Análise de dados e simulação

Versão parcial preliminar

agosto 2016

Pedro A. Morettin

Julio M. Singer

Departamento de Estatística
Universidade de São Paulo
Caixa Postal 66281
São Paulo, SP 05314-970
Brasil

Conteúdo

Capítulo 1

Introdução

1.1 Preliminares

Em praticamente todas as áreas do conhecimento, dados são coletados com o objetivo de obtenção de informação. Esses dados podem representar uma população (como o censo demográfico) ou uma parte (amostra) dessa população (como aqueles oriundos de uma pesquisa eleitoral). Eles podem ser obtidos por meio de estudos observacionais (como aqueles em que se examinam os registros médicos de um determinado hospital), de estudos amostrais (como pesquisas de opinião) ou experimentais (como ensaios clínicos). Nesse contexto, a Estatística é uma ferramenta importante para organizá-los, resumi-los, analisá-los e utilizá-los para tomada de decisões.

A abordagem estatística para o tratamento de dados envolve

i) o planejamento da forma de coleta em função dos objetivos do estudo;
ii) a organização de uma planilha para seu armazenamento eletrônico;
iii) o seu resumo por meio de tabelas e gráficos;
iv) a identificação e correção de possíveis erros de coleta e/ou digitação;
v) a proposta de modelos baseados nos objetivos do estudo para representar relações entre as características (variáveis) observadas;
vi) a avaliação do ajuste do modelo aos dados por meio de técnicas de diagnóstico e/ou simulação;
vii) a reformulação e reajuste do modelo à luz dos resultados do diagnóstico e/ou simulação;
viii) a tradução dos resultados do ajuste em termos não-técnicos.

O item i), por exemplo, pode corresponder a uma hipótese formulada por um cientista. Numa tentativa de comprovar a sua hipótese, ele identifica as variáveis de interesse e planeja um experimento (preferencialmente com o apoio de um estatístico) para a coleta dos dados que serão armazenados numa planilha. O objetivo desse livro é abordar detalhadamente os itens ii), iii), iv) e viii) com referências eventuais aos itens v) e vi).

Mais comumente, os dados envolvem valores de várias variáveis obtidos da observação de unidades de investigação que constituem uma amostra de uma população. A análise de dados amostrais possibilita que se faça

inferência sobre a distribuição de probabilidades das variáveis de interesse, definidas sobre a população da qual a amostra foi colhida.

Exemplo 1.1 Se quisermos saber se há relação entre o consumo (variável C) e renda (variável Y) de indivíduos de uma população, podemos escolher uma amostra de n indivíduos dessa população e medir essas duas variáveis nesses indivíduos, obtendo-se o conjunto $\{(Y_1, C_1), \ldots, (Y_n, C_n)\}$, que conterá os nossos dados.

No Capítulo 2 definiremos formalmente o que se chama uma amostra aleatória simples retirada de uma população. Para saber se existe alguma relação entre C e Y podemos fazer um gráfico de dispersão, colocando-se no eixo das abcissas a variável Y e no eixo das ordenadas a variável C. Obteremos uma nuvem de pontos no plano (Y, C), que pode nos dar uma ideia de um **modelo** relacionando Y e C. No Capítulo 3 trataremos da análise de duas variáveis e, no Capítulo 5, estudaremos os chamados modelos de regressão, que são apropriados para o exemplo em questão.

Exemplo 1.1 (continuação) Retomemos o Exemplo 1.1 para ilustrar o método. Quanto ao estágio (i), em Economia, sabe-se, desde Keynes, que o gasto com o consumo de pessoas é uma função da renda pessoal disponível. Denotando essas variáveis por C e Y, respetivamente, poderemos escrever

$$C = f(Y),$$

para alguma função f.

No estágio (ii), para se ter uma ideia de como é a função f para essa comunidade (estágio (iii)), podemos construir um gráfico de dispersão para Y e C, obtendo algo como a Figura 1.1 (nesse exemplo, n=20).

Suponha que seja razoável postular o modelo

$$C_i = \alpha + \beta Y_i + e_i, \quad i = 1, 2, \ldots, n, \tag{1.1}$$

em que $(Y_i, C_i),\ i = 1, \ldots, n$ são variáveis efetivamente observadas e $e_i,\ i = 1, \ldots, n$ são variáveis não observadas e chamadas erros. O parâmetro α é denominado consumo autônomo e β representa a propensão marginal a consumir. A reta representada no gráfico foi obtida usando-se métodos discutidos no Capítulo 5. Nesse caso, obtemos $\alpha = 1,48$ e $\beta = 0,71$, aproximadamente. Para diferentes comunidades (populações) poderemos ter curvas (modelos) diferentes para relacionar Y e C.

Exemplo 1.2 Os dados da Tabela 1.2 foram extraídos de um estudo realizado no Instituto de Ciências Biomédicas da Universidade de São Paulo com o objetivo de avaliar a associação entre a infecção de gestantes por malária e a ocorrência de microcefalia nos respectivos bebês. O dicionário das variáveis observadas está indicado na Tabela 1.1.

A disposição dos dados do Exemplo 1.2 no formato de uma planilha está representada na Tabela 1.2.

Figura 1.1: Relação entre renda e consumo de 20 indivíduos

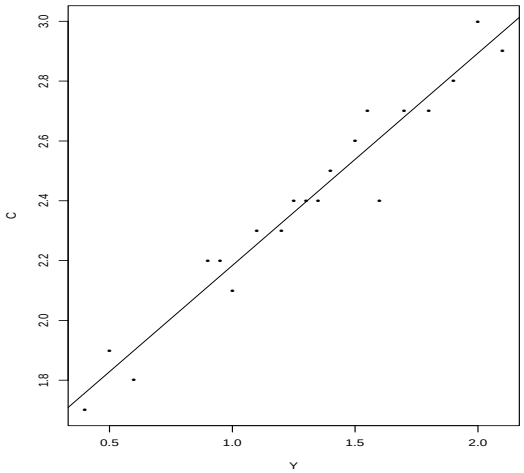


Tabela 1.1: Dicionário para as variáveis referentes ao Exemplo 1.2

Rótulos	**Variável**	**Unidade de medida**
idade	Idade da mãe	anos
nummal	Quantidade de malárias durante a gestação	número inteiro
parasita	Espécie do parasita da malária	0: não infectada
		1: P. vivax
		2: P. falciparum
		3: malária mista
		4: indeterminado
numgest	Paridade (quantidade de gestações)	Número inteiro
idgest	Idade gestacional no parto	semanas
sexorn	Sexo do recém-nascido	1: masculino
		2: feminino
pesorn	Peso do recém-nascido	g
estrn	Estatura do recém-nascido	cm
pcefal	Perímetro cefálico do recém-nascido	cm
Obs:	Observações omissas são representadas por um ponto	

Tabela 1.2: Planilha com dados referentes ao Exemplo 1.2

ident	idade	nummal	parasita	numgest	idgest	sexorn	pesorn	estrn	pcefal
1	25	0	0	3	38	2	3665	46	36
2	30	0	0	9	37	1	2880	44	33
3	40	0	0	1	41	1	2960	52	35
4	26	0	0	2	40	1	2740	47	34
5	.	0	0	1	38	1	2975	50	33
6	18	0	0	.	38	2	2770	48	33
7	20	0	0	1	41	1	2755	48	34
8	15	0	0	1	39	1	2860	49	32
9	.	0	0	.	42	2	3000	50	35
10	18	0	0	1	40	1	3515	51	34
11	17	0	0	2	40	1	3645	54	35
12	18	1	1	3	40	2	2665	48	35
13	30	0	0	6	40	2	2995	49	33
14	19	0	0	1	40	1	2972	46	34
15	32	0	0	5	41	2	3045	50	35
16	32	0	0	8	38	2	3150	44	35
17	18	0	0	2	40	1	2650	48	33.5
18	18	0	0	1	41	1	3200	50	37
19	19	0	0	1	39	1	3140	48	32
20	18	0	0	2	40	1	3150	47	35
21	27	0	0	3	40	1	4185	52	35.5
22	26	0	0	3	40	2	4070	52	35
23	.	0	0	.	40	1	3950	50	37
24	19	0	0	1	40	1	3245	51	33
25	23	0	0	.	41	1	3010	49	35
26	.	0	0	.	40	2	3260	50	33
27	20	1	1	2	40	2	3450	49	33
28	19	0	0	3	40	2	2765	48	32
29	22	0	0	4	40	1	4190	50	34
30	32	0	0	4	42	2	4035	51	34
31	33	0	0	5	39	2	3620	51	33
32	30	3	3	5	38	1	3230	48	34
33	36	0	0	7	39	2	3185	50	38
34	.	0	0	.	39	2	2950	47	33

Neste livro estaremos interessados na análise de conjuntos de dados, que poderão ser provenientes de populações, amostras ou de estudos observacionais. Para essa análise usaremos tabelas, gráficos e diversas medidas de posição (localização), variabilidade e associação, com o intuito de resumir e interpretar os dados.

A seguir, faremos algumas considerações sobre a construção de planilha de dados, tabelas e gráficos.

1.2 Planilhas de Dados

Planilhas (usualmente eletrônicas) são matrizes em que se armazenam dados com o objetivo de permitir sua análise estatística. Em geral, cada linha da matriz de dados corresponde a uma unidade de investigação (*e.g.* unidade amostral) e cada coluna, a uma variável. Uma planilha bem elaborada contribui tanto para o entendimento do processo de coleta de dados e especificação das variáveis sob investigação quanto para a proposta de uma análise estatística adequada. A primeira etapa para a construção de uma planilha de dados consiste na elaboração de um dicionário com a especificação das variáveis, que envolve

i) sua definição operacional;
ii) a atribuição de rótulos (mnemônicos com letras minúsculas e sem acentos para facilitar a digitação e leitura por pacotes computacionais);
iii) a especificação das unidades de medida ou definição de categorias; para variáveis categorizadas, convém atribuir valores numéricos às categorias com a finalidade de facilitar a digitação e evitar erros (veja a variável "Sexo do recém-nascido" na Tabela 1.1);
iv) a atribuição de um código para valores omissos (*missing*);
v) a indicação de como devem ser codificados dados abaixo do limite de detecção (*e.g.*, $< 0,05$ ou $0,025$ se considerarmos que medidas abaixo do limite de detecção serão definidas como o ponto médio entre $0,00$ e $0,05$);
vi) a especificação do número de casas decimais (correspondente à precisão do instrumento de medida). Ver Nota de Capítulo 1.

Algumas recomendações para a construção da planilha de dados são

i) não utilizar limitadores de celas (*borders*) ou cores;
ii) reservar a primeira linha para os rótulos das variáveis;
iii) não esquecer uma coluna para a variável indicadora das unidades de investigação (evitar informações confidenciais como nomes de pacientes);
iv) escolher ponto ou vírgula para separação de casas decimais[1];

[1]Embora a norma brasileira ABNT indique a vírgula para separação de casas decimais, a maioria dos pacotes computacionais utiliza o ponto com essa função; por essa razão é preciso tomar cuidado com esse detalhe na construção de planilhas a serem analisadas computacionalmente. Em geral adotaremos a norma brasileira no texto.

v) especificar o número de casas decimais (ver Nota de Capítulo 1).

Exemplo 1.3 Na Tabela 1.3 apresentamos dados provenientes de um estudo em que o objetivo é avaliar a variação do peso (kg) de bezerros submetidos a uma determinada dieta entre 12 e 26 semanas após o nascimento.

Tabela 1.3: Peso de bezerros (kg)

	Semanas após nascimento							
animal	12	14	16	18	20	22	24	26
1	54.1	65.4	75.1	87.9	98.0	108.7	124.2	131.3
2	91.7	104.0	119.2	133.1	145.4	156.5	167.2	176.8
3	64.2	81.0	91.5	106.9	117.1	127.7	144.2	154.9
4	70.3	80.0	90.0	102.6	101.2	120.4	130.9	137.1
5	68.3	77.2	84.2	96.2	104.1	114.0	123.0	132.0
6	43.9	48.1	58.3	68.6	78.5	86.8	99.9	106.2
7	87.4	95.4	110.5	122.5	127.0	136.3	144.8	151.5
8	74.5	86.8	94.4	103.6	110.7	120.0	126.7	132.2
9	50.5	55.0	59.1	68.9	78.2	75.1	79.0	77.0
10	91.0	95.5	109.8	124.9	135.9	148.0	154.5	167.6
11	83.3	89.7	99.7	110.0	120.8	135.1	141.5	157.0
12	76.3	80.8	94.2	102.6	111.0	115.6	121.4	134.5
13	55.9	61.1	67.7	80.9	93.0	100.1	103.2	108.0
14	76.1	81.1	84.6	89.8	97.4	111.0	120.2	134.2
15	56.6	63.7	70.1	74.4	85.1	90.2	96.1	103.6

Dados com essa natureza são chamados de dados longitudinais por terem a mesma característica (peso, no exemplo) medida ao longo de uma certa dimensão (tempo, no exemplo). De acordo com nossa especificação, há nove variáveis na planilha representada na Tabela 1.3, nomeadamente, Animal, Peso na 12a semana, Peso na 14a semana etc. Para efeito computacional, no entanto, esse tipo de dados deve ser disposto numa planilha com formato diferente (às vezes chamado de formato longo) como indicado na Tabela 1.4. Nesse formato apropriado para dados longitudinais (ou mais geralmente, para medidas repetidas), há apenas três variáveis, a saber, Animal, Semana e Peso. Note que a mesma unidade amostral (animal) é repetida na primeira coluna para caracterizar a natureza longitudinal dos dados. Ele é especialmente adequado para casos em que as unidades de investigação são avaliadas em instantes diferentes. Um exemplo em que o diâmetro da aorta (mm) de recém nascidos pré-termo, com peso adequado (AIG) ou pequeno (PIG) para a idade gestacional foi avaliado até a 40a semana pós-concepção está apresentado na Tabela 1.5. Note que o número de observações pode ser diferente para as diferentes unidades de investigação. Esse formato também é comumente utilizado para armazenar dados de séries temporais.

Tabela 1.4: Planilha computacionalmente adequada para os dados do Exemplo 1.3

animal	semana	peso
1	12	54.1
1	14	65.4
1	16	75.1
1	18	87.9
1	20	98.0
1	22	108.7
1	24	124.2
1	26	131.3
2	12	91.7
2	14	104.0
⋮	⋮	⋮
2	26	176.8
⋮	⋮	⋮
15	12	56.6
⋮	⋮	⋮
15	26	103.6

1.3 Construção de tabelas

A finalidade primordial de uma tabela é resumir a informação obtida dos dados. Sua construção deve permitir que o leitor entenda esse resumo sem a necessidade de recorrer ao texto. Nesta seção vamos apresentar algumas sugestões para construção de tabelas.

[1] Não utilize mais casas decimais do que o necessário para não mascarar as comparações de interesse. A escolha do número de casas decimais depende da precisão do instrumento de medida e/ou da importância prática dos valores representados. Para descrever a redução de peso após um mês de dieta, por exemplo, é mais conveniente representá-lo como 6 kg do que como 6,200 kg. Por outro lado, quanto mais casas decimais forem incluídas, mais difícil é a comparação. Por exemplo, compare a Tabela 1.6 com a Tabela 1.7.

Observe que calculamos porcentagens em relação ao total de cada linha. Poderíamos, também, ter calculado porcentagens em relação ao total de cada coluna ou porcentagens em relação ao total geral (50). Cada uma dessas maneiras pode ser útil em determinada situação; por exemplo, determinar se há alguma dependência entre as duas variáveis estado civil e bebida preferida, avaliada em 50 indivíduos.

[2] Proponha um título autoexplicativo e inclua as unidades de medida. O título deve dizer o que representam os números do corpo da tabela e, em

Tabela 1.5: Planilha com diâmetro da aorta (mm) observado em recém-nascidos pré-termo

grupo	ident	sem	diam
AIG	2	30	7.7
AIG	2	31	8.0
AIG	2	32	8.2
AIG	2	34	9.1
AIG	2	35	9.4
AIG	2	36	9.8
AIG	12	28	7.1
AIG	12	29	7.1
AIG	12	37	7.3
AIG	12	39	9.0
AIG	12	30	9.4
⋮	⋮	⋮	⋮
PIG	17	33	7.5
PIG	17	34	7.7
PIG	17	36	8.2
PIG	29	26	6.3
PIG	29	27	6.5
PIG	29	28	6.6
PIG	29	29	6.6
PIG	29	30	6.8
PIG	29	31	7.2
PIG	29	32	7.2

Tabela 1.6: Número de alunos

Estado civil	Bebida preferida			Total
	não alcoólica	cerveja	outra alcoólica	
Solteiro	19 (53%)	7 (19%)	10 (28%)	36 (100%)
Casado	3 (25%)	4 (33%)	5 (42%)	12 (100%)
Outros	1 (50%)	0 (0%)	1 (50%)	2 (100%)
Total	23 (46%)	11 (22%)	16 (32%)	50 (100%)

geral, não deve conter informações que possam ser obtidas diretamente dos rótulos de linhas e colunas. Compare o título da Tabela 1.8 com: Intenção de voto (%) por candidato para diferentes meses.

[3] Inclua totais de linhas e/ou colunas para facilitar as comparações. É sempre bom ter um padrão contra o qual os dados possam ser avaliados.

[4] Não utilize abreviaturas ou indique o seu significado no rodapé da tabela

Tabela 1.7: Número de alunos (e porcentagens com duas casas decimais)

Estado	Bebida preferida			Total
civil	não alcoólica	cerveja	outra alcoólica	
Solteiro	19 (52,78%)	7 (19,44%)	10 (27,78%)	36 (100,00%)
Casado	3 (25,00%)	4 (33,33%)	5 (41,67%)	12 (100,00%)
Outros	1 (50,00%)	0 (0,00%)	1 (50,00%)	2 (100,00%)
Total	23 (46,00%)	11 (22,00%)	16 (32,00%)	50 (100%)

(*e.g.* Desvio padrão em vez de DP); se precisar utilize duas linhas para indicar os valores da coluna correspondente.

[5] Ordene colunas e/ou linhas quando possível. Se não houver impedimentos, ordene-as segundo os valores, crescente ou decrescentemente. Compare a Tabela 1.8 com a Tabela 1.9.

Tabela 1.8: Intenção de voto (%)

Candidato	janeiro	fevereiro	março	abril
Nononono	39	41	40	38
Nananana	20	18	21	24
Nenenene	8	15	18	22

[6] Tente trocar a orientação de linhas e colunas para melhorar a apresentação. Em geral, é mais fácil fazer comparações ao longo das linhas do que das colunas.

[7] Altere a disposição e o espaçamento das linhas e colunas para facilitar a leitura. Inclua um maior espaçamento a cada grupo de linhas e/ou colunas em tabelas muito extensas.

[8] Não analise a tabela descrevendo-a, mas sim comentando as principais tendências sugeridas pelos dados. Por exemplo, os dados apresentados na Tabela 1.6 indicam que a preferência por bebidas alcoólicas é maior entre os alunos casados do que entre os solteiros; além disso, há indicações de que a

Tabela 1.9: Intenção de voto (%)

Candidato	janeiro	fevereiro	março	abril
Nananana	20	18	21	24
Nononono	39	41	40	38
Nenenene	8	15	18	22

cerveja é menos preferida que outras bebidas alcoólicas, tanto entre solteiros quanto entre casados.

1.4 Construção de gráficos

A seguir apresentamos algumas sugestões para a construção de gráficos, cuja finalidade é similar àquela de tabelas, ou seja, resumir a informação obtida dos dados; por esse motivo, convém optar pelo resumo em forma de tabela ou de gráfico.

[1] Proponha um título autoexplicativo.

[2] Escolha o tipo de gráfico apropriado para os dados.

[3] Rotule os eixos apropriadamente, incluindo unidades de medida.

[4] Procure escolher adequadamente as escalas dos eixos para não distorcer a informação que se pretende transmitir. Se o objetivo for comparar as informações de dois os mais gráficos, use a mesma escala.

[5] Inclua indicações de "quebra" nos eixos para mostrar que a origem (zero) está deslocada.

[6] Altere as dimensões do gráfico até encontrar o formato adequado.

[7] Inclua uma legenda.

[8] Tome cuidado com a utilização de áreas para comparações, pois elas variam com o quadrado das dimensões lineares.

[9] Não exagere nas ilustrações que acompanham o gráfico para não o "poluir"visualmente, mascarando seus aspectos mais relevantes.

1.5 Aspectos Computacionais

Embora muitos cálculos necessários para uma análise estatística possam ser concretizados por meio de calculadoras, o recurso a pacotes computacionais é necessário tanto para as análises mais sofisticadas quanto para análises extensas. Neste livro usaremos preferencialmente o repositório de pacotes `R`, obtido livremente em *Comprehensive R Archive Network*, `CRAN`, no sítio

`http://CRAN.R-project.org`..

Pacotes alternativos são o SPlus, Minitab, SAS, MatLab etc.
Dentre as principais programotecas do pacote `R` usados, citamos:XXXX e YYYY.

Alguns conjuntos de dados analisados são dispostos ao longo do texto; outros são apresentados em formato Excel no arquivo MorettinSingerDados.xls disponível no sítio

`http://www.ime.usp.br/ jmsinger/Dados/MorettinSingerDados.xls`.

As folhas desse arquivo Excel são rotuladas no formato "CD-dados", *e.g.* CD-Poluicao, CD-Salarios etc. (sem incluir acentos). Quando necessário, indicaremos outros sítios em que se podem obter os dados utilizados nas análises.

1.6 Notas de capítulo

1) **Ordem de grandeza, precisão e arredondamento de dados quantitativos**

A precisão de dados quantitativos contínuos está relacionada com a capacidade de os instrumentos de medida distinguirem entre valores próximos na escala de observação do atributo de interesse. O número de dígitos colocados após a vírgula indica a precisão associada à medida que estamos considerando. O volume de um certo recipiente expresso como 0,740 L implica que o instrumento de medida pode detectar diferenças da ordem de 0,001 l (= 1 mL, ou seja 1 mililitro); se esse volume for expresso na forma 0,74 L, a precisão correspondente será de 0,01 L (= 1 cL, ou seja 1 centilitro).

Muitas vezes, em função dos objetivos do estudo em questão, a expressão de uma grandeza quantitativa pode não corresponder à precisão dos instrumentos de medida. Embora com uma balança suficientemente precisa, seja possível dizer que o peso de uma pessoa é de 89,230 kg, para avaliar o efeito de uma dieta, o que interessa saber é a ordem de grandeza da perda de peso após três meses de regime, por exemplo. Nesse caso, saber se a perda de peso foi de 10,230 kg ou de 10,245 kg é totalmente irrelevante. Para efeitos práticos, basta dizer que a perda foi da ordem de 10 kg. A ausência de casas decimais nessa representação indica que o próximo valor na escala de interesse seria 11 kg, embora todos os valores intermediários com unidades de 1 g sejam mensuráveis.

Para efeitos contábeis, por exemplo, convém expressar o aumento das exportações brasileiras num determinado período como R$ 1 657 235 458,29; no entanto, para efeitos de comparação com outros períodos, é mais conveniente dizer que o aumento das exportações foi da ordem de 1,7 bilhões de reais. Note que nesse caso, as grandezas significativas são aquelas da ordem de 0,1 bilhão de reais (= 100 milhões de reais).

Nesse processo de transformação de valores expressos com uma determinada precisão para outros com a precisão de interesse é preciso arredondar os números correspondentes. Em termos gerais, se o dígito a ser eliminado for 0, 1 2, 3 ou 4, o dígito precedente não deve sofrer alterações e se o dígito a ser eliminado for 5, 6, 7, 8 ou 9, o dígito precedente deve ser acrescido de uma unidade. Por exemplo, se desejarmos reduzir para duas casas decimais números originalmente expressos com três casas decimais, 0,263 deve ser transformado para 0,26 e 0,267 para

0,27. Se desejarmos uma redução mais drástica para apenas uma casa decimal, tanto 0,263 quanto 0,267 devem ser transformados para 0,3.

É preciso tomar cuidado com essas transformações quando elas são aplicadas a conjuntos de números cuja soma seja prefixada (porcentagens, por exemplo) pois elas podem introduzir erros cumulativos. Discutiremos esse problema ao tratar de porcentagens e tabulação de dados. É interessante lembrar que a representação decimal utilizada nos EUA e nos países da comunidade britânica substitui a vírgula por um ponto. Cuidados devem ser tomados ao se fazerem traduções, embora em alguns casos, esse tipo de representação já tenha sido adotada no cotidiano (veículos com motor 2.0, por exemplo, são veículos cujo volume dos cilindros é de 2,0 L).

2) **Proporções e porcentagens**

Uma proporção é um quociente utilizado para comparar duas grandezas através da adoção de um padrão comum. Se 31 indivíduos, num total de 138, são fumantes, dizemos que a proporção de fumantes entre esses 138 indivíduos é de 0,22 (= 31/138). O denominador desse quociente é chamado de base e a interpretação associada à proporção é que 31 está para a base 138 assim como 0,22 está para a base 1,00. Essa redução a uma base fixa permite a comparação com outras situações em que os totais são diferentes. Consideremos, por exemplo, um outro conjunto de 77 indivíduos em que 20 são fumantes; embora o número de fumantes não seja comparável com o do primeiro grupo, dado que as bases são diferentes, pode-se dizer que a proporção de fumantes desse segundo grupo, 0,26 (= 20/77) é maior que aquela associada ao primeiro conjunto.

Porcentagens, nada mais são do que proporções multiplicadas por 100, o que equivale a fazer a base comum igual a 100. No exemplo acima, podemos dizer que a porcentagem de fumantes é de 22% (=100 x 31/138) no primeiro grupo e de 26% no segundo. Para efeito da escolha do número de casas decimais, note que a comparação entre essas duas porcentagens é mais direta do que se considerássemos suas expressões mais precisas (com duas casas decimais), ou seja 22,46% contra 25,97%.

A utilização de porcentagens pode gerar problemas de interpretação em algumas situações. A seguir consideramos algumas delas. Se o valor do IPTU de um determinado imóvel cobrado foi de R$ 500,00 em 1998 e de R$ 700,00 em 1999, podemos dizer que o valor do IPTU em 1999 é 140% (= 100 x 700/500) do valor em 1998, mas o aumento foi de 40% (= 100 x (700-500)/500). Se o preço de uma determinada ação varia de R$ 22,00 num determinado instante para R$ 550,00 um ano depois, podemos dizer que o aumento de seu preço foi de 2400% (=100x(550-22)/22) nesse período. É difícil interpretar porcentagens "grandes" como essa. Nesse caso é melhor dizer que o preço dessa ação

é 25 (= 550/22) vezes seu preço há um ano. Porcentagens calculadas a partir de bases de pequena magnitude podem induzir conclusões inadequadas. Dizer que 43% dos participantes de uma pesquisa preferem um determinado produto tem uma conotação diferente se o cálculo for baseado em 7 ou em 120 entrevistados. É sempre conveniente explicitar a base relativamente à qual se estão fazendo os cálculos.

Para se calcular uma porcentagem global a partir das porcentagens associadas às partes de uma população, é preciso levar em conta sua composição. Suponhamos que numa determinada faculdade, 90% dos alunos que usam transporte coletivo sejam favoráveis à cobrança de estacionamento no campus e que apenas 20% dos alunos que usam transporte individual o sejam. A porcentagem de alunos dessa faculdade favoráveis à cobrança do estacionamento só será igual à média aritmética dessas duas porcentagens, ou seja 55%, se a composição da população de alunos for tal que metade usa transporte coletivo e metade não. Se essa composição for de 70% e 30% respectivamente, a porcentagem de alunos favoráveis à cobrança de estacionamento será de 69% (= 0,9x70% + 0,20x30% ou seja, 90% dos 70% que usam transporte coletivo + 20% dos 30% que utilizam transporte individual). Para evitar confusão, ao se fazer referência a variações, convém distinguir porcentagem e ponto percentual. Se a porcentagem de eleitores favoráveis a um determinado candidato aumentou de 14% antes para 21% depois da propaganda na televisão, pode-se dizer que a preferência eleitoral por esse candidato aumentou 50% (= 100 x (21-14)/14) ou foi de 7 pontos percentuais (e não de 7%). Note que o que diferencia esses dois enfoques é a base em relação à qual se calculam as porcentagens; no primeiro caso, essa base é a porcentagem de eleitores favoráveis ao candidato antes da propaganda (14%) e no segundo caso é o total (não especificado) de eleitores avaliados na amostra (favoráveis ou não ao candidato).

Uma porcentagem não pode diminuir mais do que 100%. Se o preço de um determinado produto decresce de R$ 3,60 para R$ 1,20, a diminuição de preço é de 67% (= 100 x (3,60 - 1,20)/3,60) e não de 200% (= 100 x (3,60 - 1,20)/1,20). Aqui também, o importante é definir a base: a idéia é comparar a variação de preço (R$ 2,40) com o preço inicial do produto (R$ 3,60) e não com o preço final (R$ 1,20). Na situação limite, em que o produto é oferecido gratuitamente, a variação de preço é de R$3,60; consequentemente, a diminuição de preço limite é de 100%. Note que se estivéssemos diante de um aumento de preço de R$ 1,20 para R$ 3,60, diríamos que o aumento foi de 200% (= 100 x (3,60 - 1,20)/1,20)).

1.7 Exercícios

1) O objetivo de um estudo da Faculdade de Medicina da USP foi avaliar a associação entre a quantidade de morfina administrada a pacientes com dores intensas provenientes de lesões medulares ou radiculares e a dosagem dessa substância em seus cabelos. Três medidas foram realizadas em cada paciente, a primeira logo após o início do tratamento e as demais após 30 e 60 dias. Detalhes podem ser obtidos no documento intitulado "morfina.doc", disponível no sítio

   `http://www.ime.usp.br/ jmsinger/MAE0217/morfina.doc`.

   A planilha morfina.xls, disponível no sítio

   `http://www.ime.usp.br/ jmsinger/MAE0217/morfina.xls`,

   foi entregue ao estatístico para análise e contém resumos de características demográficas além dos dados do estudo. Organize-a, construindo tabelas apropriadas para descrever essas características demográficas e uma planilha num formato apropriado para análise estatística.

2) A Figura 1.2 foi extraída de um relatório do Centro de Estatística Aplicada do IME/USP [ver Giampaoli et al. (2008) para detalhes]. Critique-a e reformule-a para facilitar sua leitura.

3) Utilize as sugestões para construção de planilhas apresentadas na Seção 1.2 com a finalidade de preparar os dados dos diferentes conjuntos do arquivo MorettinSingerDados.xls para análise estatística.

Figura 1.2: Tabela comparativa das notas médias da avaliação das subprefeituras no modelo padronizado para a escala [0,1]

Sheet1

Subprefeitura	Observado	Ajustado	Nota SP
Aricanduva	0,586	0,588	0,396
Butantã	0,483	0,468	0,334
Campo Limpo	0,484	0,526	0,362
Casa Verde/Cachoeirinha	0,558	0,554	0,382
Cidade Tiradentes	0,543	0,540	0,369
Freguesia/Brasilândia	0,545	0,540	0,371
Ipiranga	0,593	0,539	0,368
Itaim Paulista	0,566	0,557	0,374
Itaquera	0,396	0,563	0,383
Jabaquara	0,533	0,533	0,364
M' Boi Mirim	0,566	0,552	0,368
São Mateus	0,523	0,511	0,354
São Miguel	0,601	0,583	0,395
Socorro	0,601	0,523	0,360
V, Prudente/Sapopemba	0,648	0,620	0,413

Capítulo 2

Análise de dados de uma variável

2.1 Introdução

Neste capítulo consideraremos a análise descritiva de dados provenientes da observação de uma variável. As técnicas utilizadas podem ser empregadas tanto para dados provenientes de uma população quanto para dados oriundos de uma amostra.

A ideia de uma análise exploratória de dados (AED) é tentar responder as seguintes questões:

i) qual a frequência com que cada valor (ou intervalo de valores) aparece no conjunto de dados ou seja qual a distribuição de frequências dos dados?
ii) quais são alguns valores típicos do conjunto de dados, como mínimo e máximo?
iii) qual seria um valor para representar a posição (ou localização) central do conjunto de dados?
iv) qual seria uma medida da variabilidade ou dispersão dos dados?
v) existem valores atípicos ou discrepantes (*outliers*) no conjunto de dados?
vi) os dados podem ser considerados simétricos?

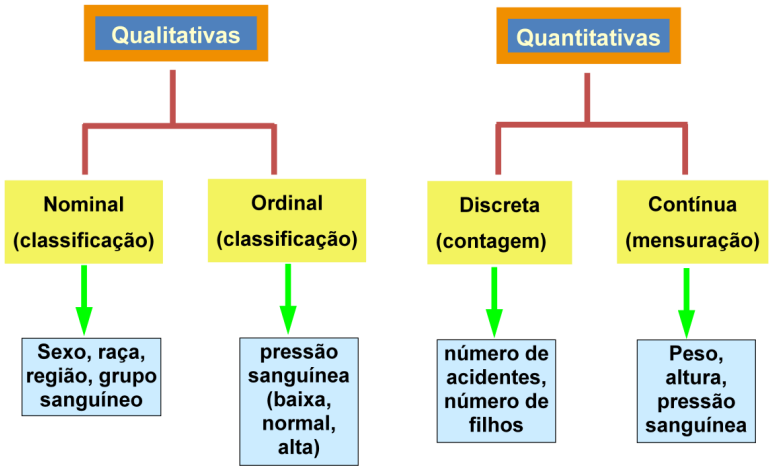
Nesse contexto, um dos objetivos da AED é organizar e exibir os dados de maneira apropriada e para isso utilizamos

i) gráficos e tabelas;
i) medidas para resumo de dados.

As técnicas empregadas na AED dependem do tipo de variáveis que compõem o conjunto de dados em questão. Uma possível classificação de variáveis está representada na Figura 2.1.

Variáveis qualitativas são aquelas que indicam um atributo da unidade de investigação (sexo, por exemplo). Elas podem ser ordinais, quando há

Figura 2.1: Classificação de variáveis



uma ordem nas diferentes categorias do atributo (tamanho de uma escola - pequena, média ou grande, por exemplo) ou nominais, quando não há essa ordem (região em que está localizada uma empresa - norte, sul, leste ou oeste, por exemplo).

Variáveis quantitativas são aquelas que exibem uma valor numérico associado à unidade de investigação (peso, por exemplo). Elas podem ser discreta, quando assume valores no conjunto dos números naturais (número de gestações de uma paciente) ou contínuas, quando assume valores no conjunto dos números reais (tempo gasto por um atleta para percorrer 100 m, por exemplo).

2.2 Distribuições de frequências

Exemplo 2.1: Consideremos um conjunto de dados como aquele apresentado na Tabela 2.1 obtido de um questionário respondido por 50 alunos de um curso ministrado na Fundação Getúlio Vargas em São Paulo.

Em geral, a primeira tarefa de uma análise estatística de um conjunto de dados consiste em resumi-los. As técnicas disponíveis para essa finalidade dependem do tipo de variáveis envolvidas, tema que discutiremos a seguir.

2.2.1 Variáveis qualitativas

Uma tabela contendo as frequências (absolutas e/ou relativas) de unidades de investigação para cada categoria do atributo avaliado por uma variável qualitativa é chamada de distribuição de frequências dessa variável. As Tabelas 2.2 e 2.3 por exemplo, representam respectivamente as distribuições de frequências das variáveis "Bebida preferida" e "Fluência em inglês" para

Tabela 2.1: Dados de um estudo realizado na FGV

ident	Salário (R$)	Fluência inglês	Anos de formado	Estado civil	Número de filhos	Bebida preferida
1	3500	fluente	12.0	casado	1	outra alcoólica
2	1800	nenhum	2.0	casado	3	não alcoólica
3	4000	fluente	5.0	casado	1	outra alcoólica
4	4000	fluente	7.0	casado	3	outra alcoólica
5	2500	nenhum	11.0	casado	2	não alcoólica
6	2000	fluente	1.0	solteiro	0	não alcoólica
7	4100	fluente	4.0	solteiro	0	não alcoólica
8	4250	algum	10.0	casado	2	cerveja
9	2000	algum	1.0	solteiro	2	cerveja
10	2400	algum	1.0	solteiro	0	não alcoólica
11	7000	algum	15.0	casado	1	não alcoólica
12	2500	algum	1.0	outros	2	não alcoólica
13	2800	fluente	2.0	solteiro	1	não alcoólica
14	1800	algum	1.0	solteiro	0	não alcoólica
15	3700	algum	10.0	casado	4	cerveja
16	1600	fluente	1.0	solteiro	2	cerveja
⋮	⋮	⋮	⋮	⋮	⋮	⋮
26	1000	algum	1.0	solteiro	1	outra alcoólica
27	2000	algum	5.0	solteiro	0	outra alcoólica
28	1900	fluente	2.0	solteiro	0	outra alcoólica
29	2600	algum	1.0	solteiro	0	não alcoólica
30	3200		6.0	casado	3	cerveja
31	1800	algum	1.0	solteiro	2	outra alcoólica
32	3500		7.0	solteiro	1	cerveja
33	1600	algum	1.0	solteiro	0	não alcoólica
34	1700	algum	4.0	solteiro	0	não alcoólica
35	2000	fluente	1.0	solteiro	2	não alcoólica
36	3200	algum	3.0	solteiro	2	outra alcoólica
37	2500	fluente	2.0	solteiro	2	outra alcoólica
38	7000	fluente	10.0	solteiro	1	não alcoólica
39	2500	algum	5.0	solteiro	1	não alcoólica
40	2200	algum	0.0	casado	0	cerveja
41	1500	algum	0.0	solteiro	0	não alcoólica
42	800	algum	1.0	solteiro	0	não alcoólica
43	2000	fluente	1.0	solteiro	0	não alcoólica
44	1650	fluente	1.0	solteiro	0	não alcoólica
45		algum	1.0	solteiro	0	outra alcoólica
46	3000	algum	7.0	solteiro	0	cerveja
47	2950	fluente	5.5	outros	1	outra alcoólica
48	1200	algum	1.0	solteiro	0	não alcoólica
49	6000	algum	9.0	casado	2	outra alcoólica
50	4000	fluente	11.0	casado	3	outra alcoólica

os dados do Exemplo 2.1.

Tabela 2.2: Distribuição de frequências para a variável Bebida preferida correspondente ao Exemplo 2.1

Bebida preferida	Frequência observada	Frequência relativa (%)
não alcoólica	23	46
cerveja	11	22
outra alcoólica	16	32
Total	50	100

Tabela 2.3: Distribuição de frequências para a variável Fluência em inglês correspondente ao Exemplo 2.1

Fluência em inglês	Frequência observada	Frequência relativa (%)	Frequência acumulada (%)
nenhuma	2	4	4
alguma	26	54	58
fluente	20	42	100
Total	48	100	

Obs: dois participantes não forneceram informação

Note que para variáveis variáveis qualitativas ordinais pode-se acrescentar uma coluna com as frequências relativas acumuladas que são úteis na sua análise. Por exemplo a partir da última coluna da Tabela 2.3 pode-se afirmar que cerca de 60% dos alunos que forneceram a informação tem no máximo alguma fluência em inglês.

O resumo exibido nas tabelas com distribuições de frequências pode ser representado por meio de gráficos de barras ou gráficos do tipo pizza (ou torta). Exemplos correspondentes às variáveis "Bebida preferida" e "Fluência em inglês" são apresentados nas Figuras 2.2, 2.3 e 2.4.

Note que na Figura 2.2 as barras podem ser colocadas em posições arbitrárias; na Figura 2.4, convém coloca-las de acordo com a ordem natural das categorias.

2.2.2 Variáveis quantitativas

Se utilizássemos o mesmo critério adotado para variáveis qualitativas na construção de distribuições de frequências de variáveis quantitativas (especialmente no caso de variáveis contínuas), em geral obteríamos tabelas com frequência muito pequena (em geral 1) para as diversas categorias, deixando de atingir o objetivo de resumir os dados. Para contornar o problema,

Figura 2.2: Gráfico de barras para Bebida preferida

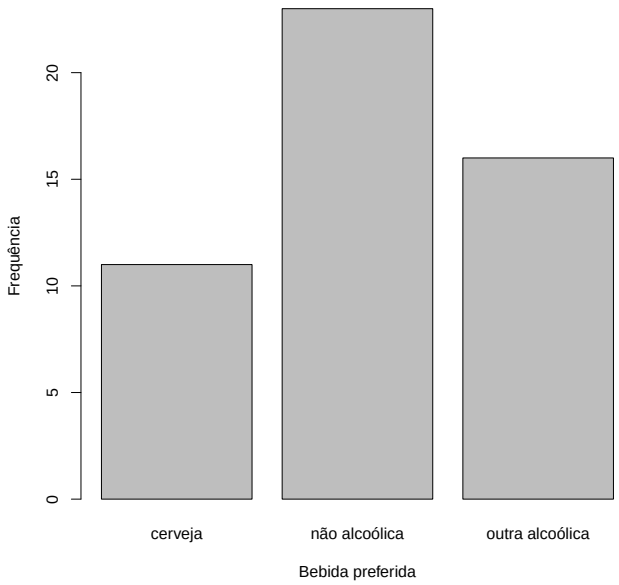
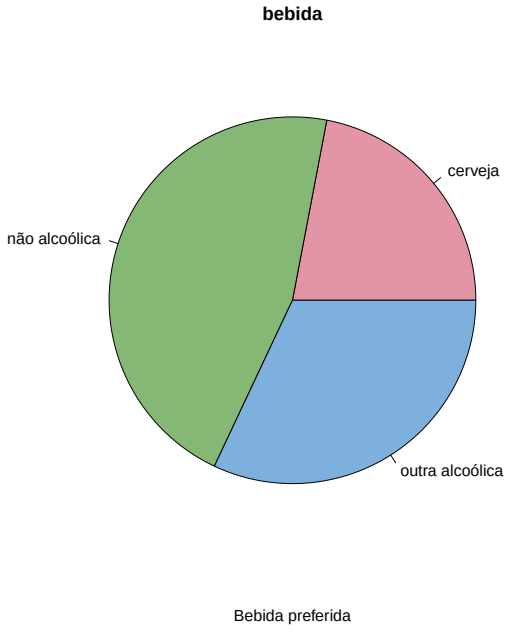


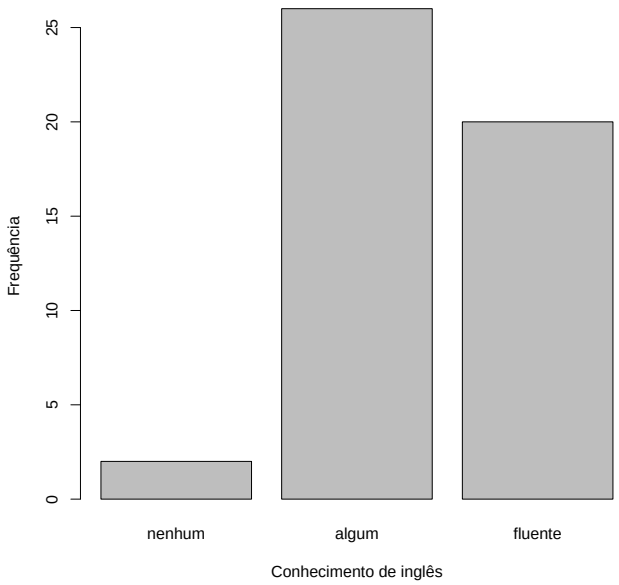
Figura 2.3: Gráfico tipo pizza para Bebida preferida



agrupam-se os valores das variáveis em classes e obtêm-se as frequências em cada classe.

Uma possível distribuição de frequências para a variável "Salário" cor-

Figura 2.4: Gráfico de barras para Fluência em inglês



respondente ao Exemplo 2.1 está apresentada na Tabela 2.4.

Tabela 2.4: Distribuição de frequências para a variável Salário correspondente ao Exemplo 2.1

Classe de salário (R$)	Frequência observada	Frequência relativa (%)	Frequência relativa acumulada (%)
$0 \vdash 1500$	6	12,2	12,2
$1500 \vdash 3000$	27	55,1	67,3
$3000 \vdash 4500$	12	24,5	91,8
$4500 \vdash 6000$	2	4,1	95,9
$6000 \vdash 7500$	2	4,1	100,0
Total	49	100,0	100,0

Obs: um dos participantes não informou o salário.

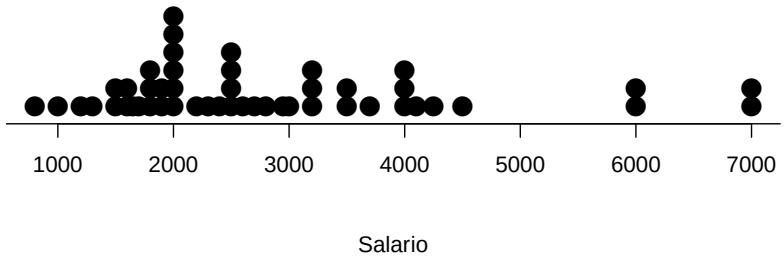
Alternativamente a tabelas com o formato da Tabela 2.4 vários gráficos podem ser utilizados para representar a distribuição de frequências de um conjunto de dados. Os mais utilizados são apresentados a seguir.

Gráfico de dispersão unidimensional (*dotplot*)

Neste gráfico representamos por pontos os valores $x_1, \ldots, x_n$ ao longo de uma reta provida de uma escala. Valores repetidos são empilhados, de modo que possamos ter uma ideia de sua distribuição. O gráfico de dispersão

unidimensional para a variável Salário do Exemplo 2.1 está representado na Figura 2.5.

Figura 2.5: Gráfico de dispersão unidimensional para a variável Salário (Exemplo 2.1)



Um procedimento alternativo para reduzir um conjunto de dados sem perder muita informação sobre eles consiste na construção de um gráfico chamado ramo-e-folhas. Não há regras fixas para construí-lo, mas a ideia básica é dividir cada observação em duas partes: o *ramo*, colocado à esquerda de uma linha vertical, e a *folha*, colocada à direita.

Considere a variável "Salário" do Exemplo 2.1. Para cada observação, podemos considerar o ramo como a parte inteira e a folha como a parte decimal. Assim fazendo, é fácil ver que obtemos muitos ramos, essencialmente tantos quantos são os dados, e não teríamos alcançado o objetivo de resumi-los. Outra possibilidade é arredondar os dados para números inteiros e considerar o primeiro dígito como o ramo e o segundo como folha, no caso de dezenas; os dois primeiros dígitos como o ramo e o terceiro como folha, para as centenas, etc. O gráfico correspondente está apresentado na Figura 2.6. Pelo gráfico podemos avaliar a forma da distribuição das observações; em particular, vemos que há quatro valores atípicos, nomeadamente, dois iguais a R$ 6000 (correspondentes aos alunos 22 e 49) e dois iguais a R$ 7000 (correspondentes aos alunos 11 e 38) respectivamente.

Histograma

O histograma é um gráfico construído a partir da distribuição de frequências dos dados e é composto de retângulos contíguos cuja área total é em geral normalizada para ter valor unitário. A **área** de cada retângulo corresponde à frequência relativa associada à classe definida pela sua base.

Um histograma correspondente à distribuição de frequências indicada na Tabela 2.4 está representado na Figura 2.7.

Mais formalmente, dados os valores $x_1, \ldots, x_n$ de uma variável quantitativa X, podemos construir uma tabela contendo

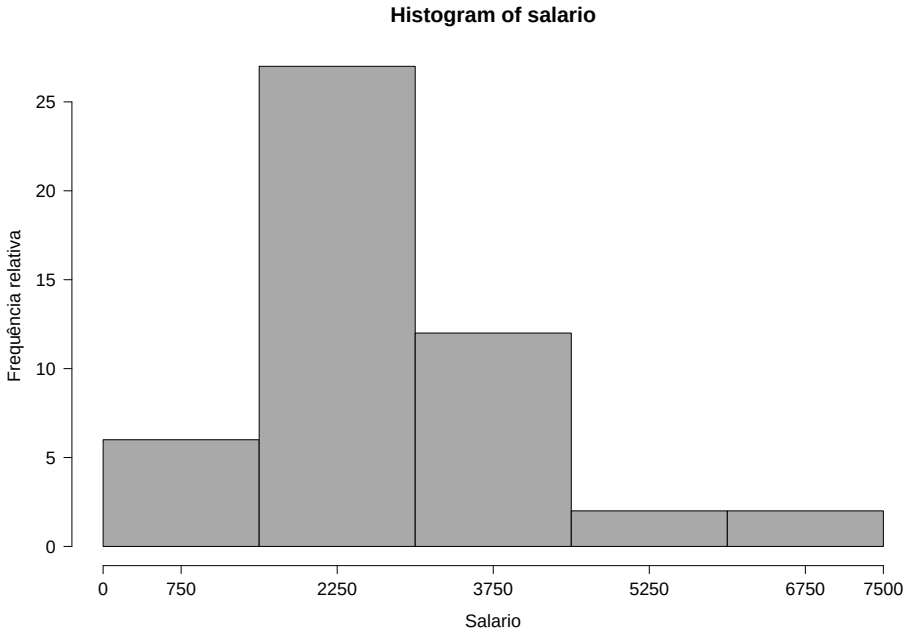
a) as frequências absolutas n_k, $k = 1, \ldots, K$, que correspondem aos números de elementos cujos valores pertencem às classes $k = 1, \ldots, K$;

Figura 2.6: Gráfico ramo-e-folhas para a variável Salário (R$)

```
1 | 2: representa 1200
unidade da folha: 100
n: 49

   0 | 8
   1 | 023
   1 | 55666788899
   2 | 000000234
   2 | 55556789
   3 | 0222
   3 | 557
   4 | 00012
   4 | 5
   5 |
   5 |
   6 | 00
   6 |
   7 | 00
```

Figura 2.7: Histograma para a variável salário (R$)



b) as frequências relativas $f_k = n_k/n,\ k = 1, \ldots, K$, que são as proporções de elementos cujos valores pertencem às classes $k = 1, \ldots, K$;

c) as densidades de frequência $d_k = f_k/h_k,\ k = 1, \ldots, K$, que representam as proporções de valores pertencentes às classes $k = 1, \ldots, K$ por unidade de comprimento h_k de cada classe.

Exemplo 2.2 Os dados correspondentes à população[1] (em 10000 habitantes) de 30 municípios brasileiros (IBGE 1996) estão dispostos na Tabela 2.5.

Tabela 2.5: População de 30 municípios brasileiros (10000 habitantes)

Município	População	Município	População
São Paulo (SP)	988.8	Nova Iguaçu (RJ)	83.9
Rio de Janeiro (RJ)	556.9	São Luis (MA)	80.2
Salvador (BA)	224.6	Maceió (AL)	74.7
Belo Horizonte (MG)	210.9	Duque de Caxias (RJ)	72.7
Fortaleza (CE)	201.5	São Bernardo do Campo (SP)	68.4
Brasília (DF)	187.7	Natal (RN)	66.8
Curitiba (PR)	151.6	Teresina (PI)	66.8
Recife (PE)	135.8	Osasco (SP)	63.7
Porto Alegre (RS)	129.8	Santo André (SP)	62.8
Manaus (AM)	119.4	Campo Grande (MS)	61.9
Belém (PA)	116.0	João Pessoa (PB)	56.2
Goiânia (GO)	102.3	Jaboatão (PE)	54.1
Guarulhos (SP)	101.8	Contagem (MG)	50.3
Campinas (SP)	92.4	São José dos Campos (SP)	49.7
São Gonçalo (RJ)	84.7	Ribeirão Preto (SP)	46.3

Ordenemos os valores da variável X = população do município $i = 1, \ldots, 30$ do menor para o maior e consideremos a primeira classe como aquela com limite inferior igual a 40 e a última com limite superior igual a 1000; para que as classes sejam disjuntas, tomemos por exemplo, intervalos semi-abertos. A Tabela 2.6 contém a distribuição de frequências para a variável X.

Observemos que as quatro primeiras classes têm amplitudes iguais a 20, a quinta tem amplitude 30, as duas seguintes, amplitudes iguais 50, a penúltima tem amplitude igual 350 e a última, amplitude igual a 400. Observemos que $K = 9$, $\sum_{k=1}^{K} n_k = n = 30$ e que $\sum_{k=1}^{K} f_k = 1$. Notemos, também, que as classes $[120, 150)$ e $[150, 200)$ têm a mesma frequência relativa $(0,067)$ mas densidades de frequência diferentes. Quanto maior for a

[1] Aqui, o termo "população"se refere ao número de habitantes e é encarado como uma variável. Não deve ser confundido com população no contexto estatístico, que se refere a um conjunto (na maioria das vezes, conceitual) de valores de uma ou mais variáveis medidas. Podemos considerar, por exemplo, a população de pesos de pacotes de feijão produzidos por uma empresa.

Tabela 2.6: Distribuição de frequências para a X = população em dezenas de milhares de habitantes

classes	h_k	n_k	f_k	$d_k = f_k/h_k$
$40 \vdash 60$	20	5	0,167	0,00835
$60 \vdash 80$	20	8	0,267	0,01333
$80 \vdash 100$	20	4	0,133	0,00665
$100 \vdash 120$	20	4	0,133	0,00665
$120 \vdash 150$	30	2	0,067	0,00223
$150 \vdash 200$	50	2	0,067	0,00134
$200 \vdash 250$	50	3	0,100	0,00200
$250 \vdash 600$	350	1	0,033	0,00009
$600 \vdash 1000$	400	1	0,033	0,00008
Total	–	30	1,000	–

densidade de frequência de uma classe, maior será a concentração de valores nessa classe.

O valor da amplitude de classes h deve ser escolhido de modo adequado. Se h for grande, teremos poucas classes e o histograma pode não mostrar detalhes importantes; por outro lado, se h for pequeno, teremos muitas classes e algumas poderão ser vazias. A escolha do número e amplitude das classes é arbitrária. Detalhes técnicos sobre a escolha do número de classes em casos específicos podem ser encontrados na Nota de Capítulo 1. Uma definição mais técnica de histograma está apresentada na Nota de Capítulo 2.

O histograma da Figura 2.8 corresponde à distribuição de frequências da variável X do Exemplo 2.2, obtido usando a função hist do R.

O gráfico de ramo-e-folhas para os dados da Tabela 2.5 está apresentado na Figura 2.9. Pelo gráfico podemos avaliar a forma da distribuição das observações; em particular, vemos que há dois valores atípicos, $556,9$ e $988,8$, correspondentes às populações do Rio de Janeiro e São Paulo, respectivamente.

Quando há muitas folhas num ramo, podemos considerar ramos subdivididos, como no exemplo a seguir.

Exemplo 2.3 Os dados da planilha intitulada "CD-Poluicao" correspondem à concentração atmosférica de poluentes ozônio O_3 e monóxido de carbono (CO) além de temperatura média e umidade na cidade de São Paulo entre 1 de janeiro e 30 de abril de 1991. O gráfico de ramo-e-folhas para a concentração de monóxido de carbono pode ser construído com dois ramos, colocando-se no primeiro folhas com dígitos de 0 a 4 e no segundo, folhas com dígitos de 5 a 9. Esse gráfico está apresentado na Figura 2.10

Figura 2.8: Histograma para a variável População (10000 habitantes)

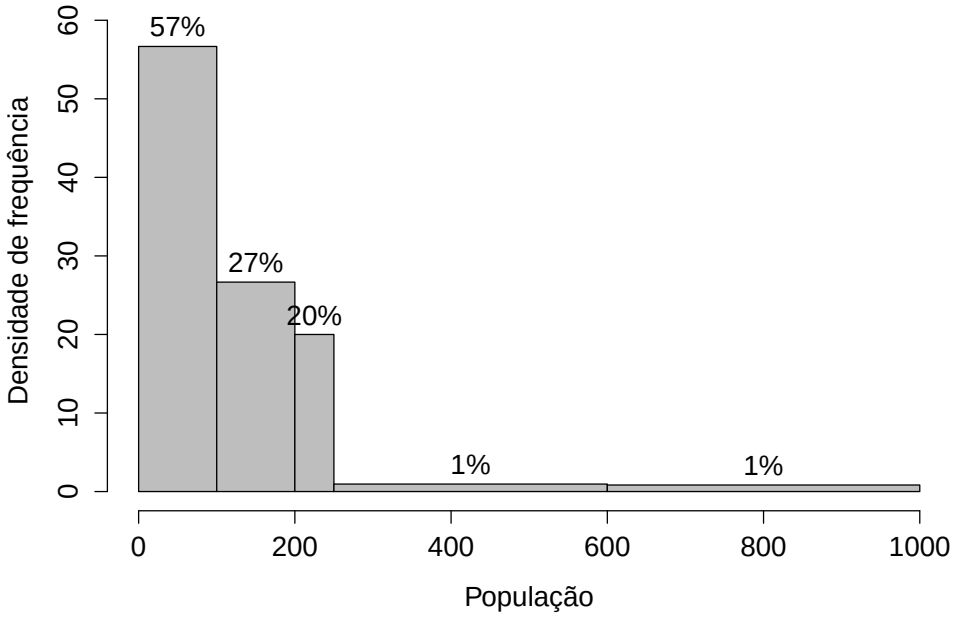


Figura 2.9: Gráfico ramo-e-folhas para a variável População (10000 habitantes)

```
1 | 2: representa 120
unidade da folha: 10
n: 30

    0 | 44555666666778889
    1 | 00112358
    2 | 012
    3 |
    4 |
    5 | 5
    6 |
    7 |
    8 |
    9 | 8
```

Figura 2.10: Gráfico ramo-e-folhas para a variável CO (ppm)

```
A separação decimal está em |

   4 | 77
   5 | 12
   5 | 55677789
   6 | 11111222222222333333444444
   6 | 5666677777899999999
   7 | 00122233444
   7 | 5566777778888899999999
   8 | 012334
   8 | 55678999
   9 | 0114
   9 | 557
  10 | 1333
  10 | 8
  11 | 4
  11 | 69
  12 | 0
  12 | 5
```

2.3 Medidas resumo

Em muitas situações deseja-se fazer um resumo mais drástico de um determinado conjunto de dados, por exemplo, por meio de um ou dois valores. A renda per capita de um país ou a porcentagem de eleitores favoráveis a um candidato são exemplos típicos. Com essa finalidade podem-se considerar as chamadas medidas de posição (localização ou de tendência central), as medidas de dispersão e medidas de assimetria, entre outras.

2.3.1 Medidas de posição

As medidas de posição mais utilizadas são a média, a mediana, a média aparada e os quantis. Para defini-las, consideremos as observações $x_1, \ldots, x_n$ de uma variável X.

A **média aritmética** (ou simplesmente média) é definida por

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2.1}$$

No caso de dados agrupados numa distribuição de frequências de um conjunto com n valores, K classes e n_k valores na classe k, $k = 1, \ldots, K$, a média pode ser calculada como

$$\overline{x} = \frac{1}{n}\sum_{k=1}^{K} n_k \tilde{x}_k = \sum_{k=1}^{K} f_k \tilde{x}_k, \tag{2.2}$$

em que $\tilde{x}_k$ é o ponto médio correspondente à classe k e $f_k = n_k/n$. Essa mesma expressão é usada para uma variável discreta, com n_k valores iguais a x_k, bastando para isso, substituir com $\tilde{x}_k$ por x_k.

A **mediana** é definida em termos das estatísticas de ordem, $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ por

$$\mathrm{md}(x) = \begin{cases} x_{(\frac{n+1}{2})}, & \text{se } n \text{ for ímpar,} \\ \frac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}], & \text{se } n \text{ for par.} \end{cases} \tag{2.3}$$

Dado um número $0 < \alpha < 1$, a **média aparada** de ordem α, $\overline{x}(\alpha)$ é definida como a média do conjunto de dados obtido após a eliminação das $100\alpha\%$ primeiras observações ordenadas e das $100\alpha\%$ últimas observações ordenadas do conjunto original. Uma definição formal é:

$$\overline{x}(\alpha) = \begin{cases} \frac{1}{n(1-2\alpha)}\{\sum_{i=m+2}^{n-m-1} x_{(i)} + (1+m-n\alpha)[x_{(m+1)} + x_{(n-m)}]\}, \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{se } m+2 \leq n-m+1 \\ \frac{1}{2}[x_{(m+1)} + x_{(n-m)}] \qquad\qquad \text{em caso contrário.} \end{cases} \tag{2.4}$$

em que m é o maior inteiro menor ou igual a $n\alpha$, $0 < \alpha < 0,5$. Se $\alpha = 0,5$, obtemos a mediana. Para $\alpha = 0,25$ obtemos a chamada a **meia média**. Observe que se $\alpha = 0$, $\overline{x}(0) = \overline{x}$.

Exemplo 2.4 Consideremos um conjunto com $n = 10$ valores de uma variável X dados por $14, 7, 3, 18, 9, 220, 34, 23, 6, 15$.

Então, $\overline{x} = 34,9$, $\text{md}(x) = (14+15)/2 = 14,5$, e $\overline{x}(0,2) = [x_{(3)} + x_{(4)} + \ldots + x_{(8)}]/6 = 14,3$. Note que se usarmos (2.4), temos $\alpha = 0,2$ e $m = 2$ obtendo o mesmo resultado. Se $\alpha = 0,25$, então de (2.4) obtemos

$$\overline{x}(0,25) = \frac{x_{(3)} + 2x_{(4)} + 2x_{(5)} + \ldots + 2x_{(7)} + x_{(8)}}{10} = 14,2$$

Observe que a média é bastante afetada pelo valor atípico 220, ao passo que a mediana e a média aparada com $\alpha = 0,2$ não o são. Dizemos que essas duas últimas são **medidas resistentes** ou **robustas**.

Uma medida diz-se resistente se ela muda pouco quando alterarmos um número pequeno dos valores do conjunto de dados. Se substituirmos o valor 220 do exemplo por 2200, a média passa para $232,9$ ao passo que a mediana e a média aparada $\overline{x}(0,20)$ não se alteram.

As três medidas consideradas acima são chamadas de posição ou localização central do conjunto de dados. Para variáveis qualitativas também é comum utilizarmos outra medida de posição que indica o valor mais frequente, denominado **moda**. Quando há duas classes com a mesma frequência máxima, a variável (ou distribuição) é dita **bimodal**. A não ser que os dados de uma variável contínua sejam agrupados em classes, caso em que se pode considerar a **classe modal**, não faz sentido considerar a moda, pois em geral, cada valor da variável tem frequência unitária.

Consideremos agora medidas de posição úteis para indicar outras posições não centrais dos dados. Informalmente, um quantil-p ou quantil de ordem p é **um valor da variável** (quando ela é contínua) ou **um valor interpolado entre dois valores da variável** (quando ela é discreta) que deixa $100p\%$ $(0 < p < 1)$ das observações à sua esquerda. Formalmente, definimos o quantil-p empírico (ou simplesmente quantil) como

$$Q(p) = \begin{cases} x_{(i)}, & \text{se } p = p_i = (i - 0,5)/n,\ i = 1, \ldots, n \\ (1 - f_i)Q(p_i) + f_i Q(p_{i+1}), & \text{se } p_i < p < p_{i+1} \\ x_{(1)}, & \text{se } 0 < p < p_1 \\ x_{(n)}, & \text{se } p_n < p < 1, \end{cases} \tag{2.5}$$

em que $f_i = (p - p_i)/(p_{i+1} - p_i)$. Ou seja, se p for da forma $p_i = (i - 0,5)/n$, o quantil-p coincide com a i-ésima observação ordenada. Para um valor p entre p_i e p_{i+1}, o quantil $Q(p)$ pode ser definido como sendo a ordenada de um ponto situado no segmento de reta determinado por $[p_i, Q(p_i)]$ e $[p_{i+1}, Q(p_{i+1})]$. Escolhemos p_i como acima (e não como i/n, por exemplo) de forma que se um quantil coincidir com uma das observações, metade dela pertencerá ao conjunto de valores à esquerda de $Q(p)$ e metade ao conjunto de valores à sua direita.

Os quantis amostrais para os dez pontos do Exemplo 2.4 estão indicados na Tabela 2.7. Com essa informação, podemos calcular outros quantis; por exemplo, $Q(0,10) = [x_{(1)} + x_{(2)}]/2 = (3 + 6)/2 = 4,5$ com $f_1 = (0,10 -$

Tabela 2.7: Quantis amostrais para os dados do Exemplo 2.4

i	1	2	3	4	5	6	7	8	9	10
p_i	0,05	0,15	0,25	0,35	0,45	0,55	0,65	0,75	0,85	0.95
$Q(p_i)$	3	6	7	9	14	15	18	23	34	220

$0,05)/(0,10) = 0,5$, $Q(0,90) = [x_{(9)} + x_{(10)}]/2 = (34 + 220)/2 = 127$, pois $f_9 = 0,5$ e $Q(0,62) = [0,30 \times x_{(5)} + 0,70 \times x_{(6)}] = (0,3 \times 14 + 0,7 \times 15 = 14,7$ pois $f_6 = (0,62 - 0,55)/0,10 = 0,7$ Note que a definição (2.5) é compatível com a definição de mediana apresentada anteriormente.

Os quantis $Q(0,25), Q(0,50)$ e $Q(0,75)$ são chamados **quartis** e usualmente são denotados por Q_1, Q_2 e Q_3, respectivamente. O quartil Q_2 é a mediana e a proporção dos dados entre Q_1 e Q_3 é 50%.

Outras denominações comumente empregadas são $Q(0,10)$: primeiro decil, $Q(0,20)$: segundo decil ou vigésimo percentil, $Q(0,85)$: octogésimo-quinto percentil etc.

2.3.2 Medidas de dispersão

Duas medidas de dispersão (ou de escala ou de variabilidade) bastante usadas são obtidas tomando-se a média dos desvios das observações em relação à sua média. Considere as observações $x_1, \ldots, x_n$, não necessariamente distintas.

A **variância** desse conjunto de valores é definida por

$$\text{var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{2.6}$$

No caso de uma tabela de frequências (com K classes), a expressão para cálculo da variância é

$$\text{var}(x) = \frac{1}{n}\sum_{k=1}^{K} n_k(\tilde{x}_k - \overline{x})^2 = \sum_{k=1}^{K} f_k(\tilde{x}_k - \overline{x})^2, \tag{2.7}$$

com a notação estabelecida anteriormente. Uma expressão equivalente que facilita os cálculos é

$$\text{var}(x) = n^{-1}\sum_{i=1}^{n} x_i^2 - \overline{x}^2. \tag{2.8}$$

Como a unidade de medida da variância é o quadrado da unidade de medida da variável correspondente, convém definir outra medida de dispersão que mantenha a unidade de medida original. Uma medida com essa propriedade é a raiz quadrada positiva da variância conhecida por **desvio padrão**, denotado $\text{dp}(x)$.

Para garantir certas propriedades estatísticas úteis para propósitos de inferência, convém modificar as definições acima. Em particular, para garantir que a variância obtida de uma amostra de dados de uma população

seja um **estimador não enviesado** da variância populacional basta definir a variância como

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{2.9}$$

em substituição à definição (2.6). Para detalhes, veja Bussab e Morettin (2014) entre outros. Em geral, S^2 é conhecida por **variância amostral**. A **variância populacional** é definida como em (2.9) com o denominador $n-1$ substituído pelo tamanho populacional N. As fórmulas de cálculo acima podem ser modificadas facilmente com essa definição; O desvio padrão amostral é denotado por S.

O **desvio médio** ou **desvio absoluto médio** definido por

$$\mathrm{dm}(x) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|. \tag{2.10}$$

Outra medida de dispersão bastante utilizada é a **distância interquartis** ou **amplitude interquartis**

$$d_Q = Q_3 - Q_1. \tag{2.11}$$

Podemos também considerar uma de dispersão definida em termos de desvios em relação à mediana. Como a mediana é uma medida robusta, nada mais natural que definir o **desvio mediano absoluto** como

$$\mathrm{dma}(x) = \mathrm{md}_{1\le i\le n}|x_i - \mathrm{md}(x)|, \tag{2.12}$$

Finalmente, uma medida correspondente à média aparada é a **variância aparada**, definida por

$$S^2(\alpha) = \begin{cases} \frac{c_\alpha}{n(1-2\alpha)}\left(\sum_{i=m+2}^{n-m-1}[x_{(i)} - \overline{x}(\alpha)]^2 + A\right), & m+2 \le n-m+1 \\ \frac{1}{2}[(x_{(m+1)} - \overline{x}(\alpha))^2 + (x_{(n-m)} - \overline{x}(\alpha))^2], & \text{em caso contrário,} \end{cases} \tag{2.13}$$

em que

$$A = (1 + m - n\alpha)[(x_{(m+1)} - \overline{x}(\alpha))^2 + (x_{(n-m)} - \overline{x}(\alpha))^2],$$

m é como em (2.4) e c_α é uma constante normalizadora que torna $S^2(\alpha)$ um estimador não enviesado para σ^2. Para n grande, $c_\alpha = 1,605$. Para amostras pequenas, veja a tabela da página 173 de Johnson e Leone (1964). Em particular, para $n = 10$, $c_\alpha = 1,46$.

A menos do fator c_α, a variância aparada pode ser obtida calculando-se a variância amostral das observações restantes, após a eliminação das $100\alpha\%$ iniciais e finais (com denominador $n-l$ em que l é o número de observações desprezadas).

Considere as observações do Exemplo 2.4. Para esse conjunto de dados as medidas de dispersão apresentadas são $S^2 = 4313,9$, $S = 65,7$, $\mathrm{dm}(x) = 37,0$, $d_Q = 23 - 7 = 16$, $S^2(0,2) = 34,3$, $S(0,20) = 5,9$, e $\mathrm{dma}(x) = 7,0$.

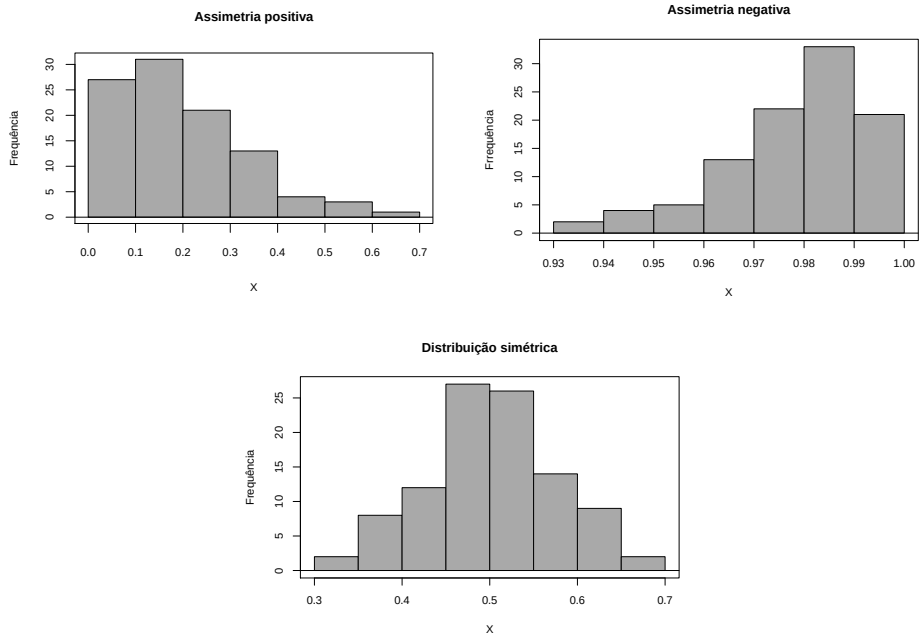
Observemos que as medidas robustas são, em geral, menores do que $\overline{x}$ e S e que $d_Q/1,349 = 11,9$. Se considerarmos que esses dados constituem uma amostra de uma população com desvio padrão σ, pode-se mostrar que, dma$/0,6745$ é um estimador não-enviesado para σ. A constante $0,6745$ é obtida por meio de considerações assintóticas. No exemplo, dma$/0,6745 = 10,4$. Note que esses dois estimadores do desvio padrão populacional coincidem. Por outro lado, S é muito maior, pois sofre bastante influencia do valor 220. Retirando-se esse valor do conjunto de dados, a média dos valores restantes é $14,3$ e o correspondente desvio padrão é $9,7$.

2.3.3 Medidas de assimetria

Embora sejam menos utilizadas na prática que as medidas de posição e dispersão, as medidas de assimetria (*skewness*) são úteis para identificar modelos probabilísticos para análise inferencial.

Na Figura 2.11 estão apresentados histogramas correspondentes a dados com assimetria positiva (ou à direita) ou negativa (ou à esquerda) e simétrico. O objetivo das medidas de assimetria é quantificar sua mag-

Figura 2.11: Histogramas com assimetria positiva e negativa e simétrico



nitude e, em geral, são baseadas na relação entre o segundo e o terceiro momentos centrados, nomeadamente

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad \text{e} \quad m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3.$$

Dentre elas as mais comuns são:

a) o coeficiente de assimetria de Fisher-Pearson: $g_1 = m_3/m_2^{3/2}$
b) o coeficiente de assimetria de Fisher-Pearson ajustado:

$$\frac{\sqrt{n(n-1)}}{n-1}]\sum_{i=1}^{n}[(x_i - \overline{x})/\sqrt{m_2}]^3.$$

As principais propriedades desses coeficientes são

i) seu sinal reflete a direção da assimetria;
ii) comparam a assimetria dos dados com aquela da distribuição Normal;
iii) valores mais afastados do zero indicam maiores magnitudes de assimetria e consequentemente, maior afastamento da distribuição Normal;
iv) a estatística indicada em b) tem um ajuste para o tamanho amostral;
v) o ajuste tem pequeno impacto em grandes amostras.

Outro coeficiente de assimetria mais intuitivo é o chamado Coeficiente de assimetria de Pearson 2,

$$Sk_2 = 3[\overline{x} - md(x)]/s.$$

A avaliação de assimetria também pode ser concretizada por meios gráficos. Em particular, o gráfico de $Q(p)$ *versus* p conhecido como **gráfico de quantis** é uma ferramenta importante para esse propósito.

A Figura 2.12 mostra o gráfico de quantis para os dados do Exemplo 2.2. Notamos que os pontos correspondentes a São Paulo e Rio de Janeiro são destacados. Além disso, se a distribuição dos dados for aproximadamente simétrica, a inclinação na parte superior do gráfico deve ser aproximadamente igual àquela da parte inferior, o que não acontece na figura em questão.

Os cinco valores $x_{(1)}, Q_1, Q_2, Q_3, x_{(n)}$, isto é, os extremos e os quartis, são medidas de localização importantes para avaliarmos a simetria dos dados.

Suponha uma distribuição simétrica (ou aproximadamente simétrica). Então,

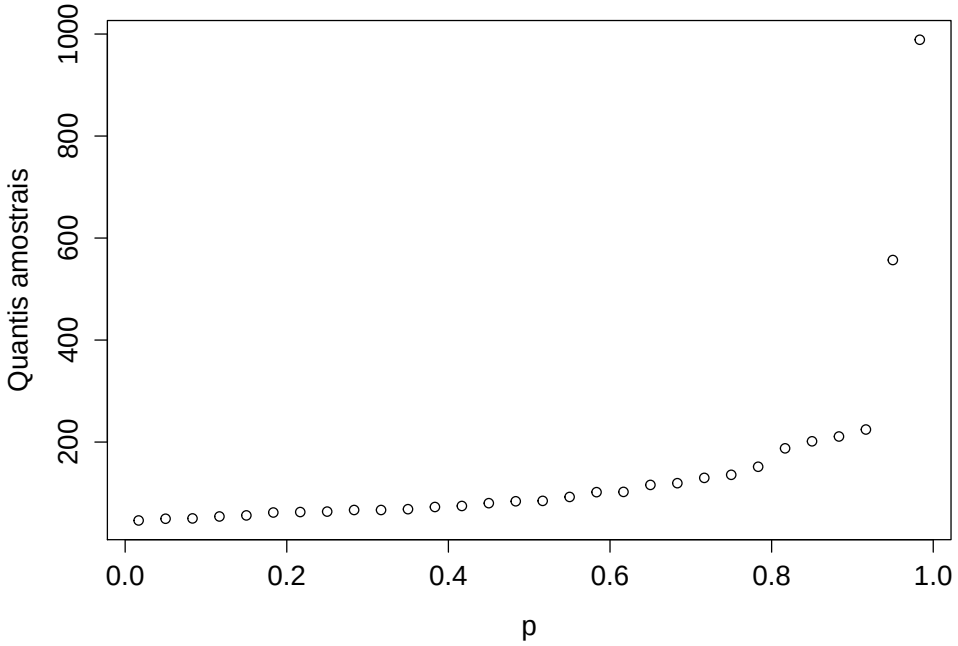
a) $Q_2 - x_{(1)} \approx x_{(n)} - Q_2$;

b) $Q_2 - Q_1 \approx Q_3 - Q_2$;

c) $Q_1 - x_{(1)} \approx x_{(n)} - Q_3$.

Para distribuições assimétricas à direita, as diferenças entre os quantis situados à direita da mediana e a mediana são maiores que as diferenças entre a mediana e os quantis situados à esquerda da mediana. A condição (a) nos diz que a **dispersão inferior** é igual (ou aproximadamente igual)

Figura 2.12: Gráfico de quantis para População (10000 habitantes)



à **dispersão superior**. Notamos, também, que se uma distribuição for (aproximadamente) simétrica, vale a relação

$$Q_2 - x_{(i)} = x_{(n+1-i)} - Q_2, \quad i = 1, \ldots, (n+1)/2, \tag{2.14}$$

em que $[x]$ representa o maior inteiro contido em x.

Chamando $u_i = Q_2 - x_{(i)},\ v_i = x_{(n+1-i)} - Q_2$, podemos considerar um **gráfico de simetria**, no qual colocamos os valores u_i como abcissas e os valores v_i como ordenadas. Se a distribuição dos dados for simétrica, os pontos (u_i, v_i) deverão estar ao longo da reta $u = v$.

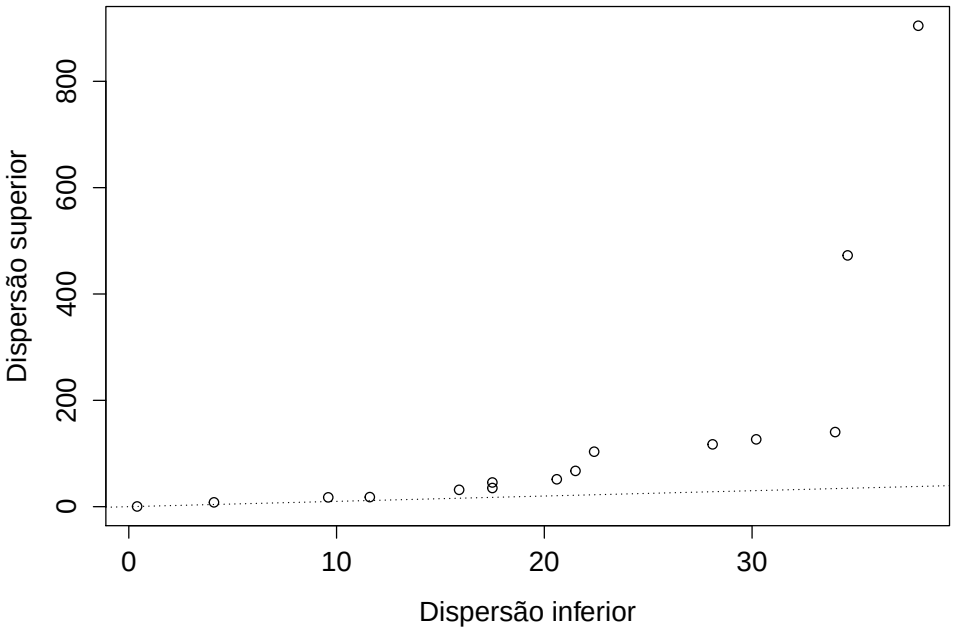
Na Figura 2.13 apresentamos o gráfico de simetria para os dados do Exemplo 2.2 no qual podemos observar que a maioria dos pontos está acima da reta $u = v$, mostrando a assimetria à direita desses dados.

2.4 *Boxplots*

O *boxplot* é um gráfico baseado nos quantis que serve como alternativa ao histograma para resumir a distribuição das dados.

Considere um retângulo, com bases determinadas por Q_1 e Q_3, como indicado na Figura 2.14. Marque com um segmento a posição da mediana. Considere dois segmentos de reta denominados bigodes (*whiskers*) colocados respectivamente acima e abaixo de Q_1 e Q_3 com limites dados, respectiva-

Figura 2.13: Gráfico de simetria para População (10000 habitantes)



mente por $\min[x_{(n)}, Q_3 + 1,5 * d_Q]$ e $\max[x_{(1)}, Q_1 - 1,5 * d_Q]$. Pontos colocados acima do limite superior ou abaixo do limite inferior são considerados **valores atípicos** ou **discrepantes** (*outliers*).

Esse gráfico permite que identifiquemos a posição dos 50% centrais dos dados (entre o primeiro e terceiro quartis), a posição da mediana, os valores atípicos, se existirem, assim como permite uma avaliação da simetria da distribuição. *Boxplots* são úteis para a comparação de vários conjuntos de dados, como veremos em capítulos seguintes.

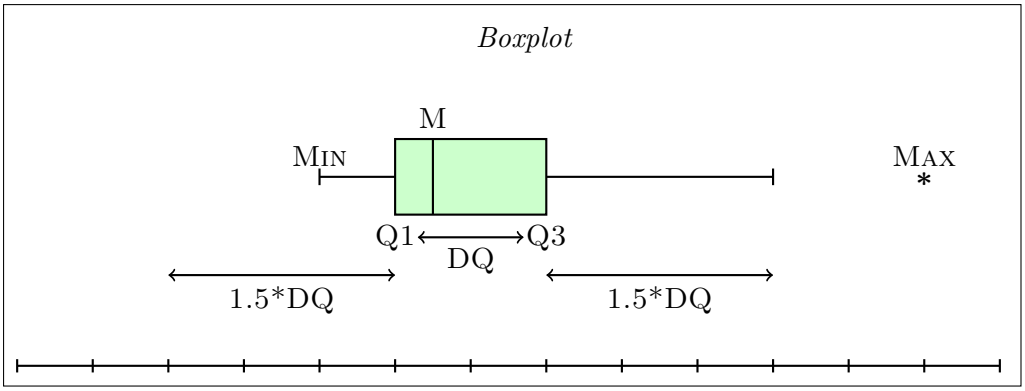
Os *boxplots* apresentados na Figura 2.15 correspondem aos dados do Exemplo 2.2 [painel (a)] e da Temperatura do Exemplo 2.3 [painel (b)]. A distribuição dos dados de Temperatura tem uma natureza mais simétrica e mais dispersa do que aquela correspondente às populações de municípios. Há valores atípicos no painel (a), representando as populações do Rio de Janeiro e de São Paulo, mas não os encontramos nos dados de temperatura.

Há uma variante dos *boxplots*, denominada ***boxplot* dentado** (*notched boxplot*) que consiste em acrescentar um dente em "v" ao redor da mediana no gráfico. O intervalo determinado pelo dente é dado por

$$Q_2 \pm \frac{1,57 d_Q}{\sqrt{n}}.$$

Para detalhes, veja McGill et al. (1978) ou Chambers et al. (1983). Na

Figura 2.14: Detalhe para a construção de *boxplots*



Q1: 1o quartil Q3: 3o quartil DQ: distância interquartis M: mediana

Figura 2.15: *Boxplots* para os dados dos Exemplos 2.2 (População) e 2.3 (Temperatura)

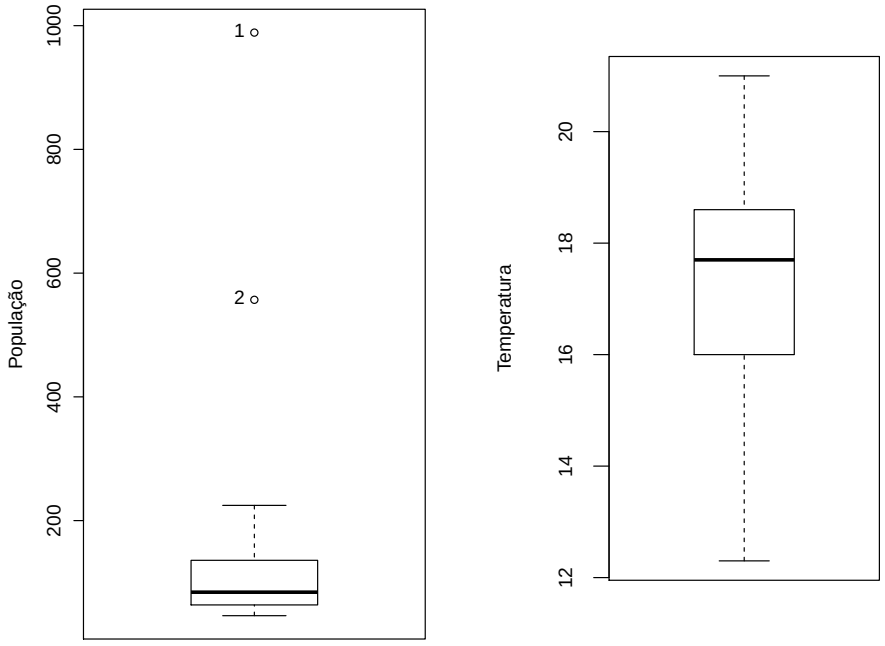
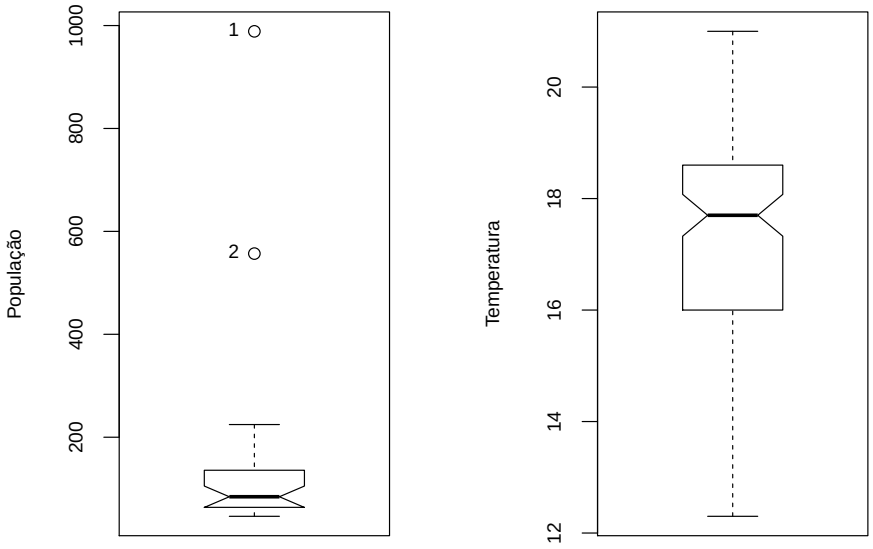


Figura 2.16 apresentamos *boxplots* correspondentes àqueles da Figura 2.15 com os dentes (*notchs*) incorporados.

2.5 Modelos probabilísticos

Um dos objetivos da Estatística é fazer inferência (ou tirar conclusões) sobre distribuições de variáveis em populações a partir de dados de uma parte dela, denominada **amostra**. A ligação entre os dados amostrais e a população

Figura 2.16: *Boxplots* dentados para os dados dos Exemplos 2.2 (População) e 2.3 (Temperatura)



depende de **modelos probabilísticos** ou seja, de modelos que representem a distribuição (desconhecida) da variável na população. Por exemplo, pode ser difícil obter informações sobre a distribuição dos salários de empregados de uma empresa com 40.000 empregados espalhados por diversos países. Nessa situação, costuma-se recorrer a uma amostra dessa população, obter as informações desejadas a partir dos valores amostrais e tentar tirar conclusões sobre toda a população a partir desses valores com base num modelo probabilístico. No exemplo acima, podemos escolher uma amostra de 400 empregados da empresa e analisar a distribuição dos salários dessa amostra. Esse procedimento é denominado **inferência estatística**.

Muitas vezes, a população para a qual se quer tirar conclusões é apenas conceitual e não pode ser efetivamente enumerada, como o conjunto de potenciais consumidores de um produto ou o conjunto de pessoas que sofrem de uma certa doença. Nesses casos, não se pode obter a correspondente distribuição de frequências de alguma característica de interesse associada a essa população e o recurso a modelos para essa distribuição faz-se necessário; esses são os chamados modelos probabilísticos e as frequências correspondentes são denominadas probabilidades. Nesse sentido, o conhecido gráfico com formato de sino associado à distribuição normal pode ser considerado como um histograma teórico. Por isso, convém chamar a média da distribuição de probabilidades (que no caso conceitual não pode ser efetivamente calculada) de **valor esperado**.

Se pudermos supor que a distribuição de probabilidades de uma variável X, definida sobre uma população possa ser representada por uma determi-

Tabela 2.8: Modelos probabilísticos para variáveis discretas

Modelo	$P(X = x)$	Parâmetros	$E(X), \mathrm{Var}(X)$
Bernoulli	$p^x(1-p)^{1-x},\ x = 0, 1$	p	$p,\ p(1-p)$
Binomial	$\binom{n}{x}p^x(1-p)^{n-x},\ x = 0, \ldots, n$	n, p	$np,\ np(1-p)$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!},\ x = 0, 1, \ldots$	λ	$\lambda,\ \lambda$
Geométrica	$p(1-p)^{x-1},\ x = 1, 2, \ldots$	p	$1/p,\ (1-p)/p^2$
Hipergeométrica	$\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}, x = 0, 1, \ldots$	N, r, n	$nr/N, n(r/N)(1-r/N)\frac{N-n}{N-1}$

nada distribuição, nosso problema reduz-se a estimar determinados parâmetros que caracterizam essa particular distribuição.

Há vários modelos probabilísticos importantes que usados em situações de interesse prático. As Tabelas 2.8 e 2.9 trazem um resumo das principais distribuições discretas e contínuas, respectivamente apresentando:

a) a **função de probabilidade** (f.p.) $p(x) = P(X = x)$, no caso discreto e a **função densidade de probabilidade** (f.d.p.), $f(x)$, no caso contínuo;
b) os parâmetros que caracterizam cada distribuição;
c) a média e a variância de cada distribuição.

Detalhes podem ser encontrados em Bussab e Morettin (2014). Para muitas dessas distribuições as probabilidades podem ser encontradas em tabelas apropriadas ou obtidas com o uso de programas de computador.

2.6 Dados amostrais

Dizer que que um conjunto de observações $x_1, \ldots, x_n$ constituem uma **amostra aleatória simples** (AAS) de tamanho n de uma variável X definida sobre uma população $\mathcal{P}$ se as variáveis $X_1, \ldots, X_n$ que geraram as observações são independentes e têm a mesma distribuição de X. Em particular, teremos que $E(X_i) = E(X), \mathrm{Var}(X_i) = \mathrm{Var}(X)$, $i = 1, \ldots, n$.

Nem sempre nossos dados representam uma AAS de uma população. Por exemplo, dados observados ao longo de um certo período de tempo são, em geral, correlacionados. Nesse caso, os dados constituem uma amostra de uma trajetória de um **processo estocástico** e a população correspondente pode

Tabela 2.9: Modelos probabilísticos para variáveis contínuas

Modelo	$f(x)$	Parâmetros	$E(X), \mathrm{Var}(X)$
Uniforme	$1/(b-a),\ a<x<b$	a, b	$(a+b)/2, (b-a)^2/12$
Exponencial	$\alpha e^{-\alpha x},\ x>0$	α	$1/\alpha,\ 1/\alpha^2$
Normal	$\frac{1}{\sigma\sqrt{2\pi}}\exp\{(\frac{x-\mu}{\sigma})^2\},\ -\infty<x<\infty$	$\mu,\ \sigma$	$\mu,\ \sigma^2$
Gama	$\frac{\alpha}{\Gamma(r)}(\alpha x)^{r-1}e^{-\alpha x},\ x>0$	$\alpha>0, r\geq 1$	$r/\alpha,\ r/\alpha^2$
Qui-quadrado	$\frac{2^{-n/2}}{\Gamma(n/2)}x^{n/2-1}e^{-x/2},\ x>0$	n	$n,\ 2n$
t-Student	$\frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}}(1+x^2/n)^{-(n+1)/2},\ -\infty<x<\infty$	n	$0,\ n/(n-2)$

ser considerada como o conjunto de todas as trajetórias de tal processo. Ver Morettin e Tolói (2006) para os conceitos necessários. Também podemos ter dados obtidos de um experimento planejado no qual uma ou mais variáveis são controladas para produzir valores de uma variável resposta. A não ser quando explicitamente indicado, para propósitos inferenciais, neste texto consideraremos os dados como provenientes de uma AAS.

Denotemos por $x_1, \ldots, x_n$ os valores efetivamente observados das variáveis $X_1, \ldots, X_n$. Denotemos por $x_{(1)}, \ldots, x_{(n)}$ esses valores observados ordenados em ordem crescente, ou seja, $x_{(1)} \leq \ldots \leq x_{(n)}$. Esses são os valores das **estatísticas de ordem** $X_{(1)}, \ldots, X_{(n)}$.

Muitas vezes não faremos distinção entre a variável e seu valor, ou seja, designaremos, indistintamente, por x a variável e um valor observado dela.

A **função distribuição acumulada** de uma variável X é definida como $F(x) = P(X \leq x)$, $x \in R$ pode ser estimada a partir dos dados amostrais, por meio da **função distribuição empírica** definida por

$$F_e(x) = \frac{n(x)}{n}, \quad \forall x \in R, \tag{2.15}$$

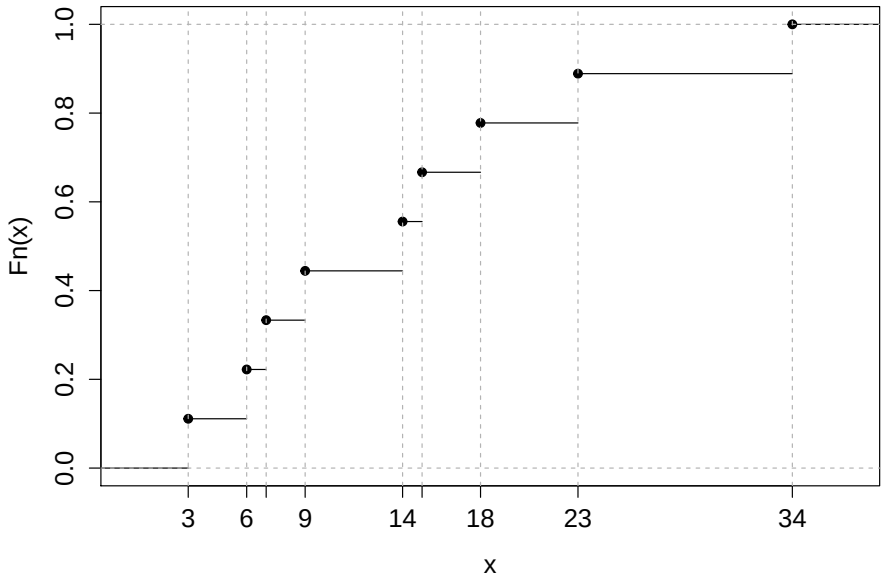
em que $n(x)$ é o número de observações amostrais menores ou iguais a x.

Considere novamente as observações do Exemplo 2.4 sem o valor 220 para efeito ilustrativo. O gráfico de F_e que é essencialmente uma função em escada, com "saltos" de magnitude $1/n$ em cada $x_{(i)}$, nomeadamente

$$F_e(x_{(i)}) = \frac{i}{n}, \quad i = 1, \ldots, n$$

está disposto na Figura 2.17.

Figura 2.17: Função distribuição empírica para os dados do Exemplo 2.4 (sem o valor 220)



2.7 Gráficos QQ

Uma das questões fundamentais na especificação de um modelo para inferência estatística é a escolha de um modelo probabilístico para representar a distribuição (desconhecida) da variável de interesse na população. Uma possível estratégia para isso é examinar o histograma dos dados amostrais e compará-lo com os histogramas teóricos associados a modelos probabilísticos candidatos. Alternativamente, os **gráficos QQ** (*QQ plots*) também podem ser utilizados com essa finalidade.

Essencialmente, gráficos QQ são gráficos cartesianos cujos pontos representam os quantis de mesma ordem obtidos das distribuições amostral e teórica. Se os dados amostrais forem compatíveis com o modelo probabilístico proposto, esses pontos devem estar sobre uma reta (com inclinação unitária se os dados forem padronizados).

Como o modelo Normal serve de base para muitos métodos estatísticos de análise, uma primeira tentativa é construir esse tipo de gráfico baseado nos quantis dessa distribuição. Os quantis Normais $Q_N(p_i)$ são obtidos da distribuição Normal padrão $[N(0,1)]$ por meio da solução da equação

$$\int_{-\infty}^{Q_N(p_i)} \frac{1}{\sqrt{2\pi}} \exp(-x^2) = p_i,\ i = 1, \ldots n,$$

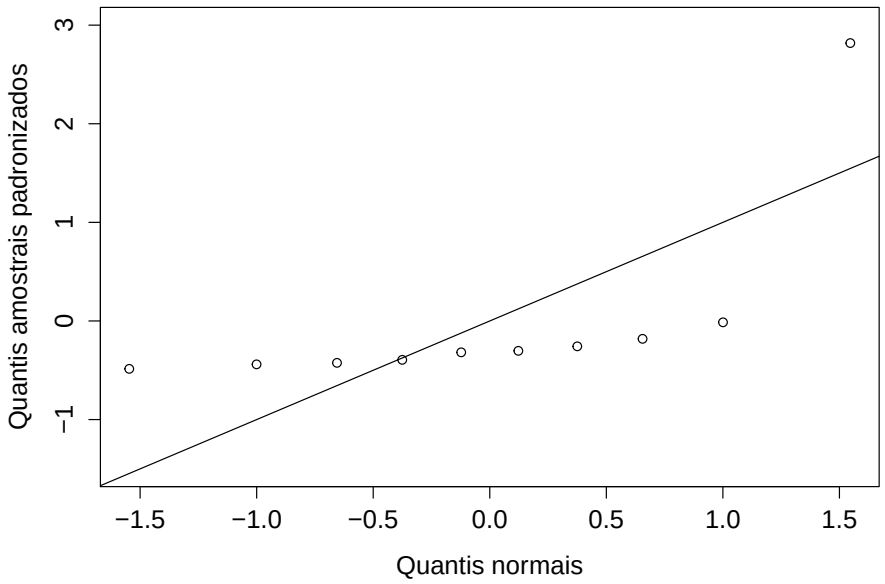
que pode ser obtida da maioria dos pacotes computacionais destinados à análise estatística. Para facilitar a comparação, convém utilizar os quantis amostrais padronizados, $Q^\star(p_i) = [Q(p_i) - (\overline{x})]/dp(x)$ nos gráficos QQ.

Consideremos novamente os dados do Exemplo 2.4. Os quantis amostrais, quantis amostrais padronizados e Normais estão dispostos na Tabela 2.7. Os correspondentes quantis Normais são exibidos na Tabela 2.10 e o correspondente gráfico QQ, na Figura 2.18.

Tabela 2.10: Quantis amostrais e Normais para os dados do Exemplo 2.4

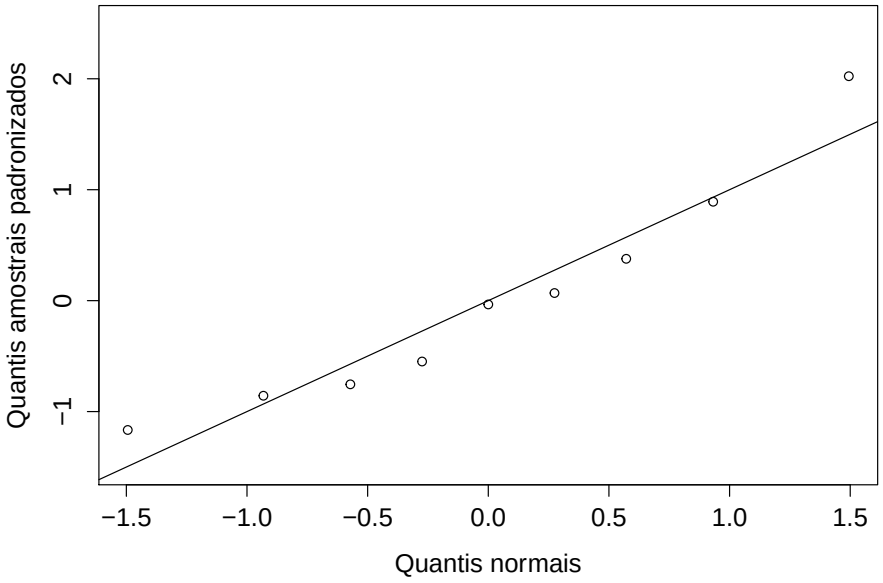
i	1	2	3	4	5	6	7	8	9	10
p_i	0,05	0,15	0,25	0,35	0,45	0,55	0,65	0,75	0,85	0.95
$Q(p_i)$	3	6	7	9	14	15	18	23	34	220
$Q^\star(p_i)$	-0,49	-0,44	-0,42	-0,39	-0,32	-0,30	-0,26	-0,18	-0,14	2,82
$Q_N(p_i)$	-1,64	-1,04	-0,67	-0,39	-0,13	0,13	0,39	0,67	1,04	1,67

Figura 2.18: Gráfico QQ Normal para os dados do Exemplo 2.4



Um exame da Figura 2.18 sugere que o modelo Normal não parece ser adequado para os dados do Exemplo 2.4. Uma das razões para isso, é a presença de um ponto atípico (220). Um gráfico QQ Normal para o conjunto de dados obtidos com a eliminação desse ponto está exibido na Figura 2.19 e indica que as evidências contrárias ao modelo Normal são menos aparentes.

Figura 2.19: Gráfico QQ Normal para os dados do Exemplo 2.4 com a eliminação do ponto 220



Um exemplo de gráfico QQ para uma distribuição amostral com 100 dados gerados a partir de uma distribuição Normal padrão está apresentado na Figura 2.20.

Embora os dados correspondentes aos quantis amostrais da Figura 2.20 tenham sido gerados a partir de uma distribuição Normal padrão, os pontos não se situam exatamente sobre a reta com inclinação de 45 graus em função de flutuações amostrais. Em geral, a adoção de um modelo probabilístico com base num exame do gráfico QQ tem uma natureza subjetiva, mas é possível incluir bandas de confiança nesse tipo de gráfico para facilitar a decisão. Essas bandas dão uma ideia sobre a faixa de variação esperada para os pontos no gráfico. Detalhes sobre a construção dessas bandas são tratados na Nota de Capítulo 4. Um exemplo de gráfico QQ com bandas de confiança para uma distribuição amostral com 100 dados gerados a partir de uma distribuição Normal padrão está apresentado na Figura 2.21.

Um exemplo de gráfico QQ em que as caudas da distribuição amostral (obtidas de uma amostra de 100 dados gerados a partir de uma distribuição t com 2 graus de liberdade) são mais pesadas que aquelas da distribuição Normal está apresentado na Figura 2.22.

Um exemplo de gráfico QQ em que a distribuição amostral (obtidas de uma amostra de 100 dados gerados a partir de uma distribuição qui-quadrado com 1 grau de liberdade) é assimétrica está apresentado na Figura 2.23.

Figura 2.20: Gráfico QQ Normal para 100 dados gerados de uma distribuição Normal padrão

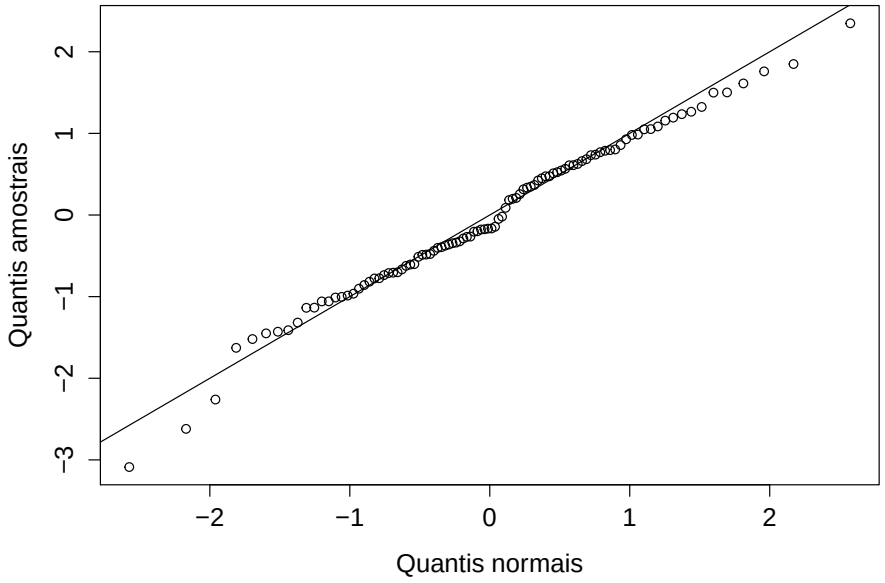


Figura 2.21: Gráfico QQ Normal para 100 dados gerados de uma distribuição Normal padrão com bandas de confiança

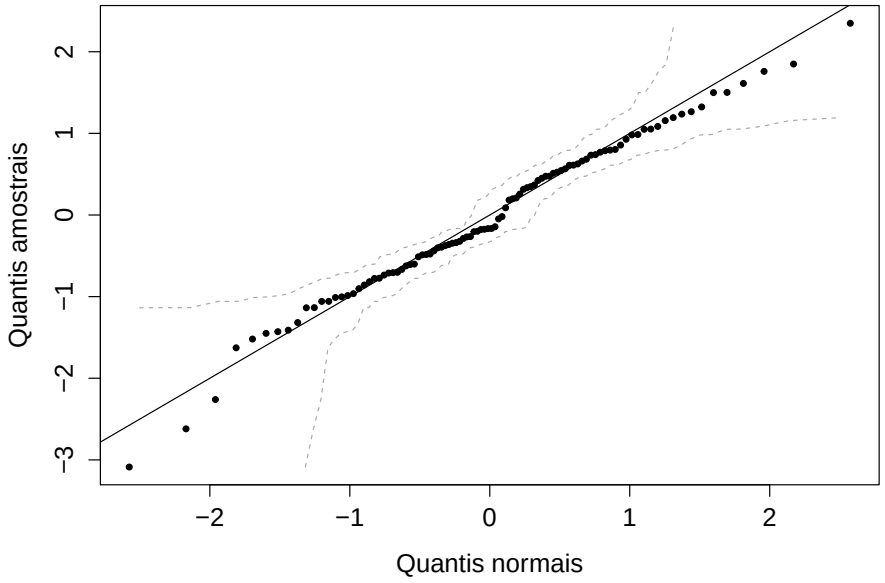


Figura 2.22: Gráfico QQ Normal para 100 dados gerados de uma distribuição t com 2 graus de liberdade

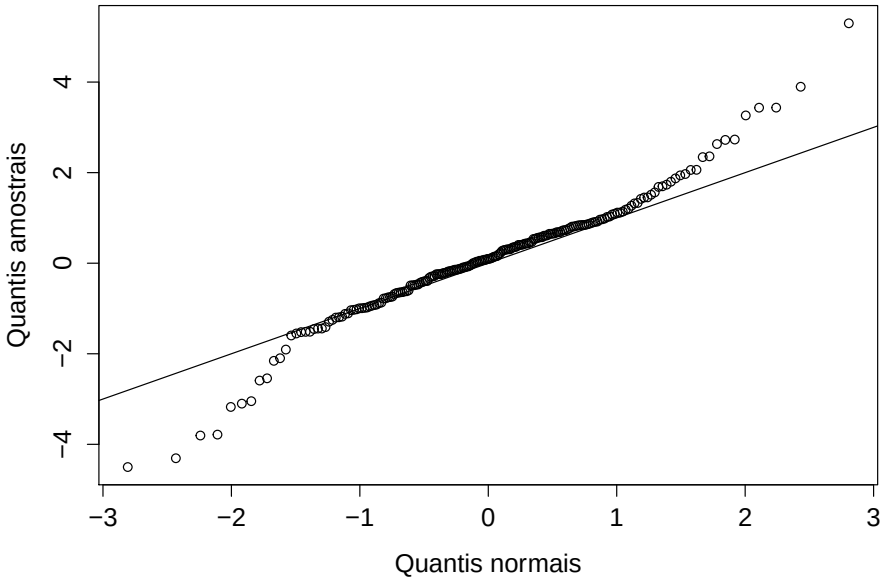
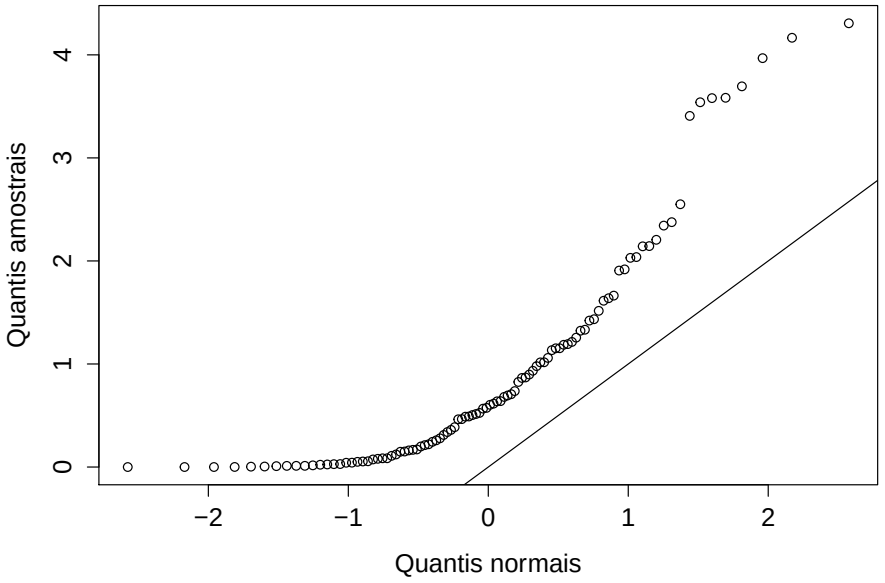


Figura 2.23: Gráfico QQ Normal para 100 dados gerados de uma distribuição qui-quadrado com 1 grau de liberdade



2.8 Transformação de variáveis

Muitos procedimentos empregados em inferência estatística são baseados na suposição de que os valores de uma (ou mais) das variáveis de interesse provêm de uma distribuição Normal, ou seja, de que os dados associados a essa variável constituem uma amostra de uma população na qual a distribuição dessa variável é Normal. No entanto, em muitas situações de interesse prático, a distribuição dos dados na amostra é assimétrica e pode conter valores atípicos, como vimos em exemplos anteriores.

Se quisermos utilizar os procedimentos talhados para análise de dados com distribuição Normal em situações nas quais a distribuição dos dados amostrais são sabidamente assimétricas, pode-se considerar uma transformação das observações com a finalidade de se obter uma distribuição "mais simétrica" e portanto, mais próxima da distribuição Normal. Uma transformação bastante usada com esse propósito é

$$y^{(p)} = \begin{cases} x^p, & \text{se } p > 0 \\ \ell n(x), & \text{se } p = 0 \\ -x^p, & \text{se } p < 0. \end{cases} \tag{2.16}$$

Essa transformação com $0 < p < 1$ é apropriada para distribuições assimétricas à direita, pois valores grandes de x decrescem mais relativamente a valores pequenos. Para distribuições assimétricas à esquerda, basta tomar $p > 1$.

Normalmente, consideramos valores de p na sequência

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

e para cada deles construímos gráficos apropriados (histogramas, *boxplots*) com os dados originais e transformados, de modo que podemos escolher o valor mais adequado de p. Hinkley (1977) sugere que para cada valor de p na sequência acima se calcule a média, mediana e um estimador de escala (desvio padrão ou algum estimador robusto) e então se escolha o valor que minimiza

$$d_p = \frac{\text{média} - \text{mediana}}{\text{medida de escala}}, \tag{2.17}$$

que pode ser vista como uma medida de assimetria; numa distribuição simétrica, $d_p = 0$.

Exemplo 2.6. Consideremos os dados do CD-? e tomemos alguns valores de p : $0, 1/4, 1/3, 1/2$. Na Figura 2.13 temos os *boxplots* para os dados transformados e na Figura 2.14 temos os respectivos histogramas. Vemos que $p = 0$ e $p = 1/3$ fornecem distribuições mais próximas de uma distribuição simétrica. Compare com as figuras 2.18(a) e 2.9.

Boxplots para os dados transformados com $p = 0, 1/4, 1/2, 1/3$

Histogramas para os dados transformados

Muitas vezes, (em Análise de Variância, por exemplo) é mais importante é transformar os dados de modo a "estabilizar" a variância do que tornar a distribuição aproximadamente Normal. Um procedimento idealizado para essa finalidade é detalhado a seguir.

Suponhamos que X seja uma variável com $E(X) = \mu$ e variância dependente da média, ou seja $\text{Var}(X) = h^2(\mu)\sigma^2$, para alguma função $h(\cdot)$. Notemos que se $h(\mu) = 1$, então $\text{Var}(X) = \sigma^2 =$ constante. Procuremos uma transformação $X \to g(X)$, de modo que $\text{Var}[g(X)] =$ constante. Com esse propósito, consideremos uma expansão de Taylor de $g(X)$ ao redor de $g(\mu)$ até primeira ordem, ou seja

$$g(X) \approx g(\mu) + (X - \mu)g'(\mu).$$

em que g' denota a derivada de g em relação a μ. Então,

$$\text{Var}[g(X)] \approx [g'(\mu)]^2\text{Var}(X) = [g'(\mu)]^2 h^2(\mu)\sigma^2.$$

Para que a variância da variável transformada seja constante, devemos tomar

$$g'(\mu) = \frac{1}{h(\mu)}.$$

Por exemplo, se o desvio padrão de X for proporcional a μ, tomamos $h(\mu) = \mu$, logo $g'(\mu) = 1/\mu$ e portanto $g(\mu) = \log(\mu)$ e devemos considerar a transformação (2.16) com $p = 0$, ou seja, $y^{(p)} = \log(x)$. Por outro lado, se a variância for proporcional à média, então usando o resultado acima, é fácil ver que a transformação adequada é $g(x) = \sqrt{x}$.

A transformação (2.19) é um caso particular das **transformações de Box-Cox** que são da forma

$$g(x) = \begin{cases} \frac{x^p - 1}{p}, & \text{se } p \neq 0 \\ \log(x), & \text{se } p = 0. \end{cases} \tag{2.18}$$

Veja Box e Cox (1964) para detalhes.

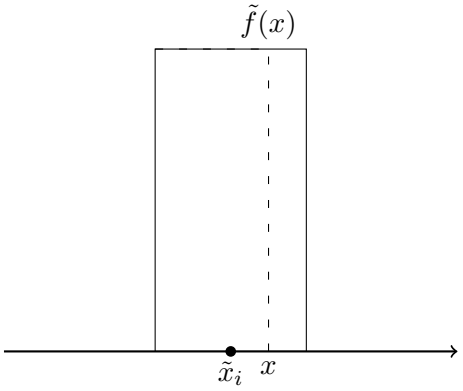
2.9 Notas de capítulo

1) Nos casos em que o histograma é obtido a partir dos dados de uma amostra de uma população com densidade $f(x)$, Freedman e Diaconis (1981) mostram que a escolha

$$h = 1,349\tilde{s}\left(\frac{\log n}{n}\right)^{1/3} \tag{2.19}$$

minimiza o desvio máximo absoluto entre o histograma e a verdadeira densidade $f(x)$. Em (2.19), $\tilde{s}$ é um estimador "robusto" do desvio padrão de X. Esse conceito será discutido adiante.

Figura 2.24: Detalhe para a construção de histogramas



2) Consideremos um exemplo com classes de comprimentos iguais a h. O número de classes a utilizar pode ser obtido aproximadamente como o quociente $(x_{(n)} - x_{(1)})/h$ em que $x_{(1)}$ é o valor mínimo e $x_{(n)}$, o valor máximo do conjunto de dados. Para que a área do histograma seja igual a 1, a altura do k-ésimo retângulo deve ser igual a f_k/h. Chamando $\tilde{x}_k$ os pontos médios dos intervalos das classes, o histograma pode ser construído a partir da seguinte função

$$\tilde{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} I(x - \tilde{x}_i; h/2), \tag{2.20}$$

em que $I(z; h)$ é a função indicadora do intervalo $[-h, h]$, ou seja,

$$I(z; h) = \begin{cases} 1, & \text{se } -h \leq z \leq h \\ 0, & \text{em caso contrário.} \end{cases}$$

Para representar essa construção, consideremos a Figura 2.24

3) Pode-se verificar que, para uma distribuição Normal,

$$d_Q = 1,349\sigma$$

com σ^2 representando a variância. Logo, um estimador do desvio padrão populacional é

$$\tilde{S} = \frac{d_Q}{1,349}.$$

Observe que nessa expressão podemos substituir $\tilde{S}$ por (2.19), de modo que

$$h \approx d_Q \left(\frac{\log n}{n} \right)^{1/3}.$$

4) Bandas de confiança para gráficos QQ.

Seja $X_1, \ldots, X_n$ uma amostra aleatória com função distribuição F desconhecida. A estatística de Kolmogorov-Smirnov [ver Wayne (1990, páginas 319-330), por exemplo], dada por

$$S = sup_x|F_n(x) - F_0(x)|$$

em que F_n é correspondente função distribuição empírica, serve para testar a hipótese $F = F_0$. A distribuição da estatística S é tabelada de forma que se pode obter o valor crítico s tal que $P(S \leq s) = 1 - \alpha$, $0 < \alpha < 1$. Isso implica que para qualquer valor x temos $|F_n(x) - F_0(x)| \leq s = 1 - \alpha$ ou seja, que com probabilidade $1 - \alpha$ temos $F_n(x) - s \leq F_0(x) \leq F_n(x) + s$. Consequentemente, os limites inferior e superior de um intervalo de confiança com coeficiente de confiança $1 - \alpha$ para F são respectivamente, $F_n(x) - s$ e $F_n(x) + s$. Essas bandas conterão a função distribuição Normal $N(\mu, \sigma^2)$ se

$$F_n(x) - s \leq \Phi[(x - \mu)/\sigma] \leq F_n(x) + s$$

o que equivale a ter uma reta contida entre os limites da banda definida por

$$\Phi^{-1}[F_n(x) - s] \leq (x - \mu)/\sigma \leq \Phi^{-1}[F_n(x) + s].$$

Para a construção do gráfico QQ esses valores são calculados nos pontos $X_1, \ldots, X_n$.

2.10 Exercícios

1. Os arquivos intitulados "rehabcardio.xls" e "rehabcardiodic.doc" disponíveis em

   `http://www.ime.usp.br/ jmsinger/MAE0217/rehabcardio.xls`

   e

   `http://www.ime.usp.br/ jmsinger/MAE0217/rehabcardio.doc`,

   respectivamente contêm informações sobre um estudo de reabilitação de pacientes cardíacos. Elabore um relatório indicando possíveis inconsistências na matriz de dados e faça uma análise descritiva de todas as variáveis do estudo.

2.

3. Calcule as medidas de posição e dispersão estudadas para os dados apresentados na Tabela 2.1.

4. Determine o valor de h dado por (2.19) para os dados do Exemplo 2.4.

5. Prove que S^2 dado por (2.9) é um estimador não-enviesado da variância populacional.

6. Considere os dados abaixo, que são medidas da velocidade do vento tomadas no aeroporto de Philadelphia (EUA), à 01:00 h, para os primeiros quinze dias de dezembro de 1974 (Graedel e Kleiner, 1985).

   $22,2 \quad 61,1 \quad 13,0 \quad 27,8 \quad 22,2 \quad 7,4 \quad 7,4 \quad 7,4 \quad 20,4 \quad 20,4 \quad 20,4 \quad 11.1$

$13,0 \quad 7,4 \quad 14,8$

Observe o valor atípico $61,1$, que na realidade ocorreu devido a forte tempestade no dia 2 de dezembro. Calcule as medidas de posição e dispersão dadas na Seção 2.3. Comente os resultados.

7. Construa gráficos ramo-e-folhas e *boxplot* para os dados do Exercício 6.
8. Usando o R, analise a variável "Temperatura" do CD-Poluicao.
9. Idem, para a variável "Salário de administradores", para o CD-Salarios.
10. Construa um gráfico ramo-e-folhas e um *boxplot* para os dados de precipitação atmosférica de Fortaleza (CD-Fortaleza).
11. Para os dados do Problema 6, faça $p = 0, 1/4, 1/3, 1/2, 3/4$ e escolha p de acordo com a medida d_p dada em (2.17).
12. Construa gráficos de quantis e simetria para os dados de manchas solares (CD-Manchas Solares).
13. Prove a relação (2.8). Como ficaria essa expressão para S^2?
14. Uma outra medida de assimetria é

$$A = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1},$$

que é igual a zero no caso de uma distribuição simétrica. Calcule A para os dados do Exercício 6.

Capítulo 3

Análise de dados de duas variáveis

3.1 Introdução

Neste capítulo trataremos da análise descritiva da associação entre duas variáveis. *Grosso modo*, dizemos que existe associação entre duas variáveis, se o conhecimento do valor de uma delas nos dá alguma informação sobre a distribuição da outra. Podemos estar interessados, por exemplo, na associação entre o grau de instrução e o salário de um conjunto de indivíduos. Nesse caso, esperamos que quanto maior seja o nível educacional de um indivíduo, maior deve ser o seu salário. Como na análise de uma única variável, também discutiremos o emprego de tabelas e gráficos para representar a distribuição conjunta das variáveis de interesse além de medidas resumo para avaliar o tipo e a magnitude da associação. Podemos destacar três casos:

i) as duas variáveis são qualitativas;
ii) as duas variáveis são quantitativas;
iii) uma variável é qualitativa e a outra é quantitativa.

As técnicas para analisar dados nos três casos acima são distintas. No primeiro caso, a análise é baseada no número de unidades de investigação (amostrais, por exemplo) em cada cela de uma tabela de dupla entrada. No segundo caso, as observações são obtidas por mensurações, e técnicas envolvendo gráficos de dispersão ou de quantis são apropriadas. Na terceira situação, podemos comparar as distribuições da variável quantitativa para cada categoria da variável qualitativa.

Aqui, é importante considerar a classificação das variáveis segundo outra característica, intimamente ligada à forma de coleta dos dados. **Variáveis explicativas** são aquelas cujas categorias ou valores são fixos, seja por planejamento ou por condicionamento. **Variáveis respostas** são aquelas cujas categorias ou valores são aleatórios.

Num estudo em que se deseja avaliar o efeito da quantidade de aditivo adicionada ao combustível no consumo de automóveis, cada um de 3 conjun-

tos de 5 automóveis (de mesmo modelo) foi observado sob o tratamento com uma de 4 quantidades de aditivo. O consumo (em km/L) foi avaliado após um determinado período de tempo. Nesse contexto, a variável qualitativa "Quantidade de aditivo" (com 4 categorias) é considerada como explicativa e a variável quantitativa "Consumo de combustível" é classificada como resposta.

Num outro cenário, em que se deseja estudar a relação entre o nível sérico de colesterol (mg/dL) e o nível de obstrução coronariana (em %), cada paciente de um conjunto de 30 selecionados de um determinado hospital foi submetido a exames de sangue e tomográfico. Nesse caso, tanto a variável "Nível sérico de colesterol" quanto a variável "Nível de obstrução coronariana" devem ser encaradas como respostas.

3.2 Duas variáveis qualitativas

Nessa situação, as classes das duas variáveis podem ser organizadas numa tabela de dupla entrada, em que as linhas correspondem aos níveis de uma das variáveis e as colunas, aos níveis da outra.

Exemplo 3.1. A planilha CD-RiscoCoronarias contém dados do projeto "Fatores de risco na doença aterosclerótica coronariana", coordenado pela Dra. Valéria Bezerra de Carvalho (INTERCOR). O arquivo contém informações sobre cerca de 70 variáveis observadas em 1500 indivíduos.

Para fins ilustrativos, consideramos apenas duas variáveis qualitativas nominais, a saber, hipertensão arterial (X) e insuficiência cardíaca (Y), ambas codificadas com os atributos 0=não tem e 1=tem observadas em 50 pacientes. Nesse contexto, as duas variáveis são classificadas como respostas. Os resultados estão dispostos na Tabela 3.1. A distribuição conjunta

Tabela 3.1: Distribuição conjunta das variáveis $X=$ hipertensão arterial e $Y=$ insuficiência cardíaca

Insuficiência cardíaca	Hipertensão arterial		Total
	Tem	Não tem	
Tem	12	4	16
Não tem	20	14	34
Total	32	18	50

das duas variáveis indica, por exemplo, que 12 indivíduos têm hipertensão arterial E insuficiência cardíaca, ao passo que 4 indivíduos não têm hipertensão E têm insuficiência. Para efeito de comparação com outros estudos envolvendo as mesmas variáveis mas com número de pacientes diferentes, convém expressar os resultados na forma de porcentagens. Com esse objetivo, podemos considerar porcentagens em relação ao total da tabela, em relação ao total das linhas ou em relação ao total das colunas. Na Tabela 3.2 apresentamos as porcentagens correspondentes à Tabela 3.1 calculadas

em relação ao seu total. Os dados da Tabela 3.2 permitem-nos concluir que

Tabela 3.2: Porcentagens para os dados da Tabela 3.1 em relação ao seu total

Insuficiência cardíaca	Hipertensão		Total
	Tem	Não tem	
Tem	24%	8%	32%
Não tem	40%	28%	68%
Total	64%	36%	100%

24% dos indivíduos avaliados têm hipertensão E insuficiência cardíaca, ao passo que 36% dos indivíduos avaliados não sofrem de hipertensão.

Também podemos considerar porcentagens calculadas em relação ao total das colunas como indicado na Tabela 3.3. Com base nessa tabela, po-

Tabela 3.3: Porcentagens com totais nas colunas

Insuficiência cardíaca	Hipertensão		Total
	Tem	Não tem	
Tem	$37,5\%$	$22,2\%$	32%
Não tem	$62,5\%$	$77,8\%$	68%
Total	$100,0\%$	$100,0\%$	$100,0\%$

demos dizer que independentemente do *status* desses indivíduos quanto à presença de hipertensão, 32% têm insuficiência cardíaca. Esse cálculo de porcentagens é mais apropriado quando uma das variáveis é considerada explicativa e a outra, considerada resposta.

No exemplo, apesar de o planejamento do estudo indicar que as duas variáveis são respostas (a frequência de cada uma delas não foi fixada *a priori*), para efeito da análise, uma delas (Hipertensão arterial) será considerada explicativa. Isso significa que não temos interesse na distribuição de frequências de hipertensos ou não dentre os 100 pacientes avaliados apesar de ainda querermos avaliar a associação entre as duas variáveis. Nesse caso, dizemos que a variável "Hipertensão arterial" é considerada explicativa **por condicionamento**. Se houvéssemos fixado *a priori* um certo número de hipertensos e outro de não hipertensos e então observado quantos dentre cada um desses dois grupos tinham ou não insuficiência cardíaca, diríamos que a variável "Hipertensão arterial" seria considerada explicativa **por planejamento**. Nesse caso, apenas as porcentagens calculadas como na Tabela 3.3 fariam sentido. Uma enfoque análogo poderia ser adotado se fixássemos as frequências de "Insuficiência cardíaca" e considerássemos "Hipertensão arterial" como variável resposta. Nesse caso, as porcentagens deveriam ser calculadas em relação ao total das linhas da tabela.

Tabelas com a natureza daquelas descritas acima são chamadas de **tabelas de contingência** ou **tabelas de dupla entrada**.. Essas tabelas são classificadas como tabelas $r \times c$ em que r é o número de linhas e c é o

número de colunas. As tabelas apresentadas acima são, portanto, tabelas 2×2. Se a variável X tiver 3 categorias e a variável Y, 4 categorias, a tabela de contingência correspondente será uma tabela 3×4.

Suponha, agora, que queiramos verificar se as variáveis X e Y são associadas. No caso da Tabela 3.2 (em que as duas variáveis são consideradas respostas), dizer que as variáveis não são associadas corresponde a dizer que essas variáveis são (estatisticamente) independentes. No caso da Tabela 3.3 (em que uma variáveis é explicativa, quer por condicionamento, quer por planejamento e a outra é considerada resposta), dizer que as variáveis não são associadas corresponde a dizer que as distribuições de frequências da variável resposta ("Insuficiência cardíaca") para indivíduos classificados em cada categoria da variável explicativa ("Hipertensão arterial") são homogêneas.

Nas Tabelas 3.2 ou 3.3, por exemplo, há diferenças, que parecem não ser "muito grandes", o que nos leva a conjecturar que **para a população de onde esses indivíduos foram extraídos**, as duas variáveis não são associadas. Para avaliar essa conjectura, pode-se construir um **teste formal** para essa **hipótese de inexistência de associação** (independência ou homogeneidade), nomeadamente

$$H \ : \ X \text{ e } Y \ \text{ são não associadas.}$$

Convém sempre lembrar que a hipótese H refere-se à associação entre as variáveis X e Y na população (geralmente conceitual) de onde foi extraída uma amostra cujos dados estão dispostos na tabela. Não há dúvidas de que na tabela, as distribuições de frequências correspondentes às colunas rotuladas por "Tem" e "Não tem" hipertensão são diferentes.

Se as duas variáveis não fossem associadas, deveríamos porcentagens iguais nas colunas da Tabela 3.3 rotuladas "Tem" e "Não tem". Podemos então calcular as frequências esperadas nas celas da tabela admitindo que a hipótese H seja verdadeira. Por exemplo, o valor $10,2$ corresponde a 32% de 32, ou ainda, $10,2 = (32 \times 16)/50$. Observe que os valores foram arredondados segundo a regra usual e que as somas de linhas e colunas são as mesmas da Tabela 3.1.

Tabela 3.4: Valores esperados das frequências na Tabela 3.3 sob H

Insuficiência	Hipertensão		Total
Cardíaca	Tem	Não Tem	
Tem	10,2	5,8	16
Não Tem	21,8	112,2	34
Total	32	18	50

Chamando os valores observados por o_i e os esperados por e_i, $\quad i = 1, 2, 3, 4$, podemos calcular os **resíduos** $r_i = o_i - e_i$ e verificar que $\sum_i r_i = 0$. Uma medida da discrepância entre o valores observados e aqueles esperados

sob a hipótese H é a chamada estatística ou **qui-quadrado** de Pearson, dada por

$$\chi^2 = \sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i}. \tag{3.1}$$

No nosso exemplo, $\chi^2 = 1,3$. Quanto maior esse valor, maior a **evidência** de que a hipótese H não é verdadeira, ou seja de que as variáveis X e Y são associadas (na população de onde foi extraída a amostra que serviu de base para os cálculos). Resta saber se o valor observado é suficientemente grande para concluirmos que H não é verdadeira. Com essa finalidade, teríamos que fazer um teste formal, que será visto no Capítulo??.

A própria estatística de Pearson poderia servir como medida da intensidade da associação mas o seu valor aumenta com o tamanho da amostra; uma alternativa para corrigir esse problema é o **coeficiente de contingência de Pearson**, dado por

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}. \tag{3.2}$$

Para o Exemplo 3.1, temos que $C = \sqrt{1,3/(1,3+50)} = 0,16$, que é um valor pequeno. Esse coeficiente tem interpretação semelhante à do coeficiente de correlação, a ser tratado na próxima seção. Mas enquanto esse último varia entre -1 e $+1$, C, como definido acima, não varia entre 0 e 1 (em módulo). O valor máximo de C depende do número de linhas, r, e do número de colunas, c, da tabela de contingência. Uma modificação de C é o coeficiente

$$T = \sqrt{\frac{\chi^2/n}{(r-1)(c-1)}}, \tag{3.3}$$

que atinge o valor máximo igual a 1 quando $r = c$. No Exemplo 3.1, $T = 0,16$.

Em estudos que envolvem a mesma característica observada sob duas condições diferentes (gerando duas variáveis, X e Y, cada uma correspondendo à observação da característica sob uma das condições), sabe-se *a priori* que elas são associadas e o interesse recai sobre a avaliação da concordância dos resultados em ambas as condições. Nesse contexto, consideremos um exemplo em que as redações de 445 alunos são classificadas por cada um de dois professores (A e B) como "ruim", "média" ou "boa" com os resultados resumidos na Tabela 3.5.

Tabela 3.5: Frequências de redações classificadas por dois professores

	Professor B		
Professor A	ruim	média	boa
ruim	192	1	5
média	2	146	5
boa	11	12	71

Se todos as frequências estivessem dispostas ao longo da diagonal principal da tabela, diríamos que a haveria completa concordância entre os dois professores com relação ao critério de avaliação das redações. Como em geral isso não acontece, é conveniente construir um índice para avaliar a magnitude da concordância. O índice

$$\kappa = \frac{\sum_{i=1}^{3} p_{ii} - \sum_{i=1}^{3} p_{i+}p_{+i}}{1 - \sum_{i=1}^{3} p_{i+}p_{+i}},$$

denominado κ **de Cohen** (1960) é o mais utilizado com esse propósito. Nessa expressão, p_{ij} representa frequência relativa associada à cela correspondente à linha i e coluna j da tabela e p_{i+} e p_{+j} representam a soma das frequências relativas associadas à linha i e coluna j, respectivamente. O numerador corresponde à diferença entre a soma das frequências relativas correspondentes à diagonal principal da tabela e a soma das frequências relativas que seriam esperadas se as avaliações dos dois professores fossem independentes. Portanto, quando há concordância completa, $\sum_{i=1}^{3} p_{ii} = 1$, o numerador é igual ao denominador e o valor do índice de Cohen é $\kappa = 1$. Quando os dois professores não concordam em nenhuma das avaliações, $\kappa < 0$. Para os dados da Tabela 3.5 temos $\kappa = 0.87$ sugerindo uma "boa" concordância entre as avaliações dos dois professores. Embora o nível de concordância medido pelo índice κ seja subjetivo e dependa da área em que se realiza o estudo gerador dos dados, há autores que sugerem modelos de classificação, como aquele proposto por Viera and Garrett (2005) e reproduzido na Tabela 3.6

Tabela 3.6: Níveis de concordância segundo o índice κ de Cohen

κ de Cohen	Nível de concordância
< 0	Menor do que por acaso
0.01–0.20	Leve
0.21– 0.40	Razoável
0.41–0.60	Moderado
0.61–0.80	Substancial
0.81–0.99	Quase perfeito

Para salientar discordâncias mais extremas como no exemplo, um professor classifica a redação como "ruim" e o outro como "boa", pode-se considerar o índice κ ponderado, definido como

$$\kappa_p = \frac{\sum_{i=1}^{3}\sum_{j=1}^{3} w_{ij}p_{ij} - \sum_{i=1}^{3}\sum_{j=1}^{3} w_{ij}p_{i+}p_{+j}}{1 - \sum_{i=1}^{3}\sum_{j=1}^{3} w_{ij}p_{i+}p_{+j}},$$

em que $w_{ij}, i, j = 1, 2, 3$ é um conjunto de pesos convenientes. Por exemplo, $w_{ii} = 1$, $w_{ij} = 1 - (i - j)/(I - 1)$ em que I é o número de categorias em que a característica de interesse é classificada. Para o exemplo, $w_{12} = w_{21} = w_{23} = w_{32} = 1 - 1/2 = 1/2$, $w_{13} = w_{31} = 1 - 2/2 = 0$.

Exemplo 3.2. (Exemplo em outra área)

Em muitas áreas do conhecimento há interesse em avaliar a associação entre um ou mais **fatores de risco** e uma variável resposta. Num estudo epidemiológico, por exemplo, pode haver interesse em avaliar a associação entre o hábito tabagista (fator de risco) e a ocorrência de algum tipo de câncer pulmonar (variável resposta). Um exemplo na área de Seguros pode envolver a avaliação da associação entre estado civil e sexo (considerados como fatores de risco) e o envolvimento em acidente automobilístico (variável resposta).

No primeiro caso, os dados obtidos de uma amostra de 50 fumantes e 100 não fumantes, por exemplo, para os quais se observa a ocorrência de câncer pulmonar após um determinado período podem ser dispostos no formato da Tabela 3.7. Esse tipo de estudo em que se fixam os níveis do fator de risco (hábito tabagista) e se observa a ocorrência do evento de interesse (câncer pulmonar) após um determinado tempo é conhecido como **estudo prospectivo**.

Tabela 3.7: Frequências de doentes observados num estudo prospectivo

Hábito	Câncer pulmonar		Total
tabagista	sem	com	
não fumante	80	20	100
fumante	35	15	50

Para a população da qual essa amostra é considerada oriunda (e para a qual se quer fazer inferência), a tabela correspondente pode ser expressa como na Tabela 3.8.

Tabela 3.8: Probabilidades de ocorrência de doença

Hábito	Câncer pulmonar		Total
tabagista	sem	com	
não fumante	$1 - \pi_0$	π_0	1
fumante	$1 - \pi_1$	π_1	1

O parâmetro π_0 corresponde à proporção (ou probabilidade) de indivíduos que contraem câncer pulmonar dentre os que SABEMOS ser não fumantes; analogamente, π_1 corresponde à proporção (ou probabilidade) de indivíduos que contraem câncer pulmonar dentre os que SABEMOS ser não fumantes.

Nesse contexto podemos definir algumas medidas de associação (entre o fator de risco e a variável resposta).

i) **Risco atribuível**: $d = \pi_1 - \pi_0$ que corresponde à diferença entre as probabilidades de ocorrência do evento de interesse para expostos e não expostos ao fator de risco.

ii) **Risco relativo**: $r = \pi_1/\pi_0$ que corresponde ao quociente entre as probabilidades de ocorrência do evento de interesse para expostos e não expostos ao fator de risco.

iii) **Razão de chances** (*Odds ratio*)[1]: $\omega = [\pi_1/(1-\pi_1)]/[\pi_0/(1-\pi_0)]$ que corresponde ao quociente entre as chances de ocorrência do evento de interesse para expostos e não expostos ao fator de risco.

No exemplo da Tabela 3.7 essas medidas de associação podem ser estimadas como

i) Risco atribuível: $d = 0,30 - 0,20 = 0,10$ (o risco de ocorrência de câncer pulmonar aumenta de 10% para fumantes relativamente aos não fumantes)

ii) Risco relativo: $r = 0,30/0,20 = 1,50$ (o risco de ocorrência de câncer pulmonar para fumantes é $1,5$ vezes o risco correspondente para não fumantes)

iii) Chances: a chance de ocorrência de câncer pulmonar para fumantes é $0,429 = 0,30/0,70$; a chance de ocorrência de câncer pulmonar para não fumantes é $0,250 = 0,20/0,80$

iiv) Razão de chances: $\omega = 0,429/0,250 = 1,72$ (a chance de ocorrência de câncer pulmonar para fumantes é $1,71$ vezes a chance correspondente para não fumantes).

Em geral, a medida de associação de maior interesse prático pela facilidade de interpretação, seja o risco relativo, a razão de chances talvez seja a mais utilizada na prática. Primeiramente, observemos que

$$\omega = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = r\frac{1-\pi_0}{1-\pi_1} \longrightarrow r, \text{ quando } \pi_0 \text{ e } \pi_1 \longrightarrow 0$$

ou seja, para eventos raros [cujas probabilidade π_1 ou π_0 são muito pequenas], a razão de chances serve como uma boa aproximação do risco relativo.

Em geral, estudos prospectivos com a natureza daquele que motivou a discussão acima não são praticamente viáveis em função do tempo decorrido até o diagnóstico da doença. Uma alternativa é a condução de **estudos**

[1]Lembremos que **probabilidade** é uma medida de frequência de ocorrência de um evento (quanto maior a probabilidade de um evento, maior a frequência com que ele ocorre) cujos valores variam entre 0 e 1 (ou entre 0% e 100%). Uma medida de frequência equivalente mas com valores entre 0 e ∞ é conhecida como **chance** (*odds*). Por exemplo, se um evento ocorre com probabilidade 0.8 (80%), a chance de ocorrência é 4 (= 80% / 20%) ou mais comumente de 4 para 1, indicando que em cinco casos, o evento ocorre em 4 e não ocorre em 1

retrospectivos em que, por exemplo, são selecionados 35 pacientes com e 115 pacientes sem câncer pulmonar e se determinam quais dentre eles eram fumantes e não fumantes. Nesse caso, os papéis das variáveis explicativa e resposta se invertem, sendo o *status* relativo à presença da moléstia encarado como variável explicativa e o hábito tabagista, como variável resposta. A Tabela 3.9 contém dados hipotéticos de um estudo retrospectivo planejado com o mesmo intuito do estudo prospectivo descrito acima, ou seja, avaliar a associação entre tabagismo e ocorrência de câncer de pulmão.

Tabela 3.9: Frequências de fumantes observados num estudo retrospectivo

Hábito	Câncer pulmonar	
tabagista	sem	com
não fumante	80	20
fumante	35	15
Total	115	35

A Tabela 3.10 representa as probabilidades pertinentes.

Tabela 3.10: Probabilidades de hábito tabagista

Hábito	Câncer pulmonar	
tabagista	sem	com
não fumante	$1-p_0$	$1-p_1$
fumante	p_0	p_1
Total	1	1

O parâmetro p_0 corresponde à proporção (ou probabilidade) de fumantes DENTRE os indivíduos que SABEMOS não ter câncer pulmonar; analogamente, p_1 corresponde à proporção (ou probabilidade) de não fumantes DENTRE os indivíduos que SABEMOS ter câncer pulmonar. Nesse caso, não é possível calcular nem o risco atribuível nem o risco relativo, pois não se conseguem estimar as probabilidades de ocorrência de câncer pulmonar, π_1 ou π_0. No entanto, pode-se demonstrar (ver Notas de Capítulo) que a razão de chances obtida por meio de um estudo retrospectivo é igual àquela que seria obtida por intermédio de um estudo prospectivo correspondente ou seja

$$\omega = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}.$$

Num estudo retrospectivo, pode-se afirmar que a chance de ocorrência do evento de interesse (câncer pulmonar, por exemplo) para indivíduos expostos ao fator de risco é ω vezes a chance correspondente para indivíduos não expostos, embora não se possa estimar quanto valem essas chances. A partir das frequências da Tabela 3.9 podemos estimar a chance de um indivíduo ser fumante dado que tem câncer pulmonar como $0,751 = 0,429/0,571$

e a chance de um indivíduo ser fumante dado que não tem câncer pulmonar como $0,437 = 0,304/0,696$; a razão de chances correspondente é $\omega = 0,751/0,437 = 1,72$. Essas chances não são aquelas de interesse pois gostaríamos de conhecer as chances de ter câncer pulmonar para indivíduos fumantes e não fumantes. No entanto a razão de chances tem o mesmo valor que aquela calculada por meio de um estudo prospectivo, ou seja, a partir da análise dos dados da Tabela 3.9, não é possível calcular a chance de ocorrência de câncer pulmonar nem para fumantes nem para não fumantes mas podemos concluir que a primeira é $1,72$ vezes a segunda.

Dados provenientes de estudos planejados com o objetivo de avaliar a capacidade de testes laboratoriais ou exames médicos para diagnóstico de alguma doença envolvem a classificação de indivíduos segundo duas variáveis; a primeira corresponde ao verdadeiro *status* relativamente à presença da moléstia (doente ou não doente) e a segunda ao resultado do teste (positivo ou negativo). Dados correspondentes aos resultados de um determinado teste aplicado a n indivíduos podem ser dispostos no formato da Tabela 3.11.

Tabela 3.11: Frequência de pacientes submetidos a um teste diagnóstico

Verdadeiro	Resultado do teste		Total
status	positivo (T+)	negativo (T-)	
doente (D)	n_{11}	n_{12}	n_{1+}
não doente (ND)	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Aqui, n_{ij} corresponde à frequência da indivíduos com o i-ésimo *status* relativo à doença ($i = 1$ para doentes e $i = 2$ para não doentes) e j-ésimo *status* relativo ao resultado do teste ($j = 1$ para resultado positivo e $j = 2$ para resultado negativo). Além disso, $n_{i+} = n_{i1} + n_{i2}$ e $n_{+j} = n_{1j} + n_{2j}$, $i, j = 1, 2$. As seguintes características associadas aos testes diagnóstico são bastante utilizadas na prática.

i) **Sensibilidade**: corresponde à probabilidade de resultado positivo para pacientes doentes $[S = P(T + |D)]$ e pode ser estimada por $s = n_{11}/n_{1+}$;

ii) **Especificidade**: corresponde à probabilidade de resultado negativo para pacientes não doentes $[E = P(T - |ND)]$ e pode ser estimada por $e = n_{22}/n_{2+}$;

iii) **Falso positivo**: corresponde à probabilidade de resultado positivo para pacientes não doentes $[FP = P(ND|T+)]$ e pode ser estimada por $fp = n_{21}/n_{+1}$;

iv) **Falso negativo**: corresponde à probabilidade de resultado negativo para pacientes doentes $[FN = P(D|T-)]$ e pode ser estimada por $fn = n_{12}/n_{+2}$;

v) **Valor preditivo positivo**: corresponde à probabilidade de que o paciente seja doente dado que o resultado do teste é positivo [$VPP = P(D|T+)$] e pode ser estimada por $vpp = n_{11}/n_{+1}$;

vi) **Valor preditivo negativo**: corresponde à probabilidade de que o paciente não seja doente dado que o resultado do teste é negativo [$VPN = P(ND|T-)$] e pode ser estimada por $vpn = n_{22}/n_{+2}$;

vii) **Acurácia**: corresponde à probabilidade de resultados corretos [$AC = P\{(D \cap T+) \cup (ND \cap T-)\}$] e pode ser estimada por $ac = (n_{11}+n_{22})/n$.

A sensibilidade de um teste corresponde à proporção de doentes identificados por seu intermédio, ou seja, é um indicativo da capacidade de o teste detectar a doença. Por outro lado, a especificidade de um teste corresponde à sua capacidade de identificar indivíduos que não têm a doença.

Quanto maior a sensibilidade de um teste, menor é a possibilidade de que indique falsos positivos. Um teste com sensibilidade de 95%, por exemplo, consegue identificar um grande número de pacientes que realmente têm a doença e por esse motivo testes com alta sensibilidade são utilizados em triagens. Quanto maior a especificidade de um teste, maior é a probabilidade de apresentar um resultado negativo para pacientes que não têm a doença. Se, por exemplo, a especificidade de um teste for de 99% dificilmente um paciente que não tem a doença terá um resultado positivo. Um bom teste é aquele que apresenta alta sensibilidade e alta especificidade, mas nem sempre isso é possível.

O valor preditivo positivo indica a probabilidade de um indivíduo ter a doença dado que o resultado do teste é positivo e valor preditivo negativo indica a probabilidade de um indivíduo não ter a doença dado um resultado negativo no teste.

Sensibilidade e especificidade são características do teste, mas tanto o valor preditivo positivo quanto o valor preditivo negativo dependem da **prevalência** (porcentagem de indivíduos doentes) da doença. Consideremos um exemplo em que o mesmo teste diagnóstico é aplicado em duas comunidades com diferentes prevalências de uma determinada doença. A Tabela 3.12 contém os dados da comunidade em que a doença é menos prevalente e a Tabela 3.13 contém os dados da comunidade em que a doença é mais prevalente.

Tabela 3.12: Frequência de pacientes submetidos a um teste diagnóstico (prevalência da doença = 15%)

Verdadeiro	Resultado do teste		Total
status	positivo (T+)	negativo (T-)	
doente (D)	20	10	30
não doente (ND)	80	90	170
Total	100	100	200

Tabela 3.13: Frequência de pacientes submetidos a um teste diagnóstico (prevalência da doença = 30%)

Verdadeiro	Resultado do teste		Total
status	positivo (T+)	negativo (T-)	
doente (D)	40	20	60
não doente (ND)	66	74	140
Total	106	94	200

Os valores estimados para a sensibilidade, especificidade, valores preditivo positivo e negativo além da acurácia estão dispostos na Tabela 3.14

Tabela 3.14: Características do teste aplicado aos dados das Tabelas 3.12 e 3.13

	População com doença	
Característica	menos prevalente	mais prevalente
Sensibilidade	67%	67%
Especifidade	53%	53%
VPP	20%	38%
VPN	90%	79%
Acurácia	55%	55%

3.3 Duas Variáveis Quantitativas

Uma das principais ferramentas para avaliar a associação entre duas variáveis quantitativas é o **gráfico de dispersão**. Consideremos um conjunto de n pares de valores (x_i, y_i) de duas variáveis X e Y, o gráfico de dispersão correspondente é um gráfico cartesiano em que os valores de uma das variáveis são colocados no eixo das abscissas e os da outra, no eixo das ordenadas.

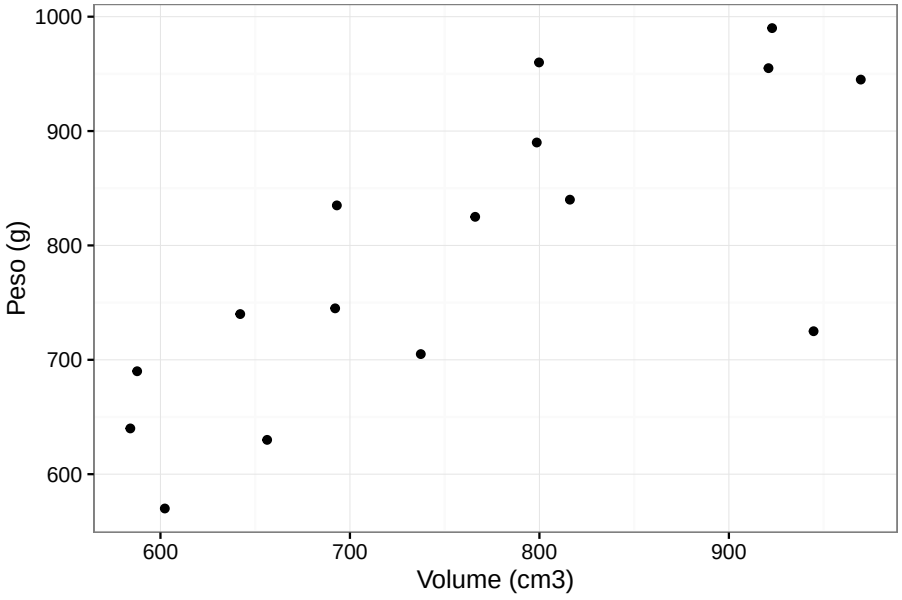
Exemplo 3.3 Os dados contidos na Tabela 3.15 correspondem a um estudo cujo objetivo principal era avaliar a associação entre o volume (cm^3) do lobo direito de fígados humanos medido ultrassonograficamente e o seu peso (g). Um objetivo secundário era avaliar a concordância de medidas ultrassonográficas do volume (Volume1 e Volume2) realizadas por dois observadores. O volume foi obtido por meio da média das duas medidas ultrassonográficas. Detalhes podem ser obtidos em Zan (2005).

O gráfico de dispersão correspondente às variáveis Volume e Peso está apresentado na Figura 3.1. Nesse gráfico pode-se notar que a valores menores para o volume correspondem valores maiores para o peso, sugerindo uma associação positiva e possivelmente linear entre as duas variáveis. Além disso, o gráfico permite identificar um possível ponto discrepante (*outlier*) correspondente à unidade amostral em que o volume é $944,7cm^3$ e o peso é $725g$. A utilização desses resultados para a construção de um modelo que

Tabela 3.15: Peso e volume do lobo direito de enxertos de fígado

Volume1 (cm^3)	Volume2 (cm^3)	Volume (cm^3)	Peso (g)
672.3	640.4	656.3	630
686.6	697.8	692.2	745
583.1	592.4	587.7	690
850.1	747.1	798.6	890
729.2	803.0	766.1	825
776.3	823.3	799.8	960
715.1	671.1	693.1	835
634.5	570.2	602.3	570
773.8	701.0	737.4	705
928.3	913.6	920.9	955
916.1	929.5	922.8	990
983.2	906.2	944.7	725
750.5	881.7	816.1	840
571.3	596.9	584.1	640
646.8	637.4	642.1	740
1021.6	917.5	969.6	945

Figura 3.1: Gráfico de dispersão entre peso e volume do lobo direito de enxertos de fígado



permita estimar o peso como função do volume é o objeto da técnica conhecida como Análise de Regressão que será considerada no Capítulo XXXX.

Dado um conjunto de n pares (x_i, y_i), a associação (linear) entre as variáveis quantitativas X e Y pode ser quantificada por meio do **coeficiente de correlação (linear)** de Pearson, definido por

$$r_P = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{[\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2]^{1/2}}. \tag{3.4}$$

Pode-se mostrar que $-1 \leq r_P \leq 1$ e, na prática, se o valor r_P estiver próximo de -1 ou $+1$, pode-se dizer que as variáveis são fortemente associadas ou linearmente correlacionadas; por outro lado, se o valor de r_P estiver próximo de zero, dizemos que as variáveis são não correlacionadas. Quanto mais próximos de uma reta estiverem os pontos (x_i, y_i), maior será a intensidade da correlação (linear) entre elas.

Não é difícil mostrar que

$$r_P = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\left[\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right)\right]^{1/2}}. \tag{3.5}$$

Essa expressão é mais conveniente que 3.4, pois basta calcular: (a) as médias amostrais $\overline{x}$ e $\overline{y}$; (b) a soma dos produtos $x_i y_i$ e (c) a soma dos quadrados dos x_i e a soma dos quadrados dos y_i.

Para os dados do Exemplo 3.3, o coeficiente de correlação de Pearson é $0,76$. Se excluirmos o dado discrepante identificado no gráfico de dispersão, o valor do coeficiente de correlação de Pearson é $0,89$, evidenciando a falta de robustez desse coeficiente relativamente a observações com essa natureza. Nesse contexto, uma medida de associação mais robusta é o coeficiente de correlação de Spearman, cuja expressão é similar à (3.4) com os valores das variáveis X e Y substituídos pelos respectivos postos.[2] Mais especificamente, o coeficiente de correlação de Spearman é

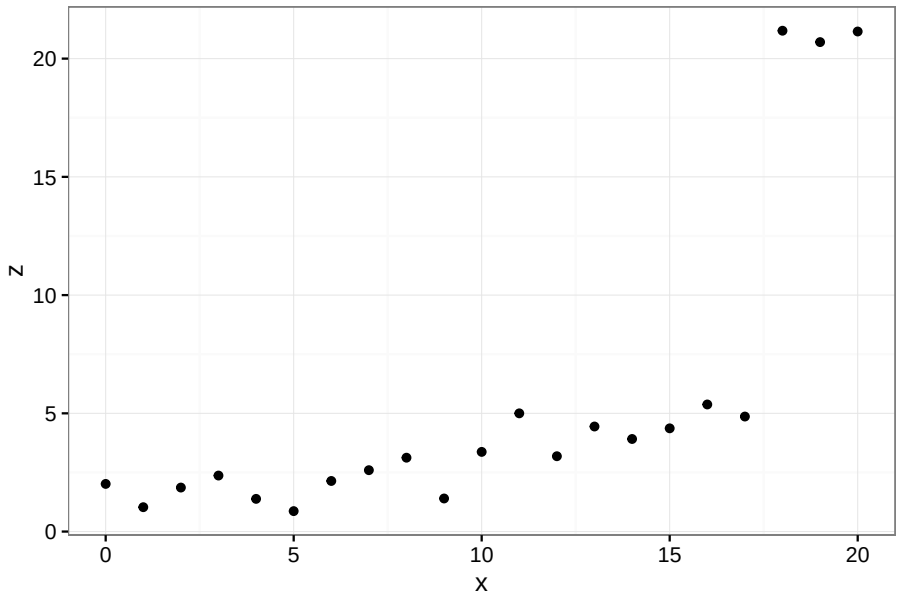
$$r_S = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{[\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})^2]^{1/2}}, \tag{3.6}$$

em que R_i corresponde ao posto da i-ésima observação da variável X entre seus valores e $\overline{R}$ à média desses postos e S_i e $\overline{S}$ têm interpretação similar para a variável Y. Para efeito de cálculo pode-se mostrar que a expressão (3.6) é equivalente a

$$r_S = 6\sum_{i=1}^{n}(R_i - S_i)^2/[n(n^2-1)]. \tag{3.7}$$

[2] O posto de uma observação x_i é o índice correspondente à sua posição no conjunto ordenado $x_1 \leq x_2 \leq \ldots \leq x_n$. Por exemplo, dado o conjunto de observações $x_1 = 4$, $x_2 = 7$, $x_3 = 5$, $x_4 = 13$, $x_5 = 6$, $x_6 = 5$, o posto correspondente à x_5 é 4. Quando há observações com o mesmo valor, o posto correspondente a cada uma delas é definido como a média dos postos correspondentes. No exemplo, os postos das observações x_3 e x_6 são iguais a $2,5 = (2+3)/2$.

Figura 3.2: Gráfico de dispersão entre valores de duas variáveis X e Z



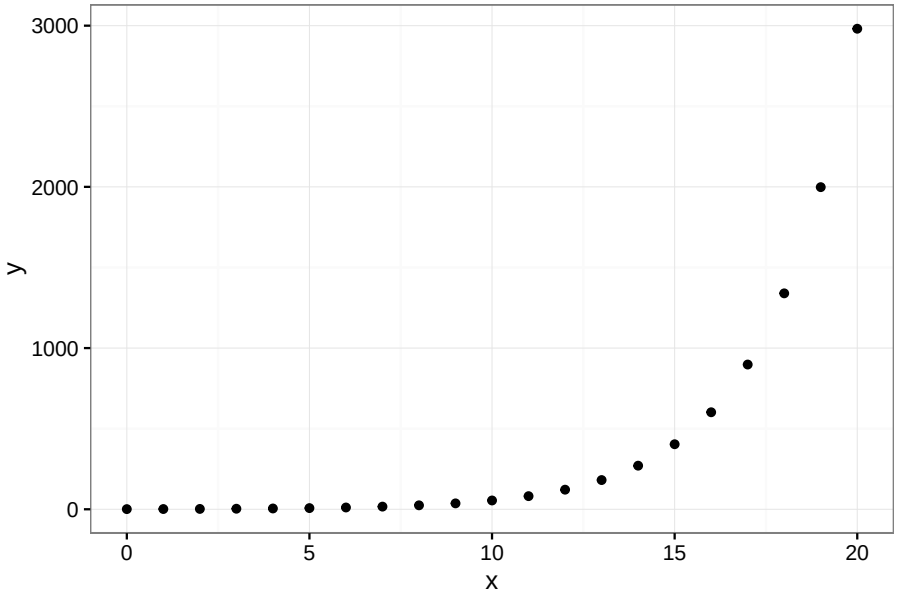
Os dados correspondentes à Figura 3.2 foram gerados a partir da expressão $z_i = 1 + 0,25x_i + e_i$ com e_i simulado a partir de uma distribuição Normal padrão e com as três últimas observações acrescidas de 15. Para esses dados obtemos $r_P = 0.73$ e $r_S = 0.90$. Eliminando as três observações com valores discrepantes, os coeficientes de correlação correspondentes são $r_P = 0.85$ e $r_S = 0.84$, indicando que o primeiro é mais afetado do que o segundo.

Além disso, o coeficiente de correlação de Spearman também é mais apropriado para avaliar associações não lineares, desde que sejam monotônicas, *i.e.*, em que os valores de uma das variáveis só aumentam ou só diminuem conforme a segunda variável aumenta (ou diminui). Os dados representados na Figura 3.3 foram gerados a partir da expressão $y_i = \exp(0.4x_i)$, $i = 1, \ldots, 20$. Nesse caso, os valores dos coeficientes de correlação de Pearson e de Spearman são, respectivamente, $r_P = 0.75$ e $r_S = 1$ indicando que apenas este último é capaz de realçar a associação perfeita entre as duas variáveis.

Uma ferramenta adequada para comparar as distribuições de uma característica observada sob duas condições diferentes é o gráfico QQ utilizado na Seção 2.7 para a comparação de uma distribuição empírica com uma distribuição teórica. Um exemplo típico é aquele referente ao objetivo secundário mencionado na descrição do Exemplo 3.3, em que se pretende avaliar a concordância entre as duas medidas ultrassonográficas do volume do lobo direito do fígado.

Denotando por X uma das medidas e por Y, a outra, sejam $Q_X(p)$ e $Q_Y(p)$ os quantis de ordem p das duas distribuições que pretendemos comparar. O gráfico QQ é um gráfico cartesiano de $Q_X(p)$ em função de

Figura 3.3: Gráfico de dispersão entre valores de duas variáveis X e Y



$Q_Y(p)$ (ou vice-versa) para diferentes valores de p. Se as distribuições de X e Y forem iguais, os pontos nesse gráfico devem estar sobre a reta $x = y$. Se uma das variáveis for uma função linear da outra, os pontos também serão dispostos sobre uma reta, porem com intercepto possivelmente diferente de zero e com inclinação possivelmente diferente de 1.

Quando o número de observações das duas variáveis for igual, o gráfico QQ é essencialmente um gráfico dos dados ordenados de X, ou seja $x_{(1)} \leq \ldots \leq x_{(n)}$, *versus* os dados ordenados de Y, nomeadamente, $y_{(1)} \leq \ldots \leq y_{(n)}$.

Quando os número de observações das duas variáveis forem diferentes, digamos $m > n$, calculam-se os quantis amostrais referentes àquela variável com menos observações utilizando $p_i = (i-0,5)/n$, $i = 1, \ldots, n$ e obtêm-se os quantis correspondentes à segunda variável por meio de interpolações como aquelas indicadas em (2.5). Consideremos, por exemplo os conjuntos de valores $x_{(1)} \leq \ldots \leq x_{(n)}$ e $y_{(1)} \leq \ldots \leq y_{(m)}$. Primeiramente, determinemos $p_i = (i - 0,5)/n$, $i = 1, \ldots, n$ para obter os quantis $Q_X(p_i)$; em seguida, devemos obter índices j tais que

$$\frac{j - 0,5}{m} = \frac{i - 0,5}{n} \text{ ou seja } j = \frac{m}{n}(i - 0,5) + 0,5.$$

Se j obtido dessa forma for inteiro, o ponto a ser disposto no gráfico QQ será $(x_{(i)}, y_{(j)})$; em caso contrário, teremos $j = [j] + f_j$ em que $[j]$ é o maior inteiro contido em j e $0 < f_j < 1$ é a correspondente parte fracionária ($f_j = j - [j]$). O quantil correspondente para a variável Y será:

$$Q_Y(p_i) = (1 - f_j)y_{([j])} + f_j y_{([j]+1)}.$$

Por exemplo, sejam $m = 45$ e $n = 30$; então, para $i = 1, \ldots, 30$ temos

$$p_i = (i - 0,5)/30 \text{ e } Q_X(p_i) = x_{(i)}$$

logo $j = 45/30(i - 0,5) + 0,5 = 1,5i - 0,25$ e $[j] = [1,5i]$ e $r = -0,25$. Consequentemente o quantil $Q_Y(p_i) = 1,25y_{([1,5i])} - 0,25y_{([1,5i]+1)}$. Os pontos correspondentes ao gráfico QQ serão

$$\begin{aligned} x_{(1)} \ &vs \ [1,25y_{(1)} - 0,25y_{(2)}], \\ x_{(2)} \ &vs \ [1,25y_{(3)} - 0,25y_{(4)}], \\ x_{(3)} \ &vs \ [1,25y_{(4)} - 0,25y_{(5)}], \\ x_{(4)} \ &vs \ [1,25y_{(6)} - 0,25y_{(7)}], \\ &\text{etc.} \end{aligned}$$

Suponha, por exemplo, que duas variáveis, X e Y, sejam tais que $Y = aX + b$, indicando que suas distribuições são iguais, exceto por uma transformação linear. Então,

$$p = P(X \leq Q_X(p)) = P(aX + b \leq aQ_X(p) + b) = P(Y \leq Q_Y(p)),$$

ou seja, $Q_Y(p) = aQ_X(p) + b$, indicando que o gráfico QQ correspondente mostrará uma reta com inclinação a e intercepto b.

Para a comparação das distribuições do volume ultrassonográfico do lobo direito do fígado medidas pelos dois observadores mencionados no Exemplo 3.3, o gráfico QQ está disposto na Figura 3.4. Os pontos distribuem-se em torno da reta $x = y$ sugerindo que as medidas realizadas pelos dois observadores tendem a ser similares. Em geral os gráficos QQ são mais sensíveis a diferenças nas caudas das distribuições, se estas forem aproximadamente simétricas e com a aparência de uma normal. Enquanto os diagramas de dispersão mostrem uma relação sistemática global entre X e Y, os gráficos QQ relacionam valores pequenos de X com valores pequenos de Y, valores medianos de X com valores medianos de Y e valores grandes de X com valores grandes de Y.

Outra ferramenta utilizada para a mesma finalidade é o **gráfico de médias/diferenças** originalmente proposto por Tukey e popularizado como **gráfico de Bland-Altman**. Essencialmente, essa ferramenta consiste num gráfico das diferenças entre as duas observações pareadas $(v_{2i} - v_{1i})$ em função das médias correspondentes $[(v_{1i} + v_{2i})/2]$, $i = 1, \ldots, n$. Esse procedimento transforma a reta com coeficiente angular igual 1 apresentada no gráfico QQ numa reta horizontal passando pelo ponto zero no gráfico de médias/diferenças de Tukey e facilita a percepção das diferenças entre as duas medidas da mesma característica.

O gráfico de médias/diferenças de Tukey (Bland-Altman) correspondente aos volumes medidos pelos dois observadores e indicados na Tabela 3.15 está apresentado na Figura 3.5. Os pontos no gráfico de médias/diferenças de Tukey distribuem-se de forma não regular em torno do valor zero e não

Figura 3.4: Gráfico QQ para avaliação da concordância de duas medidas ultrassonográficas do lobo direito do fígado

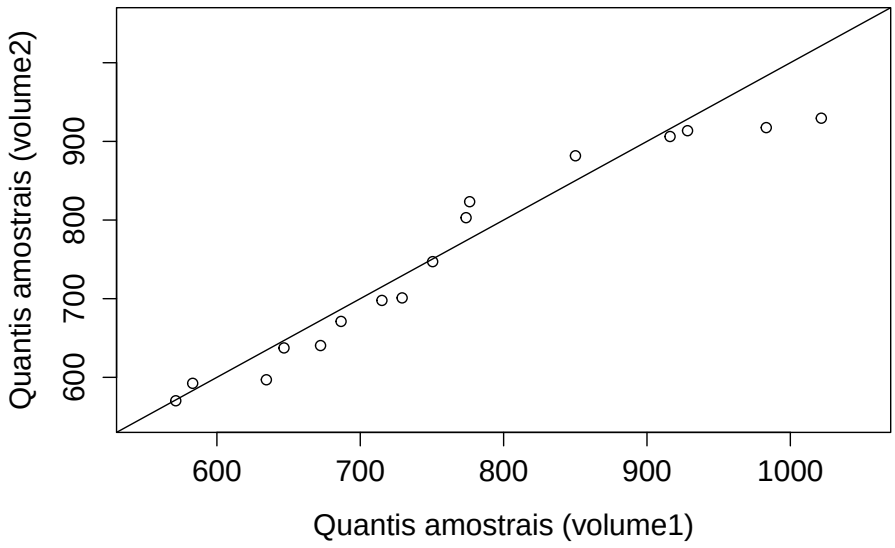
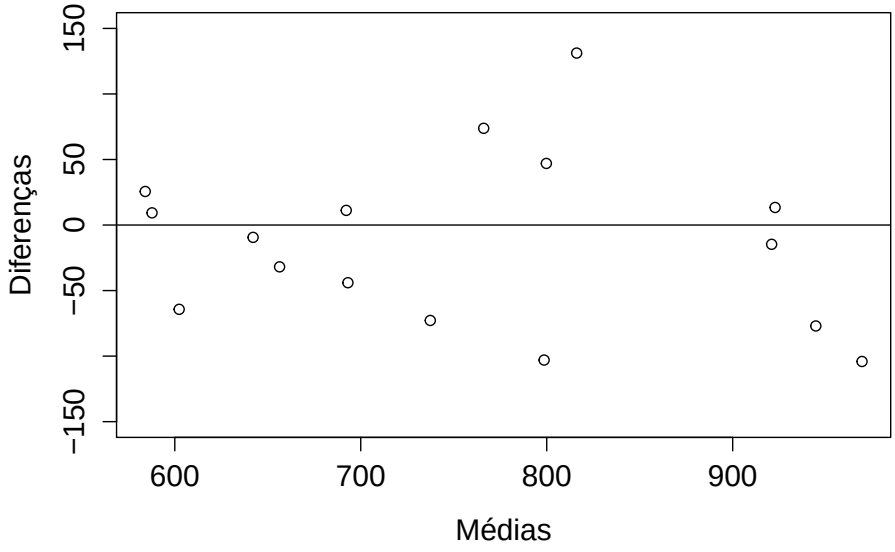


Figura 3.5: Gráfico de médias/diferenças de Tukey (Bland-Altman) para avaliação da concordância de duas medidas ultrassonográficas do lobo direito do fígado



sugerem evidências de diferenças entre as distribuições correspondentes. Por essa razão, para diminuir a variabilidade, decidiu-se adotar a média das medidas obtidas pelos dois observadores como volume do lobo direito do fígado para avaliar sua associação com o peso correspondente.

Exemplo 3.4 Os dados contidos na Tabela 3.16 foram extraídos de um estudo para avaliação de insuficiência cardíaca e correspondem à frequência cardíaca em repouso e no limiar anaeróbio de um exercício em esteira para 20 pacientes. Os gráficos QQ e de médias/diferenças de Tukey correspondentes

Tabela 3.16: Frequência cardíaca em repouso (fcrep) e no limiar anaeróbio (fclan) de um exercício em esteira

paciente	fcrep	fclan	paciente	fcrep	fclan
1	89	110	11	106	157
2	69	100	12	83	127
3	82	112	13	90	104
4	89	104	14	75	82
5	82	120	15	100	117
6	75	112	16	97	122
7	89	101	17	76	140
8	91	135	18	77	97
9	101	131	19	85	101
10	120	129	20	113	150

aos dados da Tabela 3.16 estão apresentados nas Figuras 3.6 e 3.7. Na Figura (3.6), a curva pontilhada corresponde à reta $Q_Y(p) = 1.29Q_X(p)$ sugerindo que a frequência cardíaca no limiar anaeróbio (Y) tende a ser cerca de 30% maior que aquela em repouso (X) em toda faixa de variação. Isso também pode ser observado, embora com menos evidência, no gráfico de Bland-Altman da Figura 3.7.

Exemplo 3.5. Considere o CD-Temperaturas, com os dados de temperatura para Ubatuba e Cananéia. O gráfico QQ está na Figura 3.8. Observamos que a maioria dos pontos está acima da reta $y = x$, mostrando que as temperaturas de Ubatuba são em geral maiores do que as de Cananéia para valores maiores do que 17 graus. O gráfico de Bland-Altman correspondente, apresentado na Figura 3.9, sugere que acima de 17 graus, em média Ubatuba tende a ser 1 grau mais quente que Cananéia.

Figura 3.6: Gráfico QQ para comparação das distribuições de frequência cardíaca em repouso e no limiar anaeróbio

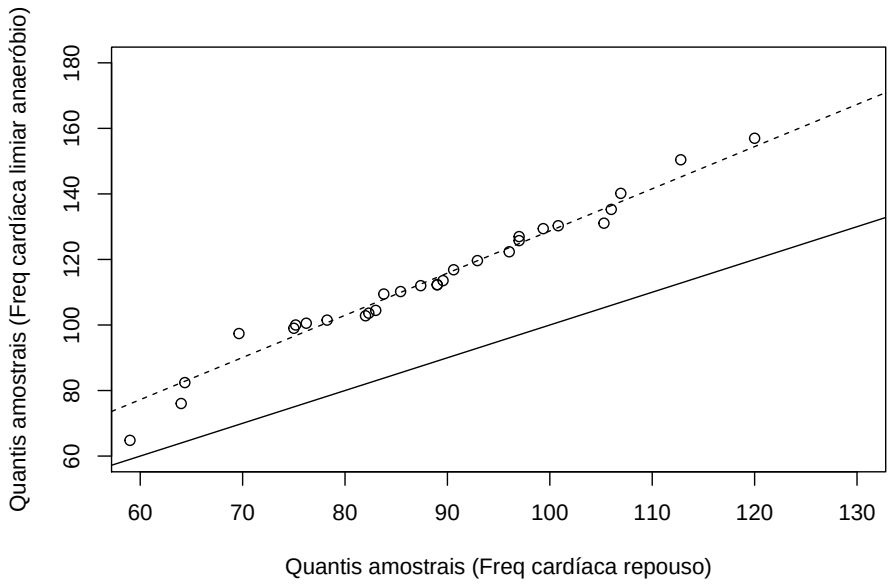


Figura 3.7: Gráfico de médias/diferenças de Tukey (Bland-Altman) para comparação das distribuições de frequência cardíaca em repouso e no limiar anaeróbio

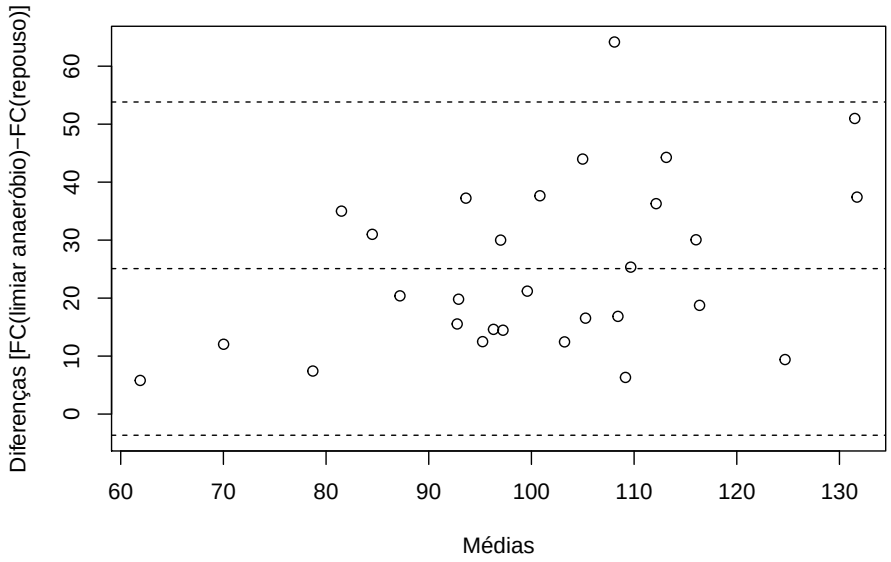


Figura 3.8: Gráfico QQ para comparação das distribuições de temperaturas de Ubatuba e Cananéia

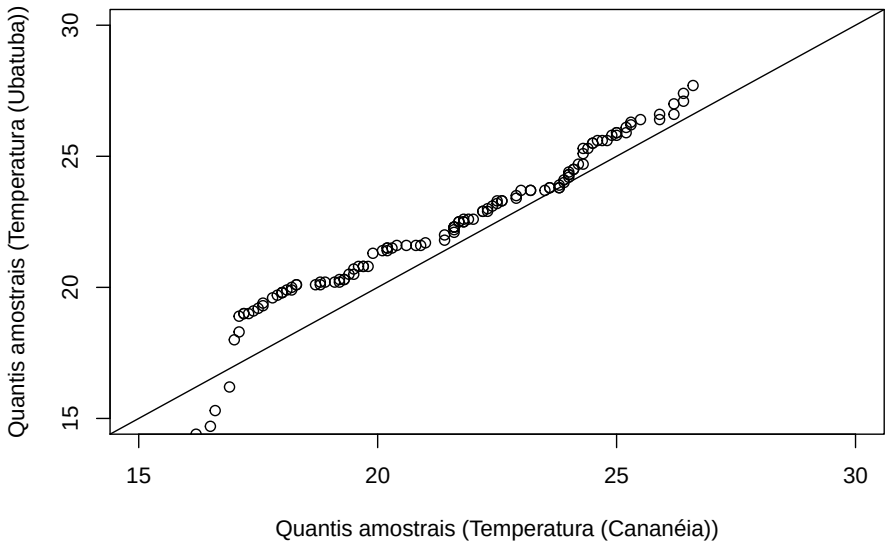
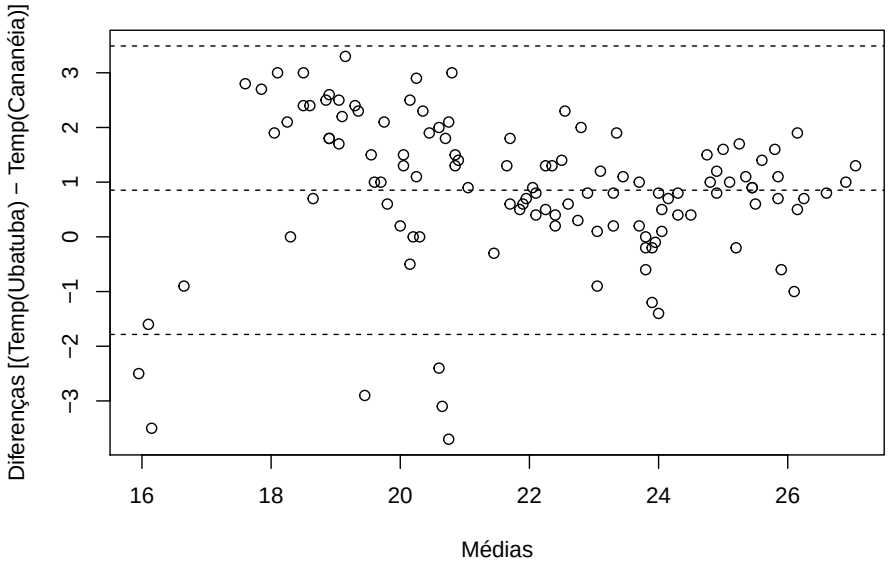


Figura 3.9: Gráfico de médias/diferenças de Tukey (Bland-Altman) para comparação das distribuições de temperaturas de Ubatuba e Cananéia



3.4 Uma Variável Qualitativa e Outra Quantitativa

Um estudo da associação entre uma variável quantitativa e uma qualitativa consiste essencialmente em comparar as distribuições da primeira nos diversos níveis da segunda. Essa análise pode ser conduzida por meio de medidas-resumo, histogramas, *boxplots* etc.

Exemplo 3.6. Num estudo coordenado pelo Laboratório de Poluição Atmosférica Experimental da USP, foram colhidos dados de concentração de vários elementos captados nas cascas de árvores em diversos pontos do centro expandido do município de São Paulo com o intuito de avaliar sua associação com a poluição atmosférica oriunda do tráfego. Os dados disponíveis em

`http://www.ime.usp.br/ jmsinger/MAE0217/arvores.xls`

foram extraídos desse estudo e contêm a concentração de Zn (ppm) em 497 árvores classificadas segundo a espécie (alfeneiro, sibipiruna e tipuana) e a localização em termos da proximidade do tipo de via (arterial, coletora, local I, local II, em ordem decrescente da intensidade de tráfego). Para efeito didático, consideramos primeiramente as 193 tipuanas. Medidas resumo para a concentração de Zn segundo os níveis de espécie e tipo de via estão indicadas na Tabela 3.17. Os resultados indicados na Tabela 3.17

Tabela 3.17: Medidas resumo para a concentração de Zn (ppm) em cascas de tipuanas

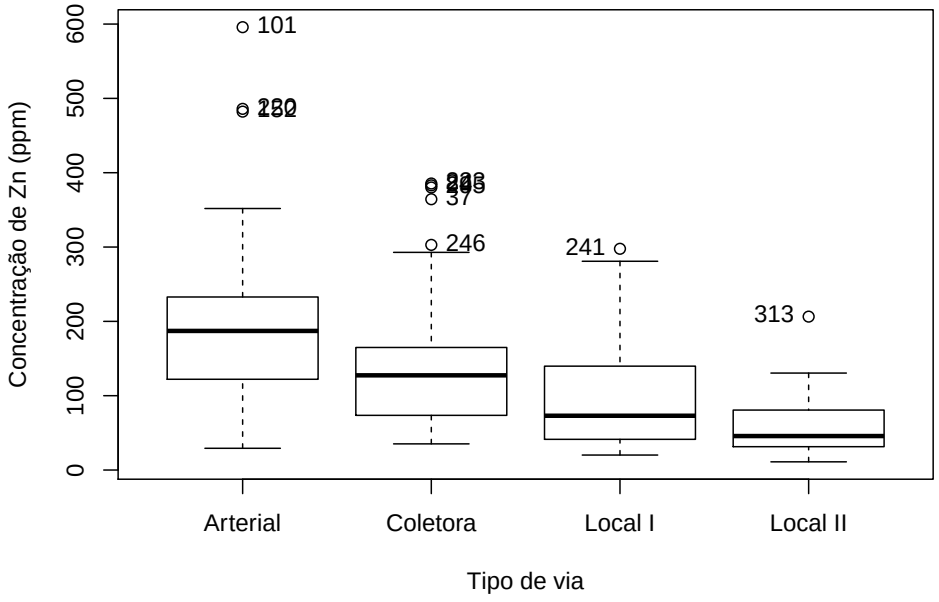
Tipo de via	Média	Desvio padrão	Min	Q1	Mediana	Q3	Max	n
Arterial	199.4	110.9	29.2	122.1	187.1	232.8	595.8	59
Coletora	139.7	90.7	35.2	74.4	127.4	164.7	385.5	52
Local I	100.6	73.4	20.1	41.9	73.0	139.4	297.7	48
Local II	59.1	42.1	11.0	31.7	45.7	79.0	206.4	34

Min: mínimo Max: máximo
Q1: primeiro quartil Q3: terceiro quartil

mostram que tanto a concentração média e mediana de Zn quanto o correspondente desvio padrão decrescem à medida que a intensidade de tráfego diminui, sugerindo que esse processo pode ser utilizado como um indicador da poluição produzida por veículos automotores. Os *boxplots* apresentados na Figura 3.10 confirmam essas conclusões e também indicam que as distribuições apresentam uma leve assimetria, especialmente para as vias coletoras e locais I além de alguns pontos discrepantes.

Outro tipo de gráfico útil para avaliar a associação entre a variável quantitativa (concentração de Zn, no exemplo) e a variável qualitativa (tipo de via, no exemplo) especialmente quando esta tem níveis ordinais (como no exemplo) é o **gráfico de perfis médios**. Nesse gráfico cartesiano as médias

Figura 3.10: *Boxplots* para comparação das distribuições da concentração de Zn nas cascas de tipuanas



(e barras representando desvios padrões, erros padrões ou intervalos de confiança)[3] da variável quantitativa são representadas no eixo das ordenadas e os níveis da variável quantitativa, no eixo das abscissas. O gráfico de perfis médios para a concentração de Zn medida nas cascas de tipuanas está apresentado na Figura 3.11. O gráfico da Figura 3.11 reflete as mesmas conclusões obtidas com as análises anteriores. No título do gráfico, deve-se sempre indicar o que representam as barras; desvios padrões são úteis para avaliar como a dispersão dos dados em torno da média correspondente varia com os níveis da variável quantitativa (e não dependem do número de observações utilizadas para o cálculo da média); erros padrões são indicados para avaliação da precisão das médias (e dependem do número de observações utilizadas para o cálculo delas); intervalos de confiança servem para comparação das médias populacionais correspondentes e dependem de suposições sobre a distribuição da variável quantitativa.

Uma análise similar para os 76 alfeneiros está resumida na Tabela 3.18, e Figuras 3.12 e 3.13.

Os resultados da Tabela 3.18, e das Figuras 3.12 e 3.13 indicam que as concentrações de Zn em alfeneiros tendem a ser maiores que aquelas encontradas em tipuanas porém são menos sensíveis à varições na intensidade de tráfego com exceção de vias locais II; no entanto, convém lembrar que

[3]Ver Nota de Capítulo 6

Figura 3.11: Gráfico de perfis médios (com barras de desvios padrões) para comparação das distribuições da concentração de Zn nas cascas de tipuanas

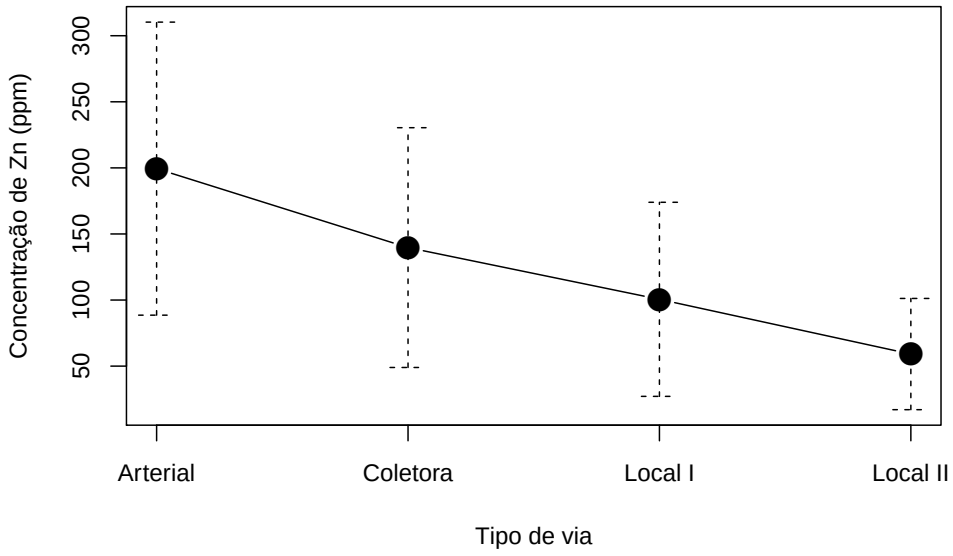


Tabela 3.18: Medidas resumo para a concentração de Zn (ppm) em cascas de alfeneiros

Tipo de via	Média	Desvio padrão	Min	Q1	Mediana	Q3	Max	n
Arterial	244.2	102.4	58.5	187.4	244.5	283.5	526.0	19
Coletora	234.8	102.7	15.6	172.4	231.6	311.0	468.6	31
Local I	256.3	142.4	60.0	154.9	187.0	403.7	485.3	19
Local II	184.4	96.4	45.8	131.1	180.8	247.6	306.6	7

Min: mínimo Max: máximo
Q1: primeiro quartil Q3: terceiro quartil

Figura 3.12: *Boxplots* para comparação das distribuições da concentração de Zn nas cascas de alfeneiros

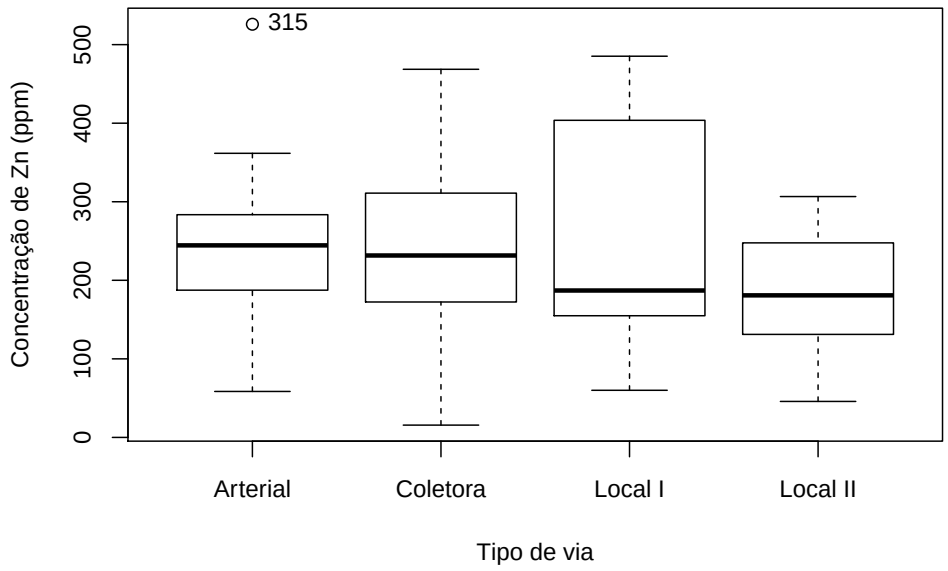
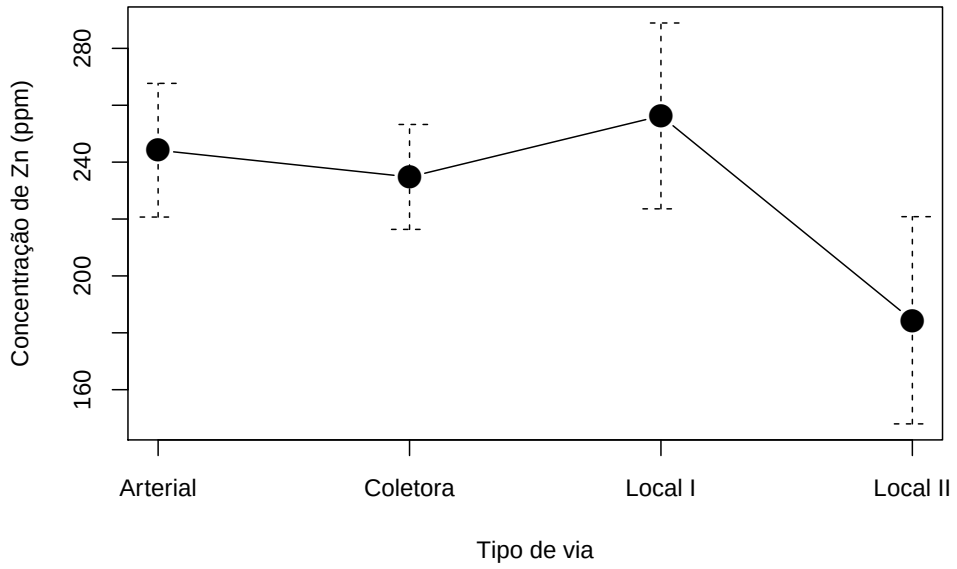


Figura 3.13: Gráfico de perfis médios (com barras de desvios padrões) para comparação das distribuições da concentração de Zn nas cascas de alfeneiros



apenas 7 alfeneiros foram avaliados nas proximidades desse tipo de via.

Exemplo 3.7. Consideremos os dados do arquivo CD-empresa referentes à informações sobre 36 funcionários de uma certa empresa. Nosso objetivo é avaliar a associação entre as variáveis "Salário" (S) expressa em número de salários mínimos e "Grau de instrução" (GI) (classificado como ensino fundamental, médio ou superior).

Medidas resumo para "Salário" em função dos níveis de "Grau de instrução"são apresentadas na Tabela 3.19.

Tabela 3.19: Medidas-resumo para a variável "Salário" (número de salários mínimos)

Grau de instrução	n	Média $\overline{S}$	Variância var(S)	Min	Q1	Q2	Q3	Max
Fundam	12	7,84	7,77	4,00	6,01	7,13	9,16	13,65
Médio	18	11,54	13,10	5,73	8,84	10,91	14,48	19,40
Superior	6	16,48	16,89	10,53	13,65	16,74	18,38	23,30
Todos	36	11,12	20,46	4,00	7,55	10,17	14,06	23,30

Min: mínimo Max: máximo
Q1: primeiro quartil Q2: mediana Q3: terceiro quartil

A leitura desses resultados sugere associação entre salários e grau de instrução: o salário tende a aumentar conforme aumenta o nível de educação. O salário médio de um funcionário é 11,12 (salários mínimos); para um funcionário com curso superior, o salário médio passa a ser 16,48, enquanto que funcionários com primeiro grau completo recebem, em média, 7,82.

Como nos casos anteriores, é conveniente poder contar com uma medida que quantifique o grau de associação entre as duas variáveis. Com esse intuito, convém observar que as variâncias podem ser usadas como insumos para construir essa medida. A variância calculada para a variável quantitativa ("Salário") para todos os dados, *i.e.*, sem usar a informação da variável qualitativa ("Grau de instrução"), mede a dispersão dos dados em torno da média global. Se as variâncias da variável "Salário" dentro de cada categoria da variável qualitativa forem pequenas (comparativamente à variância global), essa variável pode ser usada para melhorar o conhecimento da distribuição da variável quantitativa sugerindo a existência de uma associação entre ambas.

Na Tabela 3.19 pode-se observar que as variâncias do salário dentro das três categorias são menores do que a variância global e além disso, aumentam com o grau de instrução. Uma medida-resumo da variância **entre** as categorias da variável qualitativa é a média das variâncias ponderada pelo número de observações em cada categoria, ou seja,

$$\overline{\text{var}(S)} = \frac{\sum_{i=1}^{k} n_i \text{var}_i(S)}{\sum_{i=1}^{k} n_i}, \tag{3.8}$$

em que k é o número de categorias ($k = 3$ no exemplo) e $\text{var}_i(S)$ denota a variância de S dentro da categoria i, $i = 1, \ldots, k$. Pode-se mostrar que $\overline{\text{var}(S)} \leq \text{var}(S)$, de modo que podemos definir o grau de associação entre as duas variáveis como o ganho relativo na variância, obtido pela introdução da variável qualitativa. Explicitamente,

$$R^2 = \frac{\text{var}(S) - \overline{\text{var}(S)}}{\text{var}(S)} = 1 - \frac{\overline{\text{var}(S)}}{\text{var}(S)}. \tag{3.9}$$

Além disso, pode-se mostrar que $0 \leq R^2 \leq 1$. O símbolo R^2 é usual em análise de variância e regressão, tópicos a serem abordados nos Capítulos 14 e 15, respectivamente.

Para os dados do Exemplo 3.7, temos

$$\overline{\text{var}(S)} = \frac{12 \times 7,77 + 18 \times 13,10 + 6 \times 16,89}{12 + 18 + 6}. = 11,96,$$

Como $\text{var}(S) = 20,46$, obtemos $R^2 = 1 - (11,96/20,46) = 0,415$, sugerindo que $41,5\%$ da variação total do salário é **explicada** pelo grau de instrução.

3.5 Notas de capítulo

1) **Teorema de Bayes e razões de chances**

Considere a seguinte tabela 2x2

	Doente (D)	Não-doentes ($\overline{D}$)	Total
Exposto (E)	n_{11}	n_{12}	$n_{1\bullet}$
Não exposto ($\overline{E}$)	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

correspondente a um estudo em que o interesse é avaliar a associação entre a exposição de indivíduos a um certo fator de risco e a ocorrência de uma determinada moléstia. Em estudos prospectivos (prospective, follow-up, cohort) o planejamento envolve a escolha de amostras de tamanhos $n_{1\bullet}$ e $n_{2\bullet}$ de indivíduos expostos e não expostos ao fator de risco, respectivamente e a observação da ocorrência ou não da moléstia após um certo intervalo de tempo. A razão de chances é definida como:

$$\omega_1 = \frac{P(D|E)P(\overline{D}|\overline{E})}{P(\overline{D}|E)P(D|\overline{E})}.$$

Em estudos retrospectivos ou caso-controle, o planejamento envolve a escolha de amostras de tamanhos $n_{\bullet 1}$ e $n_{\bullet 2}$ de indivíduos não-doentes (controles) e doentes (casos), respectivamente e a observação retrospectiva de sua exposição ou não ao fator de risco. Nesse caso a razão de chances é definida por:

$$\omega_2 = \frac{P(E|D)P(\overline{E}|\overline{D})}{P(\overline{E}|D)P(E|\overline{D})}.$$

Pelo Teorema de Bayes [ver Bussab e Morettin (2015), por exemplo], temos

$$\begin{aligned}\omega_1 &= \frac{[P(D\cap E)/P(E)][P(\overline{D}\cap\overline{E})/P(\overline{E})]}{[P(\overline{D}\cap E)/P(E)][P(D\cap\overline{E})/P(\overline{E})} = \frac{P(D\cap E)P(\overline{D}\cap\overline{E})}{P(\overline{D}\cap E)P(D\cap\overline{E})} \\ &= \frac{[P(E|D)/P(D)][P(\overline{E}|\overline{D})/P(\overline{D})]}{[P(E|\overline{D})/P(\overline{D})][P(\overline{E}|D)/P(D)} = \omega_2\end{aligned}$$

2) **Medidas de dependência entre duas variáveis**

Dizemos que X e Y são **comonotônicas** se Y (ou X) for uma função estritamente crescente de X (ou Y) e são **contramonotônicas** se a função for estritamente decrescente.

Consideremos duas variáveis X e Y e seja $\delta(X,Y)$ uma medida de dependência entre elas. As seguintes propriedades são desejáveis para δ (Embrechts et al., 2003):

(i) $\delta(X,Y) = \delta(Y,X)$;

(ii) $-1 \leq \delta(X,Y) \leq 1$;

(iii) $\delta(X,Y) = 1$ se X e Y são comonotônicas e $\delta(X,Y) = -1$ se X e Y são contramonotônicas;

(iv) Se T for uma transformação monótona,

$$\delta(T(X),Y) = \begin{cases} \delta(X,Y), & \text{se T for crescente,} \\ -\delta(X,Y), & \text{se T for decrescente.} \end{cases}$$

(v) $\delta(X,Y) = 0$ se e somente se X e Y são independentes.

O coeficiente de correlação de Pearson entre X e Y é definido por

$$\rho_P = \frac{\text{Cov}(X,Y)}{dp(X)dp(Y)}$$

com $\text{Cov}(X,Y) = E(XY) - E(X)E(Y)$. Pode-se provar que $-1 \leq \rho_P \leq 1$ e que satisfaz as propriedades (i)-(ii). Além disso, ρ_P requer que as variâncias de X e Y sejam finitas e $\rho = 0$ não implica independência entre X e Y, a não ser que (X,Y) tenha uma distribuição normal bivariada. Também, ρ_P não é invariante sob transformações não lineares estritamente crescentes.

3) Convém reafirmar que $\rho_P(X,Y)$ mede dependência linear entre X e Y e não outro tipo de dependência. De fato, suponha que uma das variáveis possa ser expressa linearmente em termos da outra, por exemplo $X = aY + b$, e seja $d = E(|X - aY - b|^2)$. Então, pode-se provar (veja Exercício 13) que o mínimo de d ocorre quando

$$a = \frac{\sigma_X}{\sigma_Y}\rho_P(X,Y), \qquad b = E(X) - aE(Y), \tag{3.10}$$

e o mínimo é dado por

$$\min d = \sigma_X^2(1 - \rho_P(X,Y)^2). \tag{3.11}$$

Portanto, quanto maior o valor absoluto do coeficiente de correlação entre X e Y, melhor a acurácia com que uma das variáveis pode ser representada como uma combinação linear da outra. Obviamente, este mínimo anula-se se e somente se $\rho_P = 1$ ou $\rho_P = -1$. Então de (3.11) temos

$$\rho_P(X,Y) = \frac{\sigma_X^2 - \min_{a,b} E(|X - aY - b|^2)}{\sigma_X^2}, \tag{3.12}$$

ou seja, $\rho_P(X,Y)$ mede a redução relativa na variância de X por meio de uma regressão linear de X sobre Y.

4) O coeficiente de correlação não é uma medida resistente. Uma medida robusta para a associação entre duas variáveis quantitativas é construída como segue. Considere as variáveis padronizadas

$$\tilde{x}_k = \frac{x_k}{S_x(\alpha)}, \quad \tilde{y}_k = \frac{y_k}{S_y(\alpha)}, \ \ k = 1, \ldots, n,$$

em que $S_x^2(\alpha)$ e $S_y^2(\alpha)$ são as α-variâncias aparadas para os dados x_i e y_i, respectivamente. Um coeficiente de correlação robusto é definido por

$$r(\alpha) = \frac{S_{\tilde{x}+\tilde{y}}^2(\alpha) - S_{\tilde{x}-\tilde{y}}^2(\alpha)}{S_{\tilde{x}+\tilde{y}}^2(\alpha) + S_{\tilde{x}-\tilde{y}}^2(\alpha)}, \tag{3.13}$$

em que, por exemplo, $S_{\tilde{x}+\tilde{y}}^2(\alpha)$ é a α-variância aparada da soma dos valores padronizados de x_i e y_i. Pode-se mostrar que $r(\alpha) = r_P$ se $\alpha = 0$. Esse método é denominado de **somas e diferenças padronizadas**.

Exemplo 3.8. Consideremos os dados $(x_i, y_i), \ \ i = 1, \ldots, n$ apresentados na tabela abaixo.

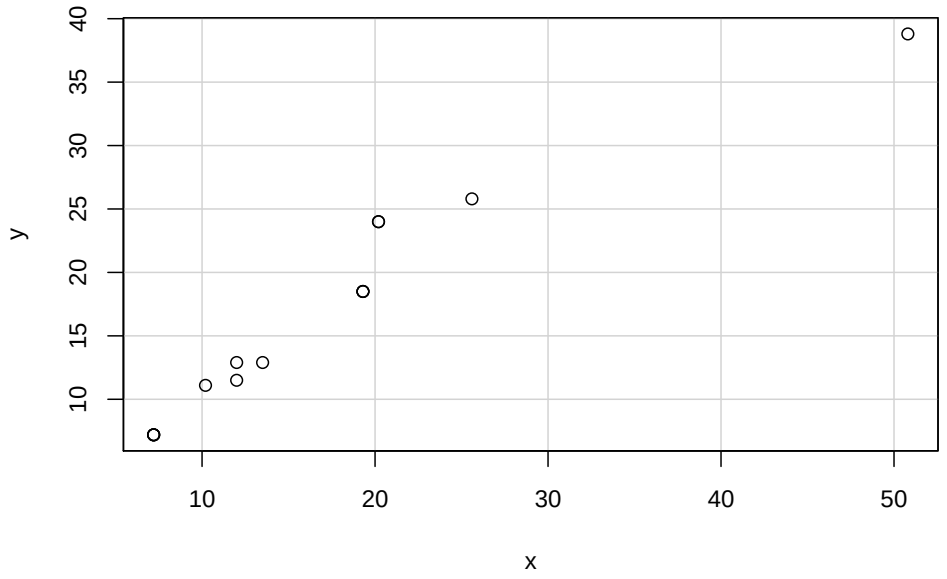
Tabela 3.20: Valores hipotéticos de duas variáveis X e Y

x	y	x	y
20,2	24,0	19,3	18,5
50,8	38,8	19,3	18,5
12,0	11,5	19,3	18,5
25,6	25,8	10,2	11,1
20,2	24,0	12,0	12,9
7,2	7,2	7,2	7,2
7,2	7,2	13,5	12,9
7,2	7,2		

Para $\alpha = 0,05$, obtemos:

$$\overline{x}(\alpha) = 14,86, \quad \overline{y}(\alpha) = 15,33, \quad S_x(\alpha) = 5,87, \quad S_y(\alpha) = 6,40,$$

Figura 3.14: Gráfico de dispersão para os dados do Exemplo 3.8



$$(\overline{\tilde{x}+\tilde{y}})(\alpha) = 4,93, \quad (\overline{\tilde{x}-\tilde{y}})(\alpha) = 0,14,$$
$$S^2_{\tilde{x}+\tilde{y}}(\alpha) = 3,93, \quad S^2_{\tilde{x}-\tilde{y}}(\alpha) = 0,054.$$

Então de 3.13 obtemos $r(\alpha) = 0,973$, o que indica uma alta correlação entre os dois conjuntos de dados. Na Figura 3.14 temos o diagrama de dispersão correspondente.

5) **Gráficos PP**

Na Figura 3.XX, observe que $p_x(q) = P(X \leq q) = F_X(q)$ e que $p_y(q) = P(Y \leq q) = F_Y(q)$, de modo que os pares $(p_x(q), p_y(q))$, para qualquer q real, formam um gráfico de probabilidades ou **gráfico PP**. Os pares $(Q_X(p), Q_Y(p))$ para $0 < p < 1$, formam um gráfico de quantis *versus* quantis (gráfico QQ).

Incluir aqui a figura com as distribuições acumuladas de X e Y

Se as distribuições de X e Y forem iguais, então $F_X = F_Y$ e os pontos dos gráficos PP e QQ se situam sobre a reta $x = y$. Em geral os gráficos QQ são mais sensíveis a diferenças nas caudas das distribuições se estas forem aproximadamente simétricas e com a aparência de uma distribuição normal. Suponha que $Y = aX + b$, ou seja, que as distribuições de X e Y são as mesmas, exceto por uma transformação linear. Então,

$$p = P(X \leq Q_X(p)) = P(aX + b \leq aQ_X(p) + b) = P(Y \leq Q_Y(p)),$$

ou seja,

$$Q_Y(p) = aQ_X(p) + b$$

e então o gráfico QQ correspondente será representado por uma reta com inclinação a e intercepto b. Essa propriedade não vale para gráficos PP.

6) Erro padrão e intervalo de confiança

3.6 Exercícios

1. Para o CD-1, analise os dados de população urbana e rural.
2. Para o mesmo conjunto de dados, CD-Empresa, considere as variáveis "Região" e "Densidade populacional". Avalie a relação entre densidade e região.
3. Construa um gráfico QQ para as variáveis "Salário de professor secundário" e "Salário de administrador"do CD-Salarios.
4. Calcule o coeficiente de correlação, r e o coeficiente de correlação robusto, $r(0,10)$, para os dados do exercício anterior.
5. Construa um gráfico QQ para os dados de temperatura de Cananéia e Ubatuba, do CD-4, desprezando os 15 dados finais de Ubatuba.
6. Para o CD-Salarios, considere a variável "Região", com as classes "América do Norte", "América Latina", "Europa"e "Outros"e a variável "Salário de professor secundário". Analise as duas variáveis.
7. Analise as variáveis "Temperatura" e "Umidade" do CD-Poluicao.
8. Analise a variável "Preço de veículos" segundo as categorias N (nacional) e I (importado) para o CD-Veículos.
9. Prove que se $\alpha = 0$, então $r(\alpha) = r$.
10. Prove que 3.1 pode ser escrita como

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}},$$

em que n_{ij} é a frequência absoluta observada na linha i e coluna j e n^*_{ij} é a respectiva frequência esperada.

11. Prove que (3.1) pode ser escrita em termos de frequências relativas como

$$\chi^2 = n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(f_{ij} - f^*_{ij})^2}{f^*_{ij}},$$

com notação similar à do problema anterior.

12. Prove que (3.6) e (3.7) são equivalentes.
13. Prove as relações (3.10)-(3.12).

Referências

Anderson, T.W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, **52**, 200–203.

Box, G.E.P. and Müller, M.E.(1958). A note on the generation of random normal deviates. *The Annals of Statistics,* 29, 610–611.

Bland, J.M. and Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet,* 327: 307–10. doi:10.1016/S0140-6736(86)90837-8

Box, G.E.P. and Cox, D.R. (1964).An analysis of transformations. *Journal of the Royal Statistical Society, Series B,* **26**: 211-252.

Boyles, (1983).

Bussab, W.O. e Morettin, P.A. (2015). *Estatística Básica, 8a Edição*. São Paulo: Saraiva.

Chambers, J.M., Cleveland, W.S., Kleiner, B and Tukey, P.A. (1983). *Graphical Methods for Data Analysis*. London: Chapman and Hall.
Cleveland, W.M. (1994). *The Elements of Graphing Data*. Summit, New Jersey: Hobart Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20: 37–46. doi:10.1177/001316446002000104

Dempster, A.P., Laird, N. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

Embrecths, P., Lindskog, F. and McNeil, A. (2003). Modelling dependence with copulas and applications to risk management. In Handbook of Heavy Tailed Distributions in Finance, ed. S. Rachev, Elsevier, Ch. 8, 329–384.

Ehrenberg, A. S. C. (1981). The problem of numeracy. *The American Statistician*, **35**, 67-71.

Fletcher, R. (1987). *Practical Methods of Optimization, Second Edition.* New York: Wiley.

Fuller, W.A. (1996). *Introduction to Statistical Time Series, Second Edition*. New York: Wiley.

Giampaoli, V., Magalhães, M.N., Fonseca, F.C. e Anorozo, N.F. (2008). Relatório de análise estatística sobre o projeto: "Avaliação e pesquisa: Investigando as dificuldades em Matemática no Ensino Fundamental da Rede Municipal da cidade de São Paulo – 2ª fase". São Paulo, IME-USP8. (RAE – CEA – 08P27).

Goldfeld, S.M., Quandt, R.E. and Trotter, H.F. (1966). Maximisation by quadratic hill-climbing. *Econometrica*, **34**, 541-551.

Graedel, T, and Kleiner, B. (1985). Exploratory analysis of atmospheric data. In *Probability, Statistics and Decision Making in Atmospheric Sciences, A.H. Murphy and R.W. Katz, eds), pp 1-43.* Boulder: Westview Press.

Hammersley, J.M. and Handscomb, D.C.(1964). *Monte Carlo Methods.* New York, Wiley.

Hinkley, B. (1977). On quick choice of probability transformations. *Applied Statsistics*, **26**, 67-69.

Johnson, N.L. and Leone, F.C. (1964). *Statistics and Experimental Design in Engineering and Physical Sciences, Vols 1, 2*. New York: Wiley.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Louis, T.A. (1982). Finding observed information using the EM algorithm. *Journal of the Royal Statistical Society, Series B,* **44**, 98-130.

McGill et al. (1978).

Meng, X.L. and Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**, 267-278.

Metropolis, N. and Ulam, S.(1949). The Monte Carlo Method. *Journal of the American Statistical Association,* 44, 335–341.

Miller, R.G. and Halpern, J.H. (1982). Regression via censored data. *Biometrika*, **69**, 521-531.

Morettin, P.A. and Tolói, C.M.C. (2006). *Análise de Séries Temporais, 2a Ed.* São Paulo: Blücher.

Nelder, J.A. and Mead, R. (1965). A simplex method for function minimi-

zation. *Computer Journal*, **7**, 308-313.

Powell, M.J.D. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Computer Journal*, **7**, 155-162.

Ross, S.(1997). *Simulation, 2nd Ed.*, New York: Academic Press.

Rubin, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, **72**, 538–543.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, **91**, 473-489.

Schmee, J. and Hahn, G.J. (1979). A simple method for regression analysis with censored data. *Technometrics,*, **21**, 417-432.

Sobol, I.M.(1976). *Método de Monte Carlo.* Moscow: Editorial MIR.

Tanner, M.A. (1996). *Tools for Statistical Inference, 3rd Ed.*. New York: Springer.

Viera, J. and Garrett, J.M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine,* 37: 360-263

von Neumann, J.(1951). Various techniques used in connection with random digits, Monte Carlo Method. *U.S. National Bureau of Standards Applied Mathematica Series*, 12, 36–38.

Wayne, D.W. (1990). *Applied Nonparametric Statistics, Second Edition* . Boston: PWS-Kent. ISBN 0-534-91976-6.

Wei, G.C.G. and Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association,* **85**, 699-704.

Wilks, S.S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *Annals of Mathematical Statistics*, **2**, 163–195

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics,* **11**, 95-103.

Zan, A.S.C.N. (2005). Ultra-sonografia tridimensional: determinação do volume do lobo hepático direito no doador para transplante intervivos. Tese de doutorado. São Paulo: Faculdade de Medicina, Universidade de São Paulo.

Índice