Igor Costa

Métodos estatísticos aplicados à psicolinguística experimental

8 de agosto de 2021

…to understand what you are seeing,
you need to know something about
how you would approach the
problem by hand…

Howell, 2010:462

1 Estatística descritiva: começando do básico

Imaginemos dois conjuntos de dados, digamos, as medidas do tempo de reação (RT, do inglês *reaction time*) para duas amostras (A e B) retiradas de populações distintas. Os dados amostrais estão na Tabela 1.1.

Sujeito	A	B
1	175.56	253.84
2	183.46	210.67
3	193.83	215.57
4	209.54	237.16
5	211.8	214.41
6	192.31	261.46
7	233.17	190.66
8	202.85	224.17
9	165.61	234.11
10	232	258.71
11	220.38	256.35
12	183.03	202.75
13	154.3	269.16
14	168.33	212.38
15	187.29	267.38

Tabela 1.1 – Valores observados para amostras A e B

Para começar, podemos fazer uma abordagem gráfica dos dados, dispondo o valor da variável independente (RT) no eixo y e os sujeitos ou as condições no eixo x, o que nos permite fazer uma abordagem mais ou menos precisa dos dados segundo nossos interesses.

O painel 1 é pouco informativo sobre os dados, já que não nos permite visualizar as condições experimentais. O painel 2 nos mostra com mais clareza a distribuição dos dados para a amostra A (laranja) e para a amostra B (azul), sugerindo que os tempos de reação

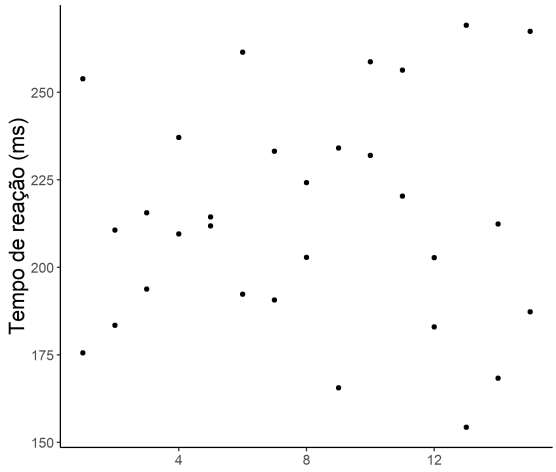

Figura 1.1 – Gráfico de dispersão (*scatterplot*) para amostras A e B

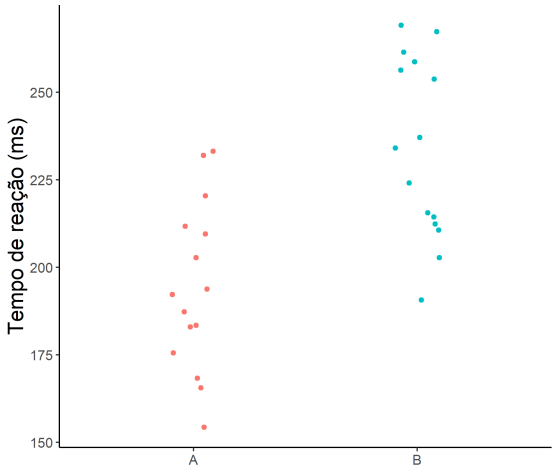

Figura 1.2 – Gráfico de dispersão com amostras A e B agrupadas

para B parecem ser maiores do que os tempos para A. O painel 3, então, agrupa os pontos para cada condição, ignorando os sujeitos, o que deixa ainda mais claro que, apesar de muitos valores sobrepostos, há a sugestão de uma leve tendência de B ser maior do que A.

Vamos, então, fazer uma abordagem desses dados considerando dois aspectos: primeiro, os pontos em torno dos quais esses dados se concentram, chamados, em estatística descritiva, *medidas da tendência central* dos dados ou *medidas de posição*; em segundo lugar, o modo como esses dados se espalham em torno desses pontos, que chamaremos de *medidas de dispersão*.

1.1 Medidas da tendência central

1.1.1 Média

O mais comum dos pontos de posição é a média aritmética dos dados, que consiste na soma dos n elementos amostrados para cada condição e a sua divisão pelo número total de observações, ou seja:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Vamos nos deter brevemente na explicação dessa fórmula, que pode parecer assombrosa para muitos, mas que simplesmente nos diz o seguinte: se x é uma variável qualquer que apresenta n valores x_i, ..., x_n, então a média de x (que vamos representar por $\bar{x}$, lido *xis barra*) é o somatório ($\sum_{i=1}^{n} x_i$) de todos os valores de x, de x_i, tal que i=1 (o primeiro valor), até x_n (o enésimo valor) multiplicado por 1 sobre n, que é o mesmo que dizer: some-se todos os valores e divida por n. Por exemplo, se x é o conjunto de dados abaixo:

$$x = \{2, 3, 5, 6, 8, 9, 2, 4, 7, 4\}$$

Então $x_1 = 2$, $x_2 = 3$, $x_3 = 5$, ...,$x_{10} = 4$. Assim, a média de x, é dada por:

$$\bar{x} = \frac{2+3+5+6+8+9+2+4+7+4}{10} = \frac{50}{10} = 5$$

Se a variável x estivesse disposta em uma tabela vertical (a mais comum para a análise de dados), poderíamos dizer que o índice subscrito a cada x seria cada uma das linhas da Tabela 1.2:

Isso pode parecer óbvio para a maioria dos leitores, mas é preciso deixar claro desde já sobre o que estamos falando, para que, mais a frente, quando estivermos lidando com fórmulas mais complexas, com mais índices subscritos, sobretudo nas fórmulas de somatórios das Análises de Variância, essa notação não seja motivo para complicação no entendimento.

Linha	x	
1	2	x_1=2
2	3	x_2=3
3	5	x_3=5
4	6	x_4=6
⋮	⋮	⋮
10	4	x_{10}=4

Tabela 1.2 – Relação entre linhas da tabela e índices de variável x

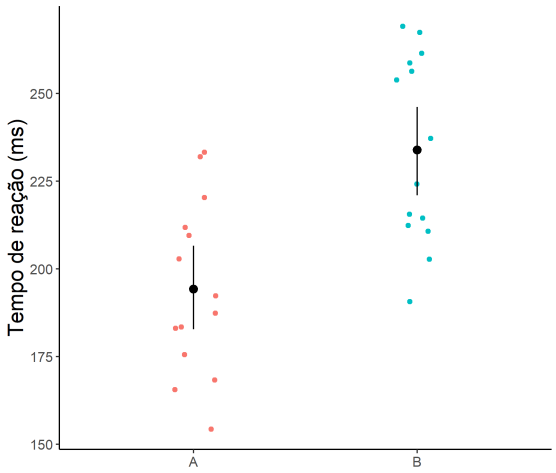

Figura 1.3 – Gráfico de dispersão com amostras agrupadas e médias de cada grupo

Isso tendo sido dito, passemos a uma análise inicial dos valores contidos nas amostras A e B. Para as amostras A e B, as médias dos tempos de reação são, respectivamente: 194.2 e 233.9 milisegundos – uma confirmação, até certo ponto, de que o valor do RT para B realmente é mais alto do que o valor para A. Essa diferença pode ser representada no gráfico de dispersão. Pontos pretos indicam as médias de cada grupo e as barras o erro padrão das médias – algo que discutiremos no capítulo sobre inferência estatística:

Repare no gráfico de pontos que a média de cada amostra é um valor mais ou menos central dos dados, indicando o que se chama de um valor típico. Isso acontece porque os dados amostrais em questão têm uma importante propriedade estatística: são normalmente distribuídos. Nesses casos, a média é um dos melhores, se não o melhor, valor para representar a amostra. Por isso, é uma das medidas mais difundidas. No entanto, observe que a média é apenas um ponto no meio de toda a multidão de dados apresentados. Há valores muito maiores e muito menores do que ela. Por isso, como veremos mais a frente, considerar apenas esse ponto como único descritor dos dados é algo que não deve ser feito, pois é muito redutor da realidade. Além disso, apesar de ser uma boa medida dos valores típicos, a média pode, em alguns casos, ser problemática, pois é facilmente influenciada

pelos valores extremos.

Por exemplo: a média da série

$$x = \{2, 3, 5, 6, 8, 9, 2, 4, 7, 4\}$$

é 5. No entanto, se acrescentarmos o valor 30 (um único item) a essa série, a média passa a ser 7.27. Se esse item (30) for trocado por 60, um valor ainda mais discrepante do restante, a média passa a ser 10, um valor nada típico da série em questão. Por ser facilmente influenciada por uma parcela pequena dos dados, a média é dita uma *medida pouco robusta* ou pouco resistente. Por isso, em alguns casos, em lugar da média, usa-se a mediana, uma medida robusta, ou seja, resistente a esses valores extremos.

1.1.2 Mediana

Tomemos, novamente, as três séries de dados usadas no último parágrafo, reapresentadas abaixo, agora como x, y e z:

$$x = \{2, 3, 5, 6, 8, 9, 2, 4, 7, 4\}$$
$$y = \{2, 3, 5, 6, 8, 9, 2, 4, 7, 4, 30\}$$
$$z = \{2, 3, 5, 6, 8, 9, 2, 4, 7, 4, 60\}$$

Para cada uma dessas séries, a mediana é o valor responsável por dividir a série ao meio. Para calculá-la, precisamos ordenar cada série em ordem crescente:

$$x = \{2, 2, 3, 4, 4, 5, 6, 7, 8, 9\}$$
$$y = \{2, 2, 3, 4, 4, 5, 6, 7, 8, 9, 30\}$$
$$z = \{2, 2, 3, 4, 4, 5, 6, 7, 8, 9, 60\}$$

Como y e z têm uma quantidade ímpar de valores (11 números), a mediana é dada simplesmente pelo valor central aos dados, o número que deixa 5 valores abaixo e 5 valores acima, ou seja, 5:

$$y = \{2, 2, 3, 4, 4, \}5, \{6, 7, 8, 9, 30\}$$
$$z = \{2, 2, 3, 4, 4, \}5, \{6, 7, 8, 9, 60\}$$

Contudo, para x, que apresenta 10 itens, isso não pode ser feito. Então, a mediana é dada pelo ponto médio entre os dois valores centrais. Os valores centrais de x são 4 e 5. A média de 4 e 5 é 4.5. Então, a mediana desses dados é 4.5.

$$x = \{2, 2, 3, 4, \}4, 5, \{6, 7, 8, 9\}$$
$$x = \{2, 2, 3, 4, 4, \}(4.5), \{5, 6, 7, 8, 9\}$$

Se você quiser uma fórmula, pode usar as seguintes:

$$Md(x) = \begin{cases} x_{\frac{n+1}{2}} & \text{para } n \text{ ímpar;} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{para } n \text{ par.} \end{cases}$$

Vamos nos deter brevemente nesses valores e comparar com a média obtida para as mesmas séries.

	x	y	z
Média	5	7.27	10
Mediana	4.5	5	5

Tabela 1.3 – Relação entre média e mediana

Como vemos da comparação, a média foi consideravelmente alterada por um único valor extremo: à medida que o valor extremo se torna maior, maior é a média dos dados. No entanto, a mediana não o foi, mantendo a representatividade da amostra mesmo nesses casos de valores extremos.

A mediana também é uma medida importante porque ela será, junto com a noção de quantil, um importante descritor da distribuição dos dados, como veremos em seguida.

1.1.3 Voltando ao exemplo inicial

Para as amostras A e B com que estamos trabalhando, as medianas e as médias estão dadas abaixo. (Não vamos calcular passo a passo a mediana para aqueles conjuntos de dados, uma vez que é uma tarefa trabalhosa e os softwares de estatística o fazem com muito mais rapidez. Se o leitor desejar, pode calcular esses valores numa planilha como a do Excel ou semelhante, ordenando os valores de cada amostra e encontrando os valores centrais).

	A	B
Média	194.2	233.9
Mediana	192.3	234.1

Tabela 1.4 – Relação entre média e mediana de A e B

Observe que, para esses dados, os valores da mediana e da média são muito próximos. Como veremos adiante, esse é um sinal de que nossos dados se distribuem simetricamente em torno dos valores centrais.

1.2 Medidas de dispersão

Como vimos acima, a média e a mediana são medidas que buscam "resumir" os dados com o auxílio de um único valor numérico, um valor "típico". No entanto, como já dissemos, esse tipo de análise é muito redutor da realidade, já que sempre existem valores muito acima e/ou muito distantes dessas medidas de posição. Por isso, precisamos olhar, também, para o quanto o conjunto total de dados sendo descritos se afasta dessas medidas de posição, ou seja, como os dados se dispersam.

1.2.1 Quantil e quartil

Um quantil é qualquer porcentagem dos dados. Normalmente, dividem-se os dados, após ordenados, em 4 partes, nos dando os quartis (1º, 2º e 3º quartis). Retomemos as séries de dados x e y para explicar esse conceito.

$$x = \{2, 3, 5, 6, 8, 9, 2, 4, 7, 4\}$$

$$y = \{2, 3, 5, 6, 8, 9, 2, 4, 7, 4, 30\}$$

O 2º quartil, o valor que divide os dados ao meio, é, obviamente, o valor da mediana. Logo, para x, $q_2 = 4,5$ e, para y, $q_2 = 5$. O mínimo e o máximo são, respectivamente, o menor e o maior valor de cada série de dados. Para x, 2 e 9; e, para y, 2 e 30.

$$x = \{\mathbf{2}, 2, 3, 4, 4, \}(4, 5)\{5, 6, 7, 8, \mathbf{9}\}$$

$$y = \{\mathbf{2}, 2, 3, 4, 4\}5\{6, 7, 8, 9, \mathbf{30}\}$$

Quanto ao 1º e 2º quartis, existem diversos métodos para calculá-los, inclusive métodos para estimar os quartis de uma população a partir de uma amostra, o que nos dá resultados diferentes. Usemos o mais fácil deles, que é simplesmente definir q_1 como o valor que divide a primeira metade dos dados (do Mínimo até q_2) ao meio; e q_3 como o valor que divide a segunda metade dos dados (de q_2 até o Máximo) ao meio. Observe os valores calculados na Tabela 1.5:

	Mínimo	q_1 25%	q_2 50%	q_3 75%	Máximo
x	2	3	4.5	7	9
y	2	3	5	8	30

Tabela 1.5 – Quartis para x e y

Tendo descoberto os quartis, temos uma visão global dos dados. Assim, conhecemos os valores centrais da amostra ou população que estamos estudando, ou seja, os valores que se encontram entre o 1º e o 3º quartis, excluindo-se, portanto, os extremos. Essa diferença

é chamada de Amplitude Interquartil (AIQ = q_3 – q_1). Essa sumarização dos dados nos dá uma visão mais global dos valores com que estamos trabalhando, mostrando como os valores encontrados se distribuem. Nos casos acima, temos que AIQ(x) = 7 – 3 = 4; e AIQ(y) = 8 – 3 = 5. Assim, y parece ter uma dispersão levemente maior do que x.

1.2.2 Valor atípico ou *outlier*

Observe, porém que, na análise de x e y, temos um problema. Isso porque y tem uma distribuição muito próxima de x – aliás, são exatamente os mesmos valores de x, não fosse um único valor de y (30), que é muito discrepante de todos os outros valores dessa série. Nesse caso, podemos verificar se 30 é o que se chama de *outlier* ou valor atípico. Um valor atípico é normalmente calculado tendo por base os quartis e a Amplitude Interquartil, e estão situados fora dos limites dos valores típicos. Esse limites são dados pelas fórmulas:

$$\text{limite inferior} = q_1 - (1.5) \times AIQ$$

$$\text{limite superior} = q_1 + (1.5) \times AIQ$$

Como considera a Amplitude Interquartil (o valor que descreve a maioria dos dados do conjunto), essa fórmula nos permite verificar aquilo que se afasta muito desses valores esperados. Assim, para x, temos que os limites inferior e superior:

$$\text{limite inferior}(x) = 3 - (1.5) \times 4 = -3$$

$$\text{limite superior}(x) = 3 + (1.5) \times 5 = 13$$

Assim, para x, qualquer valor que esteja fora do intervalo -3; 13 é considerado um *outlier* e, para y, qualquer valor que esteja fora do intervalo -4.5; 15.5 é também considerado um *outlier*. Esse é o caso, por exemplo, de 30, que está muito acima desse limite. Então, pelo menos em termos matemáticos, estamos lidando com um *outlier*.

Antes de continuar, gostaríamos de fazer um breve comentário no que diz respeito aos valores atípicos. Dissemos acima que, em *termos matemáticos*, estamos diante de um *outlier*. Isso pode não ser verdade em termos teóricos. Isso porque um valor atípico "verdadeiro" é um valor que ocorreu por um problema qualquer, como uma mensuração equivocada, um erro no programa de computador que media o RT, um sujeito distraído durante a realização de um experimento, etc. Caso a medida tenha realmente surgido nos dados, ela não é um *outlier*, mas uma realização real que precisa ser explicada pelo pesquisador.

Por exemplo: imaginemos que um nutricionista mediu a massa (em kg) de uma população qualquer de adultos e encontrou q_1 = 50 kg, e q_3 = 100 kg. Nesse caso, a AIQ é 50 kg (100 kg – 50 kg) e o limite superior é 100 + 1,5 (50) = 175 kg. Assim, qualquer valor acima desse seria considerado um *outlier*. No entanto, o pesquisador efetivamente

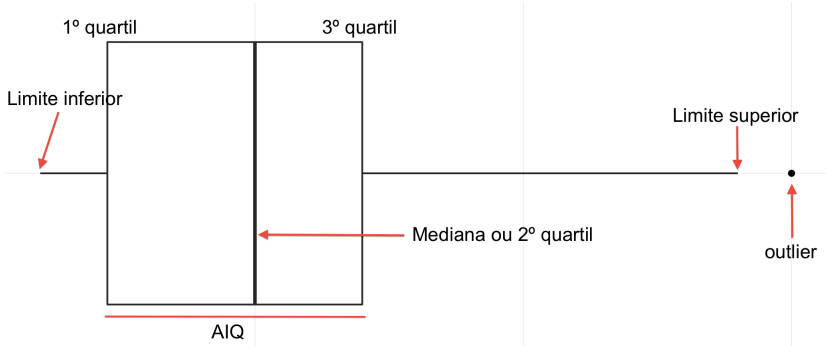

Figura 1.4 – Elementos componentes do gráfico de caixas ou *boxplot*

verificou que, nessa população, havia duas pessoas que tinham massas corporais acima desse valor. Ora, esses casos são raros, mas efetivamente ocorrem. O pesquisador não pode simplesmente excluir tais mensurações, ignorando-as. É preciso que ele as explique e, se desejar excluí-las da análise estatística, deve dar uma boa justificativa teórica para tal.

1.2.3 *Boxplots* e a representação dos quartis

Isso tendo sido colocado, podemos passar à próxima etapa, que é a apresentação e o entendimento da importância das medidas até agora apresentadas. A sumarização dos dados como proposta acima é normalmente feita com um tipo de gráfico próprio, chamado de gráfico de caixas, ou gráfico de caixa e bigodes (em inglês, *boxplot* ou *box and whisker plot*), que é a apresentação, em forma gráfica, das medidas até agora discutidas (mediana ou q_2; q_1; q_3; limite inferior; limite superior; e *outliers*), como ilustrados na imagem abaixo. A caixa, portanto, representa a Amplitude Interquartil (AIQ), ou seja, os dados mais frequentes, contidos entre o 1º e o 3º quartis e sendo cortada pela linha que representa a mediana ou 2º quartil. Os "bigodes", as linhas que saem da caixa para os extremos, vão até os limites superior e inferior, além dos quais pode ou não haver um ou mais *outliers*, representados por um ou mais pontos.

O *boxplot*, porém, não é apenas uma apresentação visual das medidas que até agora vislumbramos, mas também uma representação gráfica da curva de frequência dos dados coletados, indicando se os dados se distribuem simetricamente em torno da média e da mediana ou se os dados estão distribuídos assimetricamente (assimetria positiva – à esquerda; ou assimetria negativa – à direita), como demonstram as imagens nas páginas seguintes. Isso ocorre porque a caixa do *boxplot* mostra a concentração dos dados mais frequentes, ou seja, 50% dos dados coletados estão no intervalo delimitado pela caixa. Se a distribuição é simétrica, a caixa se encontra no centro dos "bigodes" e a mediana divide a caixa ao meio. Se a distribuição é assimétrica, a caixa encontra-se deslocada na direção em que se encontram os dados mais comuns.

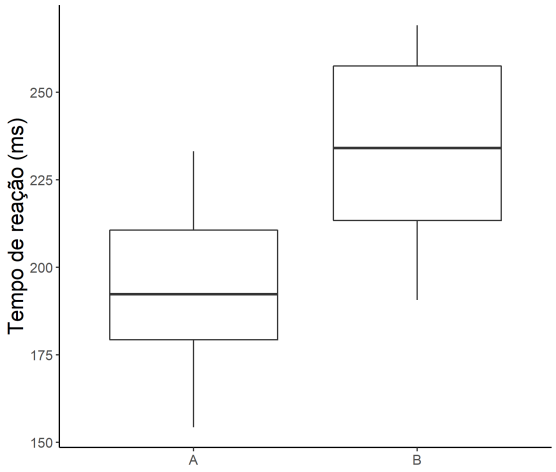

Figura 1.5 – *Boxplots* para amostras A e B

Na curva de frequências, pode-se observar, ainda, a relação entre a média (linha vermelha) e a mediana (linha azul). Nas distribuições simétricas, média e mediana coincidem. No entanto, como a média é uma medida pouco robusta, nas distribuições assimétricas, ela é "puxada" em direção à cauda mais longa, ou seja, os valores extremos na amostra ou população influenciam no valor da média. Foi por esse motivo que, quando calculamos média e mediana, algumas páginas antes, dissemos que, como elas eram próximas, já tínhamos uma noção de que nossos dados eram simétricos.

[inserir gráficos]

Com esses conceitos em mãos, podemos fazer uma análise mais precisa das nossas amostras A e B, que, até agora, tinham sido descritas apenas pela média e pela mediana. Assim:

	Mínimo	q_1	q_2	q_3	Máximo
Palavras do tipo A	154.3	179.3	192.3	210.7	233.2
Palavras do tipo B	190.7	213.4	234.1	257.5	269.2

Tabela 1.6 – Quartis para amostras A e B

Um resumo com o auxílio de gráficos de caixas também nos ajuda a ver que a amostra B parece apresentar maiores tempos de reação, não só "na média", mas também em toda a sua distribuição, sendo apenas que, visualmente, a amostra A parece ser mais homogênea (menor AIQ) do que a amostra B.

Isso parece ser confirmado pela Amplitude Interquartil de cada uma das amostras:

$$AIQ(A) = q_3 - q_1 = 210.7 - 179.3 = 31.4 milisegundos$$

$$AIQ(B) = q_3 - q_1 = 257.5 - 213.4 = 44.1 milisegundos$$

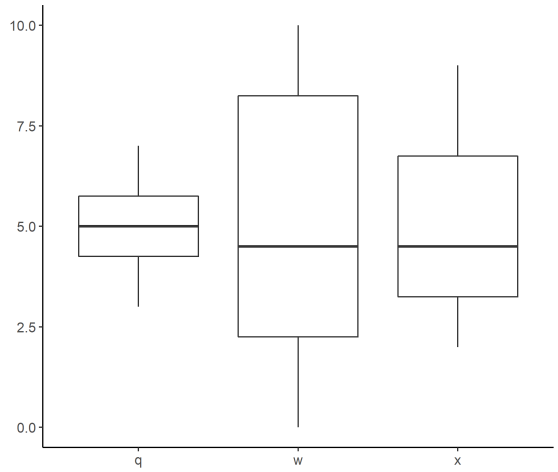

Figura 1.6 – *Boxplots* para q, w e x

Com esses dados em mãos, podemos partir para uma análise mais detalhada dessa diferença na distribuição de A e de B.

1.3 Desvios em relação à média

Tendo feito essa primeira abordagem quanto à distribuição dos dados, podemos passar a tratar de analisar a dimensão da variação dos dados em torno da média, começando com a ideia de desvios. Para isso, tomemos as séries de valores x, w e q, cujas médias são idênticas ($\bar{x} = \bar{w} = \bar{q} = 5{,}0$).

$$x = \{2, 2, 3, 4, 4, 5, 6, 7, 8, 9\}$$

$$w = \{0, 1, 2, 4, 3, 5, 6, 9, 10, 10\}$$

$$q = \{3, 3, 5, 7, 4, 5, 5, 5, 6, 7\}$$

Apesar de a média ser idêntica, os dados não têm a mesma distribuição. Isso porque w cobre quase toda a gama de inteiros de 0 a 10, exceto 7, enquanto q fica restrito ao intervalo entre 3 e 7. Por sua vez, x fica numa espécie de meio termo entre ambos, cobrindo uma gama maior de valores do que q, mas menor do que w, indo de 2 a 9. Em outras palavras, q é um conjunto mais homogêneo e w é um conjunto menos homogêneo. Isso fica explícito na comparação dos *boxplots* de cada conjunto:

Mas, seria possível mensurar essa diferença? Certamente. Um dos modos de fazer isso é calculando a Amplitude Interquartil de cada conjunto, como já vimos. Porém, existem outras. Para chegarmos a elas, vamos começar analisando como os dados de cada série se distribuem em relação à média da série, calculando o que se chama de desvios em relação à média, ou seja, simplesmente subtraindo a média da série de cada um dos valores

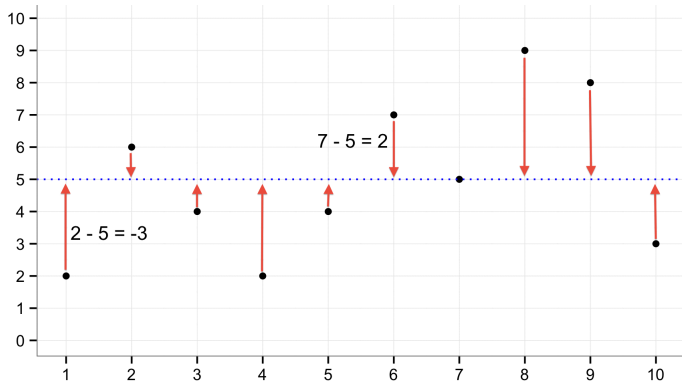

Figura 1.7 – Desvios de x em relação à média

mensurados nessa série. Assim, para o primeiro valor de x ($x_1 = 2$), o desvio é -3, ou seja, $x_i - \bar{x} = 2 - 5 =$ -3. Usaremos a expressão $x_i - \bar{x}$ para representar os desvios da variável x.

Apresentamos, na Figura 1.7, o conjunto x plotado aleatoriamente em torno da média de x. As setas vermelhas indicam os desvios de cada valor de x dessa média, ou seja, a distância que estão de $\bar{x}$. Com isso, podemos ter um vislumbre da dispersão de x em torno da média.

	x_i	$\bar{x}$	$x_i - \bar{x}$	$\lvert x_i - \bar{x}\rvert$	$(x_i - \bar{x})^2$	w_i	$\bar{w}$	$w_i - \bar{w}$	$\lvert w_i - \bar{w}\rvert$	$(w_i - \bar{w})^2$	q_i	$\bar{q}$	$\lvert q_i - \bar{q}\rvert$	$\lvert q_i - \bar{q}\rvert$	$(q_i - \bar{q})^2$
	2	5	-3	3	9	0	5	-5	5	25	3	5	-2	2	4
	2	5	-3	3	9	1	5	-4	4	16	3	5	-2	2	4
	3	5	-2	2	4	2	5	-3	3	9	5	5	0	0	0
	4	5	-1	1	1	4	5	-1	1	1	7	5	2	2	4
	4	5	-1	1	1	3	5	-2	2	4	4	5	-1	1	1
	5	5	0	0	0	5	5	0	0	0	5	5	0	0	0
	6	5	1	1	1	6	5	1	1	1	5	5	0	0	0
	7	5	2	2	4	9	5	4	4	16	5	5	0	0	0
	8	5	3	3	9	10	5	5	5	25	6	5	1	1	1
	9	5	4	4	16	10	5	5	5	25	7	5	2	2	4
Soma dos desvios			0					0					0		
...dos módulos				20					30					10	
...dos quadrados					54					122					18

Tabela 1.7 – Medidas de dispersão para x, w e q

Uma maneira de medirmos essa dispersão seria, por exemplo, calcular a soma desses desvios. Isso porque, supostamente, para as amostras com maior dispersão, a soma seria maior. No entanto, os desvios têm a propriedade de que, para qualquer conjunto de dados, a sua soma é sempre igual a zero, o que não nos permite fazer qualquer inferência sobre a distribuição dos dados.

$$\sum(x_i - \bar{x}) = -3 - 3 - 2 - 1 - 1 + 0 + 1 + 2 + 3 + 4 = 0$$

Você pode, se quiser e não confiar na Tabela 1.7, fazer a mesma conta para os desvios de w ($w_i - \bar{w}$) e para os desvios de q ($q_i - \bar{q}$). Eles sempre darão zero.

Portanto, para que a soma dos desvios possa ser realizada, precisamos eliminar seus sinais negativos, o que poderá ser feito de duas maneiras: calculando o valor absoluto dos desvios ou elevando os desvios ao quadrado, que serão usados para calcularmos duas medidas diferentes: o *desvio médio* e a *variância*. Vamos a elas.

1.4 Desvio médio

O desvio médio é calculado simplesmente somando o módulo ou valor absoluto dos desvios e dividindo esse valor pela quantidade de dados observados (n), ou seja, é uma média dos valores dos desvios. O módulo de um número, como se sabe, é esse número sem o sinal (+ ou -) que o acompanha. Assim, o módulo de -3 ou $|-3| = 3$, que é igual ao módulo e + 3 ou $|+3| = 3$.

Para as séries de dados x, w e q, os valores absolutos estão na tabela

A soma desses valores está abaixo:

$$\sum |x_i - \bar{x}| = 3+3+2+1+1+0+1+2+3+4 = 20$$

$$\sum |w_i - \bar{w}| = 5+4+3+1+2+0+1+4+5+5 = 30$$

$$\sum |q_i - \bar{q}| = 2+2+0+2+1+0+0+0+1+2 = 10$$

Observe com atenção os números calculados acima. A soma dos valores absolutos dos desvios (20, 30 e 10) nos dá uma dimensão da dispersão dos dados, mostrando que q é a série que tem dados menos "espalhados" em relação à média, enquanto w é a série que tem os dados mais "espalhados" em relação à média, o que confirma a análise visual realizada por meio do gráfico de caixas.

O desvio médio, então, seria esses valores divididos por 10 (simplesmente a média dos desvios), o que dá, respectivamente: 2, 3 e 1.

1.5 Variância

A outra maneira de analisar a dispersão dos dados, eliminando os valores negativos dos desvios, é calculando o quadrado dos desvios em relação à média $(x_i - \bar{x})^2$. Volte à nossa tabela inicial e observe esses valores para x, w e q. Podemos então somá-los e obter a Soma dos Quadrados dos Desvios:

$$\sum (x_i - \bar{x})^2 = 9+9+4+1+1+0+1+4+9+16 = 54$$

$$\sum (w_i - \bar{w})^2 = 25+16+9+1+4+0+1+16+25+25 = 122$$

$$\sum (q_i - \bar{q})^2 = 4+4+0+4+1+0+0+0+1+4 = 18$$

Observe, mais uma vez, que o valor obtido busca mensurar a variabilidade do conjunto de dados. No entanto, agora as diferenças entre eles se tornaram marcantes (18 para q e 122 para w). Lembre-se, no entanto, que estamos trabalhando agora com valores quadráticos e não na escala dos valores originais.

Da mesma forma que fizemos para os desvios originais, podemos, também para o quadrado dos desvios, calcular uma espécie de média dessa dispersão. Basta, portanto, dividir essa soma por n. A essa espécie de "média dos quadrados dos desvios" damos o nome de *variância*. Na verdade, para pequenas amostras, o ideal é que a variância seja calculada dividindo-se aquela soma por $n - 1$. Para grandes amostras, não há diferença entre os valores dos dois métodos. Se fizéssemos isso para os dados acima, teríamos que as variâncias seriam:

$$var(x) = \frac{54}{10-1} = \frac{54}{9} = 6.0$$
$$var(w) = \frac{122}{10-1} = \frac{122}{9} = 13.55$$
$$var(q) = \frac{18}{10-1} = \frac{19}{9} = 2.0$$

Como no caso do desvio médio, a variância para cada conjunto confirma a análise visual feita com o *boxplot*, já que a variância de w é a maior (13.55) e a de q é a menor (2.0). A variância, porém, é difícil de ser interpretada, já que ela não expressa a dispersão dos dados na mesma unidade em que os dados foram mensurados, mas sim em seus quadrados. Se x, w e q fossem notas de alunos, por exemplo, a variância estaria expressa em notas ao quadrado; se x, w e q fossem medidas em metros, então a variância seria em metros quadrados, e assim por diante. Para evitar esse tipo de problema, o que se faz é transformar a variância em uma medida que seja expressa na mesma unidade dos dados: o *desvio padrão*.

1.6 Desvio padrão

Como dito acima, a fim de facilitar a interpretação da dispersão dos dados, é preciso fazer com que a variância seja expressa na mesma unidade em que os dados mensurados são expressos. Ora, como a variância é expressa em quadrados da unidade padrão, para resolver o problema basta tirar a raiz quadrada da variância, o que nos dá o desvio padrão. Assim:

Como o desvio padrão está expresso na mesma unidade dos dados originais, ele é uma boa medida da dispersão dos dados e, além disso, é de fácil interpretação. Assim como o desvio médio, o desvio padrão representa a média dos desvios, ou seja, o quanto, em média, os dados se dispersam em relação à média.

Um comentário importante deve ser feito aqui. Você não deve confundir desvio padrão com erro padrão. O desvio padrão, como vimos, é uma estatística descritiva. Ele busca

	x	w	q
Variância	6	13.55	2
Desvio padrão	2.44	3.68	1.41

Tabela 1.8 – Variância e desvio padrão para x, w e q

mensurar o quanto os dados se dispersam em relação à média. O erro padrão, por outro lado, é uma estatística inferencial – sobre a qual falaremos em breve. Ele mensura o grau de confiabilidade que temos quanto a uma média amostral, nos dizendo o quanto podemos confiar nessa média como representativa da média populacional.

1.6.1 Desvio padrão e curva normal

Mais importante ainda é a relação que se pode estabelecer entre o desvio padrão e os dados normalmente distribuídos. Isso porque os dados podem ser divididos em unidades de desvio padrão em relação à média, para mais ou para menos, sendo provado que a área sob a curva normal é constante dentro dos limites de cada unidade de desvio padrão. Assim, se os dados são normalmente distribuídos, 34.13% dos dados estão contidos dentro de 1 desvio padrão em relação à média. O que nos dá que, dentro do intervalo de 1 desvio padrão para mais e para menos em relação à média, tem-se 68,26% dos dados ($34.13 + 34.13 = 68.26$).

Usando o mesmo raciocínio, pode-se provar que a área compreendida entre 1 e 2 desvios padrão contém 13.59% dos dados. Assim, sob a assunção de normalidade, pode-se afirmar com confiança que 95.44% dos dados se encontram a dois desvios padrão da média ($68.26 + 13.58 + 13.58 = 95.44$).

Continuando o raciocínio, pode-se provar também que a área sob a curva entre 2 e 3 desvios padrão contém 2.14% dos dados, o que nos dá que 99.74% dos valores estão contidos em até 3 desvios padrão da média ($95.44 + 2.14 + 2.14 = 99.74$).

Essa relação entre o desvio padrão e a curva normal é de suma importância para a *inferência estatística*.

1.7 Voltando ao exemplo

Agora que já temos uma noção inicial das medidas de dispersão, podemos voltar às nossas amostras A e B e fazermos uma descrição da variabilidade dos dados que lá estão. Até agora, tínhamos calculado, na Tabela 1.6, que repetimos abaixo, as seguintes estatísticas para aqueles dados, lembrando que q_2 = mediana:

Calculemos, então, as demais medidas de dispersão, que estão resumidas na Tabela 1.10. Observe que começamos com os valores observados (A_i e B_i) e calculamos as médias ($\bar{A}$ e $\bar{B}$). A partir de então, o nome das colunas é auto-explicativo.

	Mínimo	q_1	q_2	q_3	Máximo
A	154.3	179.3	192.3	210.7	233.2
B	190.7	213.4	234.1	257.5	269.2

Tabela 1.9 – Quartis para amostras A e B

Somas...	A_i	$\bar{A}$	$A_i - \bar{A}$	$\lvert A_i - \bar{A}\rvert$	$(A_i - \bar{A})^2$	B_i	$\bar{B}$	$B_i - \bar{B}$	$\lvert B_i - \bar{B}\rvert$	$(B_i - \bar{B})^2$
	175.56	194.23	-18.67	18.67	348.59	253.84	233.91	19.92	19.92	396.86
	183.46	194.23	-10.77	10.77	116.01	210.67	233.91	-23.25	23.25	540.5
	193.83	194.23	-0.4	0.4	0.16	215.57	233.91	-18.35	18.35	336.67
	209.54	194.23	15.31	15.31	234.38	237.16	233.91	3.24	3.24	10.51
	211.8	194.23	17.57	17.57	308.68	214.41	233.91	-19.51	19.51	380.59
	192.31	194.23	-1.92	1.92	3.69	261.46	233.91	27.54	27.54	758.53
	233.17	194.23	38.94	38.94	1516.27	190.66	233.91	-43.26	43.26	1871.31
	202.85	194.23	8.62	8.62	74.29	224.17	233.91	-9.75	9.75	95.04
	165.61	194.23	-28.62	28.62	819.14	234.11	233.91	0.19	0.19	0.04
	232	194.23	37.77	37.77	1426.52	258.71	233.91	24.79	24.79	614.61
	220.38	194.23	26.15	26.15	683.79	256.35	233.91	22.43	22.43	503.16
	183.03	194.23	-11.2	11.2	125.45	202.75	233.91	-31.17	31.17	971.49
	154.3	194.23	-39.93	39.93	1594.46	269.16	233.91	35.24	35.24	1241.95
	168.33	194.23	-25.9	25.9	670.84	212.38	233.91	-21.54	21.54	463.91
	187.29	194.23	-6.94	6.94	48.17	267.38	233.91	33.46	33.46	1119.66
...dos desvios			0					0		
...dos módulos				288.71					333.64	
...dos quadrados					7970.44					9304.83
Variância					569.32					664.63
Desvio padrão					23.86					25.78

Tabela 1.10 – Medidas de dispersão para amostras A e B

Detenha-se alguns momentos para avaliar essa tabela. Compare os valores de A com os de B apresentados na parte resumitiva final e veja como eles são descritores da variabilidade dos dados. E, mais importante, não se assuste com esse monte de números e de contas. Você não precisa saber fazê-las todas, mas precisa entendê-las. Se fizer isso, verá que a compreensão que terá dos seus próprios dados será bem maior, o que certamente o ajudará muito quando estiver com seus resultados experimentais em mãos.

Isso tendo sido feito, podemos encerrar a primeira abordagem dos dados. Com o instrumental até agora descrito, é possível fazer uma abordagem inicial dos dados, buscando neles padrões que nos sejam informativos sobre suas distribuições. Esse, todavia, é apenas o primeiro passo da análise estatística. Isso porque, até agora, apenas descrevemos aquilo que temos em mãos. A estatística, no entanto, é uma poderosa ferramenta para fazer inferências sobre aquilo que desconhecemos. Esse tópico é conhecido como *inferência estatística*.