3 MEDIDAS RESUMO

3.2 Medidas de dispersão

Silvio Sandoval Zocchi

20 de agosto de 2008

Medida de **dispersão** é um valor que quantifica a variabilidade dos dados.

Algumas medidas de dispersão:

- Amplitude
- Amplitude Interquartílica
- Desvio absoluto médio
- Variância
- Desvio padrão
- Coeficiente de variação
- Índice de diversidade de Shannon-Wiener

3.2.1 Amplitude

Amplitude ou amplitude total de uma série de dados é a diferença entre o maior e o menor valor observado.

$$\text{Amplitude} = x_{[n]} - x_{[1]}$$

Exemplo: Resistências de uvas Niágara

Produtor 7		Produtor 58	
4		4	4
5	02489	5	4
6	134479	6	128
7	0233556666	7	4478
8	0223479	8	22445566667
9	249	9	234
10	044	10	37
11	0	11	034467
12		12	288
13		13	
14		14	
15		15	
16		16	
17		17	
18	0	18	

Legenda: 5|4=0,54 N

Produtor 7: Amplitude $= 1,80 - 0,50 = 1,30$N
Produtor 58: Amplitude $= 1,28 - 0,44 = 0,84$N

Defeito: a amplitude é muito afetada por dados atípicos extremos, ou seja, é uma medida muito pouco resistente à presença desses dados.

3.2.2 Amplitude Interquartílica

Amplitude interquartílica (AIQ) é a diferença entre o terceiro e primeiro quartis.

$$AIQ = Q_3 - Q_1$$

Produtor 7:
$Q_1 = P_{25} = \frac{x_{[9]} + x_{[10]}}{2} = \frac{0,64+0,67}{2} = 0,655$N
$Q_3 = P_{75} = \frac{x_{[27]} + x_{[28]}}{2} = \frac{0,87+0,89}{2} = 0,88$N
$AIQ = Q_3 - Q_1 = 0,88 - 0,655 = 0,225$N

Produtor 58:
$Q_1 = P_{25} = x_{[9]} = 0,78\text{N}$
$Q_3 = P_{75} = x_{[26]} = 1,10\text{N}$
$AIQ = Q_3 - Q_1 = 1,10 - 0,78 = 0,320\text{N}$

Interpretação: O produtor 7 apresentou cachos de uvas mais uniformes (ou menos variáveis) quanto à resistência.

A amplitude interquartílica é uma medida de dispersão pouco afetada por dados atípicos extremos.

Veremos agora medidas que visam quantificar o grau de variabilidade das observações em relação a um valor central.

3.2.3 Desvio absoluto médio

Desvio absoluto médio (D_m) de uma série de observações é a média dos valores absolutos dos desvios das observações em relação à média aritmética das mesmas.

$$\begin{aligned} D_m &= \frac{1}{n}\left(|x_1 - \bar{x}| + \cdots + |x_n - \bar{x}|\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}| \end{aligned}$$

Exemplo: Retardantes de crescimento (Tab.7).

Tratamento	X = Diâmetro vertical maior, em cm							
Controle	75	60	70	60	57	57	65	57
Clormequat	34	34	35	39	41	35	34	34
Uniconazole	33	32	31	37	28	31	38	32
Daminozide	55	80	65	76	60	70	60	72

Controle:

$$\bar{x} = \frac{75 + \cdots + 57}{8} = 62,625 \approx 62,6\text{cm}$$

$$
\begin{aligned}
D_m &= \frac{|75 - 62,625| + \cdots + |57 - 62,625|}{8} \\
&= \frac{12,375 + \cdots + 5,625}{8} = \frac{44,25}{8} \\
&= 5,53125 \approx 5,5\text{cm}
\end{aligned}
$$

Exercício: Calcular a média e o desvio absoluto médio para os dados relativos aos outros tratamentos.

Tratamento	Média ($\bar{x}$)	D_m
Controle	62,6	5,5
Clormequat	35,8	2,1
Uniconazole	32,8	2,4
Daminozide	67,3	7,3

Interpretação: Os retardantes de crescimento Clormequat e Uniconazole induziram a plantas mais baixas e com tamanhos mais uniformes (ou menos variáveis) do que as não tratadas (Controle).

3.2.4 Variância

Variância é a média dos quadrados dos desvios das observações, em relação à média aritmética das mesmas.

População conhecida: Seja μ a média dos N dados populacionais. Então, a variância é

$$\begin{aligned}\sigma^2 &= \frac{1}{N}[(x_1-\mu)^2+\cdots+(x_N-\mu)^2] \\ &= \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2\end{aligned}$$

Como geralmente temos acesso somente uma amostra de n dados extraídos da população, tem-se que um bom estimador de σ^2, chamado variância amostral, é dado por

$$\begin{aligned}s^2 &= \frac{1}{n-1}\left[(x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2\right] \quad (1)\\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \quad (2)\\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right] \quad (3)\end{aligned}$$

Exemplo: Retardantes de crescimento

Controle:

$$\begin{aligned} s^2 &= \frac{(75-62{,}625)^2+\cdots+(57-62{,}625)^2}{8-1} \\ &= \frac{(12{,}375)^2+\cdots+(-5{,}625)^2}{7} = \frac{321{,}875}{7} \\ &= 45{,}98\text{cm}^2 \end{aligned}$$

Note que

$$\begin{aligned} \sum_{i=1}^{n} x_i &= 75+\cdots+57 = 501 \\ \sum_{i=1}^{n} x_i^2 &= 75^2+\cdots+57^2 = 31697 \end{aligned}$$

Logo, usando a expressão (3) temos que:

$$\begin{aligned} s^2 &= \frac{1}{8-1}\left(31697-\frac{501^2}{8}\right) \\ &= \frac{321{,}875}{7} = 45{,}98\text{cm}^2 \end{aligned}$$

Exercício: calcular as variâncias para os outros tratamentos.

- Controle: $s^2 = 45{,}98\text{cm}^2$
- Clormequat: $s^2 = 7{,}36\text{cm}^2$
- Uniconazole: $s^2 = 10{,}79\text{cm}^2$
- Daminozide: $s^2 = 75{,}64\text{cm}^2$

Observação: A variância não possui a mesma unidade dos dados originais!

Vejamos agora, como calcular as variâncias para **dados agrupados em tabelas de freqüências**

		Freqüência
j	x_j	f_j
1	x_1	f_1
2	x_2	f_2
$\ldots$	$\ldots$	$\ldots$
k	x_k	f_k
	Total	$n = \sum_{j=1}^{k} f_j$

Exemplo: Avaliação do desempenho de semeadoras manuais (Molin *et al*, 2001)

semeadora manual

Um grande número de pequenas propriedades rurais no Brasil utiliza semeadoras manuais para a operação de semeadura

Objetivo do trabalho: avaliar as semeadoras existentes no mercado, classificando-as quanto à *regularidade* de vazão dos seus mecanismos dosadores

Método: as semeadoras foram inicialmente reguladas de modo que a cair duas sementes de milho por golpe e em seguida, usando-se um sistema mecanizado, simularam o efeito de uma pessoa operando cada semeadora 150 vezes. Na seqüência, anotaram o *número de sementes por golpe*.

Resultados: Freqüências observadas do número de sementes caídas por golpe para a semeadora manual A

	Nº de sementes por golpe	Freqüência
j	x_j	f_j
1	0	0
2	1	19
3	2	103
4	3	18
5	4	3
$k = 6$	5	7
	Total	n =150

Fórmula:

$$
\begin{aligned}
s^2 &= \frac{1}{n-1} \sum_{j=1}^{k} f_j (x_j - \bar{x})^2 \\
&= \frac{1}{n-1} \left[\sum_{j=1}^{k} f_j \; x_j^2 - \frac{\left(\sum_{j=1}^{k} f_j \; x_j \right)^2}{n} \right]
\end{aligned}
$$

Exemplo:

$$\begin{aligned}
\sum_{j=1}^{k} f_j \; x_j &= f_1 x_1 + f_2 x_2 + \cdots + f_k x_k \\
&= 0 \times 0 + 19 \times 1 + \cdots + 7 \times 5 = 326 \\
\sum_{i=1}^{n} f_j \; x_i^2 &= f_1 x_1^2 + f_2 x_2^2 + \cdots + f_k x_k^2 \\
&= 0 \times 0^2 + 19 \times 1^2 + \cdots + 7 \times 5^2 \\
&= 816
\end{aligned}$$

$$\bar{x} = \frac{326}{150} = 2{,}173 \text{ sementes}$$

$$\begin{aligned}
s^2 &= \frac{1}{150-1}\left(816 - \frac{326^2}{150}\right) = \frac{107{,}49333}{149} \\
&= 0{,}721 \text{ sementes}^2
\end{aligned}$$

Exercício: calcular a média e a variância dos números de sementes por golpe para as semeadoras B, C e D (Tabela 11)

Nº de sementes por golpe	Semeadora			
	A	B	C	D
0	0	2	14	7
1	19	26	21	19
2	103	70	82	82
3	18	48	27	42
4	3	4	5	0
5	7	0	1	0
Total	150	150	150	150

Exemplo: Presença em sala de aula no 1º semestre do curso, em %, segundo a intenção do estudo (somente para passar, ou não)

Presença (%)	Ponto médio	Estudou só para passar?	
		Não	Sim
]60; 70]	65	0	11
]70; 80]	75	7	30
]80; 90]	85	22	26
]90;100]	95	40	25
Total		69	92

Calcular a média e a variância da presença em sala de aula segundo a intenção do estudo

Presença (%)	Estudou só para passar?	
	Não	Sim
Média ($\bar{x}$)	89,78	82,07
Variância (s^2)	45,91	100,08

3.2.5 Desvio padrão

Desvio padrão é a raiz quadrada da variância.

$$s = \sqrt{s^2}$$

Tem a vantagem de possuir a mesma unidade dos dados originais.

Exemplo: semeadora manual A

Variância	$s^2 = 0{,}721$ sementes^2
Desvio padrão	$s = \sqrt{0{,}721} = 0{,}849$ sementes

Nota: Para dados com distribuição unimodal simétrica com formato de sino, espera-se que a grande maioria (aproximadamente 95%) do dados pertençam ao intervalo

$$[\bar{x} - 2s; \bar{x} + 2s]$$

3.2.6 Coeficiente de variação

$$CV = \frac{s}{\overline{x}} 100\%$$

É uma medida de dispersão relativa.

Serve para comparar as dispersões de diferentes variáveis.

Exemplo: Qual variável apresenta maior variabilidade? Altura ou Peso?

	Altura (cm)	Peso (kg)
Média	172,8	68,1
Desvio padrão	12,1	9,82
CV (%)	7,0	14,4

3.2.7 Índice de diversidade de Shannon-Wiener

É uma medida de dispersão adequada para variáveis qualitativas. É dada por:

$$H' = \log n - \frac{1}{n} \sum_{j=1}^{k} f_j \ \log f_j$$

sendo: k = número de categorias com freqüências não nulas; f_j = freqüência da j-ésima categoria e n = número total de observações.

Exemplo: Moradia de 162 alunos segundo o sexo.

	Sexo	
Moradia	F	M
Com até 2 colegas	7	27
Com 3 colegas ou mais	21	59
Família	5	15
CEU ou Vila	7	6
Outros	2	13
Total geral	42	120

Sexo feminino:

$$\begin{aligned} H' &= \log 42 - \frac{1}{42}(7\log 7 + \cdots + 2\log 2) \\ &= 0,583 \end{aligned}$$

Sexo masculino:

$$\begin{aligned} H' &= \log 120 - \frac{1}{120}(27\log 27 + \cdots + 13\log 13) \\ &= 0,580 \end{aligned}$$

Como o valor máximo que H' pode atingir é $H'_{max} = \log k$, podemos utilizar para a comparação de variáveis com diferentes números de categorias, o índice H' padronizado, também chamado *índice de eqüitabilidade*, dado por

$$J = \frac{H'}{\log k}$$

Exemplo:
Sexo feminino: $J = \frac{0,583}{\log 5} = 0,833$

Sexo masculino: $J = \frac{0,580}{\log 5} = 0,830$

Exercício: Comparar as diversidades de gêneros de nematóides antes e depois do plantio.

Tabela 10. Números de nematóides de solo, segundo o gênero e a época do ano.

	Época do ano		
Gênero	Antes do plantio	Depois da colheita	Total
Helicotylenchus	13972	29997	43969
Pratylenchus	1312	1924	3236
Mesocriconema	2044	1164	3208
Trichodorus	1088	970	2058
Scutellonema	720	980	1700
Heterodera	16	332	348
Meloidogyne	48	64	112
Aphelenchoides	108	0	108
Rotylenchulus	0	56	56
Xiphinema	8	0	8
Total	19316	35487	54803

Exercício. Qual variedade apresenta intensidade de odor menos variável?

Tabela. Distribuição de freqüências da intensidade do odor de pedaços de abacaxi, emitidas por 40 provadores, segundo a variedade.

Intensidade do odor	Variedade de abacaxi	
	Smooth Cayenne	Pérola
muito fraco	2	5
fraco	9	14
médio	15	12
forte	10	8
muito forte	4	1
Total	40	40

Resposta:

	Variedade	
Estatística	Cayenne	Pérola
H'	0,621	0,609
J	0,889	0,872

Apresentam dispersões semelhantes

Considerações adicionais

Como detectar dados atípicos de uma maneira simples e rápida?

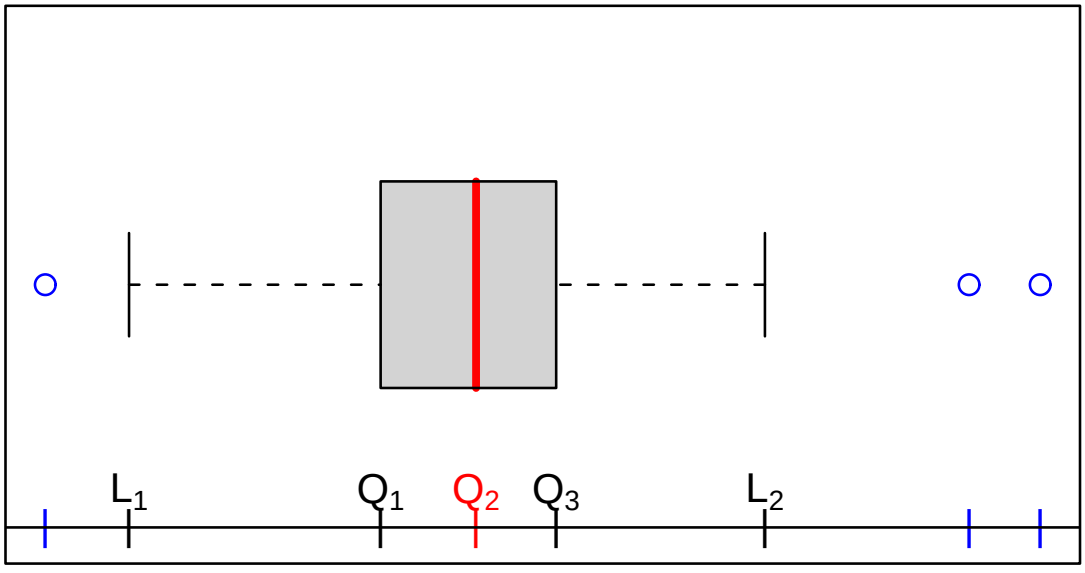
Como representar as informações fundamentais de conjuntos de dados por meio de um desenho?

Uma solução, proposta por Tukey (1970, 1979), é construir o chamado **gráfico de caixa** (ou "box-plot", em inglês), que fornece as seguintes informações:

- presença ou não de valores atípicos
- medida de posição: mediana
- medidas de dispersão: amplitude interquartílica e amplitude dos dados sem considerar os dados atípicos
- classificação da distribuição quanto à simetria

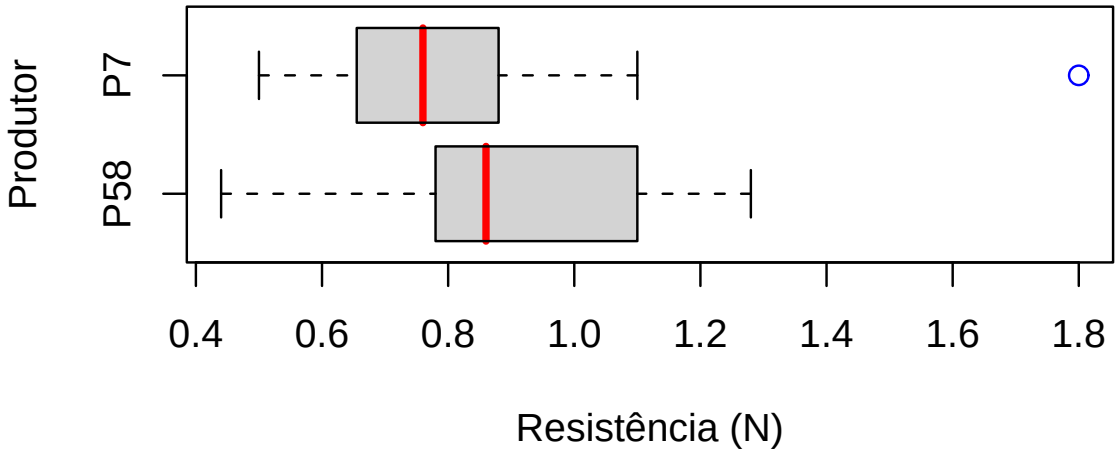
Procedimento para a construção do gráfico de caixa:

1. Calcular os quartis: Q_1, $Q_2 = Md$ e Q_3 e a amplitude interquartílica $AIQ = Q_3 - Q_1$
2. Verificar se há dados atípicos, isto é, se há valores não pertencentes ao intervalo $[Q_1 - 1,5 \times AIQ; Q_3 + 1,5 \times AIQ]$
3. Calcular o menor (L_1) e maior (L_2) valores **sem considerar os atípicos**
4. Construir o gráfico de caixa seguindo o esquema abaixo



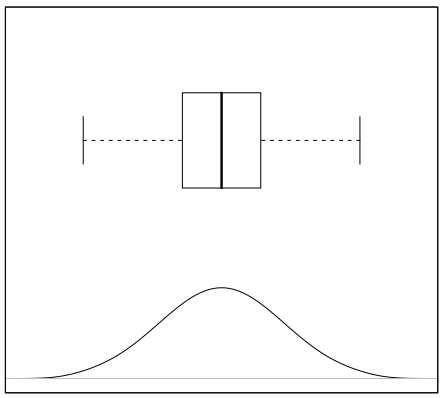
Exemplo: Resistências de uvas Niágara

Estatísticas	Produtor P7	Produtor P58
1º Quartil (Q_1)	0,655	0,780
Mediana (Md)	0,760	0,860
3º Quartil (Q_3)	0,880	1,100
Ampl. interq. (AIQ)	0,225	0,320
$Q_1 - 1,5 \times AIQ$	0,3175	0,300
$Q_3 + 1,5 \times AIQ$	1,2175	1,580
Dados atípicos	1,80	não há
L_1	0,50	0,44
L_2	1,10	1,28

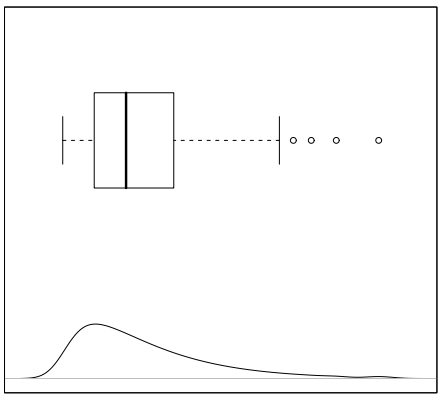


Classificação da distribuição quanto ao formato

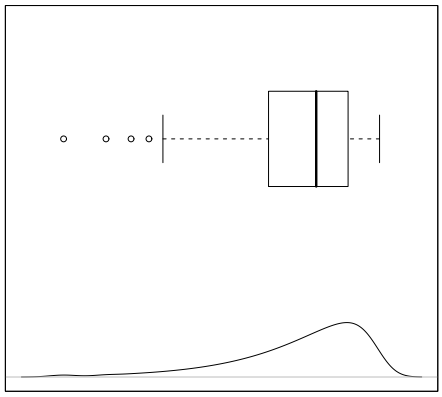
Simétrica



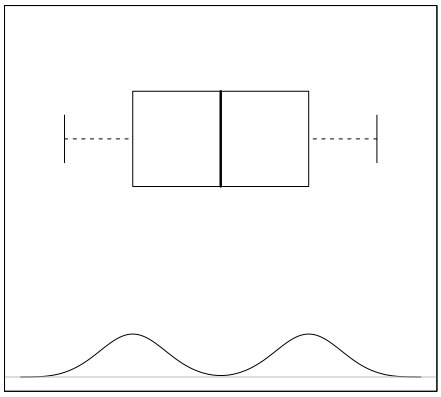
Assimétrica à direita



Assimétrica à esquerda



Bimodal



Exemplo: Densidade do solo em função do manejo