The McGraw-Hill Companies

3ª Edição
ISBN 85-8680493-2

Tradução do original em espanhol *Metodología de la Investigación*
Copyright © 2003, 1998, 1991 da 3ª edição em espanhol por The McGraw-Hill/Interamericana Editores, S. A. de C.V., uma subsidiária da The McGraw-Hill Companies, Inc.
ISBN da obra original: 970-10-3632-8

Diretor-geral: *Adilson Pereira*
Editora de Desenvolvimento: *Ada Santos Seles*
Preparação de Texto: *Carla Montagner*
Editoração Eletrônica: *Virtual Laser*
Criação de Capa: *Megaart Design*

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Sampieri, Roberto Hernández
Metodologia de pesquisa / Roberto Hernández Sampieri, Carlos Fernández Collado, Pilar Baptista Lucio ; tradução Fátima Conceição Murad, Melissa Kassner, Sheila Clara Dystyler Ladeira ; revisão técnica e adaptação Ana Gracinda Queluz Garcia, Paulo Heraldo Costa do Valle. -- 3. ed. -- São Paulo : McGraw-Hill, 2006.

Título original: Metodología de la investigación.
Bibliografia. ISBN 85-8680493-2

1. Pesquisa 2. Pesquisa - Metodologia I. Collado, Carlos Fernández. II. Lucio, Pilar Baptista . III. Título.

06-7701 CDD-001.42

Índices para catálogo sistemático:

1. Metodologia da pesquisa 001.42
2. Pesquisa : Metodologia 001.42

Se você tem dúvidas, críticas ou sugestões, entre em contato pelo endereço eletrônico *mh_brasil@mcgraw-hill.com.* Em Portugal, use o endereço *serviço_clientes@mcgraw-hill.com.*

Aula 4

Andrea Guimarães

Dedicatória

Aos meus pais, Pola e Roberto, por sua orientação e exemplo maravilhosos. À minha esposa, Elisa Costa Aizcorbe, pelo seu amor e apoio. À memória do meu irmão, meus avós e a Raúl Durán Reveles.

ROBERTO HERNÁNDEZ SAMPIERI

À memória do meu pai, Telésforo Fernández. Aos meus filhos, Íñigo e Alonso, com amor cada vez maior e profundo.

CARLOS FERNÁNDEZ COLLADO

Aos meus alunos.

PILAR BAPTISTA LUCIO

final dos oito alunos de uma turma (4, 5, 5, 6, 6, 7, 7 e 8), podemos calcular a média aritmética por:

$$\frac{4+5+5+6+6+7+7+8}{8} = 6$$

De modo geral, dado um conjunto de n valores de uma certa variável X, podemos definir a **média aritmética** por:

$$\overline{X} = \frac{\sum X}{n}$$

onde $\sum X$ representa a soma dos valores da variável X. Em geral, a média aritmética é bastante informativa. Se, por exemplo, na primeira avaliação de uma disciplina, a média das notas dos alunos foi igual a 7,0, e na segunda avaliação foi igual a 9,0, podemos dizer que, em geral, os alunos tiveram melhor aproveitamento na segunda avaliação, mesmo sem nos referirmos às notas de cada aluno individualmente. Mas devemos sempre ter em mente que a média é um resumo dos dados e, por isso, pode esconder informações relevantes.

Exemplo 6.1 Vamos considerar a comparação de três turmas de estudantes em termos de suas notas (veja a Tabela 6.1 e Figura 6.1).

Tabela 6.1 Notas finais de três turmas de estudantes e as respectivas médias.

Turma	Notas dos alunos								Média da turma
A	4	5	5	6	6	7	7	8	6,00
B	1	2	4	6	6	9	10	10	6,00
C	0	6	7	7	7	7,5	7,5		6,00

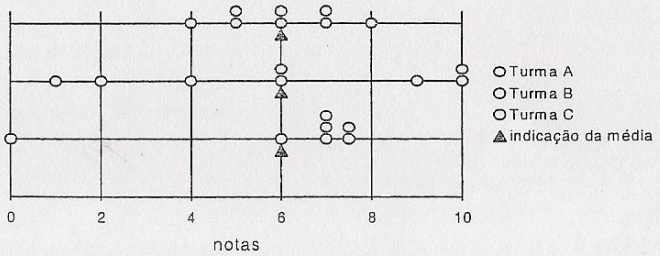


Figura 6.1 Representação das distribuições das notas de três turmas e as correspondentes posições das médias aritméticas.

Observando a Figura 6.1, percebemos que em cada diagrama de pontos a média aritmética representa, num certo sentido, a posição central dos valores. Mais especificamente, podemos dizer que a média aritmética indica o *centro* de um conjunto de valores, considerando o conceito físico de *ponto de equilíbrio* ou *centro de gravidade*. Se imaginarmos os pontos como pesos sobre uma tábua, a *média* é a posição em que um suporte equilibraria a tábua.

A média aritmética *resume* o conjunto de dados em termos de uma *posição central* ou *valor típico*, mas, em geral, não fornece informação sobre outros aspectos da distribuição.

Observamos, na Figura 6.1, que os três conjuntos de valores, apesar de estarem distribuídos sob diferentes formas, apontam para uma mesma média. Comparando as notas da Turma A com as notas da Turma B, verificamos que as notas da Turma B são bem mais *dispersas*, indicando que essa turma é mais heterogênea. Na Turma C, observamos um ponto discrepante dos demais, uma nota extremamente baixa. Com isso, a média fica abaixo da maioria das notas da turma.[1]

Para melhorar o resumo dos dados, podemos apresentar, ao lado da média aritmética, uma medida de dispersão, como a variância ou o desvio padrão.

A variância e o desvio padrão

Tanto a variância quanto o desvio padrão são medidas que fornecem informações complementares à informação da média aritmética. Estas medidas avaliam a *dispersão* do conjunto de valores em análise. Para calcularmos a variância ou o desvio padrão, devemos considerar os desvios de cada valor em relação à média aritmética. Depois, construímos uma espécie de média desses desvios. Ilustramos, a seguir, as etapas de cálculo, usando as notas da Turma A.

Descrição	notação	resultados numéricos							
Valores (notas dos alunos)	X	4	5	5	6	6	7	7	8
Média	$\overline{X}$				6				
Desvios	$X-\overline{X}$	-2	-1	-1	0	0	1	1	2
Desvios quadráticos	$(X-\overline{X})^2$	4	1	1	0	0	1	1	4

[1] Podemos observar no diagrama de pontos referente à Turma C que a presença de um valor discrepante *arrasta* a média para o seu lado. Assim, a média deixa de representar propriamente um *valor típico* do conjunto de dados. Um tratamento mais adequado para dados que contenham valores discrepantes será visto na Seção 6.3.

Para evitar o problema dos desvios negativos, vamos trabalhar com os desvios quadráticos, $(X-\overline{X})^2$. A variância é definida como a média aritmética dos desvios quadráticos. Por conveniência, vamos calcular esta média, usando como denominador $n-1$ no lugar de n.[2] Assim, definimos a **variância** de um conjunto de valores pela expressão:

$$S^2 = \frac{\sum(X-\overline{X})^2}{n-1}$$

onde $\Sigma(X-\overline{X})$ é a soma dos desvios quadráticos. Em relação ao conjunto de notas da Turma A, a variância é

$$S^2 = \frac{4+1+1+0+0+1+1+4}{8-1} = 1{,}71$$

Como a variância de um conjunto de dados é calculada em função dos desvios quadráticos, sua unidade de medida equivale à unidade de medida dos dados ao quadrado. Nesse contexto, é mais comum se trabalhar com a *raiz quadrada positiva* da variância. Esta medida é conhecida como *desvio padrão*, o qual é expresso na mesma unidade de medida dos dados em análise. Então, o **desvio padrão** de um conjunto de valores pode ser calculado por:

$$S = \sqrt{\frac{\sum(X-\overline{X})^2}{n-1}}$$

Em termos do conjunto de notas da Turma A, temos o seguinte desvio padrão: $S = \sqrt{1{,}71} = 1{,}31$.

Ao compararmos os desvios padrões de vários conjuntos de dados, podemos avaliar quais dados se distribuem de forma mais (ou menos) dispersa. O desvio padrão será sempre *não negativo* e será tão maior quanto mais dispersos forem os valores em análise. A Tabela 6.2 mostra o desvio padrão das notas de cada uma das três turmas de alunos, referente aos dados do Exemplo 6.1.

Tabela 6.2 Medidas descritivas das notas finais dos alunos de três turmas.

Turma	Número de alunos	Média	Desvio padrão
A	8	6,00	1,31
B	8	6,00	3,51
C	7	6,00	2,69

[2] Muitos autores costumam diferenciar a fórmula da variância quando os dados se referem a uma população ou a uma amostra. Quando os dados representam uma população de N elementos, a variância é definida com o denominador N. Quando os dados se referem a uma amostra de n elementos devemos usar o denominador $n-1$. Por simplicidade, vamos considerar sempre o segundo caso.

Ao analisarmos a Tabela 6.2, verificamos, através das médias, que os alunos das três turmas *tenderam* a ter as notas em torno de seis, mas, pelos desvios padrões, concluímos que os alunos da Turma A obtiveram notas relativamente próximas umas das outras, quando comparados aos alunos das outras turmas. Por outro lado, as notas dos alunos da Turma B foram as que se apresentaram mais heterogêneas.[3]

O desvio padrão fornece informação sobre a dispersão (variância ou heterogeneidade) dos valores.

Exercícios

1) Faça os cálculos dos desvios padrões das notas dos alunos das turmas B e C (Tabela 6.1). Verifique se os resultados conferem com os apresentados na Tabela 6.2.
2) Admita que todos os alunos de uma Turma D obtiveram notas iguais a *sete*. Qual o valor da média aritmética? E qual o valor do desvio padrão?
3) A tabela seguinte mostra os resultados dos cálculos das médias e desvios padrões das taxas de crescimento demográfico dos municípios de duas microrregiões catarinenses. Quais as conclusões que você pode tirar desta tabela?

Medidas descritivas das taxas de crescimento demográfico de duas microrregiões de Santa Catarina, 1970-80.

Microrregião	Nº de municípios	Média	Desvio padrão
Serrana	12	-0,36	0,67
Litoral de Itajaí	8	3,55	2,47

6.2 Fórmulas para o cálculo de $\overline{X}$ e S

Ao calcular o desvio padrão nos casos em que a média, $\overline{X}$, acusar um valor fracionário, os desvios, $X-\overline{X}$, acumularão erros de arredondamento, que poderão comprometer o resultado final. Para evitar este inconveniente, podemos usar a seguinte fórmula para o cálculo do

[3] Observe, pela Figura 6.1, que as notas da turma C estão mais concentradas do que as da turma A. Porém, o valor discrepante, além de deslocar a média, aumenta o desvio padrão. Se o valor discrepante fosse desconsiderado, o desvio padrão das notas da turma C seria o menor de todos – a média seria 7 e o desvio padrão 0,55.

desvio padrão, que é matematicamente equivalente àquela apresentada no tópico anterior:

$$S = \sqrt{\frac{\sum X^2 - n\overline{X}^2}{n-1}}$$

onde: $\sum X^2$ é a soma dos valores quadráticos;
$\overline{X}^2$ é a média elevada ao quadrado; e
n é o número de valores.

Ilustraremos o uso desta nova formulação com as notas obtidas pelos alunos da Turma A (Exemplo 6.1).

Valores (notas) X: 4 5 5 6 6 7 7 8 ($\sum X = 48$ e $\overline{X} = 6$)

Valores ao quadrado X^2: 16 25 25 36 36 49 49 64 ($\sum X^2 = 300$)

Assim,

$$S = \sqrt{\frac{300 - 8(6)^2}{7}} = \sqrt{\frac{300 - 288}{7}} = \sqrt{\frac{12}{7}} = 1{,}31$$

Como era de se esperar, chegamos ao mesmo resultado encontrado anteriormente.

Ponderando pelas freqüências

Outro aspecto relativo ao cálculo da média e do desvio padrão refere-se à soma de valores repetidos. Por exemplo, ao calcularmos a média das notas da Turma A, fizemos a seguinte soma:

$$\sum X = 4 + 5 + 5 + 6 + 6 + 7 + 7 + 8,$$

que é equivalente a: $4\times1 + 5\times2 + 6\times2 + 7\times2 + 8\times1 = \sum (X \cdot f)$

onde consideramos apenas os valores distintos de X e ponderamos pelas respectivas freqüências, f. Analogamente, podemos calcular a soma quadrática dos valores de X por

$$\sum (X^2 \cdot f) = 4^2 + 5^2 \times 2 + 6^2 \times 2 + 7^2 \times 2 + 8^2 =$$

Com esta nova notação, as formulações de média e desvio padrão são apresentadas a seguir.

$$\overline{X} = \frac{\sum (X \cdot f)}{n} \quad \text{e} \quad S = \sqrt{\frac{\sum (X^2 \cdot f) - n \cdot \overline{X}^2}{n-1}}$$

A Tabela 6.3 mostra a seqüência de cálculos para a obtenção da média e do desvio padrão, usando as notas finais dos alunos da Turma A.

Tabela 6.3 Cálculos auxiliares para a obtenção de $\overline{X}$ e S.

Nota X	Freqüência *f*	*X·f*	$X^2 \cdot f$
4	1	4	16
5	2	10	50
6	2	12	72
7	2	14	98
8	1	8	64
Total	8	48	300

Assim, $\overline{X} = \frac{48}{8} = 6$ e $S = \sqrt{\frac{300 - 8 \cdot (6)^2}{7}} = 1{,}31$

Os cálculos usando as freqüências facilitam bastante quando existirem muitas repetições de valores.

Dados grupados em classes

Quando os dados estão grupados em classes, os cálculos de $\overline{X}$ e S somente poderão ser feitos de forma aproximada, usando o *ponto médio* de cada classe para representar os valores que ocorreram nessa classe (veja Exemplo 6.2).[4]

Exemplo 6.2 Cálculo aproximado de $\overline{X}$ e S dos valores da taxa de alfabetização, relativos a uma amostra aleatória de municípios brasileiros, ano 2000.

Classes da taxa de alfabetização	Ponto médio X	Freqüência de municípios f	$X \cdot f$	$X^2 \cdot f$
40 |— 50	45	1	45	2.025
50 |— 60	55	5	275	15.125
60 |— 70	65	8	520	33.800
70 |— 80	75	6	450	33.750
80 |— 90	85	12	1.020	86.700
90 |— 100	95	8	760	72.200
Total	–	40	3.070	243.600

[4] Ao buscarmos dados em fontes secundárias, muitas vezes já os encontramos grupados em distribuições de freqüências, donde os cálculos de $\overline{X}$ e S somente poderão ser feitos de forma aproximada.

Donde:[5]

$$\overline{X} = \frac{3.070}{40} = 76,75 \quad \text{e} \quad S = \sqrt{\frac{243.600 - (40)\cdot(76,75)^2}{n-1}} = 14,30$$

Média ponderada

O cálculo da média e do desvio padrão com ponderação pela freqüência é um caso particular de média e desvio padrão ponderados. Em geral, a ponderação é feita sempre que precisamos dar mais importância a um caso do que a outro. Por exemplo, a média aritmética simples dos valores do Índice de Desenvolvimento Humano (IDH) dos municípios da Microrregião da Grande Florianópolis, embora seja um valor central do IDH desses municípios, não corresponde ao IDH da Microrregião, porque temos municípios mais importantes (mais populosos) que outros. Para se ter o IDH da Grande Florianópolis, precisamos ponderar pela população do município, como segue:

Município	População p	IDH X	$X \cdot p$
Antônio Carlos	6.434	0,83	5.320,9
Biguaçu	48.077	0,82	39.327,0
Florianópolis	342.315	0,88	299.525,6
Governador Celso Ramos	11.598	0,79	9.162,4
Palhoça	102.742	0,82	83.837,5
Paulo Lopes	5.924	0,76	4.496,3
Santo Amaro da Imperatriz	15.708	0,84	13.241,8
São José	173.559	0,85	14.7351,6
São Pedro de Alcântara	3.584	0,80	2.849,3
Soma	709.941	7,37	605.112,5

Média simples: $\overline{X} = \frac{\sum X}{n} = \frac{7,37}{9} = 0,82$

Média ponderada: $\overline{X}_p = \frac{\sum (X \cdot p)}{\sum p} = \frac{605.112,5}{709.941} = 0,85$

[5] Se tivéssemos feito os cálculos diretamente com os 40 valores da taxa de alfabetização (ver capítulo anterior), encontraríamos $\overline{X} = 76,89$ e $S = 13,41$.

Exercícios

4) Dado o seguinte conjunto de dados: {7, 8, 6, 10, 5, 9, 4, 12, 7, 8}, calcule:
   a) a média e
   b) o desvio padrão.

5) Calcule a média e o desvio padrão da seguinte distribuição de freqüências:

Distribuição de freqüências do tamanho da família, numa amostra de 40 famílias do Conjunto Residencial Monte Verde, Florianópolis, SC, 1988.

Tamanho da família	Freqüência de famílias	Percentagem de famílias
1	1	2,5
2	3	7,5
3	6	15,0
4	13	32,5
5	11	27,5
6	4	10,0
7	0	0,0
8	2	5,0

6) Faça um histograma para a distribuição de freqüências da Tabela 6.4 e indique o valor da média aritmética no gráfico.

7) Considerando os dados do anexo do Capítulo 2, obtenha a média e o desvio padrão dos valores do índice de desempenho do aluno (item 5 do questionário), considerando:
   a) os dados do anexo do Capítulo 2 (cálculo exato);
   b) a tabela de distribuição de freqüências construída no Exercício 5 do capítulo anterior, (cálculo aproximado).

8) Sejam os dados do anexo do Capítulo 2.
   a) Calcule as médias e os desvios padrões das respostas dos itens 3(a) a 3(g) do questionário.
   b) Apresente os resultados numa tabela.
   c) Interprete, considerando os objetivos 1 e 3 da pesquisa (Seção 2.4, Capítulo 2).

9) Sejam os dados do anexo do Capítulo 4.
   a) Calcule a renda familiar média em cada uma das três localidades.
   b) Calcule o desvio padrão da renda familiar em cada localidade.
   c) Apresente esses resultados numa tabela.
   d) O que você pode concluir a partir desses resultados?

6.3 Medidas baseadas na ordenação dos dados

A média e o desvio padrão são as medidas mais usadas para avaliar a posição central e a dispersão de um conjunto de valores. Contudo, essas medidas são fortemente influenciadas por *valores discrepantes*. Por

exemplo, nas notas da Turma C (Exemplo 6.1), o valor discrepante 0 (zero) *puxa* a média para baixo, como ilustra a Figura 6.2. Apesar de a média aritmética ser 6 (seis), o diagrama de pontos sugere que o valor 7 (sete) seja um valor *mais típico* para representar as notas da turma, pois, além de ser o valor *mais freqüente*, ele é o *valor do meio*, deixando metade das notas abaixo dele e metade acima.

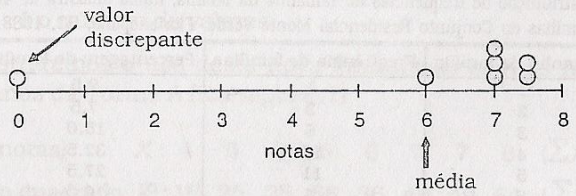


Figura 6.2 A influência de um valor discrepante no cálculo da média aritmética.

Nesta seção apresentaremos algumas medidas que são menos afetadas por valores discrepantes e, em conseqüência, são mais recomendadas para a análise de dados que possam conter valores discrepantes.

A mediana

A mediana avalia o centro de um conjunto de valores, sob o critério de ser o valor que divide a distribuição ao meio, deixando os 50% menores valores de um lado e os 50% maiores valores do outro lado. Por exemplo, o conjunto de valores {2, 3, 4, 5, 8} tem como mediana o valor 4 (quatro), porque a quantidade de valores com magnitude inferior a 4 é a mesma do que a quantidade de valores com magnitude superior a 4. Mais precisamente:

> Dado um conjunto de n valores, definimos ***mediana*** como o valor, M_d, que ocupa a posição $\frac{n+1}{2}$, considerando os dados ordenados crescente ou decrescentemente. Se $\frac{n+1}{2}$ for fracionário, toma-se como mediana a média dos dois valores de posições mais próximas a $\frac{n+1}{2}$.

Exemplos:

a) Conjunto de notas da Turma C: {0; 6; 7; 7; 7; 7,5 7,5}

$\Rightarrow$ posição: $\frac{n+1}{2} = 4 \Rightarrow M_d = 7.$

b) {5, 3, 2, 8, 4}

Ordenando: 2, 3, 4, 5, 8 $\Rightarrow$ posição: $\frac{n+1}{2} = 3 \Rightarrow M_d = 4.$

c) {3, 5, 6, 7, 10, 11} $\Rightarrow$ posição: $\frac{n+1}{2} = 3{,}5$ (3ª e 4ª) $\Rightarrow M_d = \frac{6+7}{2} = 6{,}5$

Comparação entre média e mediana

A Figura 6.3 mostra os valores da média e da mediana num diagrama de pontos. Note que o valor discrepante 62 *puxa* mais a média do que a mediana.

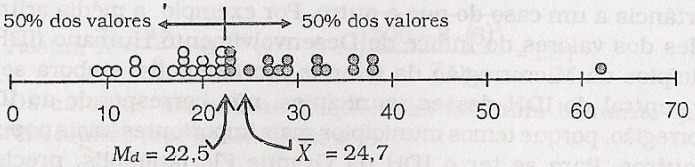


Figura 6.3 Posição da média e da mediana no diagrama de pontos das taxas de mortalidade infantil dos municípios da Microrregião Oeste de Santa Catarina, 1982.

A Figura 6.4 mostra as posições da média e da mediana em distribuições com diferentes formas: uma simétrica e outra assimétrica. No primeiro caso, a média e a mediana são iguias. Em distribuições assimétricas, a média tende a se deslocar para o lado da cauda mais longa.

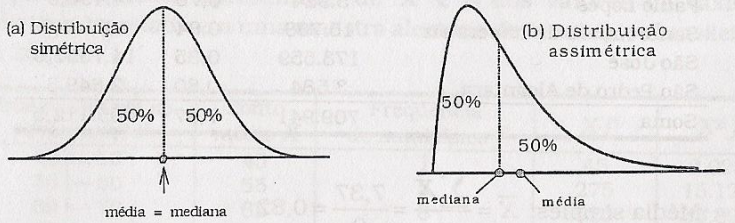


Figura 6.4 Posições da média e mediana, segundo a forma (simétrica ou assimétrica) da distribuição.

Em geral, dado um conjunto de valores, a média é a medida de posição central mais adequada, quando se supõe que estes valores tenham uma distribuição razoavelmente simétrica, enquanto que a mediana surge como uma alternativa para representar a posição central em distribuições

muito assimétricas.[6] Muitas vezes, calculam-se ambas as medidas para avaliar a posição central sob dois enfoques diferentes, como também para se ter uma primeira avaliação sobre a assimetria da distribuição.

Quartis e extremos

Na maioria dos casos práticos, o pesquisador tem interesse em conhecer outros aspectos relativos ao conjunto de valores, além de um valor central, ou valor típico. Algumas informações relevantes podem ser obtidas através do conjunto de medidas: *mediana, extremos e quartis*, como veremos a seguir.

Chamamos de *extremo inferior*, E_I, ao menor valor dos dados em análise. De *extremo superior*, E_S, ao maior valor. Por exemplo, dado o conjunto de valores {5, 3, 6, 11, 7}, temos $E_I = 3$ e $E_S = 11$.

Chamamos de *primeiro quartil* ou *quartil inferior*, Q_I, ao valor que delimita os 25% menores valores. De *terceiro quartil* ou *quartil superior*, Q_S, o valor que separa os 25% maiores valores. O *segundo quartil*, ou *quartil do meio*, é a própria mediana, que separa os 50% menores dos 50% maiores valores. Veja a Figura 6.5.

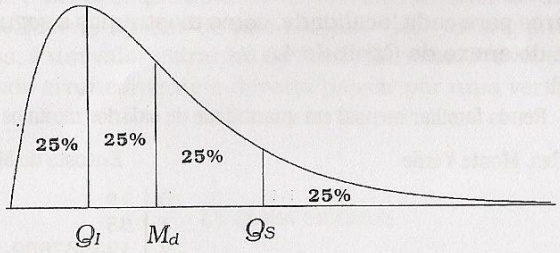


Figura 6.5 Os quartis dividem a distribuição em quatro partes iguais.

Dado um conjunto de valores ordenados, podemos obter, de forma aproximada, o quartil inferior, Q_I, como a mediana dos valores de posições menores ou iguais à posição da mediana. A mediana dos valores de posições maiores ou iguais à posição da mediana corresponde ao quartil superior, Q_S.[7] Se a mediana coincidir com um valor do conjunto de valores, vamos convencionar em considerá-la tanto no cômputo de Q_I como de Q_S.

[6] Mesmo para variáveis que supostamente tenham distribuições razoavelmente simétricas, a média e a mediana podem não se igualar, porque, em geral, estamos observando apenas alguns valores (amostras) dessas variáveis. Para variáveis com distribuições razoavelmente simétricas, a média é a medida de posição central mais adequada, porque usa o máximo de informações dos dados. A média é calculada usando a magnitude dos valores, enquanto a mediana utiliza somente a ordenação dos valores.

Exemplos:

a) Dados: 2, 0, 5, 7, 9, 1, 3, 4, 6, 8.

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

$Q_I = 2$ $M_d = 4,5$ $Q_S = 7$

b) Dados (já ordenados): 2 3 4 4 5 5 5 7 8 9 10

$Q_I = 4$ $M_d = 5$ $Q_S = 7,5$

Exemplo 6.7 Obtenção da mediana num *ramo-e-folhas*: valores referentes às taxas de alfabetização de quarenta municípios brasileiros, ano 2000.[8]

(1)	4	5
(2)	5	4
(6)	5	7789
(9)	6	444
(14)	6	56789
(17)	7	123
(20)	7	567
(20)	8	1112344
(13)	8	56789
(8)	9	01244
(3)	9	555

Unidade = 1
4|5 = 45

$n = 40 \Rightarrow$ posição: $\frac{n+1}{2} = 20,5$ (20ª e 21ª) $\Rightarrow M_d = \frac{77+81}{2} = 79$.

[7] Dado um conjunto de valores, nem sempre conseguimos dividi-lo exatamente em quatro partes iguais. O procedimento exposto oferece uma solução aproximada, mas bastante satisfatória quando a quantidade de valores for grande e com poucas repetições.

[8] No *ramo-e-folhas*, construído na seção 5.7, incluímos uma coluna à esquerda com as freqüências acumuladas. Essas freqüências foram acumuladas das extremidades até o centro (mediana) da distribuição, o que facilita a contagem das freqüências para a obtenção da mediana e quartis.

Para os quartis: $n' = 20 \Rightarrow$ posição 10,5 (10ª e 11ª). Daí:

$$Q_I = 65{,}5 \text{ e } Q_S = 87{,}5.$$

Podemos considerar o valor $M_d = 79$ como o *valor típico* das taxas de alfabetização dos quarenta municípios em estudo, pois metade dos municípios acusa taxa de alfabetização inferior a 79 e a outra metade tem níveis mais elevados de alfabetização. Com os quartis, podemos dizer que os 50% dos municípios mais típicos, em termos de alfabetização, acusam taxas variando de 65,5 a 87,5. Podemos dizer, também, que 25% desses municípios têm taxas de alfabetização não superiores a 65,5; enquanto 25% de municípios têm taxas iguais ou superiores a 87,5.

Esquema de cinco números

O esquema de cinco números é uma forma de apresentação da mediana, quartis e extremos, como mostramos ao lado. Através desses cinco números podemos ter informações sobre a posição central, dispersão e assimetria da distribuição de freqüências, como ilustra a Figura 6.6.

$n = 40$

M_d	79	
Q	65,5	87,5
E	45	95

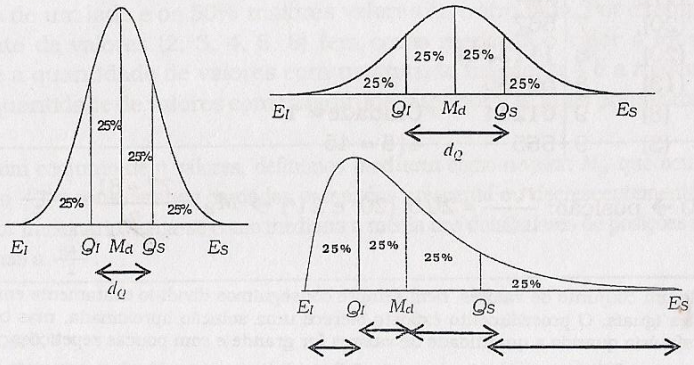


Figura 6.6 Posições da mediana, quartis e extremos em distribuições diferentes quanto à dispersão e assimetria.

O desvio entre quartis, $d_Q = Q_S - Q_I$, é muitas vezes usado como uma medida de dispersão. Veja na Figura 6.6 que, quanto mais dispersa a distribuição, maior será o valor de d_Q. Em distribuições mais dispersas, os valores dos quartis (e dos extremos) ficam mais distantes. Em distribuições simétricas, a distância entre o quartil inferior e a mediana é igual à distância entre a mediana e o quartil superior, enquanto que em distribuições assimétricas isto não acontece.

Uma regra muitas vezes usada para detectar valores discrepantes é verificar se existe algum valor do conjunto de dados que se afasta mais do que $(1{,}5)\cdot d_Q$ do quartil superior (ou inferior). No Exemplo 6.3, temos:

$$d_Q = Q_S - Q_I = 87{,}5 - 65{,}5 = 22$$
$$Q_I - (1{,}5)\cdot d_Q = 65{,}5 - (1{,}5)\cdot(22) = 32{,}5$$
$$Q_S + (1{,}5)\cdot d_Q = 87{,}5 + (1{,}5)\cdot(22) = 120{,}5$$

Como nenhum valor está fora do intervalo [32,5; 120,5], não temos valor suspeito de ser discrepante.

Exemplo 6.4 Com o objetivo de comparar as distribuições da renda familiar em duas localidades, construímos um *ramo-e-folhas* e um esquema de cinco números para cada localidade, como mostramos a seguir. Os dados fazem parte do anexo do Capítulo 4.

Renda familiar mensal em quantidade de salários mínimos

Conj. Res. Monte Verde		Encosta do Morro
1	1	0 | 19
2	1446	1 | 38
3	9	2 | 123367889
4	168	3 | 599999
5	11588	4 | 224569
6	8	5 | 188
7	12577	6 | 4
8	4469	7 | 19
9	6	
10	3349	
11		
12	25999	
13		
14		
15	4	

Unidade = 0,1
1|1 = 1,1

Discrepantes:
18|6 e 19|3

Discrepantes:
11|1, 11|4, 13|9 e 25|7

$n = 40$

M_d	7,7	
Q	4,95	10,35
E	1,1	19,3

$n = 37$

M_d	3,9	
Q	2,7	5,1
E	0,1	25,7

Notamos, inicialmente, que o nível de renda no Conjunto Residencial Monte Verde (mediana de 7,7 salários mínimos) é maior do que na Encosta do Morro (mediana de 3,9 salários mínimos). No Monte Verde, 50% das famílias mais típicas, em termos de renda, estão na faixa de 4,95 a 10,35 salários mínimos mensais; já na Encosta do Morro, as rendas familiares estão na faixa de 2,7 a 5,1 salários mínimos mensais.

A distribuição de renda na Encosta do Morro é mais concentrada em torno de um valor típico. Esta característica pode ser observada pelo desvio entre os quartis, d_Q, que é menor na Encosta do Morro do que no Monte Verde. O desvio entre extremos é maior na Encosta do Morro, mas tal desvio deve ser observado com cautela, pois em ambas as distribuições os extremos superiores são valores discrepantes em relação à maioria dos outros valores.

As duas distribuições são razoavelmente simétricas, quando observadas próximas de suas medianas, pois, em ambas as distribuições, as distâncias entre Q_I e M_d são próximas das distâncias entre M_d e Q_S. Contudo, fora do intervalo entre os quartis temos uma cauda mais longa do lado direito, mostrando que existem algumas poucas famílias com renda relativamente alta em relação ao típico destas localidades. O valor 0,1 salários mínimos, que aparece no extremo inferior da distribuição da Encosta do Morro, apesar de não ser um valor discrepante em termos estatísticos, é um valor estranho de renda familiar. Provavelmente tenha sido coletado erroneamente e deveria passar por uma verificação.

Diagrama em caixas

Uma maneira de apresentar aspectos relevantes de uma distribuição de freqüências é através do chamado *diagrama em caixas* ou *desenho esquemático*. Traçamos dois retângulos: um representando o espaço entre o quartil inferior e a mediana, e o outro entre a mediana e o quartil superior. Esses retângulos, em conjunto, representam a faixa dos 50% dos valores mais típicos da distribuição. Entre os quartis e os extremos traçamos uma linha. Caso existam valores discrepantes [valores inferiores a $Q_I - 1{,}5 \cdot d_Q$ ou superiores a $Q_S + 1{,}5 \cdot d_Q$], a linha é traçada até o último valor não discrepante; e os valores discrepantes são indicados por pontos (veja a Figura 6.7).

A Figura 6.8 mostra a forma do *diagrama em caixas* para uma distribuição simétrica e para uma distribuição assimétrica. Note as diferenças e imagine como ficaria um *diagrama em caixas* se tivéssemos uma distribuição mais dispersa.

A Figura 6.9 apresenta os *diagramas em caixas* das duas distribuições de renda do Exemplo 6.4. Compare esta representação com os *ramos-e-folhas* vistos anteriormente.

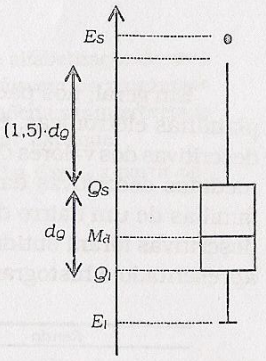


Figura 6.7 Esquema para construção de um diagrama em caixas.

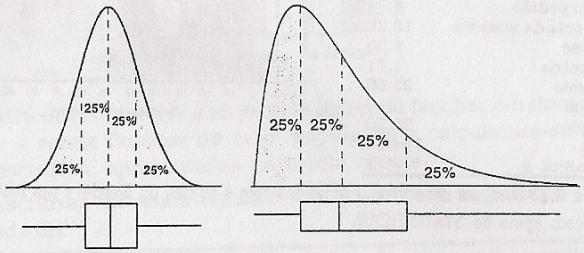


Figura 6.8 *Diagrama em caixas* e a forma da distribuição.

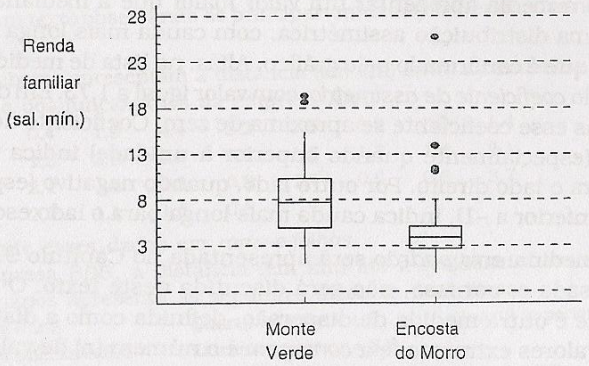


Figura 6.9 Representação em *diagramas em caixas* das distribuições de renda do Exemplo 6.4.

Uso do computador

Em geral, nos pacotes computacionais de estatística, ou mesmo em planilhas eletrônicas, é bastante simples obter um conjunto de medidas descritivas dos valores de uma variável quantitativa. A Figura 6.10 apresenta medidas descritivas da renda, em salários mínimos, de uma amostra de famílias de um bairro de Florianópolis (anexo do Capítulo 4). As medidas descritivas foram obtidas através da planilha eletrônica *Excel*®. Ao lado é apresentado o histograma de freqüências para facilitar a interpretação.[9]

Renda	
Média	6.34
Erro padrão	0.37
Mediana	5.40
Moda	3.90
Desvio padrão	4.03
Variância da amostra	16.26
Curtose	4.55
Assimetria	1.71
Intervalo	25.60
Mínimo	0.10
Máximo	25.70
Soma	754.50
Contagem	119

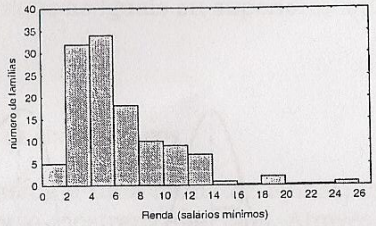


Figura 6.10 Medidas descritivas calculadas com o auxílio do Excel® e um histograma feito com apoio do *STATISTICA*®.

Em termos de posição central, temos a *média*, a *mediana* e a *moda*. Esta última medida apresenta o valor mais freqüente do conjunto de dados. O fato de a média apresentar um valor maior que a mediana e a moda, sugere uma distribuição assimétrica, com cauda mais longa para o lado direito, o que é confirmado pelo gráfico. Aliás, na lista de medidas, aparece o chamado *coeficiente de assimetria*, com valor igual a 1,73. Em distribuições simétricas esse coeficiente se aproxima de zero. Coeficiente de assimetria positivo (especialmente quando superior à unidade) indica cauda mais longa para o lado direito. Por outro lado, quando negativo (especialmente quando inferior a –1), indica cauda mais longa para o lado esquerdo.

A medida *erro padrão* será apresentada no Capítulo 9. A *curtose* é pouco usada e, por isso, não será discutida neste texto. O *intervalo* ou *amplitude* é outra medida de dispersão, definida como a distância entre os dois valores extremos; e a *contagem* é o número (n) de valores usados no cálculo das medidas descritivas.

[9] Sobre o uso do Excel, ver *Excel.doc* em www.inf.ufsc.br/~barbetta/livro1.htm. O histograma foi construído com o apoio do *STATISTICA*®. Ver www.statsoft.com.br.

6.4 Orientação para análise exploratória de dados

Na análise exploratória de grandes conjuntos de dados, é comum, inicialmente, construirmos uma distribuição de freqüências para cada variável, verificando os valores ou categorias típicas, possíveis casos discrepantes, etc. É a descrição ou caracterização dos dados em estudo. Lembramos que a construção da distribuição e a representação gráfica dependem do tipo de variável em estudo, em termos do nível de mensuração (ver Figuras 6.11).

Numa fase seguinte, é comum buscarmos possíveis relações (associações ou correlações) entre as variáveis em estudo. Os procedimentos também dependem do tipo das variáveis (ver Figura 6.12).

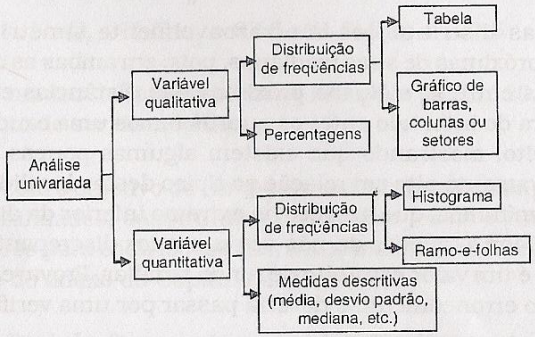


Figura 6.11 Esquema para análise de cada variável individualmente.

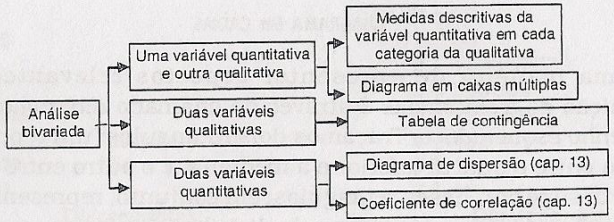


Figura 6.12 Esquema para análise entre pares de variáveis.

Exercícios

10) Calcule a mediana e os quartis dos seguintes dados:

a) {15, 9, 7, 20, 18, 19, 23, 32,14, 10, 11}

b) {15, 9, 7, 20, 18, 19, 23, 32,14, 10, 11, 16}

11) Obtenha a mediana e os quartis da distribuição de freqüências do Exercício 5 (Seção 6.2).

12) Considere o anexo do Capítulo 2:
   a) Obtenha a mediana, os quartis e os extremos dos valores do índice de desempenho do aluno (item 5 do questionário) e interprete. Sugestão: apresente, inicialmente, os dados num *ramo-e-folhas*.
   b) Comparando o valor da mediana com o valor que você obteve para a média aritmética no Exercício 7 (igual a 2,311), o que você diria sobre a simetria da distribuição desses valores?

13) A tabela abaixo mostra a distribuição de freqüências do número de filhos dos pais de alunos da UFSC, considerando uma amostra de 212 estudantes, entrevistados pelos alunos do Curso de Ciências Sociais, UFSC, 1990. Obtenha os extremos, a mediana e os quartis.

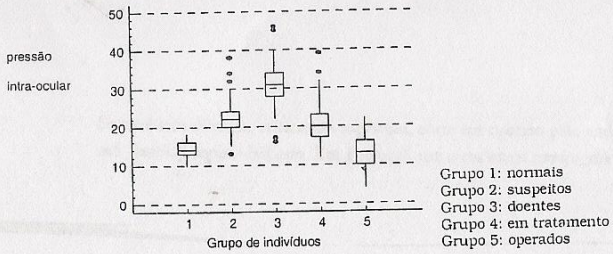
Nº de filhos	1	2	3	4	5	6	7	8	9	10	11	12
freqüência	10	45	32	50	23	23	9	7	6	2	3	2

14) A tabela seguinte é composta de medidas descritivas, calculadas a partir de quatro conjuntos de valores, oriundos de uma amostra de 212 estudantes da UFSC. Os estudantes foram indagados acerca do número de filhos que planejam ter, do número de filhos de seus pais, do número de filhos de seus avós maternos e do número de filhos de seus avós paternos.

Medidas descritivas	número de filhos			
	planejados	dos pais	dos avós maternos	dos avós paternos
média	2,06	4,23	6,35	6,15
desvio padrão	1,26	2,29	3,21	3,12
extremo inferior	0	1	1	1
quartil inferior	1	2	4	4
mediana	2	4	6	6
quartil superior	2	5	8	8
extremo superior	12	12	18	16

Faça uma redação comparando os quatro conjuntos de valores, tomando por base as medidas descritivas apresentadas na tabela.

15) A figura seguinte apresenta cinco distribuições de freqüências representadas por *diagramas em caixas*. São dados de pressão intra-ocular de uma amostra de 243 indivíduos, divididos em cinco grupos, segundo a condição clínica da doença *glaucoma*. Descreva as principais informações oriundas desta análise.

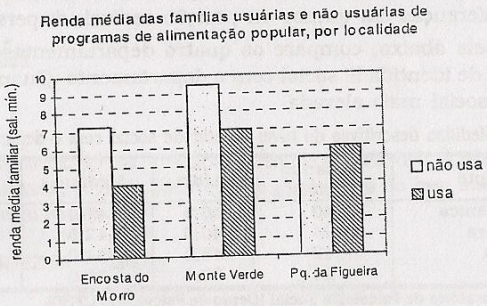


Grupo 1: normais
Grupo 2: suspeitos
Grupo 3: doentes
Grupo 4: em tratamento
Grupo 5: operados

Exercícios complementares

16) No Exemplo 6.2, calculamos a média aritmética da taxa de alfabetização de uma amostra de municípios brasileiros. Se esses municípios fossem os municípios de uma Unidade da Federação, o valor da média (76,75) poderia ser interpretado como a taxa de alfabetização dessa Unidade da Federação? Explique.

17) O gráfico seguinte foi construído com o auxílio da planilha *Excel*, a partir dos dados do anexo do Capítulo 4. Interprete.



18) Com o objetivo de comparar a distribuição da renda familiar em duas cidades, levantou-se a renda familiar de cada população e calcularam-se algumas medidas descritivas, apresentadas na tabela abaixo.

Medidas descritivas da renda familiar, em quantidade de salários mínimos, em duas cidades.

Cidade	média	desvio padrão	quartil inferior	mediana	quartil superior
A	4,8	3,2	3,4	4,9	6,5
B	4,9	6,2	3,0	3,8	9,0

Descreva um texto observando as principais informações verificadas nos dados da tabela.

19) Os dados abaixo apresentam a distância (em km) entre a residência e o local de trabalho dos funcionários da empresa AAA.

1,8	2,5	0,4	1,9	4,4	2,2	3,5	0,2	0,9	1,4
1,1	1,7	1,2	2,3	1,9	0,8	1,5	1,7	1,4	2,1
3,2	15,1	2,1	1,4	0,5	0,9	1,7	0,5	0,8	3,7
1,4	1,8	2,0	1,1	1,0	0,8				

   a) Apresente esses dados em *ramo-e-folhas*.
   b) Na empresa BBB, a distância (em km) até a residência dos seus 300 funcionários apresenta as seguintes medidas descritivas:
   Mediana = 2,8 Quartil inferior = 1,6 Quartil superior = 4,2
   Extremo inferior = 0,4 Extremo superior = 8,8
   Quais as principais diferenças entre as empresas AAA e BBB em termos da distância entre a residência e o local de trabalho dos funcionários?

20) Apresentamos, abaixo, algumas medidas descritivas da distribuição de salários, em R$, de três empresas de um certo ramo.

Empresa	média	desvio padrão	extremo inferior	quartil inferior	mediana	quartil superior	extremo superior
A	300	100	100	200	302	400	510
B	400	180	100	250	398	550	720
C	420	350	100	230	300	650	10.000

O que se pode dizer sobre a distribuição dos salários nas três empresas? Quais as diferenças em termos da posição central, dispersão e assimetria?

21) Dada a tabela abaixo, compare os quatro departamentos da UFSC quanto aos escores de identidade social com o departamento. Quanto maior o escore, identidade social mais elevada.

Medidas descritivas do nível identidade social com o departamento.

Depto	Tamanho da amostra	Média	Mediana	Desvio padrão
Eng. Mecânica	40	46,9	47,0	2,1
Arquitetura	24	40,8	42,5	5,9
Psicologia	19	42,5	44,0	5,4
História	21	38,4	39,0	5,4

Fonte: Laboratório de Psicologia Social (Depto de Psicologia/UFSC).

PARTE III

Modelos de probabilidade

COMO USAR MODELOS DE PROBABILIDADE PARA ENTENDER MELHOR OS FENÔMENOS ALEATÓRIOS