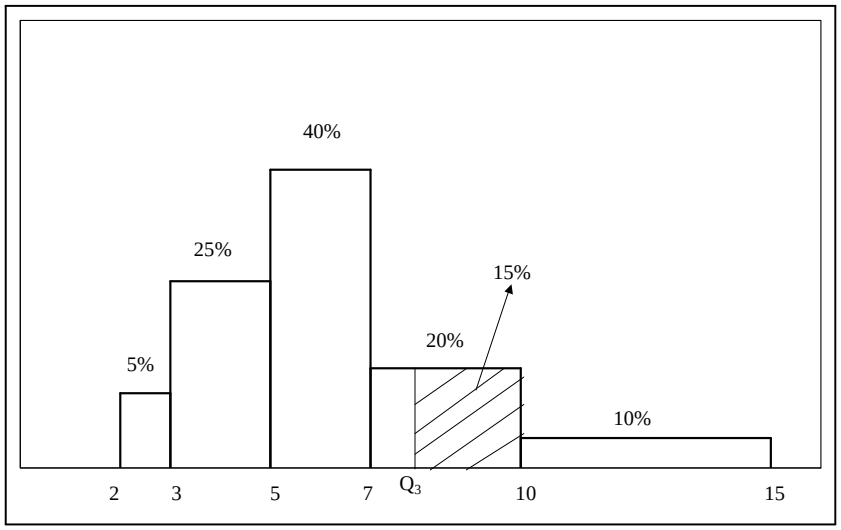
O terceiro quartil, então, é dado por

$$Q_3 = 10 - \frac{90 - 75}{20} \times 3 = 7,75$$

resultante da seguinte relação entre áreas (ver histograma da Figura 3.18):

$$\frac{(10 - Q_3) \times h}{90 - 75} = \frac{(10 - 7) \times h}{20}$$

Figura 3.18: Cálculo do terceiro quartil



3.7.5 Moda

Como visto, a moda de um conjunto de dados é o valor mais freqüente. Para dados agrupados em classes, uma definição análoga seria a de *classe modal*, que é a classe de maior freqüência. No exemplo da Tabela 3.15, a classe modal é a classe $5 \vdash 7$. Por ser o ponto médio o representante da classe, podemos definir a moda dos dados como sendo o ponto médio da classe modal; essa é a definição de *moda bruta*. Então, para o exemplo anterior, a moda bruta é $x^* = 6$.

Existem, no entanto, alguns métodos que permitem obter uma estimativa mais refinada da moda. Todos esses métodos buscam, na classe modal, um ponto (valor) que seja representativo da moda dos dados.

Os métodos que veremos baseiam-se no seguinte raciocínio intuitivo: as classes vizinhas à classe modal "puxam" a moda, como numa brincadeira de cabo de guerra. Quanto maior a "força" da classe, mais próxima dela estará a moda, ou seja, quanto maior a "força", menor a distância da moda à classe vizinha. Podemos representar esquematicamente essa situação da seguinte forma:

↑ força	↓ distância

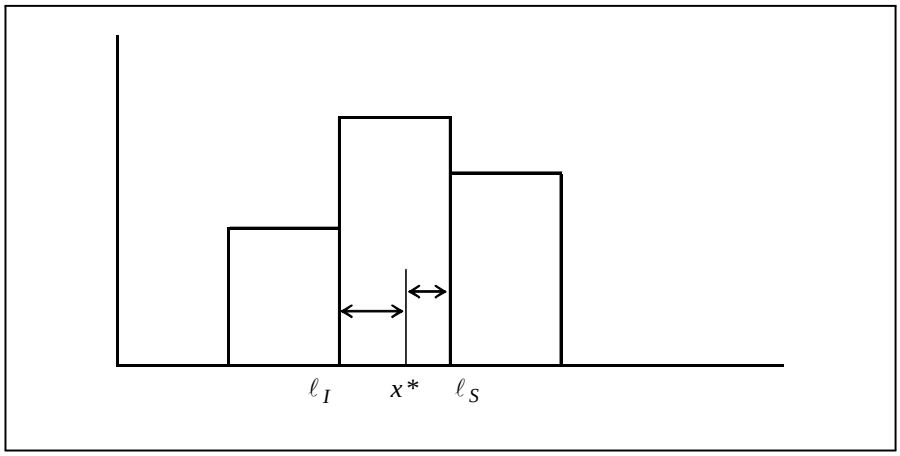
Na Figura 3.19 representa-se a idéia geral dos métodos de cálculo da moda, com as distâncias sendo $x^* - \ell_I$ e $\ell_S - x^*$, onde ℓ_I e ℓ_S são, respectivamente, os limites inferior e superior da classe modal.

Método de King

O método de King baseia-se na influência das freqüências das classes adjacentes à classe modal, ou seja, a "força"de cada classe vizinha é proporcional à sua freqüência; assim, a moda estará mais próxima da classe com maior freqüência ou, equivalentemente, quanto maior a freqüência, menor a distância da moda à classe vizinha. Sendo assim, existe uma proporcionalidade inversa entre as distâncias e as freqüências das classes vizinhas. Completando o esquema anterior, temos

↑ força	↑ freqüência	↓ distância

Figura 3.19: Ilustração dos métodos de cálculo de moda



Em termos matemáticos, isso significa que

$$\begin{aligned} x^* - \ell_I &= \frac{k}{n_I} \\ \ell_S - x^* &= \frac{k}{n_S} \end{aligned}$$

onde k é a constante de proporcionalidade e n_I e n_S são, respectivamente, as freqüências das classes vizinhas inferior e superior. Dividindo ambas as equações termo a termo, obtemos que

$$\frac{x^* - \ell_I}{\ell_S - x^*} = \frac{n_S}{n_I} \tag{3.37}$$

Desenvolvendo a equação (3.37) resulta que

$$\frac{x^* - \ell_I}{\ell_S - x^*} = \frac{n_S}{n_I} \Rightarrow n_I x^* - n_I \ell_I = n_S \ell_S - n_S x^* \Rightarrow (n_I + n_S) x^* = n_I \ell_I + n_S \ell_S \Rightarrow$$

e daí obtém-se uma outra fórmula geral da moda de King:

$$x^* = \frac{n_I}{n_I + n_S} \times \ell_I + \frac{n_S}{n_I + n_S} \times \ell_S \tag{3.38}$$

onde

- ℓ_I limite inferior da classe modal (ou limite superior da classe anterior);
- ℓ_S limite superior da classe modal (ou limite inferior da classe posterior);
- n_I freqüência absoluta da classe anterior à classe modal;
- n_S freqüência absoluta da classe posterior à classe modal;

Da equação (3.38), podemos ver que a moda é uma média ponderada dos extremos da classe modal, ℓ_I e ℓ_S, onde os pesos são definidos pelas freqüências das classes vizinhas.

Método de Czuber

No método de King, a freqüência da classe modal não tem qualquer influência; assim, diferentes freqüências modais poderiam levar à mesma moda, desde que as classes vizinhas fossem iguais. Uma maneira de introduzir a freqüência da classe modal é através do método de Czuber, em que a "força" de cada classe vizinha é definida pela diferença entre a freqüência da classe modal e a sua própria freqüência. No entanto, essa diferença é inversamente proporcional à "força"da classe, ou seja, quanto menor a diferença entre as freqüências, maior a "força" e vice-versa. O esquema de proporcionalidades para esse método é

↑ força	↓ diferença	↓ distância

ou seja, em termos de distâncias e medida de força, temos, agora, uma proporcionaldiade direta, o que nos leva à seguinte equação:

$$\begin{aligned} x^* - \ell_I &= k(n_m - n_I) \\ \ell_S - x^* &= k(n_m - n_S) \end{aligned}$$

Dividindo termo a termo, temos que

$$\frac{x^* - \ell_I}{\ell_S - x^*} = \frac{n_m - n_I}{n_m - n_S} \tag{3.39}$$

onde n_m é a freqüência da classe modal e os outros termos são como antes.

Vamos adotar a seguinte notação:

$$\begin{aligned} \delta_I &= n_m - n_I \\ \delta_S &= n_m - n_S \end{aligned}$$

Desenvolvendo a equação (3.39) temos que

$$\frac{x^* - \ell_I}{\delta_I} = \frac{\ell_S - x^*}{\delta_S} \Rightarrow \delta_S x^* - \delta_S \ell_I = \delta_I \ell_S - \delta_I x^* \Rightarrow (\delta_I + \delta_S) x^* = \delta_S \ell_I + \delta_I \ell_S$$

o que resulta na fórmula geral da moda de Czuber:

$$x^* = \frac{\delta_S}{\delta_I + \delta_S} \ell_I + \frac{\delta_I}{\delta_I + \delta_S} \ell_S \tag{3.40}$$

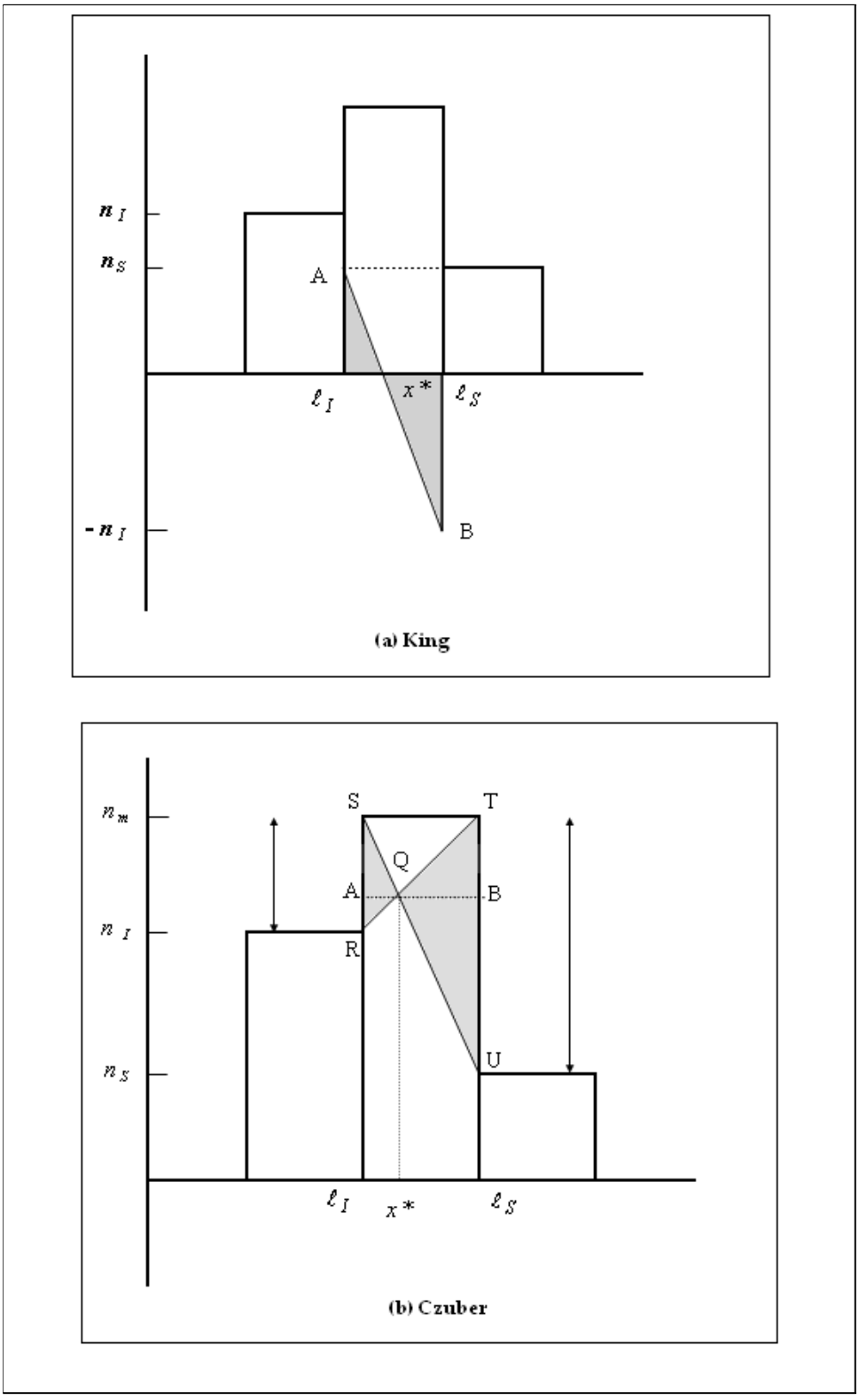
Como no método de King, a moda de Czuber também é uma média ponderada dos extremos da classe modal, mas, agora, os pesos são definidos em termos das diferenças entre as freqüências modal e vizinhas.

Interpretação geométrica dos métodos de King e Czuber

Ambos os métodos de King e Czuber podem ser deduzidos a partir de argumentos de semelhança de triângulos, conforme ilustrado na Figura 3.20.

Para o método de King, considere o retângulo correspondente à classe modal. No lado inferior, marcamos o ponto A na altura igual à freqüência da classe posterior à classe modal. No lado superior, mas na parte inferior, marca-se o ponto B , de modo que sua altura seja igual à freqüência

Figura 3.20: Interpretação geométrica dos métodos de King e Czuber



da classe anterior à classe modal. Os triângulos sombreados $A\ell_I x^*$ e $B\ell_S x^*$ são semelhantes. Então, resulta a seguinte proporcionalidade entre os lados:

$$\frac{\overline{\ell_I x^*}}{\overline{\ell_S x^*}} = \frac{\overline{A\ell_I}}{\overline{B\ell_S}}$$

Pela construção desses triângulos, isso significa que:

$$\frac{x^* - \ell_I}{\ell_S - x^*} = \frac{n_S}{n_I}$$

a mesma igualdade obtida anteriormente.

Para o método de Czuber, traça-se o segmento $\overline{SU}$ ligando o extremo superior do lado inferior do retângulo modal ao extremo superior do lado inferior do retângulo da classe posterior à classe modal e o segmento $\overline{RT}$ ligando o extremo superior do lado superior do retângulo modal ao extremo superior do lado superior do retângulo da classe anterior à classe modal. Obtêm-se os triângulos sombreados RQS e TQU, que são semelhantes. Portanto, vale a seguinte proporcionalidade entre lados e alturas:

$$\frac{\overline{AQ}}{\overline{RS}} = \frac{\overline{BQ}}{\overline{TU}}$$

ou equivalentemente

$$\frac{x^* - \ell_I}{n_m - n_I} = \frac{\ell_S - x^*}{n_m - n_S}$$

a mesma proporção obtida anteriormente.

3.7.6 Médias geométrica e harmônica

Embora não muito usual, o cálculo das médias geométrica e harmônica para dados agrupados será apresentado principalmente por aspectos didáticos, visando sua aplicação no estudo de números índices.

Suponhamos, então, que temos n_1 valores iguais a x_1, n_2 iguais a x_2 , ..., n_k iguais a x_k. Os valores x_i podem ou não ser pontos médios das classes de uma tabela de freqüências; o que importa é a repetição de cada um deles. Seja $n = n_1 + n_2 + \cdots + n_k$ o número total de observações.

A média geométrica, por definição, é:

$$\begin{aligned}\overline{x}_g &= \sqrt[n]{x_1 \times \cdots \times x_1 \times x_2 \times \cdots \times x_2 \times \cdots \times x_k \times \cdots \times x_k} = \sqrt[n]{x_1^{n_1} \times x_2^{n_2} \times \cdots \times x_k^{n_k}} = \\ &= \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}} = \prod_{i=1}^{k} x_i^{f_i}\end{aligned}$$

Para a média harmônica, temos que:

$$\begin{aligned}\overline{x}_h &= \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_2} + \cdots + \frac{1}{x_k} + \cdots + \frac{1}{x_k}} = \\ &= \frac{n}{n_1 \times \frac{1}{x_1} + n_2 \times \frac{1}{x_2} + \cdots + n_k \times \frac{1}{x_k}} = \\ &= \frac{1}{\frac{n_1}{n} \times \frac{1}{x_1} + \frac{n_2}{n} \times \frac{1}{x_2} + \cdots + \frac{n_k}{n} \times \frac{1}{x_k}} \Rightarrow\end{aligned}$$

$$\overline{x}_h = \frac{1}{f_1 \times \frac{1}{x_1} + f_2 \times \frac{1}{x_2} + \cdots + f_k \times \frac{1}{x_k}} = \frac{1}{\sum_{i=1}^{k} \frac{f_i}{x_i}}$$

Essa última expressão será muito útil quando for apresentado o índice de Paasche.

Exemplo 3.1 *Consideremos novamente o exercício resolvido 6 da Seção 3.2 em que se considerou a compra de folhas de cartolina por uma escola a cada trimestre do ano de 2004. Naquele exercício, a quantia gasta a cada trimestre era constante. Suponhamos, agora, que essas quantias sejam variáveis, de acordo com o esquema mostrado na Tabela 3.16.*

Tabela 3.16: Preço da cartolina para o Exercício Resolvido 6 da Seção 3.2

Trimestre	Preço (R$)	Quantia (R$)
jan-mar	0,35	500,00
abr-jun	0,45	500,00
jul-set	0,50	450,00
out-dez	0,52	400,00

Nesse caso, o preço médio é dado por

$$\begin{aligned} p_m &= \frac{500+500+450+400}{\frac{500}{0,35}+\frac{500}{0,45}+\frac{450}{0,50}+\frac{400}{0,52}} = \frac{1850}{\frac{500}{0,35}+\frac{500}{0,45}+\frac{450}{0,50}+\frac{400}{0,52}} \\ &= \frac{1}{\frac{500}{1850}\times\frac{1}{0,35}+\frac{500}{1850}\times\frac{1}{0,45}+\frac{450}{1850}\times\frac{1}{0,50}+\frac{400}{1850}\times\frac{1}{0,52}} \\ &= 0,4395 \end{aligned}$$

que nada mais é que a média harmônica dos preços ponderada pelas quantias gastas.

3.7.7 Exercícios resolvidos da Seção 3.7

1. Para os dados da Tabela 3.17, calcule a média, o desvio padrão, a mediana, o intervalo interquartil, o oitavo decil e a moda pelos métodos de King e Czuber.

Tabela 3.17: Exercício Resolvido 1 da Seção 3.5

Classe	n_i
$[0,1)$	15
$[1,2)$	26
$[2,3)$	21
$[3,4)$	10
$[4,5)$	8

Solução:

Para facilitar a solução, vamos completar a tabela dada, acrescentando as colunas de freqüências relativas simples e acumuladas e também as colunas $f_i x_i$ e $f_i x_i^2$ necessárias para o cálculo da média e do desvio padrão. O resultado está na Tabela 3.18.

$$\overline{x} = \sum_i f_i x_i = 0,09375 + 0,48750 + \cdots + 0,45000 = 2,12500$$

Tabela 3.18: Solução do Exercício Resolvido 1 da Seção 3.5

Classe	x_i	n_i	f_i	N_i	F_i	$f_i x_i$	$f_i x_i^2$
$[0,1)$	0,5	15	0,1875	15	0,1875	0,09375	0,046875
$[1,2)$	1,5	26	0,3250	41	0,5125	0,48750	0,731250
$[2,3)$	2,5	21	0,2625	62	0,7750	0,65625	1,640625
$[3,4)$	3,5	10	0,1250	72	0,9000	0,43750	1,531250
$[4,5)$	4,5	8	0,1000	80	1,0000	0,45000	2,025000
Soma		80	1,0000			2,12500	5,975000

$$\begin{aligned}\sigma^2 &= \sum_i f_i x_i^2 - \overline{x}^2 = 0,046875 + 0,731250 + \cdots + 2,02500 - (2,12500)^2 = \\ &= 5,975000 - 4,515625 = 1,459375\end{aligned}$$

$$\sigma = \sqrt{1,459375} = 1,208046$$

A classe mediana é a classe $[1,2)$, onde acumula 51,25% da freqüência e cuja freqüência simples é 32,5%. A regra de três que define a mediana, baseada no subretângulo inferior, é, pois:

$$\frac{0,5000 - 0,1875}{Q_2 - 1} = \frac{0,32500}{2-1} \Rightarrow Q_2 = 1 + \frac{0,3125}{0,325} \Rightarrow Q_2 = 1,961538$$

Baseada no subretângulo superior, temos:

$$\frac{2 - Q_2}{51,25 - 50,00} = \frac{2-1}{32,50} \Rightarrow \frac{2 - Q_2}{1,25} = \frac{1}{32,5} \Rightarrow Q_2 = 2 - \frac{1,25}{32,50} \Rightarrow Q_2 = 1,961538$$

Note que podemos fazer os cálculos com as freqüências absolutas ou relativas, desde que trabalhemos com apenas uma delas de cada vez! No entanto, como estamos lidando com as áreas, que são representativas das freqüências relativas, é melhor uniformizar os procedimentos, utilizando sempre as freqüências relativas (multiplicadas por 100 ou não).

O primeiro quartil também está na classe $[1,2)$. A regra de três que o define é:

$$\frac{51,25 - 25,00}{2 - Q_1} = \frac{32,50}{2-1} \Rightarrow Q_1 = 2 - \frac{26,25}{32,50} \Rightarrow Q_1 = 1,1923077$$

O terceiro quartil está na classe $[2,3)$; trabalhando com o subretângulo superior, a regra de três que o define é:

$$\frac{77,5 - 75,0}{3 - Q_3} = \frac{26,25}{3-2} \Rightarrow Q_3 = 3 - \frac{2,5}{26,25} \Rightarrow Q_3 = 2,9047619$$

Assim, o intervalo interquartil é

$$IQ = 2,904762 - 1,1923076 = 1,7124543$$

A classe modal é a classe $[1,2)$. A moda pelo método de King é calculada através da seguinte proporção:

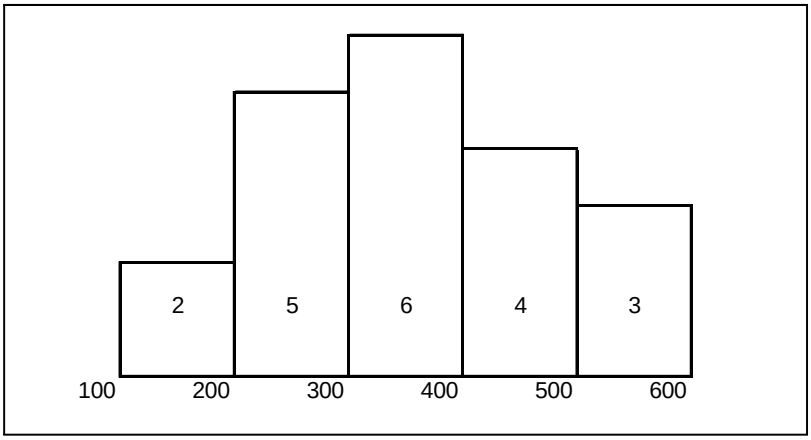
$$\frac{2 - x^*}{x^* - 1} = \frac{15}{21} \Rightarrow 15x^* - 15 = 42 - 21x^* \Rightarrow 36x^* = 57 \Rightarrow x^* = 1,583333$$

A moda pelo método de Czuber é calculada através da seguinte proporção:

$$\frac{2 - x^*}{x^* - 1} = \frac{26 - 21}{26 - 15} \Rightarrow \frac{2 - x^*}{x^* - 1} = \frac{5}{11} \Rightarrow 22 - 11x^* = 5x^* - 5 \Rightarrow 16x^* = 27 \Rightarrow x^* = 1,6875$$

2. Considere o histograma da Figura 3.21, onde, no interior dos retângulos, temos anotada a freqüência absoluta das classes. Calcule a média, o desvio padrão, a mediana refinada dos dados, a moda usando os métodos de King e Czuber e o terceiro e sétimo decis.

Figura 3.21: Histograma para o Exercício Resolvido 2 da Seção 3.5



Solução:

Considere a Tabela 3.19, construída para auxiliar a solução do exercício.

Tabela 3.19: Solução do Exercício Resolvido2 da Seção 3.5

Classe	Ponto médio	Freqüência. simples		Freqüência acumulada		Cálculo da média	Cálculo da variância
	x_i	n_i	f_i	N_i	F_i	$f_i x_i$	$f_i x_i^2$
$[100, 200)$	150	2	0,10	2	0,10	15,0	2250
$[200, 300)$	250	5	0,25	7	0,35	62,5	15625
$[300, 400)$	350	6	0,30	13	0,65	105,0	36750
$[400, 500)$	450	4	0,20	17	0,85	90,0	40500
$[500, 600)$	550	3	0,15	20	1,00	82,5	45375
Soma		20	1,00			355,0	140500

Como $\overline{x} = \sum_i f_i x_i$, resulta que $\overline{x} = 355,0$. A variância é calculada como

$$\sigma^2 = \sum_i f_i x_i^2 - \overline{x}^2 = 140500 - (355)^2 = 140500 - 126025 = 14475$$

e o desvio padrão como

$$\sigma = \sqrt{\sigma^2} = \sqrt{14475} = 120,3121$$

A mediana se encontra na terceira classe, $[300, 400)$, cujas freqüências relativas simples e acumulada são 0,30 e 0,65 respectivamente. Logo,

$$\frac{400 - Q_2}{0,65 - 0,50} = \frac{400 - 300}{0,30} \Rightarrow Q_2 = 350$$

A classe modal é a terceira classe, $[300, 400)$ e a moda pelo método de King é calculada através da seguinte proporção:

$$\frac{x^* - 300}{400 - x^*} = \frac{4}{5} \Rightarrow 5x^* - 1500 = 1600 - 4x^* \Rightarrow 9x^* = 3100 \Rightarrow x^* = 344,444$$

O método de Czuber resulta na seguinte proporção:

$$\frac{x^* - 300}{400 - x^*} = \frac{6 - 5}{6 - 4} \Rightarrow 2x^* - 600 = 400 - x^* \Rightarrow 3x^* = 1000 \Rightarrow x^* = 333,333$$

O terceiro decil está na classe $[200, 300)$, que tem freqüência simples igual a 0,25 e acumulada igual a 0,35. Logo,

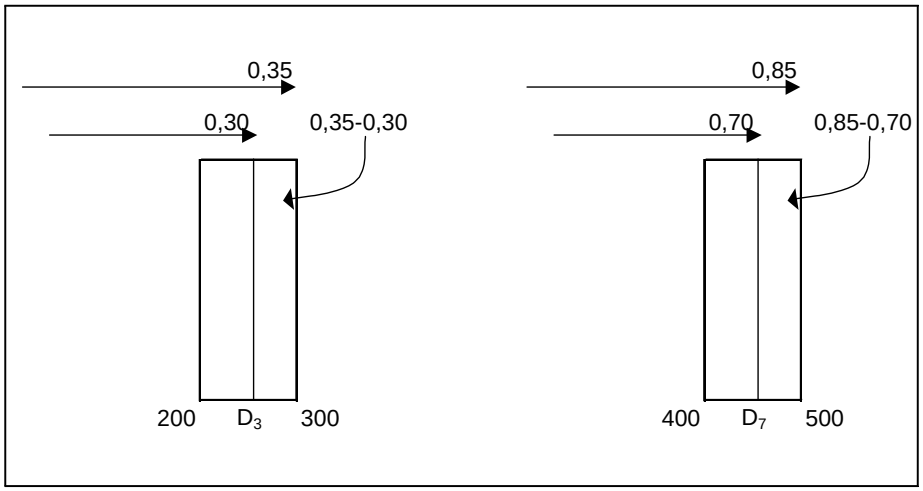
$$\frac{300 - D_3}{0,35 - 0,30} = \frac{300 - 200}{0,25} \Rightarrow D_3 = 280$$

O sétimo decil está na classe $[400, 500)$ que tem freqüência simples igual a 0,20 e acumulada igual a 0,85. Logo,

$$\frac{500 - D_7}{0,85 - 0,70} = \frac{500 - 400}{0,20} \Rightarrow D_3 = 425$$

O cálculo desses decis está ilustado na Figura 3.22.

Figura 3.22: Cálculo dos decis para o Exercício Resolvido2 da Seção 3.5



3.7.8 Exercícios propostos da Seção 3.7

3.11 *A idade média dos candidatos a um determinado curso de aperfeiçoamento oferecido por uma empresa foi sempre baixa, da ordem de 22 anos. Como esse curso foi preparado para todas as idades, decidiu-se fazer uma campanha de divulgação. Para verificar se a campanha foi ou não eficiente, fêz-se um levantamento da idade dos candidatos à última promoção, obtendo-se os resultados da Tabela 3.20.*

(a) Baseando-se nesses resultados, você diria que a campanha surtiu o efeito desejado?

(b) Um outro pesquisador decidiu usar o seguinte critério: se a diferença $\overline{X} - 22$ fosse maior que o valor $\alpha = \frac{2\sigma}{\sqrt{n}}$, então a campanha teria sido efetiva. Qual a conclusão dele?

Tabela 3.20: Idade dos candidatos de um curso de aperfeiçoamento

Idade	Freq.	%
$18 \vdash 20$	18	36,0
$20 \vdash 22$	12	24,0
$22 \vdash 26$	10	20,0
$26 \vdash 30$	8	16,0
$30 \vdash 36$	2	4,0
Total	50	

3.12 *Para os dados da tabela construída no Exercício 2.5 do Capítulo 2, calcule a média e o desvio padrão.*

3.13 *Com base na tabela de freqüência construída na solução do Exercício 2.6 do Capítulo 2, calcule a média e o desvio padrão. Compare com o resultado obtido no Exercício 3.6.*

3.14 *Para os dados da tabela construída no Exercício2.5 do Capítulo 2, calcule a moda utilizando os métodos de King e Czuber.*

3.15 *Com base na tabela de freqüência construída na solução do Exercício 2.6 do Capítulo 2, calcule a moda utilizando os métodos de King e Czuber.*

3.16 *Em uma granja foi observada a distribuição dos frangos com relação ao peso apresentada na Tabela 3.21.*

(a) Qual é a média da distribuição?

(b) Qual é a variância da distribuição?

(c) Construa o histograma.

(d) Queremos dividir os frangos em 4 categorias, com relação ao peso, de modo que

- *os 20% mais leves sejam da categoria D;*
- *os 30% seguintes sejam da categoria C;*
- *os 30% seguintes sejam da categoria B;*
- *os 20% restantes sejam da categoria A.*

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

(e) O granjeiro decide separar deste lote os animais com peso inferior a dois desvios padrões abaixo da média para receberem ração reforçada e também separar os animais com peso superior a um e meio desvio padrão acima da média para usá-los como reprodutores. Qual a porcentagem de animais que serão separados em cada caso?

3.17 *Para os dados da tabela construída no Exercício 2.5 do Capítulo 2, calcule a mediana e o terceiro decil.*

3.18 *Com base na tabela de freqüência construída na solução do Exercício 2.6 do Capítulo 2, calcule a mediana e o intervalo interquartil.*

Tabela 3.21: Peso dos frangos

Peso (gramas)	n_i
$960 \vdash 980$	60
$980 \vdash 1000$	160
$1000 \vdash 1020$	280
$1020 \vdash 1040$	260
$1040 \vdash 1060$	160
$1060 \vdash 1080$	80

3.8 Covariância e Correlação

Vimos que o diagrama de dispersão é um instrumento bastante útil na análise de duas variáveis quantitativas, pois exibe possíveis relações entre essas variáveis Na Tabelas 3.22 a 3.24 temos três conjuntos de dados, cujos diagramas de dispersão se encontram nas Figuras 3.23 a 3.25. Nesses gráficos, as linhas pontilhadas estão passando pelo ponto central do conjunto, isto é, pelo ponto $(\overline{x}, \overline{y})$.

Tabela 3.22: Variação diária das Bolsas de Valores - Junho 1993

Dia	Variação percentual		Dia	Variação percentual	
	Bovespa	BVRJ		Bovespa	BVRJ
1	4,9935	6,9773	17	-4,6706	-6,2360
2	5,5899	6,1085	18	0,6629	2,6259
3	3,8520	2,4847	21	1,1651	0,8728
4	0,9984	-0,1044	22	3,2213	4,8243
7	2,4872	2,4942	23	-2,7226	-4,7266
8	0,0142	0,1239	24	1,2508	-0,4985
9	-1,7535	-0,4221	25	7,1845	6,6798
11	8,1764	9,5148	28	2,5674	1,2299
14	0,6956	-1,7350	29	-1,3235	-3,0375
15	1,6164	2,2749	30	1,6685	1,2303
16	7,5829	15,4173			

Fonte: Folha de São Paulo (índice de fechamento)

Analisando esses gráficos , pode-se ver que as relações entre as variáveis envolvidas mudam; na Figura 3.23 existe uma tendência crescente entre as variáveis, isto é, quando o índice da Bovespa aumenta, o índice da BVRJ também tende a aumentar. Na Figura 3.24 essa relação se inverte, ou seja, aumentando a latitude, a temperatura tende a diminuir. Já na Figura 3.25 não é possível estabelecer nenhuma relação entre as variáveis, contrariando a superstição de que linhas da vida longas indicam maior longevidade.

3.8.1 Covariância

Vamos estudar, agora, uma medida de associação entre variáveis, que está relacionada ao tipo mais simples de associação: a linear. Então, tal medida irá representar o quanto a "nuvem" de dados em um diagrama de dispersão se aproxima de uma reta.

Tabela 3.23: Latitude e temperatura média de 15 cidades dos EUA

Latitude	Temperatura (^{o}F)
34	56,4
32	51,0
39	36,7
39	37,8
41	36,7
45	18,2
41	30,1
33	55,9
34	46,6
47	13,3
44	34,0
39	36,3
41	34,0
32	49,1
40	34,5

Fonte: Dunn e Clark (1974) p. 250

Tabela 3.24: Idade ao morrer e comprimento da "linha da vida"

Idade (anos)	Linha da vida (cm)	Idade (anos)	Linha da vida (cm)	Idade (anos)	Linha da da vida (cm)
19	9,75	65	8,85	74	8,85
40	9,00	65	9,75	74	9,60
42	9,60	66	8,85	75	6,45
42	9,75	66	9,15	75	9,76
47	11,25	66	10,20	75	10,20
49	9,45	67	9,15	76	6,00
50	11,25	68	7,95	77	8,85
54	9,00	68	8,85	80	9,00
56	7,95	68	9,00	82	9,75
56	12,00	69	7,80	82	10,65
57	8,10	69	10,05	82	13,20
57	10,20	70	10,50	83	7,95
58	8,55	71	9,15	86	7,95
61	7,20	71	9,45	88	9,15
62	7,95	71	9,45	88	9,75
62	8,85	72	9,45	94	9,00
65	8,25	73	8,10		

Figura 3.23: Variação diária das Bolsas de Valores - dados originais

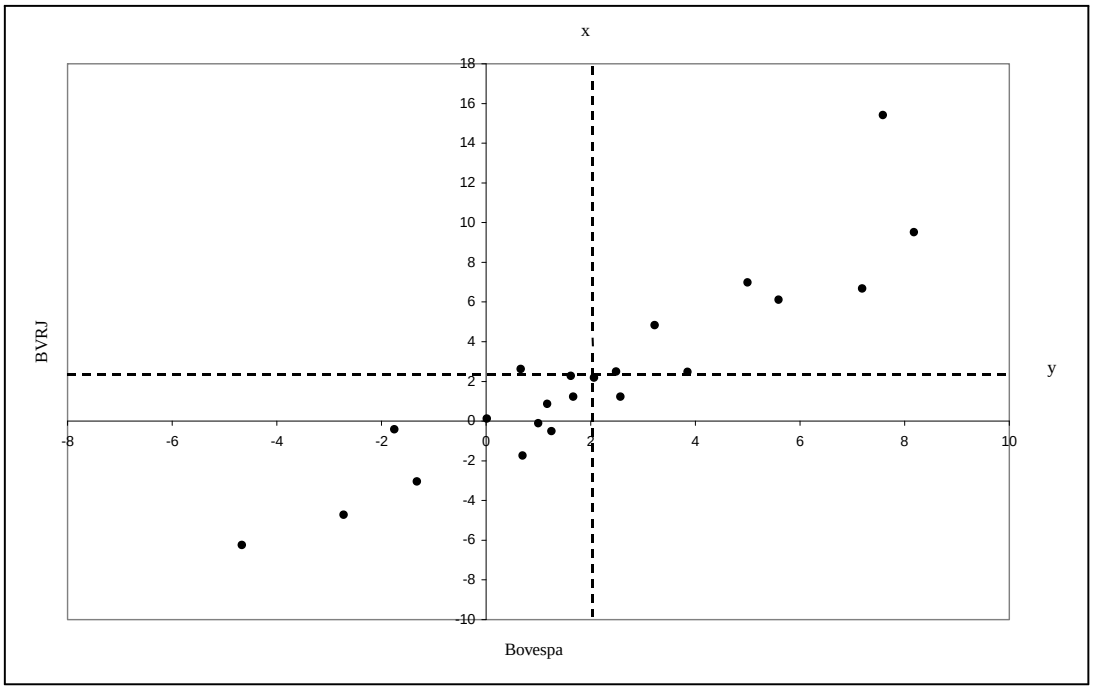


Figura 3.24: Latitude e temperatura média de 15 cidades dos EUA - dados originais

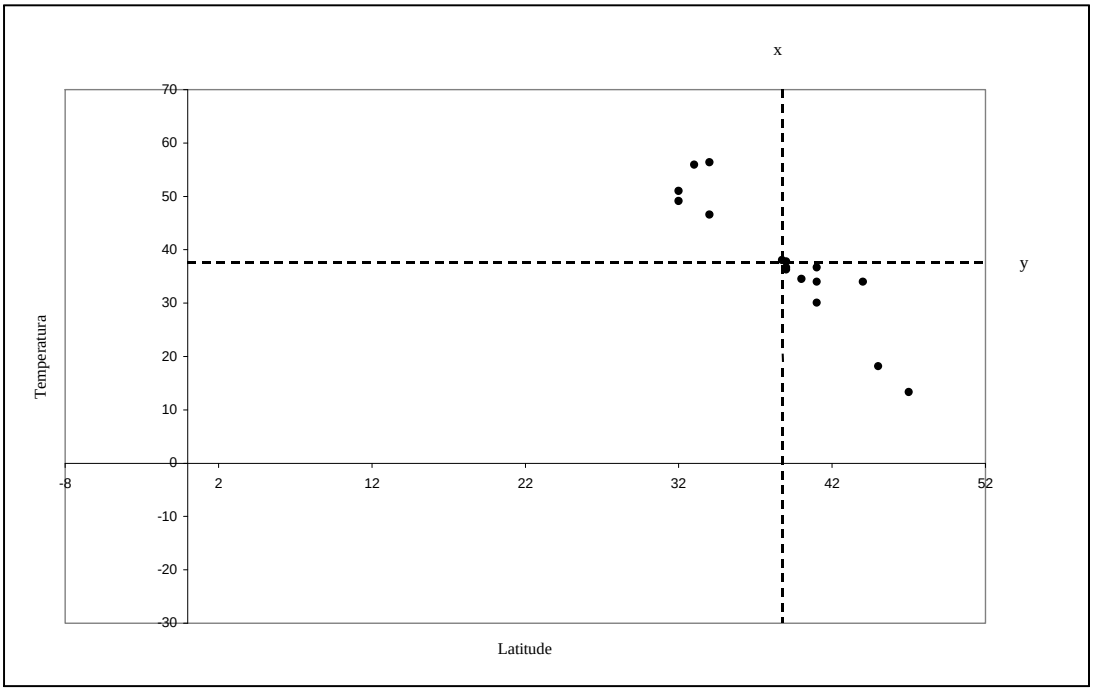
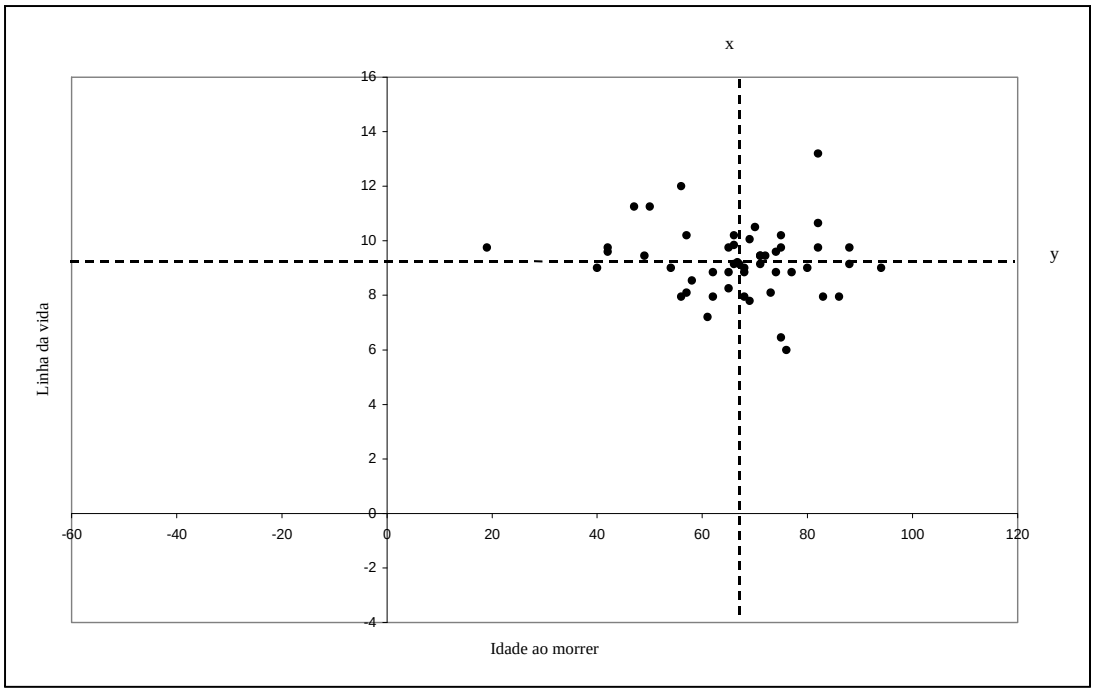


Figura 3.25: Diagrama de dispersão do comprimento da linha da vida e idade ao morrer - dados originais



Para diferenciar as três situações ilustradas nos gráficos anteriores, um primeiro ponto que devemos observar é o fato de as três "nuvens" de pontos estarem centradas em pontos diferentes, representados pela interseção dos eixos em linha pontilhada; note que esse é o ponto $(\overline{x}, \overline{y})$. Para facilitar comparações, é interessante uniformizar a origem, colocando as três nuvens centradas na origem $(0, 0)$. Lembrando as propriedades da média aritmética, sabe-se que a transformação $x_i - \overline{x}$ resulta em um conjunto de dados com média zero. Então, para quantificar as diferenças entre os gráficos anteriores, o primeiro ponto a considerar é a centralização da nuvem: em vez de trabalharmos com os dados originais (x_i, y_i),vamos trabalhar com os dados transformados $(x_i - \overline{x}, y_i - \overline{y})$. Nas Figuras 3.26 a 3.28 estão representados os diagramas de dispersão para essas variáveis transformadas, mantendo-se a mesma escala anterior.

Analisando esses três últimos gráficos, pode-se ver que, para o primeiro conjunto de dados, onde a tendência entre as variáveis é crescente, a maioria dos pontos está no primeiro e terceiro quadrantes, enquanto que, no segundo gráfico, onde a relação é decrescente, a maioria dos pontos está no segundo e quarto quadrantes.

O primeiro e terceiro quadrantes se caracterizam pelo fato de as abscissas e ordenadas terem o mesmo sinal e, portanto, seu produto é positivo; já no segundo e quarto quadrantes, as abscissas e ordenadas têm sinais opostos e, portanto, seu produto é negativo. Então, para diferenciar esses gráficos, podemos usar uma medida baseada no produto das coordenadas $x_i - \overline{x}$ e $y_i - \overline{y}$. Como no caso da variância ou desvio médio absoluto, para considerar todos os pares possíveis e descontar o número de observações, vamos tomar o valor médio desses produtos.

Figura 3.26: Variação diária das Bolsas de Valores - dados centrados na média

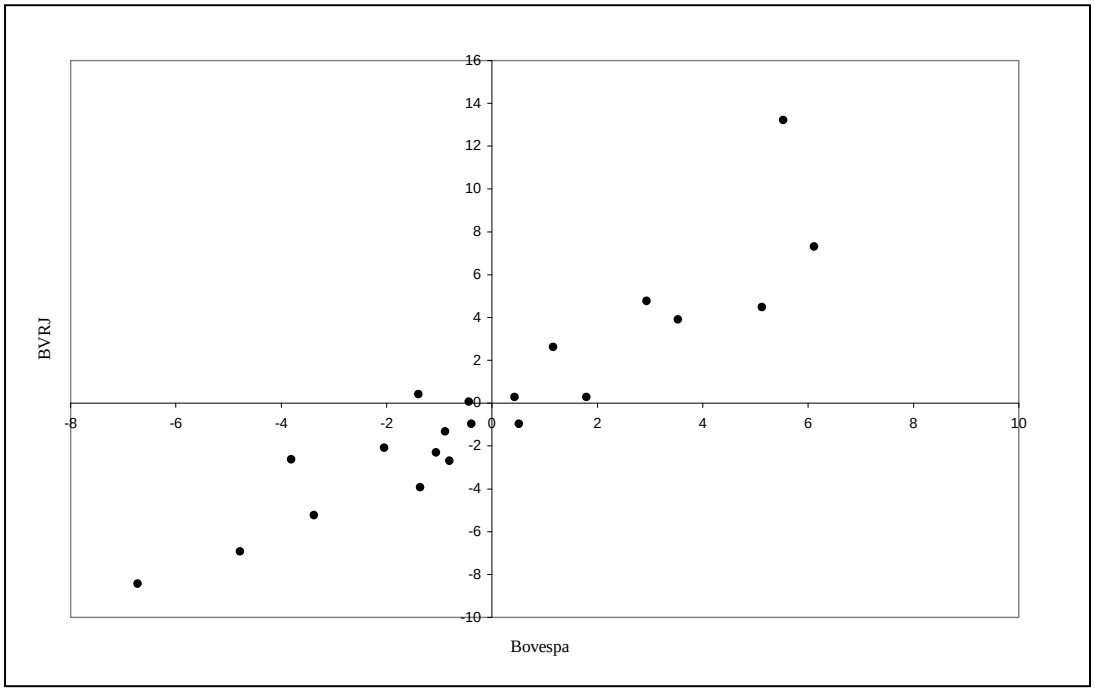


Figura 3.27: Latitude e temperatura média de 15 cidades dos EUA - dados centrados na média

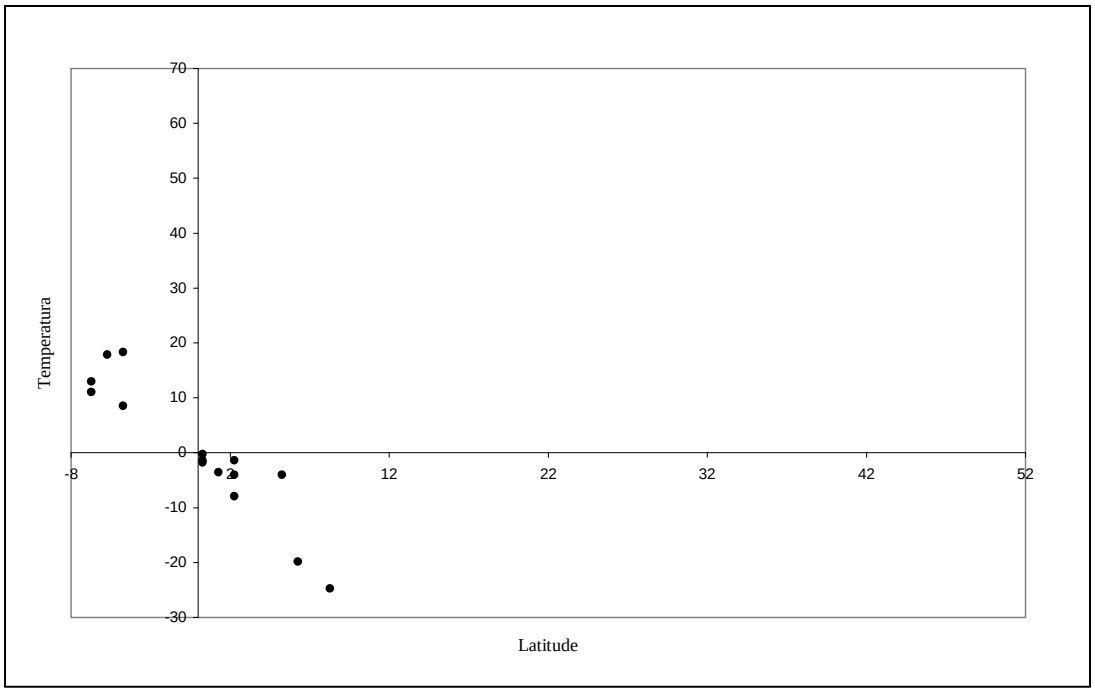
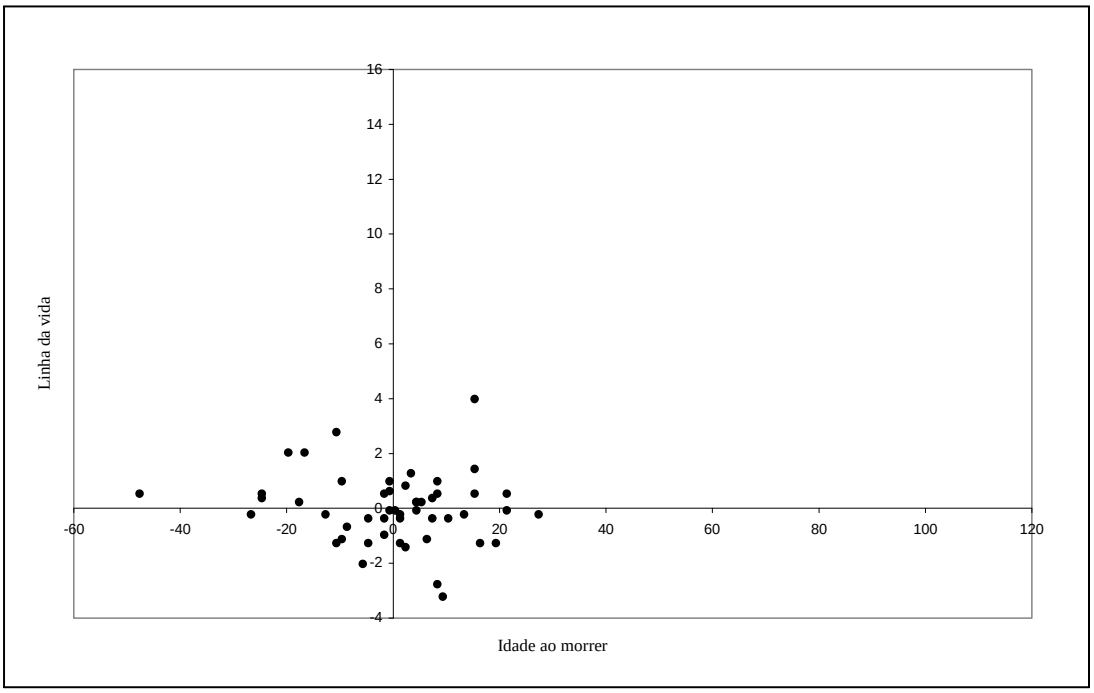


Figura 3.28: Diagrama de dispersão do comprimento da linha da vida e idade ao morrer - dados centrados na média



Definição 3.16 *A covariância entre as variáveis X e Y é definida por*

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{3.41}$$

onde x_i e y_i são os valores observados.

No gráfico 3.28, os pontos estão espalhados nos quatro quadrantes e, assim, essa média tende a ser nula, ou melhor, próxima de zero.

De maneira análoga à desenvolvida para a variância, a fórmula acima não é conveniente para fazer cálculos em máquinas de calcular mais simples. Assim, vamos desenvolver uma expressão alternativa. Note que:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) &= \sum_{i=1}^{n}(x_i y_i - x_i\overline{y} - y_i\overline{x} + \overline{x}\,\overline{y}) = \\
&= \sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i - \overline{x}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\overline{x}\,\overline{y} = \\
&= \sum_{i=1}^{n} x_i y_i - \overline{y}n\overline{x} - \overline{x}n\overline{y} + n\overline{x}\,\overline{y} = \\
&= \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}
\end{aligned}$$

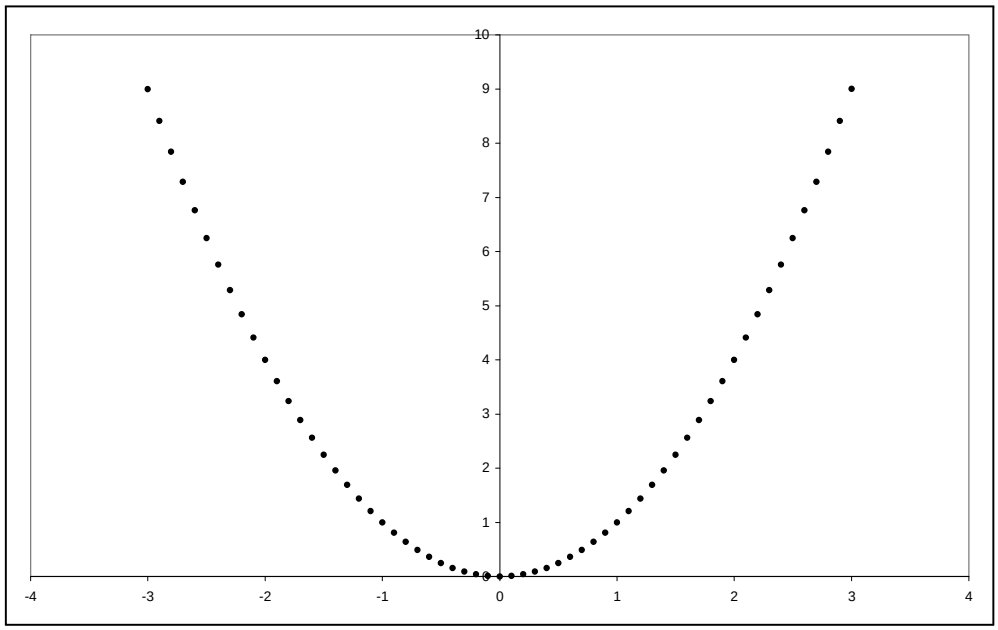
Logo,

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - n\overline{x\,y}\right) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y} \tag{3.42}$$

Da fórmula (3.42) podemos ver que a covariância é a "média dos produtos menos o produto das médias". Resulta também que a covariância entre X e X é a variância de X, isto é: $\text{Cov}(X, X) = \text{Var}(X)$.

É bastante importante salientar a interpretação da covariância: ela mede o grau de *associação linear* entre variáveis. Considerando o diagrama de dispersão da Figura 3.29, pode-se ver que existe uma associação quadrática perfeita entre as variáveis; no entanto, a covariância entre elas é nula!

Figura 3.29: Conjunto de dados com covariância nula



3.8.2 Coeficiente de correlação

Um dos problemas da covariância é a sua dependência da escala dos dados, o que faz com que seus valores possam variar de $-\infty$ a $+\infty$. Note que sua unidade de medida é dada pelo produto das unidades de medida das variáveis X e Y envolvidas. Então, fica difícil comparar situações como as ilustradas nos gráficos das Figuras 3.30 e 3.31; para a primeira, temos que $\text{Cov}(X, Y) = 304,51$ e para a segunda, $\text{Cov}(X, Y) = 609,02$. No entanto, os valores de X no primeiro conjunto variam de $-4,6706$ a $8,1764$ com um desvio padrão de $3,2757$ e no segundo conjunto de dados, variam de $-9,3412$ a $16,3528$, com um desvio padrão de $6,5514$.

Para uniformizar as escalas dos dados, iremos trabalhar com as variáveis padronizadas, isto é, $\dfrac{x_i - \overline{x}}{\sigma_x}$ e $\dfrac{y_i - \overline{y}}{\sigma_y}$. Como já visto, cada um dos conjuntos de dados assim tansformados tem desvio padrão igual a 1. Nas Figuras 3.32 a 3.34 temos o diagrama de dispersão para os dados transformados, novamente mantendo-se as escalas originais para facilitar a comparação.

Figura 3.30: Influência da escala na covariância - parte (a)

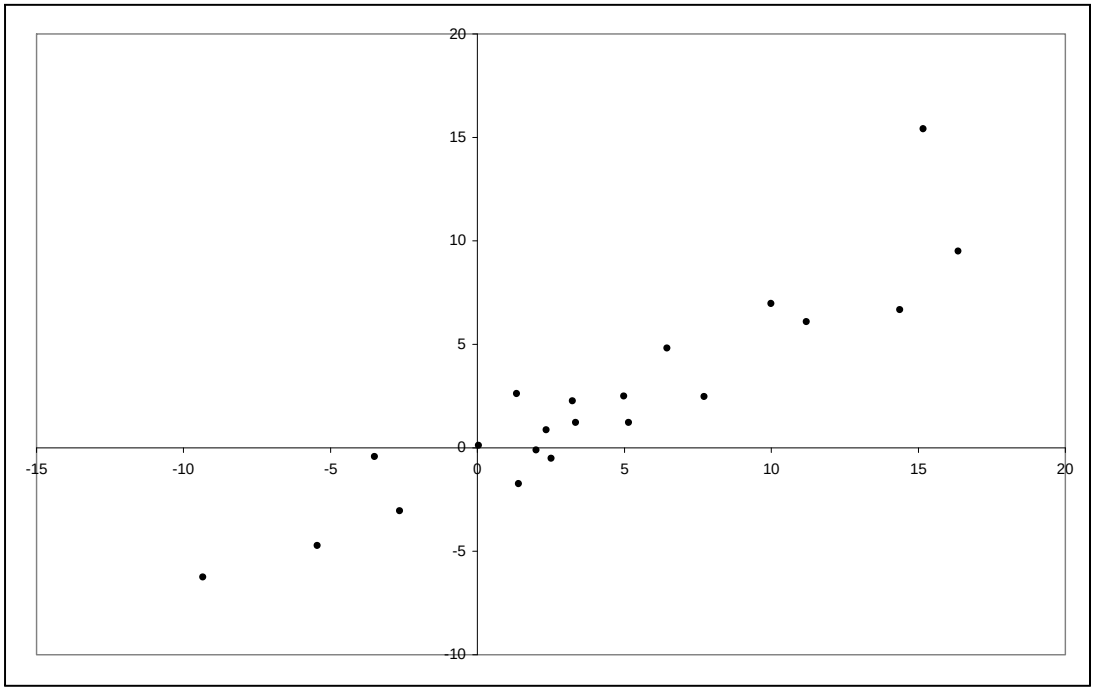


Figura 3.31: Influência da escala na covariância - parte (b)

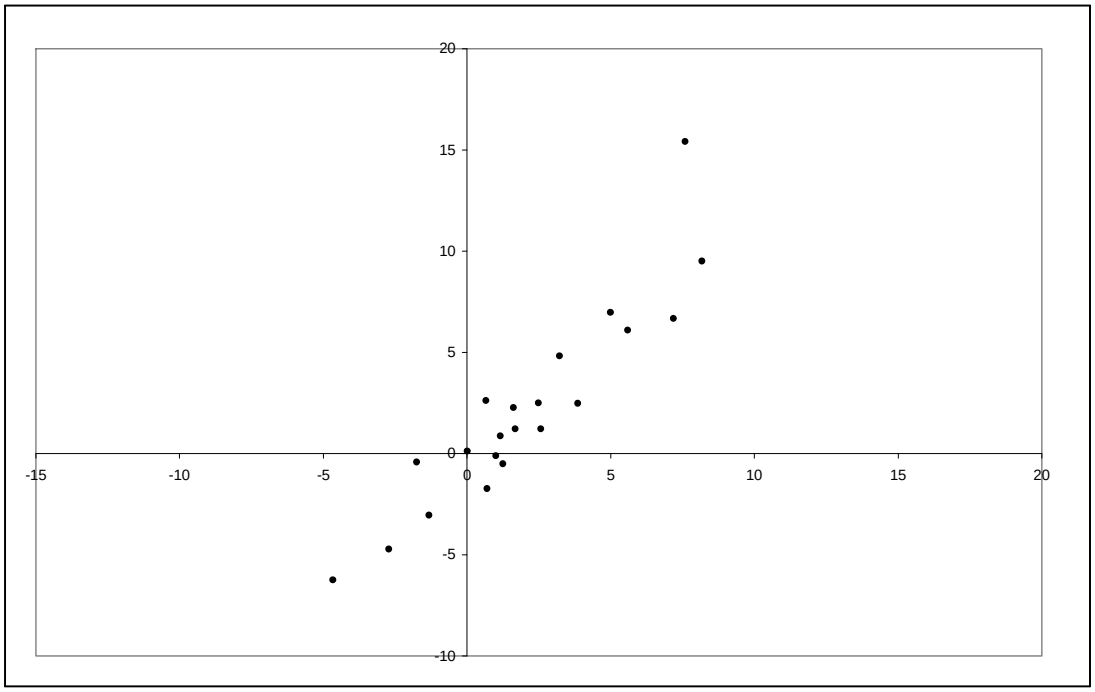


Figura 3.32: Variação diária nas Bolsas de Valores - dados padronizados

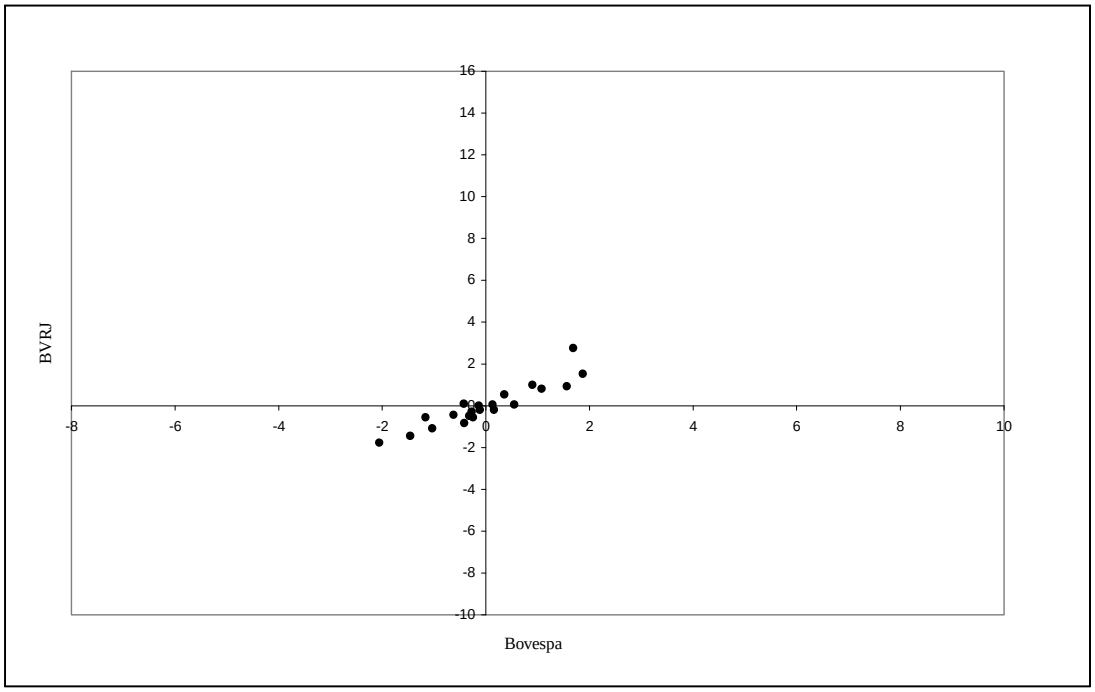


Figura 3.33: Latitude e temperatura média de 15 cidades dos EUA - dados padronizados

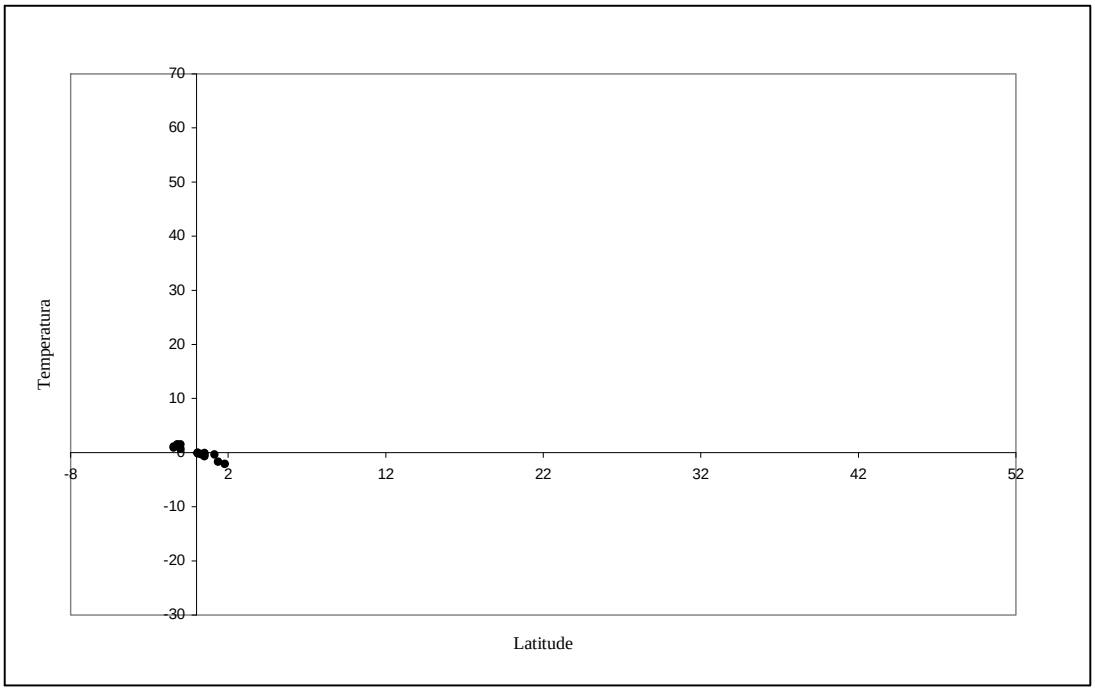
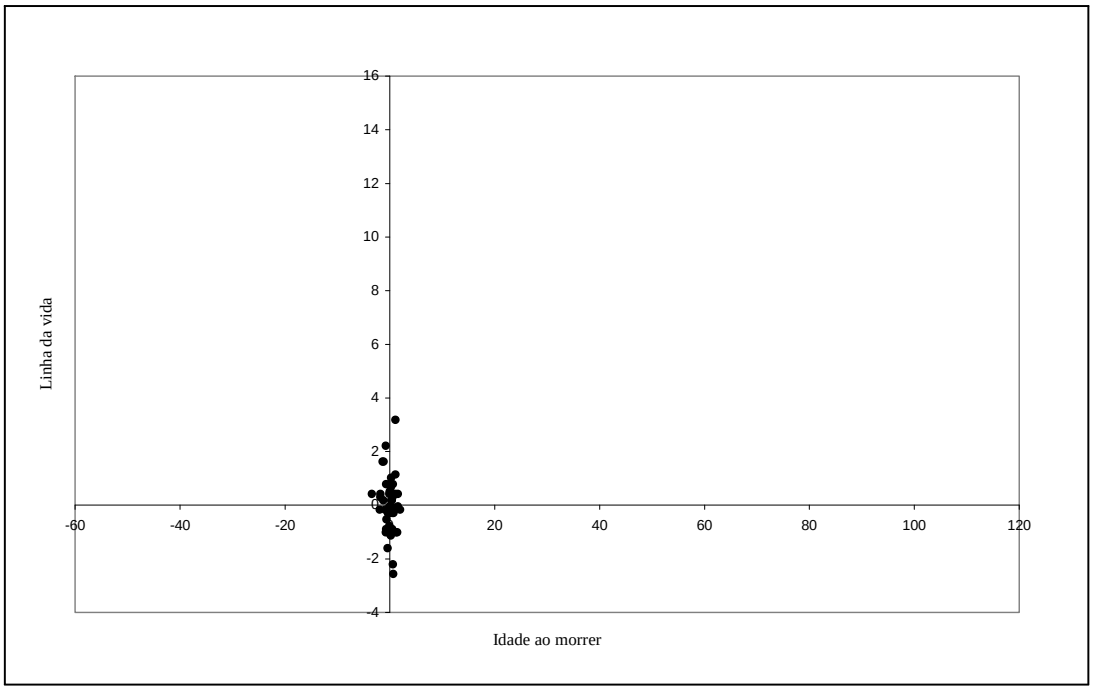


Figura 3.34: Diagrama de dispersão da idade ao morrer e comprimento da "linha da vida" - dados padronizados



Definição 3.17 *O coeficiente de correlação entre as variáveis X e Y é definido como*

$$\text{Corr}(X,Y) = \rho(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma_x}\right)\left(\frac{y_i - \overline{y}}{\sigma_y}\right) = \frac{\text{Cov}(X,Y)}{\sigma_x\,\sigma_y} \tag{3.43}$$

Os dois conjuntos de dados das Figuras 3.30 e 3.31 têm, ambos, o mesmo coeficiente de correlação igual a $0,9229$.

3.8.3 Propriedades da covariância e do coeficiente de correlação

Note que o coeficiente de correlação é adimensional! Além disso, ele tem uma propriedade bastante interessante, que é a seguinte:

$$-1 \leq \rho(X,Y) \leq 1 \tag{3.44}$$

Assim, valores do coeficiente de correlação próximos de 1 indicam uma forte associação linear crescente entre as variáveis, enquanto valores próximos de -1 indicam uma forte associação linear decrescente. Já valores próximos de zero indicam fraca associação linear (isso não significa que não exista algum outro tipo de associação; veja o caso da Figura 3.29). A demonstração da propriedade (3.44) é dada no Anexo 2 no final do capítulo.

Vamos ver agora o que acontece com a covariância e o coeficiente de correlação quando somamos uma constante aos dados e/ou multiplicamos os dados por uma constante. Vamos mostrar que

$$\text{Cov}(aX+b, cY+d) = ac\,\text{Cov}(X,Y) \tag{3.45}$$

e

$$\text{Corr}(aX+b, cY+d) = \frac{ac}{|ac|}\,\text{Corr}(X,Y) \tag{3.46}$$

De fato: fazendo $U = aX + b$ e $V = cY + d$, sabemos que $\overline{u} = a\overline{x} + b$ e $\overline{v} = c\overline{y} + d$ e $\sigma_u = |a|\,\sigma_x$ e $\sigma_v = |c|\,\sigma_y$ Logo,

$$\begin{aligned}
\mathrm{Cov}(U,V) &= \frac{1}{n}\sum_{i=1}^{n}(u_i - \overline{u})(v_i - \overline{v}) = \\
&= \frac{1}{n}\sum_{i=1}^{n}(ax_i + b - a\overline{x} - b)(cy_i + d - c\overline{y} - d) = \\
&= \frac{1}{n}\sum_{i=1}^{n}(ax_i - a\overline{x})(cy_i - c\overline{y}) = \\
&= \frac{ac}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \\
&= ac\,\mathrm{Cov}(X,Y)
\end{aligned}$$

Para o coeficiente de correlação, temos que

$$\begin{aligned}
\mathrm{Corr}(aX + b, cY + d) &= \mathrm{Corr}(U,V) = \frac{\mathrm{Cov}(U,V)}{\sigma_u\sigma_v} = \\
&= \frac{ac\,\mathrm{Cov}(X,Y)}{|c|\,\sigma_x.\,|d|\,\sigma_y} = \frac{ac}{|ac|}\,\mathrm{Corr}(X,Y)
\end{aligned}$$

Logo,

$$\mathrm{Corr}(aX + b, cY + d) = \begin{cases} \mathrm{Corr}(X,Y) & \text{se } ac > 0 \\ -\mathrm{Corr}(X,Y) & \text{se } ac < 0 \end{cases}.$$

3.8.4 Exercícios resolvidos da Seção 3.8

1. Considere novamente os dados sobre consumo de cigarros e mortes por câncer de pulmão, reproduzidos a seguir para facilitar a apresentação. Calcule o coeficiente de correlação entre as variáveis.

Tabela 3.25: Consumo de cigarros (X) e morte por câncer de pulmão (Y)

País	X	Y	País	X	Y
Islândia	240	63	Holanda	490	250
Noruega	255	100	Suiça	180	180
Suécia	340	140	Finlândia	1125	360
Dinamarca	375	175	Grã-Bretanha	1150	470
Canadá	510	160	Estados Unidos	1275	200
Austrália	490	180			

Solução:

Na tabela a seguir temos os detalhes dos cálculos a serem feitos no caso de se estar utilizando

uma calculadora mais simples.

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	240	63	57600	3969	15120
	255	100	65025	10000	25500
	340	140	115600	19600	47600
	375	175	140625	30625	65625
	510	160	260100	25600	81600
	490	180	240100	32400	88200
	490	250	240100	62500	122500
	180	180	32400	32400	32400
	1125	360	1265625	129600	405000
	1150	470	1322500	220900	540500
	1275	200	1625625	40000	255000
Soma	6430	2278	5365300	607594	1679045

A covariância de X e Y é a média do produto menos o produto das médias, ou seja:

$$\text{Cov}\,(X,Y) = \frac{1679045}{11} - \frac{6430}{11} \times \frac{2278}{11} = \frac{18469495 - 14647540}{121} = \frac{3821955}{121} = 31586,404959$$

A variância de cada variável é a "média dos quadrados menos o quadrado da média", ou seja:

$$\text{Var}\,(X) = \frac{5365300}{11} - \left(\frac{6430}{11}\right)^2 = \frac{59018300 - 41344900}{121} = \frac{17673400}{121} = 146061,157025$$

$$\text{Var}\,(Y) = \frac{607594}{11} - \left(\frac{2278}{11}\right)^2 = \frac{6683534 - 5189284}{121} = \frac{1494250}{121} = 12349,173554$$

Os desvios padrões são:

$$\sigma_x = 382,179483 \qquad \sigma_y = 111,126835$$

e, assim, o coeficiente de correlação é:

$$\rho\,(X,Y) = \frac{31586,404959}{382,179483 \times 111,1268354} = 0,743728$$

Essa correlação parece indicar que há um aumento no número de mortes por câncer do pulmão à medida que aumenta o número de cigarros consumidos.

Note como os cálculos foram feitos! Trabalhando com o denominador comum, reduz-se o número de divisões nos cálculos!

2. Calcule o coeficiente de correlação entre o preço de venda e a área das casas, cujos dados encontram-se na Tabela 2.49.

   Solução:

   Para esses dados, temos:

$$n = 59 \qquad \sum_{i=1}^{59} x_i = 10472 \qquad \sum_{i=1}^{59} y_i = 14433 \qquad \sum_{i=1}^{59} x_i y_i = 2667287$$

$$\sum_{i=1}^{59} x_i^2 = 1976810 \qquad \sum_{i=1}^{59} y_i^2 = 3736397$$

$$\text{Cov}\,(X,Y) = \frac{2667287}{59} - \frac{10472}{59} \times \frac{14433}{59} = \frac{157369933 - 151142376}{3481} = 1789,013789$$

$$\text{Var}\,(X) = \frac{1976810}{59} - \left(\frac{10472}{59}\right)^2 = \frac{116631790 - 109662784}{3481} = 2002,01264005$$

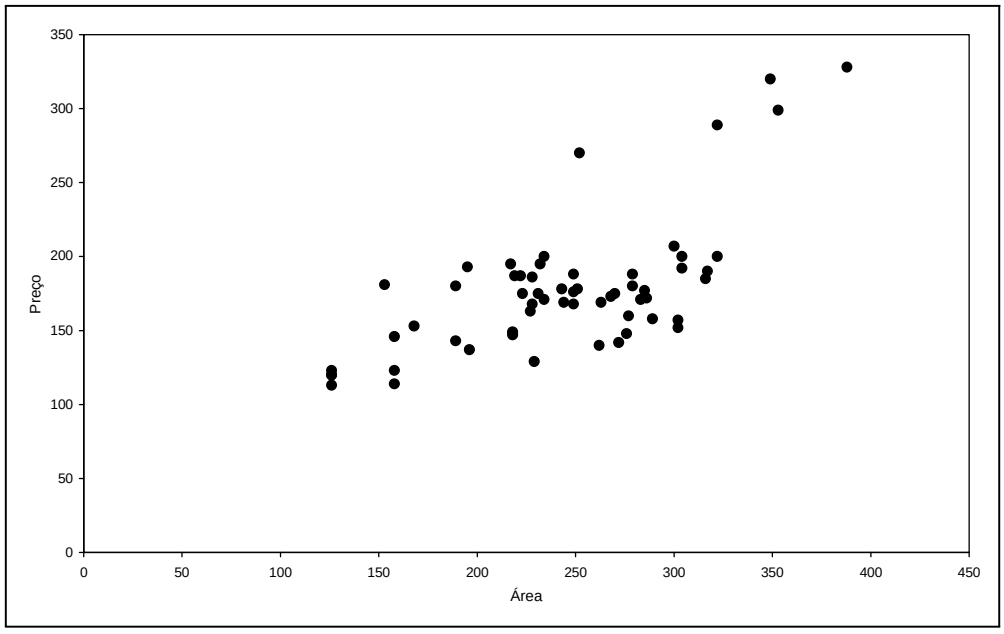
$$\text{Var}\,(Y) = \frac{3736397}{59} - \left(\frac{14433}{59}\right)^2 = \frac{220447423 - 208311489}{3481} = 3486,335536$$

$$\sigma_x = 44,74385589 \qquad \sigma_y = 59,04519909$$

$$\rho\,(X,Y) = \frac{1789,013789}{44,74385589 \times 59,04519909} = 0,677166849$$

Uma correlação positiva, bastante forte, indica que o preço de venda de uma casa aumenta à medida que aumenta a área útil da casa, conforme ilustrado na Figura 3.35

Figura 3.35: Preço de venda e área das casas de Boulder para o Exercício Resolvido2 da Seção 3.8



3.9 Exercícios Complementares

3.19 *Os dados da Tabela 3.26 representam as notas finais de 54 alunos da turma C1 de Estatística II no segundo semestre de 1992. Calcule a nota média, a nota mediana, a nota modal e o 1º quartil.*

3.20 *Segundo o critério de avaliação adotado pelo Departamento de Estatística, cada aluno será submetido a 2 provas, a primeira tendo peso 2 e a segunda tendo peso 3. Para ser aprovado, a média nas 2 provas tem que ser 6. Se um aluno tirar 5,5 na primeira prova, quanto deverá tirar na segunda prova para não ter que fazer Verificação Suplementar? E se as provas tivessem o mesmo peso?*

Tabela 3.26: Notas de 54 alunos para o Exercício 3.19 do Capítulo 3

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

3.21 *As notas de 1000 calouros na prova de Matemática da UFF estão apresentadas na Tabela 3.27.*

(a) Qual é a nota média? E a variância?

(b) Calcule o desvio médio absoluto.

(c) Os alunos com notas superiores a $\overline{x} + 1,5DP$ *(nota média mais 1,5 desvio padrão) serão convidados a participar de um programa especial de Iniciação Científica. Quantos alunos serão convidados?*

(d) Os alunos com 30% das notas mais baixas serão obrigados a assistir um curso de Cálculo Zero. Qual a menor nota necessária para o aluno não ter que frequentar esse curso especial?

Tabela 3.27: Notas de calouros para o Exercício 3.21 do Capítulo 3

Notas	Número de alunos
$0 \vdash 2$	55
$2 \vdash 3$	65
$3 \vdash 4$	172
$4 \vdash 5$	254
$5 \vdash 6$	278
$6 \vdash 7$	76
$7 \vdash 8$	75
$8 \vdash 10$	25

3.22 *Esboce um histograma de uma distribuição de dados com mesma média e mediana. Existe alguma classe de histogramas que apresente sempre essa característica?*

3.23 *Em 1993, o New York Mets teve o seu pior desempenho na Liga Principal de Beisebol (Estados Unidos). Eles foram bem pagos mas jogaram mal. Na Tabela 3.28 temos os salários anuais dos jogadores do Mets, em milhares de dólares.*[3]

(a) Calcule a média e o desvio padrão dos salários. Obs.: $\sum_{i=1}^{27} x_i = 38639$ *e* $\sum_{i=1}^{27} x_i^2 = 135079221$.

(b) Calcule a mediana e o intervalo interquartil IQ.

(c) Usando o critério $1,5 \times IQ$*, liste os possíveis outliers.*

(d) Com base nos resultados anteriores, qual das medidas você usaria para representar o salário dos jogadores do Mets?

3.24 *No controle de qualidade da produção de cigarros, o peso é uma característica importante. Na Tabela 3.29 temos a distribuição de freqüências acumuladas para o peso (em miligramas) dos cigarros de um lote inspecionado.*

[3] Dados extraídos de Moore e McCabe (1999).

Tabela 3.28: Salários dos Mets para o Exercício 3.23 do Capítulo 3

6200	5917	4000	3375	3000	2312	2300	2150	2100
1500	1012	850	650	635	500	475	220	205
195	195	158	145	109	109	109	109	109

(a) Construa a tabela de freqüências completa, com colunas auxiliares para o cálculo da média e do desvio padrão.

(b) Calcule o peso médio e o desvio padrão do peso, não esquecendo de indicar a unidade de medida dessas estatísticas.

(c) Calcule o peso modal, usando os métodos de King e Czuber; indique a unidade de medida.

(d) Calcule o peso mediano; indique a unidade de medida.

(e) Usando a regra $1,5 \times IQ$*,você diria que alguns cigarros têm pesos discrepantes neste lote? Em caso afirmativo, estime essas percentagens.*

Tabela 3.29: Pesos de cigarros para o Exercício 3.24 do Capítulo 3

Classes de peso (mg)	Freq. Acum. N_i
$760 \vdash 780$	4
$780 \vdash 800$	47
$800 \vdash 820$	165
$820 \vdash 840$	333
$840 \vdash 860$	450
$860 \vdash 880$	489
$880 \vdash 900$	500

3.25 *Os 4 conjuntos de dados apresentados na Tabela 3.30 constam de Anscombe(1973). Para cada um deles construa o diagrama de dispersão e calcule a média, o desvio padrão e o coeficiente de correlação. Comente os resultados obtidos.*

Tabela 3.30: Dados de Anscombe para o Exercício 3.25 do Capítulo 3

Conjunto 1		Conjunto 2		Conjunto 3		Conjunto 4	
X	Y	X	Y	X	Y	X	Y
10,0	9,14	8,0	6,58	10	8,04	10,0	7,46
8,0	8,14	8,0	5,76	8	6,95	8,0	6,77
13,0	8,74	8,0	7,71	13	7,58	13,0	12,74
9,0	8,77	8,0	8,84	9	8,81	9,0	7,11
11,0	9,26	8,0	8,47	11	8,33	11,0	7,81
14,0	8,10	8,0	7,04	14	9,96	14,0	8,84
6,0	6,13	8,0	5,25	6	7,24	6,0	6,08
4,0	3,10	19,0	12,50	4	4,26	4,0	5,39
12,0	9,13	8,0	5,56	12	10,84	12,0	8,15
7,0	7,26	8,0	7,91	7	4,82	7,0	6,42
5,0	4,74	8,0	6,89	5	5,68	5,0	5,73

3.26 *Muitas vezes a determinação da capacidade de produção instalada para certo tipo de indústria em certos tipos de localidades é um processo difícil e custoso. Como alternativa, pode-se estimar a capacidade de produção através de uma outra variável de medida mais fácil, que esteja linearmente relacionada com ela. Suponha que foram observados os valores, dados na Tabela 3.31, para as variáveis capacidade de produção instalada, potência instalada e área construída. Com base num critério estatístico, qual das variáveis você escolheria para estimar a capacidade de produção instalada?*

Tabela 3.31: Dados de capacidade da produção para o Exercício 3.26 do Capítulo 3

X: capacidade de produção instalada (ton)									
4	5	4	5	8	9	10	11	12	12
Y: potência instalada (1000 kW)									
1	1	2	3	3	5	5	6	6	6
Z: área construída (100 m^2)									
6	7	10	10	11	9	12	10	11	14

Anexo 1: Relação entre as médias aritmética, geométrica e harmônica

Sejam $x_1, x_2, \ldots, x_n$ números reais positivos. Vamos mostrar que, nesse caso, é válida a seguinte relação entre as médias aritmética ($\overline{x}$), geométrica ($\overline{x}_g$) e harmônica ($\overline{x}_h$):[4]

$$\overline{x}_h \leq \overline{x}_g \leq \overline{x} \tag{3.47}$$

Para isso vamos usar a seguinte desigualdade, cuja demonstração apresentaremos posteriormente: sejam $p_1, p_2, \ldots, p_n$ números reais não negativos tais que $\sum_{i=1}^{n} p_i = 1$. Então

$$\prod_{i=1}^{n} x_i^{p_i} \leq \sum_{i=1}^{n} p_i x_i \tag{3.48}$$

ou equivalentemente (tomando logaritmo, que é uma função estritamente crescente),

$$\sum_{i=1}^{n} p_i \ln x_i \leq \ln\left(\sum_{i=1}^{n} p_i x_i\right) \tag{3.49}$$

Fazendo $p_i = \dfrac{1}{n}$ em (3.48), obtém-se que:

$$\prod_{i=1}^{n} x_i^{\frac{1}{n}} \leq \sum_{i=1}^{n} \frac{1}{n} x_i$$

ou

$$\boxed{\overline{x}_g \leq \overline{x}} \tag{3.50}$$

Como o resultado (3.50) vale para quaisquer números reais positivos, vale em particular para $y_i = \dfrac{1}{x_i}$, isto é:

$$\begin{aligned} \overline{y}_g &\leq \overline{y} \Leftrightarrow \left(\frac{1}{x_1} \times \frac{1}{x_2} \times \cdots \times \frac{1}{x_n}\right)^{\frac{1}{n}} \leq \frac{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}{n} \Leftrightarrow \\ \frac{1}{(x_1 \times x_2 \times \cdots \times x_n)^{\frac{1}{n}}} &\leq \frac{1}{\frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}} \Leftrightarrow \frac{1}{\overline{x}_g} \leq \frac{1}{\overline{x}_h} \Leftrightarrow \boxed{\overline{x}_g \geq \overline{x}_h} \end{aligned}$$

o que prova a segunda parte da desigualdade (3.47).

Para completar a prova do resultado, temos que demonstrar que é válida a relação (3.49). Essa demonstração será feita por indução em n, o número de valores.

Para $n = 2$, o resultado segue da concavidade da função logaritmo. Note que a derivada segunda da função $f(x) = \ln(x)$ é $f''(x) = -\dfrac{1}{x^2} < 0 \ \forall x > 0$.

A definição de função côncava é a seguinte: uma função $f : I \to \mathbb{R}$, definida num intervalo I, é côncava quando, para $a < x < b$ arbitrários em I, o ponto $(x, f(x))$ do gráfico está situado acima do segmento de reta que liga os pontos $(a, f(a))$, $(b, f(b))$. Na Figura 3.36 temos a ilustração dessa definição. Usando a forma paramétrica da equação de um segmento de reta que liga dois pontos quaisquer, essa condição de concavidade da função logaritmo nos diz que

$$t \ln(x_1) + (1-t)\ln(x_2) \leq \ln[t x_1 + (1-t)x_2]$$

[4] Demonstração apresentada aos autores pelo Prof. Hamilton Prado Bueno (Ph.D.) - UFMG

o que prova o resultado para $n = 2$, já que $t + (1 - t) = 1$.

Suponhamos a relação válida para n; vamos provar que vale para $n + 1$.De fato:

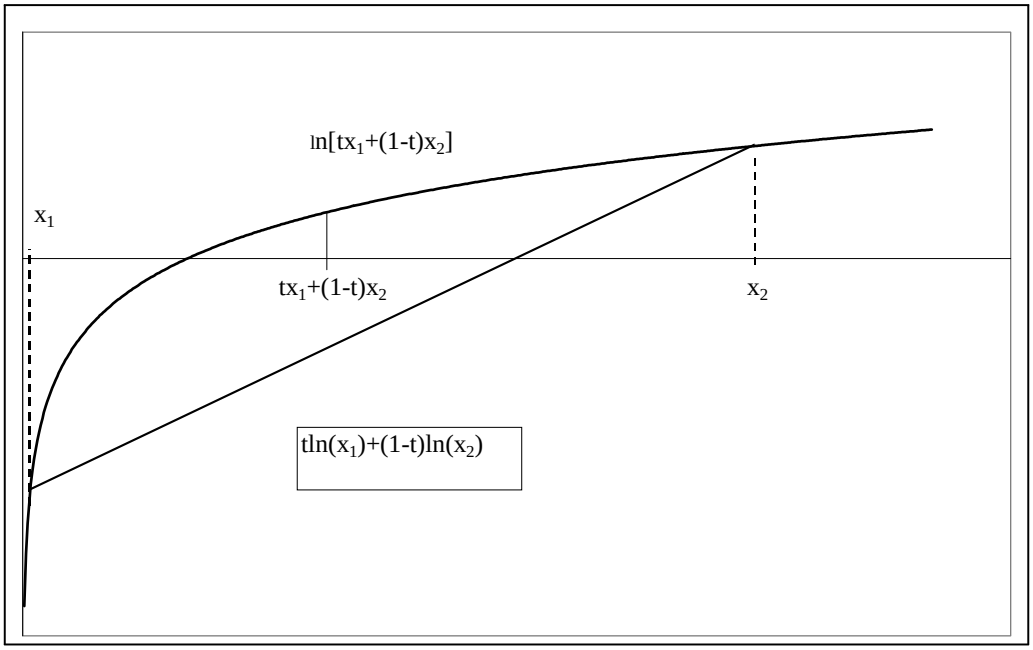
$$\begin{aligned} & \ln(p_1x_1 + \cdots + p_nx_n + p_{n+1}x_{n+1}) \\ = & \ln\left[(p_1 + \cdots + p_n)\left(\frac{p_1x_1}{p_1 + \cdots + p_n} + \cdots + \frac{p_nx_n}{p_1 + \cdots + p_n}\right) + p_{n+1}x_{n+1}\right] \geq \\ \geq & (p_1 + \cdots + p_n)\ln\left[\left(\frac{p_1x_1}{p_1 + \cdots + p_n} + \cdots + \frac{p_nx_n}{p_1 + \cdots + p_n}\right)\right] + p_{n+1}\ln(x_{n+1}) \end{aligned}$$

Essa desigualdade segue do fato de que o resultado vale para $n = 2$. Aplicando a hipótese de indução no primeiro termo do membro direito da desigualdade obtemos que:

$$\begin{aligned} & \ln(p_1x_1 + \cdots + p_nx_n + p_{n+1}x_{n+1}) \\ \geq & (p_1 + \cdots + p_n)\ln\left[\left(\frac{p_1x_1}{p_1 + \cdots + p_n} + \cdots + \frac{p_nx_n}{p_1 + \cdots + p_n}\right)\right] + p_{n+1}\ln(x_{n+1}) \\ \geq & (p_1 + \cdots + p_n)\left[\frac{p_1}{p_1 + \cdots + p_n}\ln(x_1) + \cdots + \frac{p_n}{p_1 + \cdots + p_n}\ln(x_n)\right] + p_{n+1}\ln(x_{n+1}) \\ = & p_1\ln(x_1) + \cdots + p_n\ln(x_n) + p_{n+1}\ln(x_{n+1}) = \sum_{i=1}^{n+1} p_ix_i \end{aligned}$$

e isso completa a prova.

Figura 3.36: Ilustração da concavidade da função logaritmo



Anexo 2: Demonstração da propriedade (3.44)

Para demonstrar a propriedade (3.44), precisamos de algumas definições referentes a vetores no espaço euclidiano $\mathbb{R}^n$.

Definição 3.18 *Dados dois vetores* $\mathbf{u} = (u_1, \ldots, u_n)$ *e* $\mathbf{v} = (v_1, \ldots, v_n)$ *em* $\mathbb{R}^n$ *define-se:*

1. *o produto interno dos vetores* $\mathbf{u}$ *e* $\mathbf{v}$, *representado por* $\langle \mathbf{u}, \mathbf{v} \rangle$, *como*

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^{n} u_i v_i \quad ;$$

2. *a norma de um vetor* $\mathbf{u}$, *representada por* $\|\mathbf{u}\|$,*como*

$$\|\mathbf{u}\| = \sqrt{\sum_{i=1}^{n} u_i^2} = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle} \quad .$$

A demonstração da propriedade é uma conseqüência direta da desigualdade de Cauchy-Schwarz, que estabelecemos a seguir.

Teorema 3.1 *Dados dois vetores* $\mathbf{u} = (u_1, \ldots, u_n)$ *e* $\mathbf{v} = (v_1, \ldots, v_n)$ *em* $\mathbb{R}^n$,*então*

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \, \|\mathbf{v}\| \tag{3.51}$$

Demonstração: Se $\mathbf{u} = \mathbf{0}$ ou $\mathbf{v} = \mathbf{0}$, onde $\mathbf{0}$ é o vetor nulo, a desigualdade é trivial pois resulta $0 \leq 0$. Precisamos considerar, então, apenas a situação em que $\mathbf{u} \neq \mathbf{0}$ e $\mathbf{v} \neq \mathbf{0}$.

Dados dois números reais x e y, é verdade que

$$0 \leq (x - y)^2 = x^2 - 2xy + y^2$$

Logo,

$$2xy \leq x^2 + y^2$$

Fazendo

$$x = \frac{|u_i|}{\|\mathbf{u}\|} \qquad \text{e} \qquad y = \frac{|v_i|}{\|\mathbf{v}\|}$$

resulta que

$$2\frac{|u_i|}{\|\mathbf{u}\|}\frac{|v_i|}{\|\mathbf{v}\|} \leq \frac{|u_i|^2}{\|\mathbf{u}\|^2} + \frac{|v_i|^2}{\|\mathbf{v}\|^2}$$

Mas $|u_i| \, |v_i| = |u_i v_i|$ e $|u_i|^2 = u_i^2$.Logo,

$$2\frac{|u_i v_i|}{\|\mathbf{u}\| \, \|\mathbf{v}\|} \leq \frac{u_i^2}{\|\mathbf{u}\|^2} + \frac{v_i^2}{\|\mathbf{v}\|^2}$$

Somando membro a membro para $i = 1, 2, \ldots, n$, resulta

$$2\frac{\sum_{i=1}^{n} |u_i v_i|}{\|\mathbf{u}\| \, \|\mathbf{v}\|} \leq \frac{\sum_{i=1}^{n} u_i^2}{\|\mathbf{u}\|^2} + \frac{\sum_{i=1}^{n} v_i^2}{\|\mathbf{v}\|^2}$$

Logo,

$$2\frac{\sum_{i=1}^{n}|u_iv_i|}{\|\mathbf{u}\|\,\|\mathbf{v}\|} \leq \frac{\|\mathbf{u}\|^2}{\|\mathbf{u}\|^2} + \frac{\|\mathbf{v}\|^2}{\|\mathbf{v}\|^2} = 2$$

e, então:

$$\frac{\sum_{i=1}^{n}|u_iv_i|}{\|\mathbf{u}\|\,\|\mathbf{v}\|} \leq 1$$

ou

$$\sum_{i=1}^{n}|u_iv_i| \leq \|\mathbf{u}\|\,\|\mathbf{v}\|$$

Como $\left|\sum_{i=1}^{n}u_iv_i\right| \leq \sum_{i=1}^{n}|u_iv_i|$, segue que

$$|\langle u, v\rangle| \leq \|\mathbf{u}\|\,\|\mathbf{v}\|$$

ou equivalentemente,

$$\left|\sum_{i=1}^{n}u_iv_i\right| \leq \sqrt{\sum_{i=1}^{n}u_i^2} \times \sqrt{\sum_{i=1}^{n}v_i^2} \tag{3.52}$$

Fazendo, em (3.52),

$$u_i = \frac{x_i - \overline{x}}{\sigma_X} \qquad \text{e} \qquad v_i = \frac{y_i - \overline{y}}{\sigma_Y}$$

obtém-se que

$$\left|\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{\sigma_X}\right)\left(\frac{y_i-\overline{y}}{\sigma_Y}\right)\right| \leq \sqrt{\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{\sigma_X}\right)^2} \times \sqrt{\sum_{i=1}^{n}\left(\frac{y_i-\overline{y}}{\sigma_Y}\right)^2}$$

ou

$$\left|\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{\sigma_X}\right)\left(\frac{y_i-\overline{y}}{\sigma_Y}\right)\right| \leq \sqrt{\frac{1}{\sigma_X^2}\sum_{i=1}^{n}(x_i-\overline{x})^2} \times \sqrt{\frac{1}{\sigma_Y^2}\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

ou

$$\left|\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{\sigma_X}\right)\left(\frac{y_i-\overline{y}}{\sigma_Y}\right)\right| \leq \sqrt{\frac{1}{\sigma_X^2}n\sigma_X^2} \times \sqrt{\frac{1}{\sigma_Y^2}n\sigma_Y^2}$$

ou

$$\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{\sigma_X}\right)\left(\frac{y_i-\overline{y}}{\sigma_Y}\right)\right| \leq 1$$

ou finalmente

$$|\mathrm{Corr}(X,Y)| \leq 1$$

como queríamos demonstrar.

Capítulo 4

Solução dos Exercícios

4.1 Capítulo 2

Seção 2.2

1. Podemos pensar em uma variável Bens (qualitativa) com categorias Máquina de lavar, TV, Geladeira e outra variável Serviços (qualitativa) com categorias Rede de água, Rede de esgoto, Telefone. Uma outra forma é olhar cada um dos bens e serviços como uma variável qualitativa com categorias Sim/Não. Na verdade, essa é a forma da pergunta no questionário da pesquisa. Uma outra variável envolvida é a Renda, que neste caso é uma variável qualitativa, já que só aparecem duas classes.

2. As variáveis são: Tipo de estabelecimento (Público ou Privado); Nível de ensino (Pré-escolar, 1º grau, 2º grau, Superior), Número de estabelecimentos e Número de alunos matriculados. As duas primeiras são qualitativas, sendo que a segunda tem uma escala ordinal. As duas últimas são variáveis quantitativas discretas.

Seção 2.3

3. Ver tabela 4.1.

Tabela 4.1: Solução do Exercício 3 do Capítulo 2

Notas	Freqüência simples		Freqüência acumulada	
	Absoluta	Relativa (%)	Absoluta	Relativa (%)
1	3	0,06	3	0,06
2	4	0,08	7	0,14
3	4	0,08	11	0,22
4	5	0,10	16	0,32
5	8	0,16	24	0,48
6	8	0,16	32	0,64
7	5	0,10	37	0,74
8	5	0,10	42	0,84
9	5	0,10	47	0,94
10	3	0,06	50	1,00

4. O menor valor é 0,7 e o maior é 7472. Vamos arredondar o menor valor para 0, o que resulta em uma amplitude de 7472. Para trabalhar com comprimentos de classe inteiros, aproximamos a amplitude para o próximo múltiplo do número de classes, o que dá 7475. Resulta, então, um comprimento de classe igual a 1495. Ver tabela 4.2.

Tabela 4.2: Solução do Exercício 4 do Capítulo 2

Quantidade de ovos (milhões)	Freqüência simples		Freqüência acumulada	
	Absoluta	Relativa (%)	Absoluta	Relativa (%)
$0 \vdash 1495$	37	74,0	37	74,0
$1495 \vdash 2990$	5	10,0	42	84,0
$2990 \vdash 4485$	4	8,0	46	92,0
$4485 \vdash 5980$	3	6,0	49	98,0
$5980 \vdash 7475$	1	2,0	50	100,0

5. Ver Tabela 4.3. Note que, como não sabemos o número de famílias, não é possível calcular as freqüências absolutas.

Tabela 4.3: Solução do Exercício 5 do Capítulo 2

Consumo de leite (litros)	Freqüência simples Relativa (%)	Freqüência acumulada Relativa (%)
$0 \vdash 1$	20,0	20,0
$1 \vdash 2$	50,0	70,0
$2 \vdash 3$	20,0	90,0
$3 \vdash 5$	10,0	100,0

6. Ver Tabela 4.4. A variável de estudo é número de empregados.

Tabela 4.4: Solução do Exercício 6 do Capítulo 2

Número de empregados	Freqüência simples		Freqüência acumulada	
	Absoluta	Relativa (%)	Absoluta	Relativa (%)
$152 \vdash 6277$	51	63,75	51	63,75
$6277 \vdash 12402$	21	26,25	72	90,00
$12402 \vdash 18527$	4	5,00	76	95,00
$18527 \vdash 24652$	3	3,75	79	98,75
$24652 \vdash 30777$	1	1,25	80	100,0

7. Ver Tabela 4.5.

Seção 2.4

8. Para o Exercício 2.3: Figura 4.1

   Para o Exercício 2.4: Figura 4.2

   Para o Exercício 2.5: Figura 4.3

   Para o Exercício 2.6: Figura 4.4

   Para o Exercício 2.7 é: Figura 4.5

Tabela 4.5: Solução do Exercício 7 do Capítulo 2

Situação	Freqüência simples		Freqüência acumulada	
	Absoluta	Relativa (%)	Absoluta	Relativa (%)
Reprovado	9	16,6667	9	16,6667
VS	23	42,5926	32	59,2593
Aprovado	22	40,7407	54	100,0000
Total	54	100,0000		

Figura 4.1: Solução do Exercício 2.3 do Capítulo 2

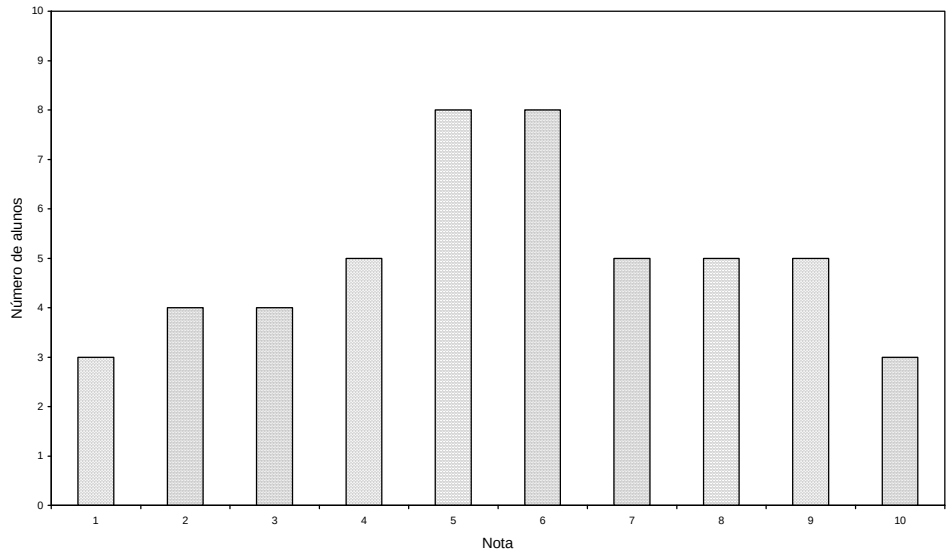


Figura 4.2: Solução do Exercício 2.4 do Capítulo 2

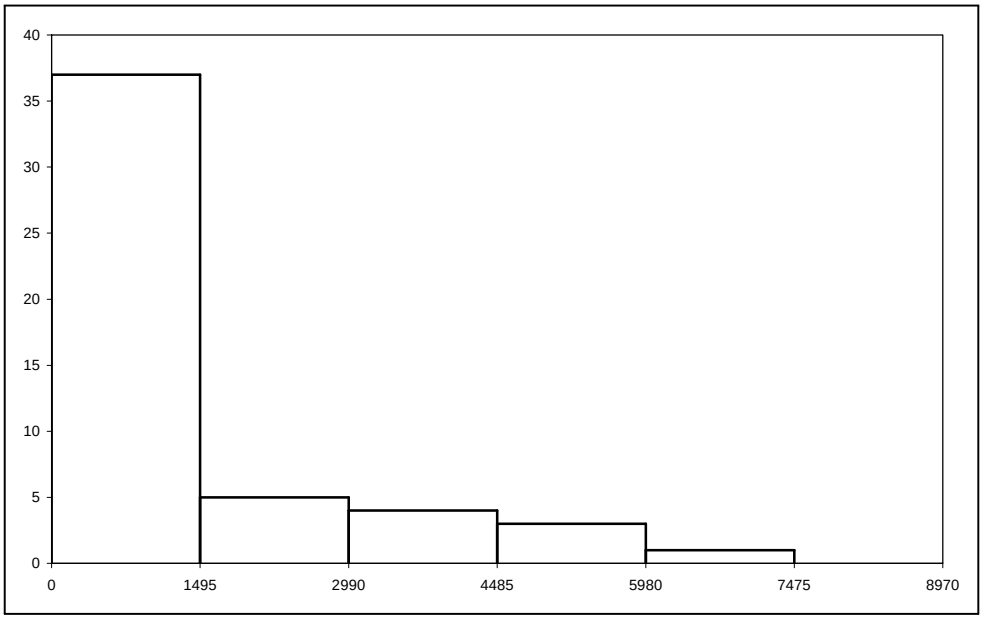


Figura 4.3: Solução do Exercício 2.5 do Capítulo 2

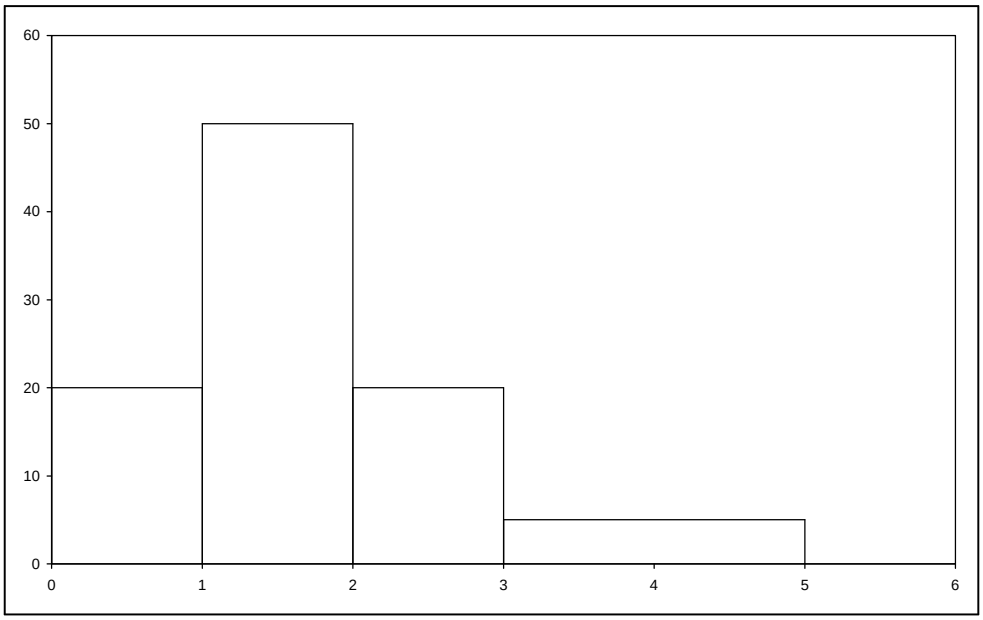


Figura 4.4: Solução do Exercício 2.6 do Capítulo 2

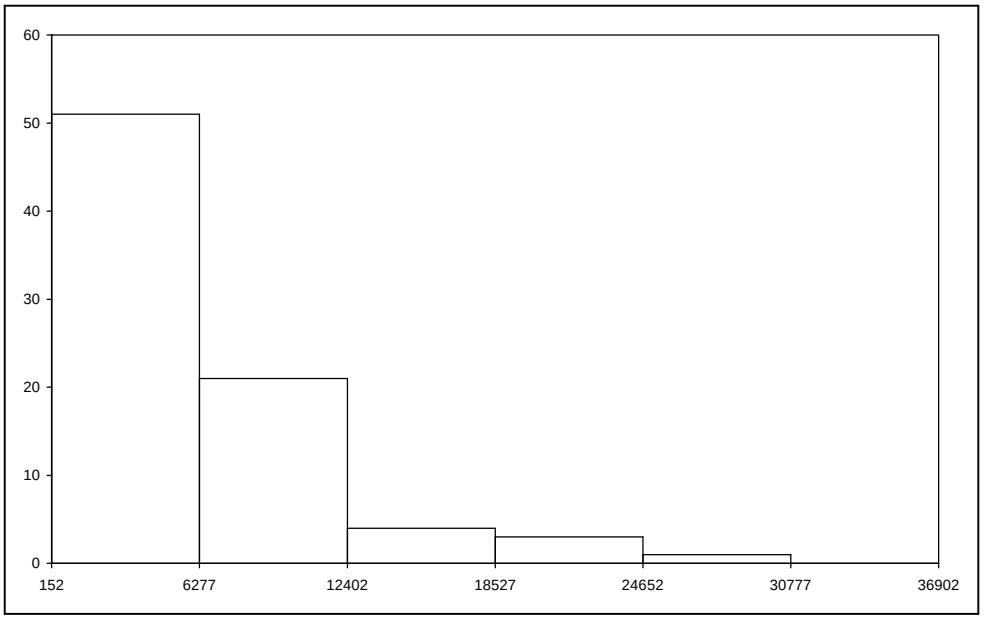
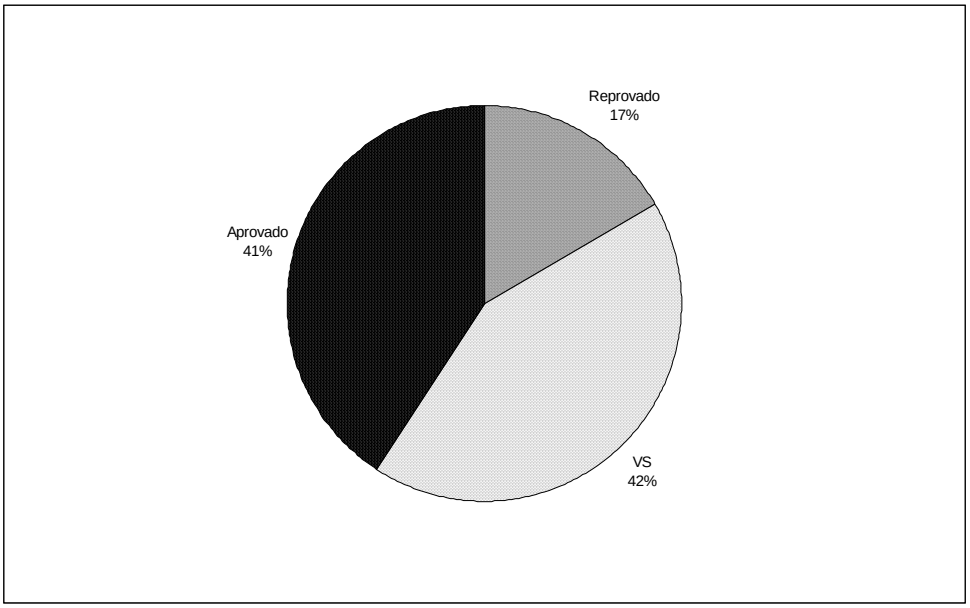


Figura 4.5: Solução do Exercício 2.7 do Capítulo 2



Exercícios Complementares

9. Os valores mínimo e máximo são 75 kg e 95 kg; logo, a amplitude total é de 20 kg. Para grupar os dados em 5 classes, mantendo os limites como números inteiros, temos que mudar a amplitude para 25 e aí o comprimento de cada classe será de 5 kg. Distribuindo a diferença nas duas caudas da distribuição, os limites de classe podem ser:

   $73 \vdash 78$ $78 \vdash 83$ $83 \vdash 88$ $88 \vdash 93$ $93 \vdash 98$ ou

   $72 \vdash 77$ $77 \vdash 82$ $82 \vdash 87$ $87 \vdash 92$ $92 \vdash 97$.

10. Os valores mínimo e máximo são 1500 e 3150 u.m.. Logo, a amplitude total é de 1650 u.m., que é um múltiplo exato de 6. Então, para definir os limites como números inteiros, temos que redefinir a amplitude como 1656 e, nesse caso, o comprimento de cada classe será $\frac{1656}{6} = 276$. Distribuindo igualmente as 6 unidades, os limites de classe são:

    $1497 \vdash 1773$ $1773 \vdash 2049$ $2049 \vdash 2325$

    $2325 \vdash 2601$ $2601 \vdash 2877$ $2877 \vdash 3153$

11. A tabela e os gráficos são apresentados a seguir (Tabela 4.6 e Figuras 4.6 e 4.7).

Tabela 4.6: Solução do Exercício 2.11 do Capítulo 2

	Freqüência Simples		Freqüência Acumulada	
Notas	Absoluta	Relativa	Absoluta	Relativa
$20 \vdash 30$	1	0,02	1	0,02
$30 \vdash 40$	2	0,04	3	0,06
$40 \vdash 50$	2	0,04	5	0,10
$50 \vdash 60$	3	0,06	8	0,16
$60 \vdash 70$	12	0,24	20	0,40
$70 \vdash 80$	14	0,28	34	0,68
$80 \vdash 90$	12	0,24	46	0,92
$90 \vdash 100$	4	0,08	50	1,00
Total	50	1,00		

12. Como a amplitude exata é múltiplo de 5, vamos trabalhar com o próximo múltiplo, que é 116.990. A definição das classes foi feita distribuindo duas unidades extras na cauda inferior e três na cauda superior. (Tabela 4.7 e Figuras 4.8 e 4.9)

Tabela 4.7: Solução do Exercício 2.12 do Capítulo 2

	Freqüência Simples		Freqüência Acumulada	
Notas	Absoluta	Relativa	Absoluta	Relativa
$1813 \vdash 25211$	63	$0,63$	63	$0,63$
$25211 \vdash 48609$	17	$0,17$	80	$0,80$
$48609 \vdash 72007$	9	$0,09$	89	$0,89$
$72007 \vdash 95405$	8	$0,08$	97	$0,97$
$95405 \vdash 118803$	3	$0,03$	100	$1,00$
Total	100	$1,00$		

Figura 4.6: Solução do Exercício 2.11 do Capítulo 2

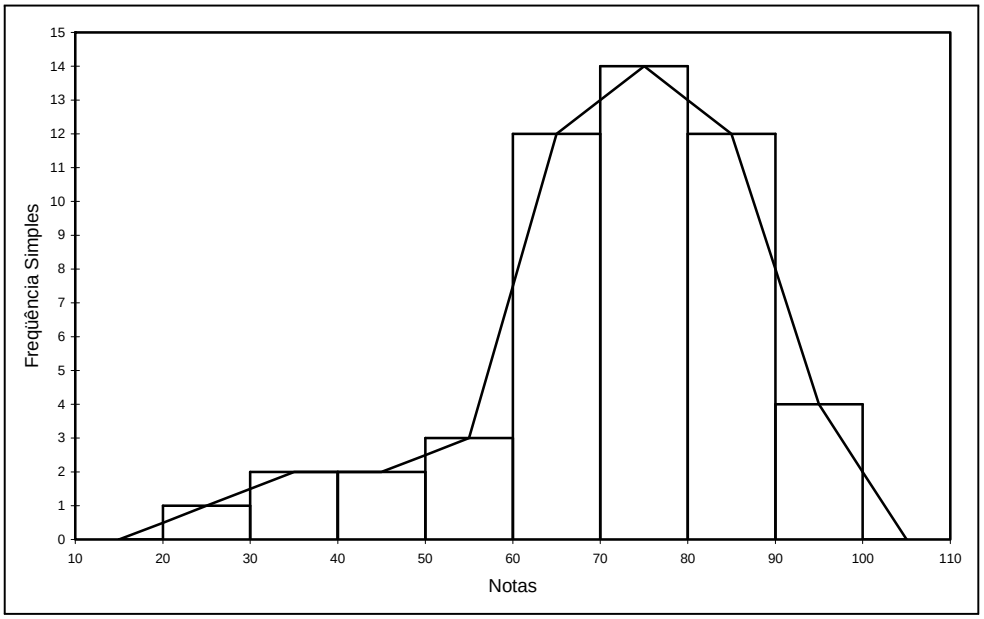


Figura 4.7: Solução do Exercício 2.11 do Capítulo 2

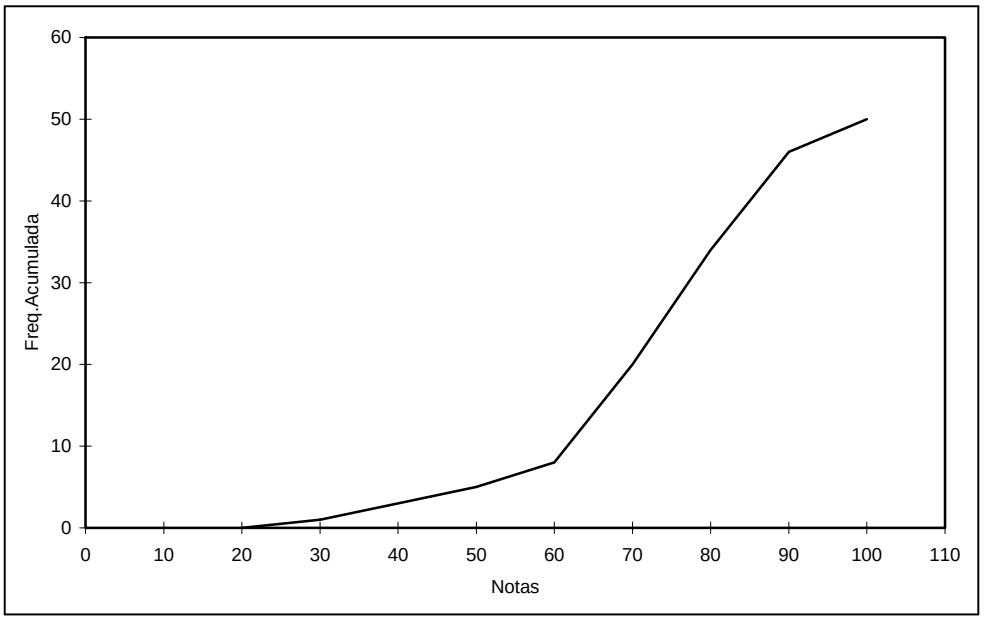


Figura 4.8: Solução do Exercício 2.12 do Capítulo 2

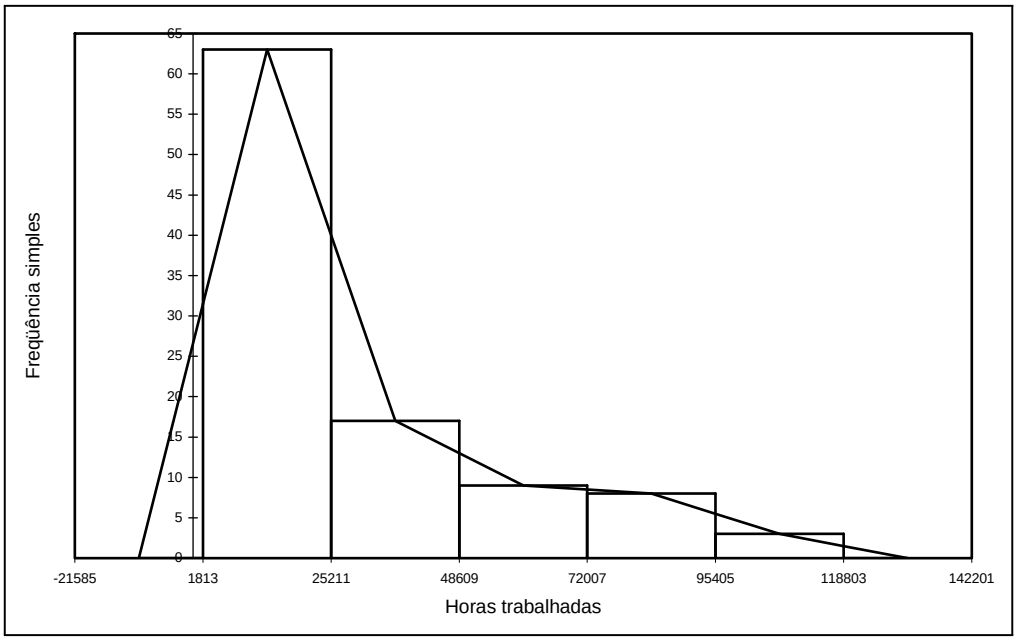


Figura 4.9: Solução do Exercício 2.12 do Capítulo 2

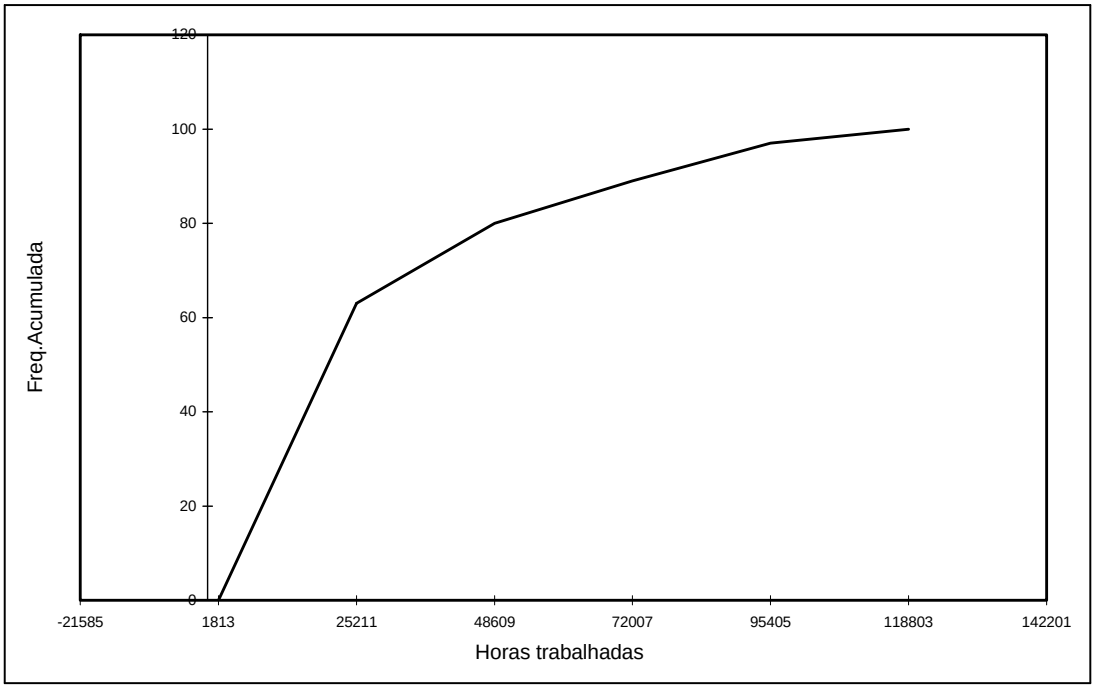


Figura 4.10: Solução do Exercício 2.13 do Capítulo 2

0	2 2 3 4 4 5 5 6 6
1	2 5 7
2	4
3	1 7
4	8
5	1 3 7
6	1 8
7	
8	1 1
9	
10	2
11	
12	
13	
14	9
32	**8**
35	**3**

13. As folhas serão definidas pelo algarismo da unidade e cada ramo representará uma dezena. (Figura 4.10).

14. Como as classes são desiguais, temos que trabalhar com o conceito de densidade. A solução apresentada considera as seguintes classes: [50,60), [60,70), [70,80), [80,90), [90,100), [100,200), [200,300), [300,400), [400,500), [500,600) e exclui Belo Horizonte. Aqui estamos usando a densidade definida em termos da freqüência absoluta (Tabela 4.8 e Figura 4.11).

Tabela 4.8: Solução do Exercício 2.14 do Capítulo 2

	Freqüência Simples		Freqüência Acumulada		Densidade
População (milhares)	Absoluta	Relativa	Absoluta	Relativa	Absoluta
$50 \vdash 60$	7	11,86	7	11,86	0,70
$60 \vdash 70$	12	20,34	19	32,20	1,20
$70 \vdash 80$	11	18,64	30	50,85	1,10
$80 \vdash 90$	3	5,08	33	55,93	0,30
$90 \vdash 100$	4	6,78	37	62,71	0,40
$100 \vdash 200$	13	22,03	50	84,75	0,13
$200 \vdash 300$	4	6,78	54	91,53	0,04
$300 \vdash 400$	2	3,39	56	94,92	0,02
$400 \vdash 500$	1	1,69	27	96,61	0,01
$500 \vdash 600$	2	3,39	59	100,00	0,02
Total	57	100,00			

15. O gráfico apropriado é um gráfico tipo linha, que mostra a evolução dos dados ao longo do tempo. (Figura 4.12).

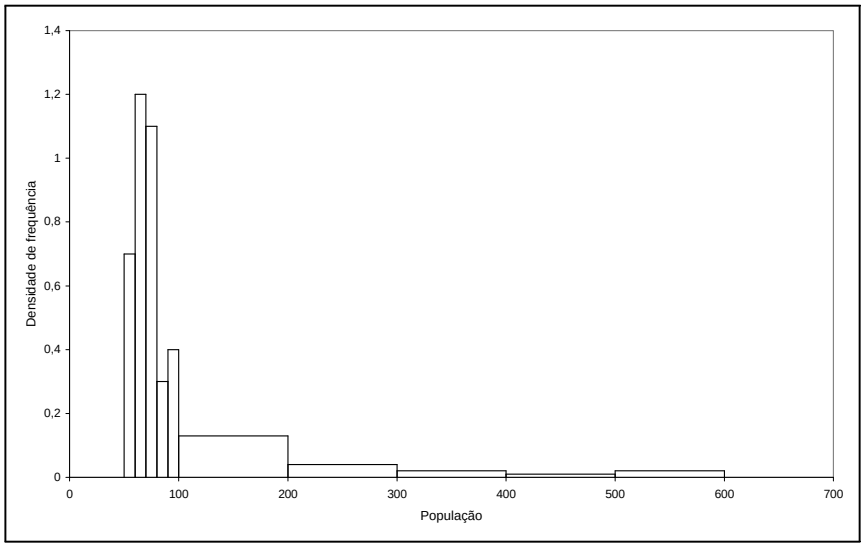
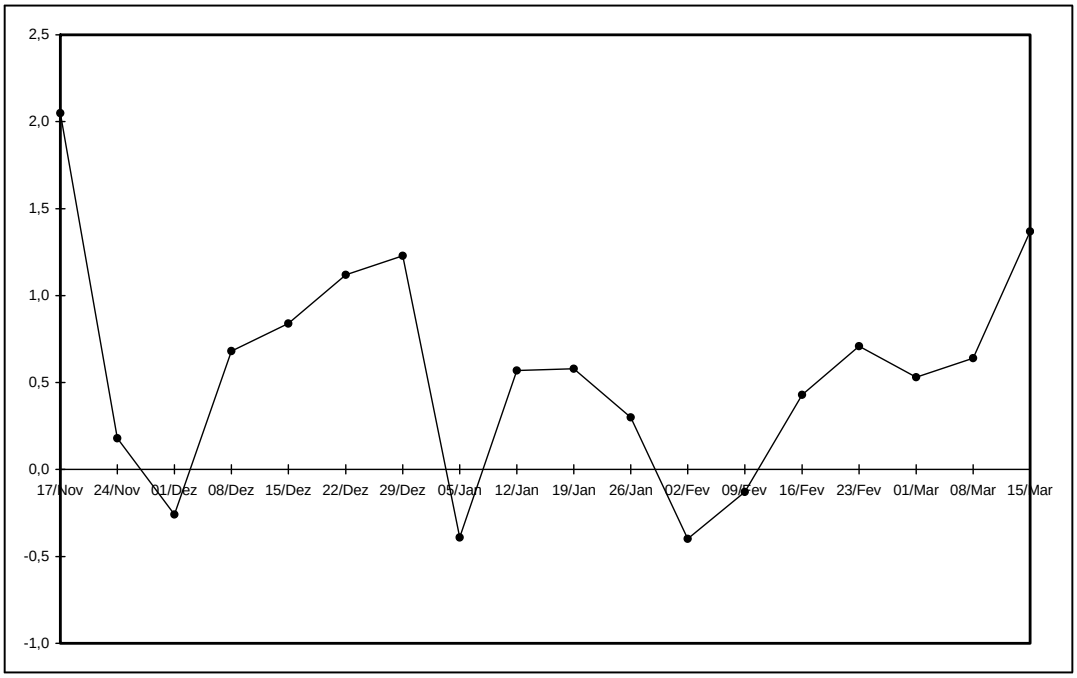


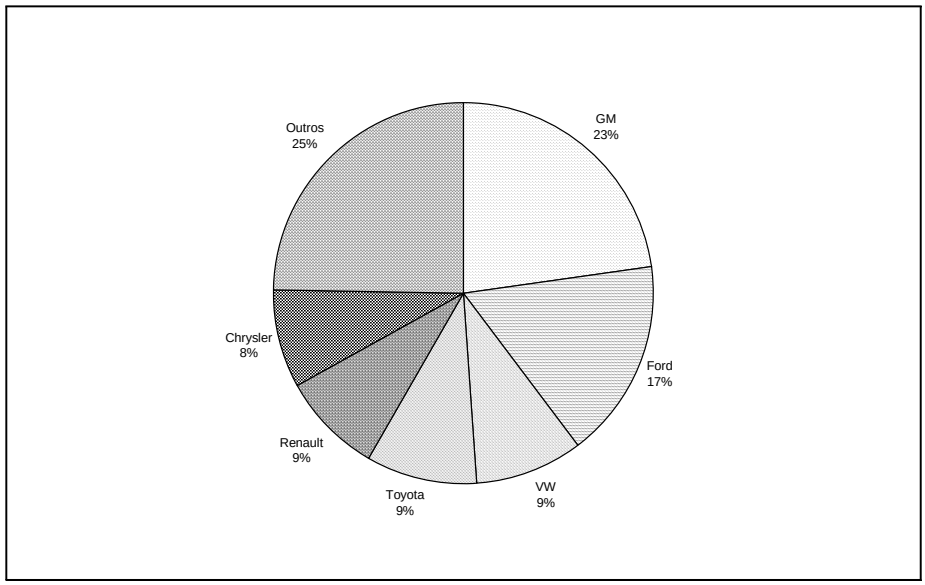
Figura 4.11:

Figura 4.12: Solução do Exercício 2.15 do Capítulo 2



16. O gráfico apropriado é um gráfico tipo setores. Havendo disponibilidade, esse gráfico deve ser construído de modo que as "fatias" sejam diferenciadas por cores. (Figura 4.13)

Figura 4.13: Solução do Exercício 2.16 do Capítulo 2



17. Um gráfico apropriado é o tipo barras, onde os sexos são representados em colunas adjacentes. (Figura 4.14)

18. Novamente, o gráfico apropriado é o tipo linha para mostrar a evolução ao longo do tempo; poderia ser usado também um gráfico tipo barras. (Figura 4.15)

19. Ver Tabela 4.9 e Figuras 4.16 e .

Tabela 4.9: Solução do Exercício 19 do Capítulo 2

Número de	Freqüência simples		Freqüência acumulada	
empregados	Absoluta	Relativa (%)	Absoluta	Relativa (%)
$110 \vdash 154$	17	28,8135	17	28,8135
$154 \vdash 198$	33	55,9322	50	84,7457
$198 \vdash 242$	4	6,7797	54	91,5254
$242 \vdash 286$	1	1,6949	55	93,2203
$286 \vdash 330$	4	6,7797	59	100,0000

20. Ver Figura 4.18. Esse é o tipo de gráfico utilizado pelas companhias de eletricidade (LIGHT, AMPLA, etc) nas contas de luz para ilustrar o consumo dos clientes. Poderia ser feito também um gráfico de linnhas.

21. Ver Tabela 4.10

    O número mediano de sinistros é 0 e o $90^{\underline{o}}$ percentil é 1.

Figura 4.14: Solução do Exercício 2.17 do Capítulo 2

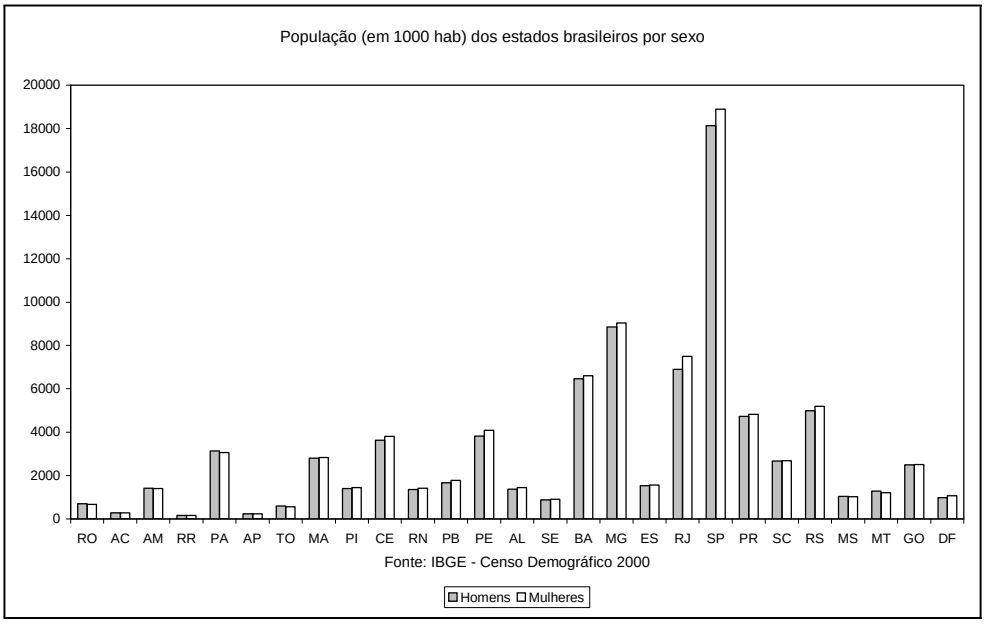


Figura 4.15: Solução do Exercício 2.18 do Capítulo 2

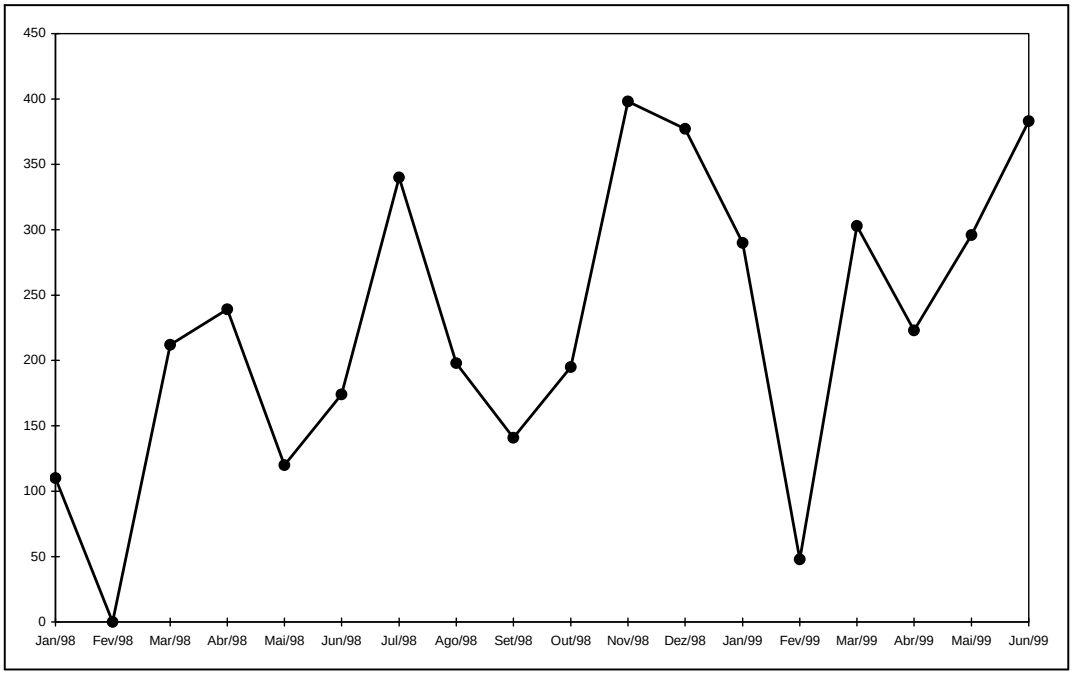


Figura 4.16: Ramo e folhas para área das casas de Boulder, Colorado - Exercício 2.19 do Capítulo 2

12	6 6 6 6
13	
14	
15	3 8 8 8
16	8
17	
18	9 9
19	5 6
20	
21	7 8 8 9
22	2 3 7 8 8 9
23	1 2 4 4
24	3 4 9 9 9
25	1 2
26	2 3 8
27	0 2 6 7 9 9
28	3 5 6 9
29	
30	0 2 2 4 4
31	6 7
32	2 2
33	
34	9
35	3
36	
37	
38	8

Figura 4.17: Diagrama de dispersão para preço e área das casas de Boulder, Colorado Exercício - Exercício 2.19 do Capítulo 2

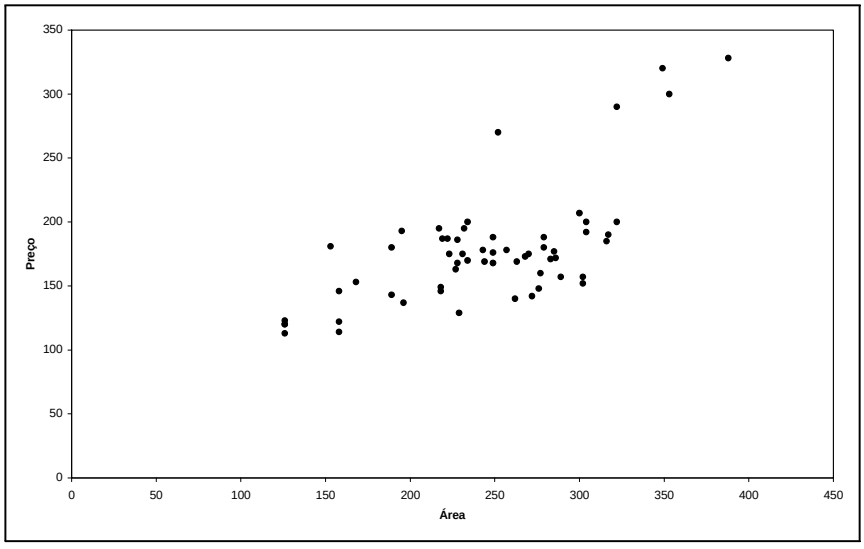


Figura 4.18: Solução do Exercício 2.20 do Capítulo 2

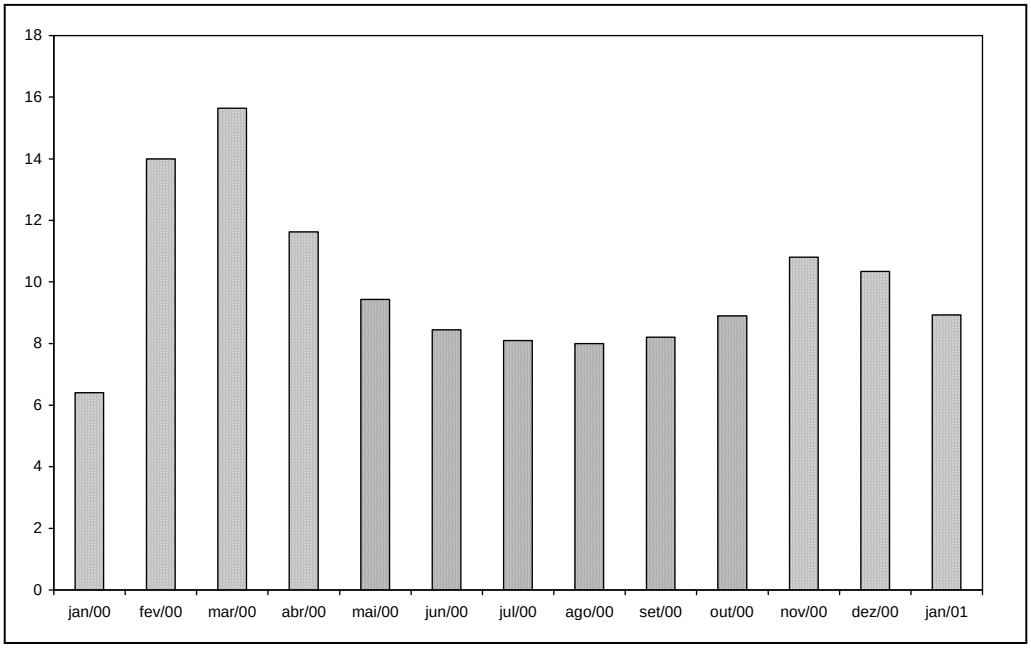


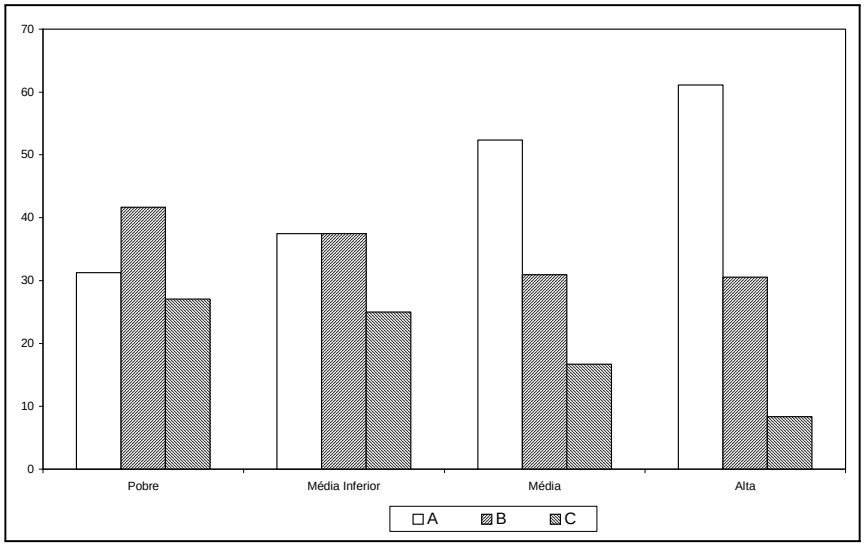
Tabela 4.10: Solução do Exercício 21 do Capítulo 2

Número de sinistros	Freqüência simples		Freqüência acumulada	
	Absoluta	Relativa (%)	Absoluta	Relativa (%)
0	2913	58,26	2913	58,26
1	1587	31,74	4500	90,00
2	326	6,52	4826	96,52
3	102	2,04	4928	98,56
4	72	1,44	5000	100,00

22. A variável independente é classe social e o jornal preferido é a variável dependente. Veja Tabela **??** e Figura 4.19. Podemos ver que nas duas classes superiores há uma maior preferência pelo jornal A, enquanto na classe Pobre, o jornal preferido é o B. Em todas as classes, o jornal C é o menos lido.

	Classe Social				
Jornal	Pobre	Média Inferior	Média	Alta	Total
A	31,25	37,50	52,38	61,11	45,00
B	41,67	37,50	30,95	30,56	35,00
C	27,08	25,00	16,67	8,33	20,00
Total	100,00	100,00	100,00	100,00	100,00

Figura 4.19: Solução do Exercício 22



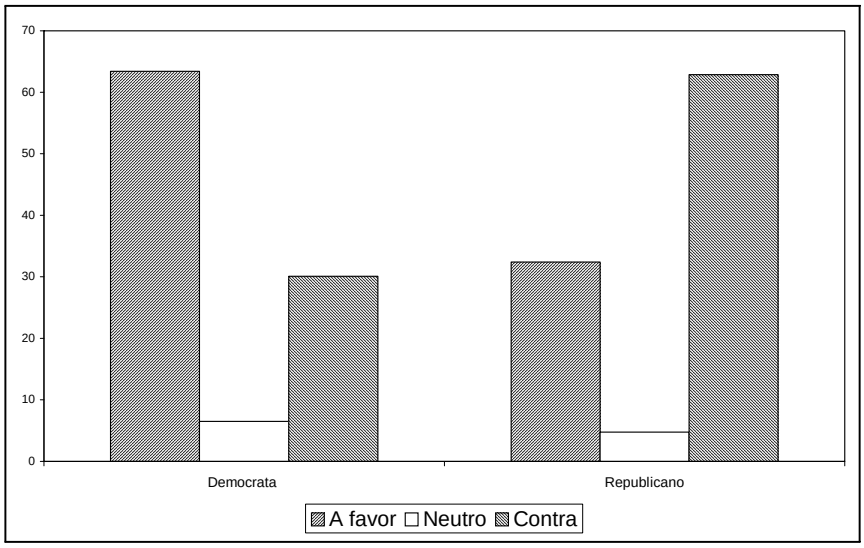
23. Aqui queremos ver se há diferença entre republicanbos e deo=mocratas em relação ao aborto. Podemos pensar no partido como variável independente e na opinião sobre o aborto como a variável dependente.. Veja Tabela **??** e Figura 4.20. Podemos ver que há uma inversão nos dois partidos entre aqueles que são contra ou a favor.

Opinião sobre	Partido		Total
o aborto	Democrata	Republicano	
A favor	63,41	32,38	49,12
Neutro	6,50	4,76	5,70
Contra	30,08	62,86	45,18
Total	100,00	100,00	100,00

4.2 Capítulo 3

Seção 3.2

Figura 4.20: Solução do Exercício 23



1. A propriedade básica a ser usada na solução deste exercício é que a média de um conjunto de dados é sempre maior que o valor mínimo. Se o peso médio é 81 kg, o peso total dos 11 jogadores é $11 \times 81 = 891$. Se um jogador pesa 95 kg, o peso dos 10 outros é de $891 - 95 = 796$, que dá um peso médio de 79,6, ainda maior que o valor mínimo de 72 kg. Se dois pesam 95, o peso dos 9 restantes é de $891 - 190 = 701$, com média de 77,89. Continuando com esse raciocíno, chega-se à seguinte conclusão: se 5 jogadores pesarem 95 kg, o peso médio dos 6 restantes é de 69,33, menor que 72 kg, o que não é possível. Logo, no máximo 4 jogadores podem pesar 95 kg. O peso médio dos 7 restantes é de 73 kg.

2. $$\overline{x} = \frac{2 + 4 + \cdots + 2 + 2 + 1}{20} = \frac{52}{20} = 206 \text{ apólices/dia}$$

   Dados ordenados:

   $$0 \quad 1 \quad 1 \quad 1 \quad 1 \quad 2 \quad 2 \quad 2 \quad 2 \quad 2 \quad 2 \quad 2 \quad 3 \quad 3 \quad 4 \quad 4 \quad 4 \quad 5 \quad 5 \quad 6$$

   $$Q_2 = \frac{x_{(10)} + x_{(11)}}{2} = \frac{2 + 2}{2} = 2 \text{ apólices/dia}$$

   $$x^* = 2 \text{ apólices/dia}$$

3. Como já visto, a média é sensível a valores extremos. Certamente há grandes empresas que fazem parte do índice NASDAQ, que "puxam" a média para cima. A mediana não é alterada pela presença de valores extremos. Sendo assim, ela é bem menor que a média.

4. $$\overline{x} = \frac{517462}{80} = 6468,275 \qquad Q_2 = \frac{x_{(40)} + x_{(41)}}{2} = 4916$$

   Novamente a média é influenciada pelas poucas empresas que têm um grande número de empregados.

5. A inflação acumulada até novembro é:

$$1,007 \times 1,0105 \times \cdots \times 1,0095 = 1,08290$$

Como queremos a inflação anual no máximo de 9%, temos que ter

$$1,08290 \times i_{12} \leq 1,09 \Rightarrow i_{12} \leq \frac{1,09}{1,08290} = 1,006556$$

que equivale a uma taxa máxima de 0,66%.

6. O crescimento global nos três dias foi de $\frac{9200}{2500} = 3,68$; logo, o percentual médio de crescimento foi de $100 \times \left(\sqrt[3]{3,68} - 1\right) = 100 \times (1,543889 - 1) = 54,39\%$. Aqui você tem que usar a média geométrica porque as novas bactérias também se reproduzem; é como se tivéssemos um regime de capitalização composta.

Seção 3.3

7. Na Tabela 4.11 temos os dados necessários para os cálculos.

Tabela 4.11: Solução do exercício 7 do Capítulo 3

Número de apólices x_i	n_i	$n_i x_i$	$n_i x_i^2$	$n_i \lvert x_i - \overline{x} \rvert$
0	1	0	0	2,6
1	4	4	4	6,4
2	7	14	28	4,2
3	2	6	18	0,8
4	3	12	48	4,2
5	2	10	50	4,8
6	1	6	36	3,4
	20	52	184	26,4

A média, como já visto, é 2,6. As medidas de dispersão são:

$$\Delta = 6 - 1 = 5$$

$$\sigma^2 = \frac{184}{20} - \left(\frac{52}{20}\right)^2 = 9,2 - 6,76 = 2,44 \Rightarrow \sigma = 1,56205$$

$$DMA = \frac{26,4}{20} = 1,32$$

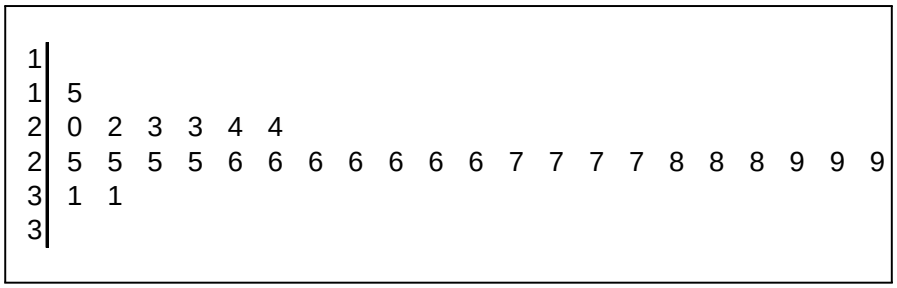
8. (a) Ver Figura 4.21

(b)

$$Q_2 = \frac{x_{(15)} + x_{(16)}}{2} = 26 \qquad Q_1 = x_{(8)} = 25 \qquad Q_3 = x_{(15+8)} = x_{(23)} = 28 \qquad IQ = 28 - 25 = 3$$

(c) $Q_1 - 1,5 \times IQ = 25 - 4,5 = 20,5 \Rightarrow$ Veículos taxados: Mercedez-Bens S420 e Rolls Royce Silver Stone.

Figura 4.21: Solução do Exercício 8 do Capítulo 3



9.

$$\begin{array}{lcllll} \overline{x}_A & = & 55,7222 & \sigma_A = 7,4596 & CV_A = 13,3871 \\ \overline{x}_B & = & 55,4286 & \sigma_B = 3,0949 & CV_B = 5,5835 \end{array}$$

Rendimentos médios semelhantes mas dispersão da corretora A é muito maior. A corretora B parece ter um comportamento mais estável.

10.

$$\begin{array}{lcllll} \overline{x}_{tarde} & = & 4,8 & \sigma_{tarde} = 1,61 & CV_{tarde} = 33,8\% \\ \overline{x}_{noite} & = & 5,6 & \sigma_{noite} = 1,84 & CV_{noite} = 32,89\% \end{array}$$

Média da turma da noite é maior, com menor variabilidade relativa.

Seção 3.7

11. Na Tabela 4.12 temos os dados necessários para os cálculos.

Tabela 4.12: Solução do Exercício 11 do Capítulo 3

Classe	x_i	f_i	$f_i \times x_i$	$f_i \times x_i^2$
$18 \vdash 20$	19	0,36	6,84	129,96
$20 \vdash 22$	21	0,24	5,04	105,84
$22 \vdash 26$	24	0,20	4,80	115,20
$26 \vdash 30$	28	0,16	4,48	125,44
$30 \vdash 36$	33	0,04	1,32	43,56
Soma		1,00	22,48	520,00

(a) A média é 22,48, maior que a média obtida anteriormente de 22 anos. Assim, parece ter havido alguma melhora.

(b) O desvio padrão dos dados é (note que na tabela já estamos multiplicando pelas freqüências relativas!):

$$\sqrt{520 - 22,48^2} = 3,8275$$

A regra do pesquisador se baseia em $\alpha = \dfrac{2 \times 3,8275}{\sqrt{50}} = 1,0826$. Como $\overline{x} - 22 = 0,48 < 1,0826$, ele decide que a diferença não foi grande o suficiente.

Tabela 4.13: Solução do Exercício 12 do Capítulo 3

Classe	x_i	f_i	$f_i \times x_i$	$f_i \times x_i^2$
$0 \vdash 1$	0,5	0,20	0,10	0,050
$1 \vdash 2$	1,5	0,50	0,75	1,125
$2 \vdash 3$	2,5	0,20	0,50	1,250
$3 \vdash 5$	4,0	0,10	0,40	1,600
Soma		1,00	1,75	4,025

12. Ver Tabela 4.13

    A média é $\overline{x} = 1,75$ litros e o desvio padrão é $\sigma = \sqrt{4,025 - 1,75^2} = 0,9811$ litros.

13. Ver Tabela 4.14

Tabela 4.14: Solução do Exercício 13 do Capítulo 3

Classe	x_i	f_i	$f_i \times x_i$	$f_i \times x_i^2$
$152 \vdash 6277$	3214,5	0,6375	2049,24375	6587294,03438
$6277 \vdash 12402$	9339,5	0,2625	2451,61875	22896893,31563
$12402 \vdash 18527$	15464,5	0,0500	773,22500	11957538,01250
$18527 \vdash 24652$	21589,5	0,0375	809,60625	17478994,13438
$24652 \vdash 30777$	27714,5	0,0125	346,43125	9601168,87813
Soma		1,0000	6430,125	68521888,37500

    A média é 6430,125 empregados e o desvio padrão é $\sigma = \sqrt{68521888,37500 - 6430,125^2} = 5213,001$ empregados.

14. Classe modal $1 \vdash 2$. As classes vizinhas têm a mesma freqüência; logo, ambos os métodos darão a moda igual ao ponto médio. De fato, pelo método de King temos::

$$\frac{x^* - 1}{2 - x^*} = \frac{0,2}{0,2} \Rightarrow x^* = 1,5$$

    e pelo método de Czuber:

$$\frac{x^* - 1}{2 - x^*} = \frac{0,3}{0,3} \Rightarrow x^* = 1,5$$

15. A classe modal é a primeira classe $152 \vdash 6277$. O método de King, então, resulta no extremo superior (não tem "ninguém puxando" pelo lado inferior). De fato:

$$\frac{6277 - x^*}{x^* - 152} = \frac{0}{21} \Rightarrow x^* = 6277$$

    Pelo método de Czuber temos:

$$\frac{6277 - x^*}{x^* - 152} = \frac{51 - 21}{51 - 0} \Rightarrow 30x^* - 4560 = 320127 - 51x^* \Rightarrow x^* = 4008,48$$

16. (a) $\overline{x} = 1020,8$
    (b) $\sigma^2 = 691,36 \qquad \sigma = 26,2937$
    (c) Histograma usual, com classes de igual comprimento.

(d) O limite superior da classe D é o 20º percentil; o da classe C é o 50º percentil, o da classe B é o 80º percentil e, obviamente, o da classe A é o valor máximo, 1080.

O 20º percentil está na classe $980 \vdash 1000$, onde acumula 22% da distribuição e a regra de proporcionalidade que o define é:

$$\frac{1000 - P_{20}}{0,22 - 0,20} = \frac{1000 - 980}{0,16} \Rightarrow P_{20} = 997,2$$

O 50º percentil (mediana) é o limite superior da terceira classe (note que nessa classe temos 50% da distribuição acumulada). O 80º percentil está na classe $1040 \vdash 1060$, onde acumula 92% da distribuição:

$$\frac{1060 - P_{80}}{0,92 - 0,80} = \frac{1000 - 980}{0,16} \Rightarrow P_{80} = 1045$$

As classes de peso são, pois: [960, 997,5); [997,5; 1020); [1020; 1045); >=1045.

(e) Ração reforçada: $\overline{x} - 2\sigma = 1020,8 - 2 \times 26,2937 = 968,2125$. Podemos estimar a percentagem de frangos por uma regra de três análoga à utilizada para determinar qualqeur separatriz. A diferençca é que agora temos a separatriz e queremos a freqüência.

$$\frac{968,2125 - 960}{x} = \frac{980 - 960}{0,06} \Rightarrow x = 0,0246 \text{ ou } 2,46\%$$

Rerpodutores: $\overline{x} + 1,5 \times \sigma = 1020,8 + 1,5 \times 26,2937 = 1060,2406$.

$$\frac{1080 - 1060,2406}{x} = \frac{1080 - 1060}{0,08} \Rightarrow x = 0,079 \text{ ou } 7,90\%$$

17. A mediana está na classe $1 \vdash 2$ onde temos 50% da distribuição e 70% da distribuição acumulada. Logo,

$$\frac{2 - Q_2}{0,7 - 0,5} = \frac{2 - 1}{0,5} \Rightarrow Q_2 = 1,6$$

O terceiro decil também está na classe $1 \vdash 2$. Logo,

$$\frac{2 - D_3}{0,7 - 0,3} = \frac{2 - 1}{0,5} \Rightarrow D_3 = 1,2$$

18. A mediana e o primeiro quartil estão ambos na primeira classe, onde temos 63,75% da distribuição.

$$\frac{6277 - Q_1}{0,6375 - 0,25} = \frac{6277 - 152}{0,6375} \Rightarrow Q_1 = 2553,9608$$

$$\frac{6277 - Q_2}{0,6375 - 0,5} = \frac{6277 - 152}{0,6375} \Rightarrow Q_2 = 4955,9216$$

O terceiro quartil está na segunda classe:

$$\frac{12402 - Q_1}{0,90 - 0,75} = \frac{12402 - 6277}{0,2625} \Rightarrow Q_3 = 8902,001$$

e o intervalo interquartil é $IQ = 8902,001 - 2553,9608 = 6348,0402$

19. $\overline{x} = \dfrac{0 + 0 + \cdots + 8,4 + 8,5}{54} = 2,411$

    $Q_2 = \dfrac{x_{(27)} + x_{(28)}}{2} = \dfrac{1,0 + 1,2}{2} = 1,1$

    $x^* = 0,0$

    $Q_1 = x_{(14)} = 0,0$

20. Seja x a nota do aluno na segunda prova. Então, temos, para a média ponderada:

$$\frac{2 \times 5,5 + 3 \times x}{5} \geq 6 \Rightarrow x \geq 6,33$$

    Se as provas tiverem peso igual, temos:

$$\frac{5,5 + x}{2} \geq 6 \Rightarrow x \geq 6,5$$

21. Na Tabela 4.15 temos a versão completa, para facilitar a solução do exercício.

Tabela 4.15: Solução do Exercício 21 do Capítulo 3

	Ponto	Freqüência simples		Freqüência acumulada	
	médio	Absoluta	Relativa	Absoluta	Relativa
$0 \vdash 2$	1,0	55	0,055	55	0,055
$2 \vdash 3$	2,5	65	0,065	120	0,120
$3 \vdash 4$	3,5	172	0,172	292	0,292
$4 \vdash 5$	4,5	254	0,254	546	0,546
$5 \vdash 6$	5,5	278	0,278	824	0,824
$6 \vdash 7$	6,5	76	0,076	900	0,900
$7 \vdash 8$	7,5	75	0,075	975	1,000
$8 \vdash 10$	9,0	25	0,025	1000	

    (a) $\overline{x} = 0,055 \times 1 + 0,065 \times 2,5 + \cdots + 0,025 \times 9 = \mathbf{4,773}$
    $\sigma^2 = 0,055 \times 1^2 + 0,065 \times 2,5^2 + \cdots + 0,025 \times 9^2 - 4,773^2 = \mathbf{2,794471} \Rightarrow \boldsymbol{\sigma} = \mathbf{1,67166713}$

    (b) $d_m = 0,055 \times |1 - 4,773| + 0,065 \times |2,5 - 4,773| + \cdots + 0,025 \times |9 - 4,773| = \mathbf{1,287116}$

    (c) $\overline{x} + 1,5\sigma = \mathbf{7,2805007}$. Logo, os alunos com nota maior que 7,28 terão bolsa de Iniciação Científica. Usando uma regra de três podemos estimar o número de alunos com nota entre 7,28 e 8 notando que a classe $7 \vdash 8$, de comprimento 1, tem 75 alunos. Logo, a classe $7 \vdash 7,28$ terá x alunos onde

$$\frac{75}{1} = \frac{x}{0,28} \Rightarrow x = 0,28 \times 75 = 21$$

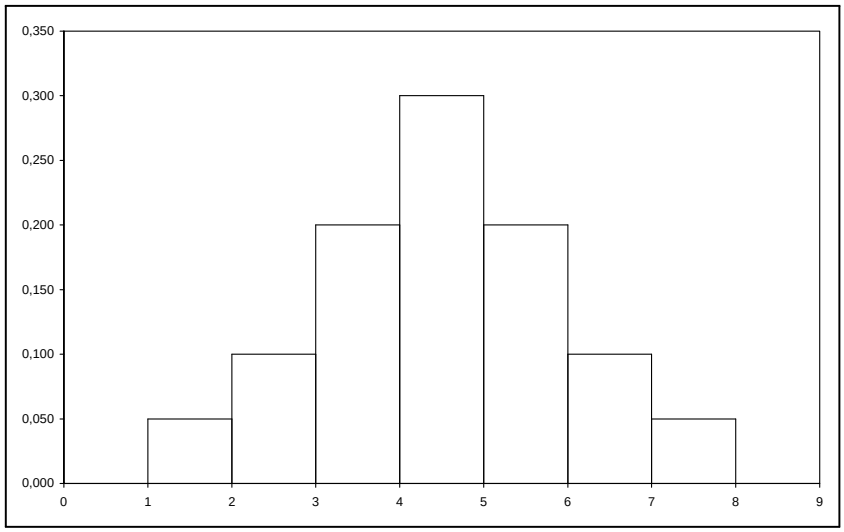
    e, assim, o número de alunos que terão bolsa de Iniciação Científica é de $21 + 25 = 46$.

    (d) Temos que calcular o terceiro decil, que está na classe $4 \vdash 5$. Usando a proporcionalidade das áreas dos retângulos envolvidos, tem-se que:

$$\frac{5 - D_3}{5 - 4} = \frac{0,546 - 0,3}{0,254} \Rightarrow D_3 = 4,0315$$

    Logo, para não ter que fazer o curso de Cálculo Zero, o aluno tem que tirar nota maior que 4,03.

Figura 4.22: Solução do Exercício 22 do Capítulo 3



22. As distribuições simétricas têm média e mediana iguais. (ver Figura 4.22)

23. (a)

$$\overline{x} = \frac{38639}{27} = 1431,074 \qquad \sigma = \sqrt{\frac{135079221}{27} - \left(\frac{38639}{27}\right)^2} = \sqrt{2954961,106} = 1719,000$$

(b) Os dados estão ordenados decrescentemente. A mediana, como observação central, pode ser calculada contando de baixo para cima (do menor para o maior) ou do maior para o menor; ela é a $14^{\underline{a}}$ observação em qualquer direção. $Q_2 = 635$.
Tirando a mediana sobram 13 observações em cada metade. Logo, os outros quartis são as observações $x_{(7)}$ e $x_{(14+7)}$. O terceiro quartil pode, então, ser calculado como a sétima observação, indo do maior para o menor, e o primeiro quartil é a sétima observação indo do menor para o maior.

$$Q_1 = 158 \qquad Q_3 = 2300 \qquad IQ = 2142$$

(c) $Q_1 - 1,5 \times IQ = -3055 \qquad Q_3 + 1,5 \times IQ = 5513$
Não há outliers inferiores mas os dois maiores salários são outliers superiores.

(d) Dada a presença de outliers, a mediana seria mais adequada para representar o salário típico do time.

24. (a) Ver Tabela 4.16.

(b) $\overline{x} = 830,48$ mg $\qquad \sigma = \sqrt{690255,2 - 830,48^2} = 23,6256$ mg.

(c) Classe modal: $820 \vdash 840$
King:

$$\frac{x^* - 820}{840 - x^*} = \frac{117}{118} \Rightarrow 118x^* - 96760 = 98280 - 117x^* \Rightarrow x^* = 829,96 \text{ mg}$$

Tabela 4.16: Solução do Exercício 24 do Capítulo 3

Classes de	Médio	Freq.Simples		Freq.Acumulada		Cálc. de $\overline{x}$	Cálc. de σ^2
peso (mg)	x_i	Abs. n_i	Rel. f_i	Abs. N_i	Rel. F_i	$f_i \times x_i$	$f_i \times x_i^2$
$760 \vdash 780$	770	4	0,008	4	0,008	6,16	4743,2
$780 \vdash 800$	790	43	0,086	47	0,94	67,94	53672,6
$800 \vdash 820$	810	118	0,236	165	0,330	191,16	154839,6
$820 \vdash 840$	830	168	0,336	333	0,666	278,88	231470,4
$840 \vdash 860$	850	117	0,234	450	0,900	198,90	169065,0
$860 \vdash 880$	870	39	0,078	489	0,978	67,86	59038,2
$880 \vdash 900$	890	11	0,022	500	1,000	19,58	17426,2
Soma		500	1,000			830,48	690255,2

Czuber:

$$\frac{x^* - 820}{840 - x^*} = \frac{168 - 118}{168 - 117} \Rightarrow x^* = 829,90 \text{ mg}$$

(d) Classe mediana: $820 \vdash 840$; aí temos 0,336 de freqüência e 0,666 da freqüência acumulada.

$$\frac{840 - Q_2}{0,666 - 0,5} = \frac{840 - 820}{0,336} \Rightarrow Q_2 = 830,119 \text{ mg}$$

(e) $Q_1 : 800 \vdash 820 \qquad f = 0,236 \qquad F = 0,33$

$$\frac{820 - Q_1}{0,33 - 0,25} = \frac{820 - 800}{0,236} \Rightarrow Q_1 = 813,220 \text{ mg}$$

$Q_3 : 840 \vdash 860 \qquad f = 0,234 \qquad F = 0,90$

$$\frac{860 - Q_3}{0,90 - 0,75} = \frac{860 - 840}{0,234} \Rightarrow Q_3 = 847,179\text{mg}$$

Outliers inferiores: $Q_1 - 1,5 \times IQ = 762,2815$

$$\frac{762,2815 - 760}{x} = \frac{780 - 760}{0,008} \Rightarrow x = 0,0009126 \text{ ou } 0,09\%$$

Outliers superiores: $Q_3 + 1,5 \times IQ = 898,1175$

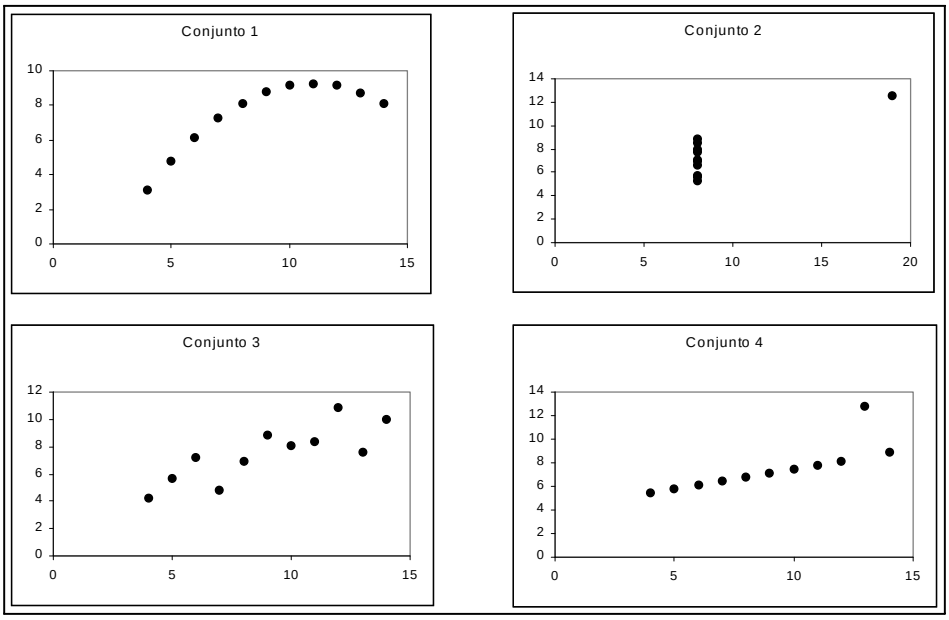
$$\frac{900 - 898,1175}{x} = \frac{900 - 880}{0,022} \Rightarrow x = 0,0021 \text{ ou } 0,21\%$$

25. Para os 4 conjuntos, temos que as médias de X e Y são as mesmas, assim como o coeficiente de correlação.

$$\begin{aligned} \overline{X} &= 9 & \overline{Y} &= 7,50091 \\ \sigma_X &= 3,16228 & \sigma_Y &= 1,93711 \\ \rho(X,Y) &= 0,816 \end{aligned}$$

No entanto, os conjuntos são completamente diferentes, conforme ilustrado pelos diagramas de dispersão da Figura 4.23. Então, uma análise de dados não deve se basear em apenas uma medida descritiva; é importante que diferentes aspectos sejam analisados, inclusive através de representações gráficas adequadas.

Figura 4.23: Dados de Anscombe - Solução do Exercício 25 do Capítulo **??**



26. A idéia é usar como *proxy* a variável mais fortemente associada com a variável de interesse, que é capacidade da produção instalada. Vamos, então, calcular os coeficientes de correlação entre essa variável e as duas "candidatas". Usando os valores dados, temos que:

$$\rho(X,Y) = \frac{361 - \frac{80\times 38}{10}}{\sqrt{736 - \frac{80^2}{10}}\sqrt{182 - \frac{38^2}{10}}} = 0,9487$$

$$\rho(X,Z) = \frac{848 - \frac{80\times 100}{10}}{\sqrt{736 - \frac{80^2}{10}}\sqrt{1048 - \frac{100^2}{10}}} = 0,7071$$

Logo, a variável a ser utilizada como *proxy* deverá ser Potência Instalada, que apresenta maior correlação com a variável de interesse.

Bibliografia

[1] Anscombe, F.J. (1974), Graphs in statistical analysis, *The American Statistician,* 27(1973), pp. 17-21.

[2] Barbetta. P.A. (1994) *Estatística Aplicada às Ciências Sociais*, Florianópolis: Editora da UFSC.

[3] Bussab, W.O. e Morettin, P.A. (1987) *Estatística Básica*, São Paulo: Editora Atual .

[4] Dunn, O.J. e Clark, V.A. (1974) *Applied Statistics: Analysis of Variance and Regression*, Nova York: John Wiley & Sons.

[5] Legrain, M. e Magain, D. (1992) *Estudo de Mercado*, São Paulo: Makron Books.

[6] Lopes, P..A.(1999) *Probabilidades e Estatística, Rio de Janeiro: Reichmann & Affonso Editores.*

[7] Moore, D.S. e McCabe, G.P. (1998) *Introduction to the Practice of Statistics*, 3ª ed., Nova York: W.H. Freeman and Company.

[8] Murteira, B.J.F. e Black, G.H.J. (1983) *Estatística Descritiva*, Lisboa: McGraw-Hill de Portugal.

[9] Soares, J.F., Farias, A.A. e Cesar, C.C. (1991) *Introdução à Estatística*, Rio de Janeiro: Guanabara Koogan.

[10] Tukey, J.W. (1977) *Exploratory Data Analysis (EDA)*, Addison-Wesley.

[11] Velleman, P.F. e Hoaglin, D.C. (1981) *Applications, Basics and Computing of Exploratory Data Analysis (ABC of EDA)*, Massachusetts: Duxbury Press.