Igor Costa

Modelos lineares mistos aplicados à psicolinguística experimental

16 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

Sumário

1 Da descrição à inferência

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 da Figura 1.1 nos mostra com mais clareza a distribuição dos dados para as amostras A e B, sugerindo que os tempos de reação para B parecem ser maiores do que os tempos para A. Vamos, então, fazer uma abordagem desses dados considerando dois aspectos: primeiro, os pontos em torno dos quais esses dados se concentram, chamados,

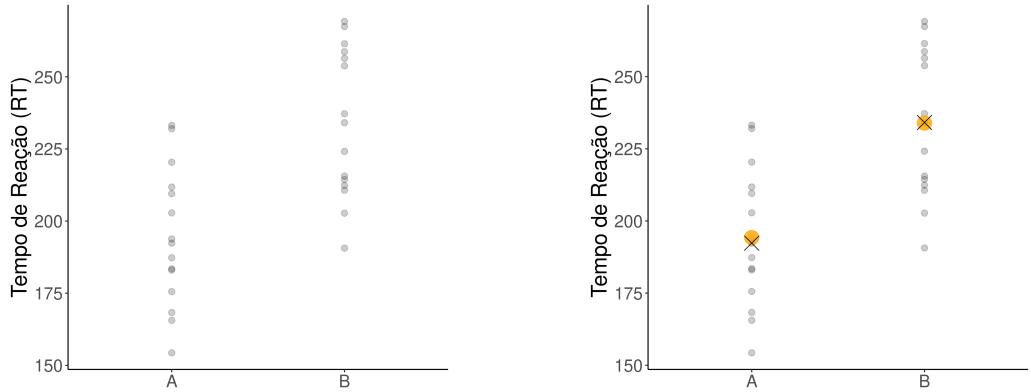

Figura 1.1 – **Painel 1:** Gráfico de dispersão (*scatterplot*) para amostras A e B. A dispersão dos dados nos parece indicar que palavras do tipo A são lidas mais rapidamente do que palavras do tipo B, mas há vários pontos que se sobrepõem. **Painel 2:** Médias (pontos laranjas) e medianas (marcacção em formato de "X") para cada um dos tipos de palavras. OBS: Todas as imagens deste capítulo estão em alta resolução. Se estiver com dificuldade de enxergá-las, basta "dar" um *zoom* que elas não perderão qualidade.

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

$$x_1 = 2, x_2 = 3, x_3 = 5, \ldots, x_{10} = 4$$

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:

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

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.

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, *em média*, do que o valor para A. Essa diferença pode ser representada no gráfico de dispersão (Painel 2 da Figura 1.1), onde os pontos laranjas indicam as médias de cada grupo e a marcação em formato de "X" indica a mediana – algo que discutiremos daqui a pouco.

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:

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

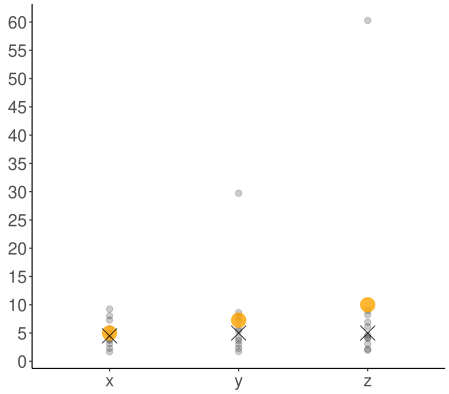

Figura 1.2 – Comparação entre a média (pontos laranjas) e a medida (marcação em "X") para os conjuntos de dados apresentados, mostrando que a mediana é uma medida robusta, estável, pouco influencidada pelos valores extremos enquanto a média, por sua vez, é uma medida pouco robusta, muito influenciada por esses valores discrepantes. Em distribuições simétricas, sem valores extremos, média e mediana tendem a coincidir.

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

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:

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

	x	y	z
Média	5	7.27	10
Mediana	4.5	5	5

Tabela 1.3 – Relação entre média e mediana

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

Como vemos na tabela acima e na Figura 1.2, 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 na Tabela 1.4. (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). Esses mesmos valores já foram mostrados no Painel 2 da Figura 1.1.

	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 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 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.

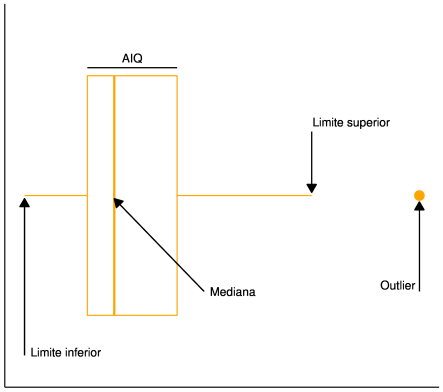

Figura 1.3 – Elementos componentes do gráfico de caixas ou *boxplot*. As caixas são delimitadas pelo primeiro quartil e pelo terceiro quartil, compreendendo, portanto, os dados mais comuns.

1.2.2.1 *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.

Na curva de frequências, pode-se observar, ainda, a relação entre a média (linha pontilhada) e a mediana (linha cheia). 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

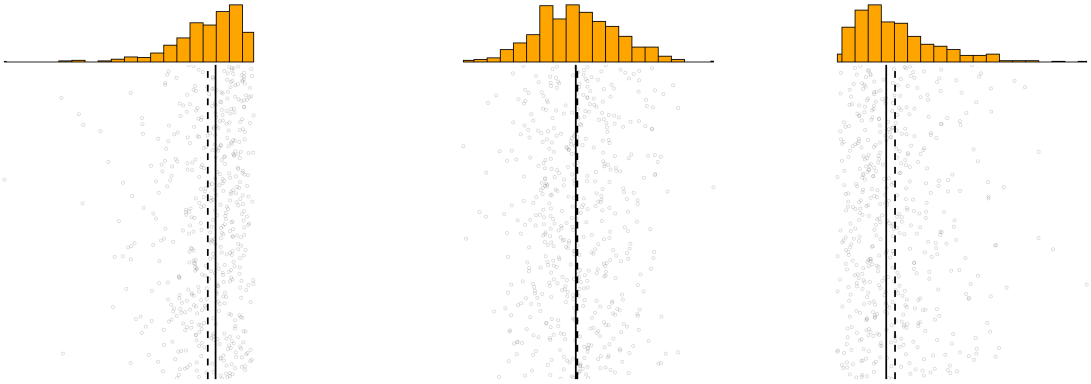

Figura 1.4 – Os três paineis mostram a relação entre as medidas da tendência central (média e mediana) com diferentes tipos de distribuição de frequência: em distribuições simétricas (como a normal - **painel central**), essas medidas tendem a coincidir; em distribuições assimétricas (como a beta - **paineis laterais**), a média é influenciada pelos valores extremos enquanto a mediana se mantém como boa descritora da maioria dos dados.

calculamos média e mediana, algumas páginas antes, dissemos que, como elas eram próximas, já tínhamos uma noção de que nossos dados eram simétricos.

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:

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

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

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

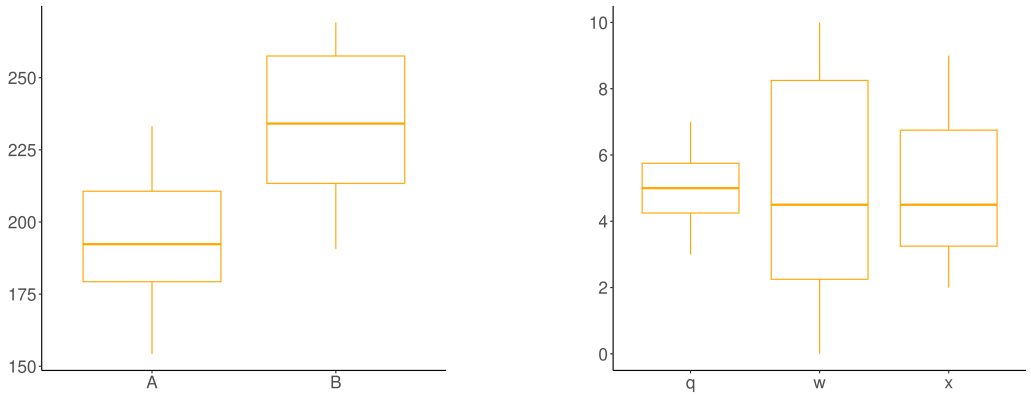

Figura 1.5 – **Painel 1:** Boxplots para as amostras A e B, mostrando a distribuição de ambos os conjuntos de dados. As caixas estão nos meios dos bigodes, demonstrando que as distribuições são simétricas e B parece ser levemente mais disperso do que A, mas uma diferença provavelmente irrelevante. **Painel 2:** Boxplots para as amostras q, w e x, mostrando a diferença de homogeneidade entre as amostras.

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 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 1.6, 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

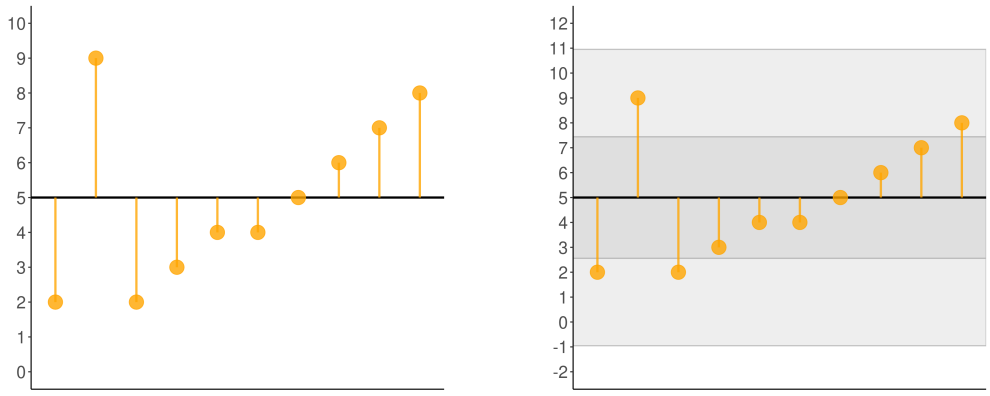

Figura 1.6 – **Painel 1:** Desvios de x em relação à média de x, ou seja, os desvios mostram o quanto cada ponto se afasta do centro dos dados, mensurando a variabilidade desses dados. **Painel 2:** A mancha mais escura mostra o desvio padrão de $x = 2.44$, ou seja, uma espécie de média da variação desses dados; a mancha mais clara mostra a variância de $x = 6$, simplesmente o quadrado do desvio padrão.

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:

	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

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

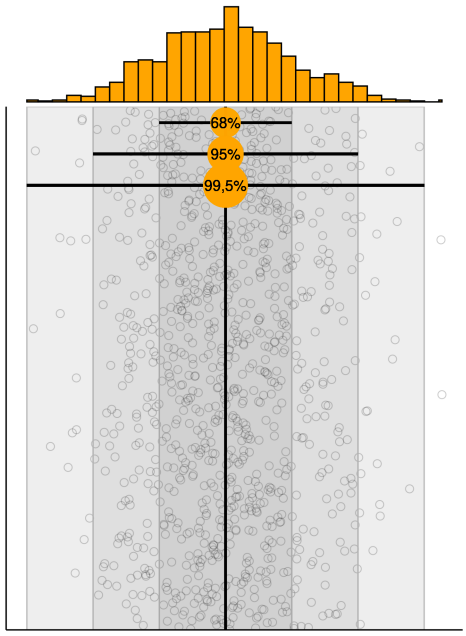

Figura 1.7 – Relação entre o desvio padrão e a curva normal para uma amostra normalmente distribuída com 1000 observações: cerca de 68% dos valores observados estão a até 1 desvio padrão da média; cerca de 95% estão a até dois desvios padrão da média; e cerca de 99.5% estão a até 3 desvios padrão da média. Essa relação entre o desvio padrão e a curva normal é de fundamental importância para a inferência estatística, como veremos adiante.

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 1.9, que repetimos abaixo, as seguintes estatísticas para aqueles dados, lembrando que q_2 = mediana:

	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

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.

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.

1.8 Inferência estatística

Até agora olhamos para os nossos dados e descobrimos uma série de informações – que descrevemos em valores numéricos – sobre os dados obtidos em nossos experimentos: descobrimos que, em média, os nossos 15 sujeitos são mais rápidos lendo palavras do tipo A do que palavras do tipo B (a média de A é menor do que a de B) e que eles são mais consistentes lendo palavras do tipo A do que do tipo B (a variância e, logo, o desvio padrão de A são menores do que os de B). Mas como saber se essa diferença é real?

Em primeiro lugar, vamos então esclarecer o que estamos querendo dizer com *ser uma diferença real*. Obviamente os valores são diferentes. Nós de fato fizemos um experimento, coletamos os dados, calculamos as médias e as outras estatísticas e elas são diferentes. Isso é verdade, obviamente. Mas pergunte-se: se fizéssemos esse experimento mais uma vez, com todo o controle necessário, será que obteríamos esse mesmo resultado? E se o repetíssemos várias e várias vezes, será que ainda assim teríamos esses mesmos valores?

A resposta óbvia a essa pergunta parece ser *não*. Os resultados variariam de experimento a experimento. Agora se pergunte: eles variariam muito ou pouco? Se eles continuassem, a cada novo experimento, muito próximo do que descobrimos até agora, diríamos que temos uma diferença real. Do contrário, se eles fossem muito diferentes, diríamos que essa diferença não é real. Mais uma vez, pare para refletir um pouco sobre a primeira situação: por que motivo, no nosso caso, os valores não mudariam (muito) entre os experimentos? Uma resposta óbvia seria: o efeito do tipo de palavra na leitura é real: palavras do tipo A de fato são lidas mais rapidamente, não só pelos 15 sujeitos que estão fazendo meu experimento, mas por toda pessoa que ler palavras desse tipo.

Esse é o princípio fundamental da *inferência estatística*. Quando fazemos um experimento, não queremos saber se os valores das **amostras** são diferentes – obviamente que o são; nós fomos lá, medimos e calculamos esses valores e eles o foram; nós queremos saber *se essa diferença é significativa*, ou seja, se ela representa uma diferença real na **população** da qual a retiramos[1].

[1] Essa descrição de diferença significativa não é unanimidade, estando, na verdade, vinculada ao chamado paradigma frequentista. De fato, ela é apenas um modo de olhar para a relação entre a amostra que coletamos e a população da qual essa amostra foi extraída. Outras correntes da estatística, como o paradigma bayesiano, olhariam para essa relação de outra maneira.

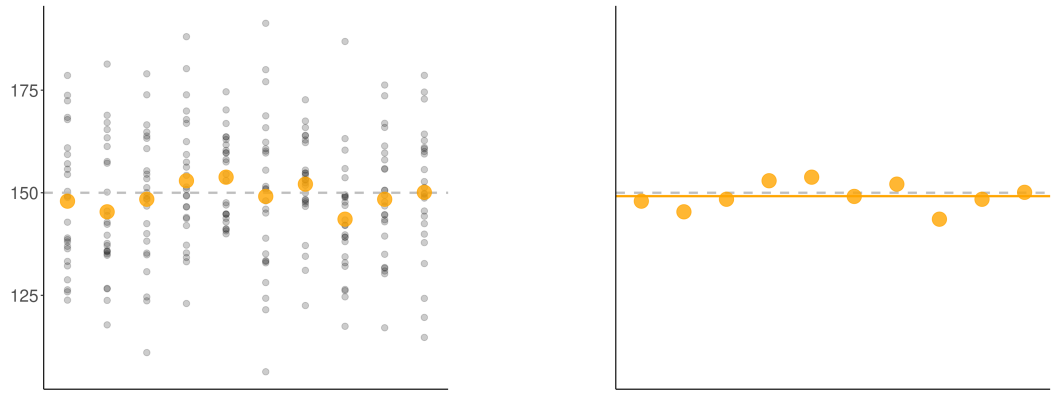

Figura 1.8 – **Painel 1:** Amostras aleatórias retiradas de uma população normalmente distribuída com média 150 e desvio padrão 15. Pontos laranjas indicam as médias de cada amostra e linha pontilhada a média da população. **Painel 2:** A média das amostras anteriores, denominada *distribuição amostral de médias*, é indicada pelos pontos laranjas. A média desses pontos (149.17), pela linha laranja.

Um outro modo de dizer isso é falando que nós calculamos as *estatísticas* das amostras a fim de estimar os *parâmetros populacionais*. A grande questão aqui, porém, é que nós não podemos realizar um monte de experimentos e comparar os valores obtidos com cada um deles a fim de verificar se são consistentes ou não. Experimentos são caros, trabalhosos, demandam grande preparação, equipamentos, tempo, participantes, horas de laboratório, etc. Para fazer essa estimativa, contamos apenas com o nosso experimento, o único que conseguimos realizar. A partir dele, temos que "adivinhar" se estamos próximos dos parâmetros populacionais ou não. Mas a nossa adivinhação é rigorosa: ela vai se valer das estatísticas que até agora calculamos, ou seja, vai se valer do fato de que sabemos os pontos em torno dos quais os dados se organizam e, talvez mais importante, como eles se dispersam em torno desses pontos.

1.8.1 Erro padrão da média

Para começarmos a falar sobre inferência estatística, vamos iniciar nossa discussão com um experimento mental. Imagine que retiremos uma quantidade enorme de amostras de uma população, cuja média (que vamos chamar de μ – a letra grega *mu*) e variância (σ^2, a letra grega *sigma* elevada ao quadrado[2]) conhecemos. Para cada uma dessas amostras, calculamos uma média. Parece óbvio que cada uma dessas médias não é exatamente igual ou mesmo próxima da média populacional (μ). Algumas estarão mais próximas, outras mais distantes, como mostra o Painel 1 da Figura 1.8.

Agora vamos imaginar que calculemos uma média de todas essas médias ($\bar{x}$), como mostra a linha laranja no Painel 2 da Figura 1.8. Se pensarmos bem, essa nova média estará bem mais próxima da média populacional (μ) do que cada uma – ou pelo menos a maioria – das médias individuais de cada amostra. Isso ocorre porque, ao calcularmos a

[2] Uma dica importante: é convencional usarmos letras gregas para nos referirmos a *parâmetros populacionais* e letras latinas para *estatísticas*, ou seja, valores amostrais.

nova média, eliminamos valores extremos, ou seja, diminuímos a variabilidade dos dados. Daí nossa precisão ser maior. Isso significa que, empiricamente, para infinitas amostras, a média das médias ($\bar{x}$) dessas amostras é igual à média populacional (μ). Em outras palavras, no limite, a média de uma distribuição amostral de médias ($\bar{x}$) é igual à média populacional μ.

Do mesmo modo como se pode demonstrar o que foi dito acima empiricamente (a Figura 1.8 ilustra esse exatamente isso.), pode-se demonstrar, também, que a variância dessa distribuição amostral de médias (as várias amostras que retiramos e para as quais calculamos uma média) é igual à variância da população dividida por n.

$$var = \frac{\sigma^2}{n}$$

Então vamos estender o nosso exercício mental para o seguinte caso. Imagine agora que nós não saibamos a média populacional (μ), mas saibamos a variância populacional (σ^2). Logo, apesar de não sabermos a média, sabemos a variabilidade dessa população. Com isso, podemos calcular o desvio padrão dessa população, que será dado pela raiz quadrada da variância:

$$\text{desvio padrão} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Perceba que, com esse desvio padrão, nós podemos ter uma estimativa de o quanto podemos confiar na média das nossas amostras como estimativa da média populacional (μ), ou seja, já que sabemos a variabilidade da população, podemos saber o quanto estamos errados quanto à média dessa mesma população. Por isso, esse desvio padrão de uma distribuição amostral de médias é chamado de *erro padrão da média*.

O problema é que, quando estamos fazendo um experimento na vida real não temos nem (i) a variância da população e nem (ii) uma quantidade infinita de amostras dessa população. Na verdade, não temos sequer uma quantidade grande de amostras. Temos apenas uma, aquela que colhemos com nosso experimento. No entanto, a partir dessa amostra, nós podemos calcular uma estimativa da variância populacional (σ^2), que é a variância amostral (s^2). Logo, se temos uma estimativa de quanto é a variabilidade na população, podemos estimar o quão precisa é a média amostral para estimar a média populacional, ou seja, podemos ter uma estimativa do erro padrão da média[3]. Esse é dado pela fórmula abaixo, em que n é o tamanho da nossa amostra:

$$\text{erro padrão} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$$

[3] Esse passo pode parecer um grande salto especulativo, mas, pense um pouco, se você coletou uma amostra aleatória de uma população, representativa dessa população, de tamanho suficientemente grande, por que diabos a variância dessa amostra seria diferente da variância populacional? Pode até ser que em alguns casos isso ocorra, mas, em geral, não há muitos motivos para isso ocorrer.

Como n (o tamanho da amostra) está no denominador, então, quanto maior for nossa amostra, menor será o erro padrão e, portanto, mais confiança podemos ter na nossa média amostral ($\bar{x}$) como estimativa da média populacional (μ). O tamanho da amostra, portanto, é um importante regulador da nossa precisão. Pense, por exemplo, no caso de coletarmos medidas de uma amostra tão grande que ela seja quase do tamanho da população. Nesse caso, nosso erro seria bem pequeno.

1.8.1.1 Voltando ao exemplo

Para os dados do nosso experimento, os desvios padrões (retomados da Tabela 1.10) e erros padrões (calculados abaixo) estão resumidos na Tabela 1.11. Na Figura 1.9, faz-se uma comparação entre o erro padrão e o desvio padrão.

$$\text{desvio padrão de A} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}} = \frac{23.86}{\sqrt{15}} = 6.16$$

$$\text{desvio padrão de B} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}} = \frac{25.78}{\sqrt{15}} = 6.65$$

	Média	Desvio padrão	Erro padrão
A	194.23	23.86	6.16
B	233.91	25.78	6.65

Tabela 1.11 – Erro padrão da média para amostras A e B

Vamos pensar um pouco sobre essa informação. Se o erro padrão diz o quanto podemos confiar no valor das médias amostrais como representativas das médias populacionais, ou seja, o quanto estamos seguros de estarmos "acertando" as médias, então, parece de fato que nossas médias são diferentes *na população* e que palavras do tipo A são lidas mais rapidamente do que palavras do tipo B. Isso ocorre porque, se as barras delimitam os limites do nosso erro, então é provável que, se replicássemos esse experimento, as médias não seriam idênticas a essas que obtivemos, mas não escapariam dos limites das barras. Ora, como as barras estão bem distantes umas das outras, não parece que estejamos correndo o risco de, em uma replicação, as médias estarem muito mais próximas ou mesmo invertidas. Mas quão seguros podemos estar quanto a essa distância? Qual seria a probabilidade de estarmos cometendo um erro na nossa estimativa? Para saber isso, vamos introduzir o conceito de *intervalo de confiança*, diretamente relacionado ao erro padrão.

1.8.2 Intervalo de confiança

Para ilustrar o conceito de intervalo de confiança, vamos usar os resultados dos nossos 15 sujeitos realizando o experimento de leitura de palavras com o qual até agora temos

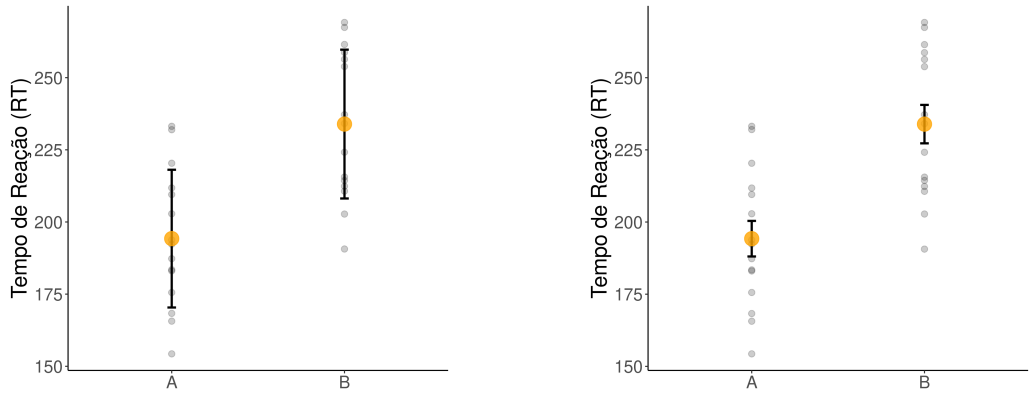

Figura 1.9 – Comparação entre o desvio padrão, estatística que mensura a variabilidade da amostra (**Painel 1**) e o erro padrão da média (**Painel 2**), que mensura o quão errado estamos, com base nessa amostra que colhemos, quanto às médias populacionais. As médias amostrais são representadas pelos pontos laranjas.

trabalhado. Nós calculamos que a média dos sujeitos lendo palavras do tipo A é de 194.23 ms. Mas o quanto essa média representa de fato a média da população lendo palavras do tipo A?

Podemos estimar a nossa precisão usando o desvio padrão que calculamos para essa amostra, que era de 23.86ms. Como falamos antes, o desvio padrão tem uma relação direta com a curva normal. Assumindo que os nossos dados foram retirados de uma população normalmente distribuída, sabemos que qualquer dado que esteja distante da média 1.96 vezes o desvio padrão é um dado raro, que ocorre apenas 5% das vezes. Vamos assumir também que a variância da nossa amostra (e, portanto, o desvio padrão) seja idêntica ou aproximadamente idêntica à variância da população de onde a amostra foi retirada (o mesmo que fizemos para o cálculo do erro padrão).

Aqui ainda cabe uma última coisa: como sabemos o quão boa é a nossa amostra? Imagine que tenhamos feito um experimento com mil participantes e um com 15, como é o nosso caso. Qual deles deve ter resultados mais precisos? Obviamente que aquele com mais participantes. Logo, precisamos considerar, também, o tamanho da nossa amostra para dizermos quão precisa é nossa estimativa (isso, mais uma vez, é o mesmo que fizemos para o cálculo do erro padrão).

Se fizéssemos isso, poderíamos ter alguma confiança de que dados que estejam distantes da média, por exemplo, que estejam a mais de 1.96 desvios padrões da média, seriam raros. Em outras palavras, se sabemos que a distribuição populacional é normalmente distribuída e se temos uma estimativa da variância populacional a partir da amostra, então podemos saber quais são os valores distantes da média a até 1.96 desvios padrão. Para isso, basta usar a fórmula:

$$\text{Intervalo de confiança para A} = \frac{1.96 \times \text{desvio padrão}}{\sqrt{n}} = \frac{1.96 \times 23.86}{\sqrt{15}} = \frac{46.76}{3.87} = 12.08$$

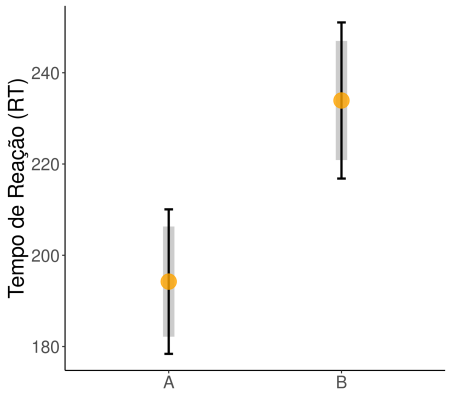

Figura 1.10 – Comparação entre os intervalos de confiança com grau de confiança de 95% (barras cinzas) e 99% (linhas pretas), mostrando que, quanto mais confiança temos, maior o nosso intervalo de valores possíveis para a média.

$$\text{Intervalo de confiança para B} = \frac{1.96 \times \text{desvio padrão}}{\sqrt{n}} = \frac{1.96 \times 25.78}{\sqrt{15}} = \frac{50.52}{3.87} = 13.04$$

Olhe para essa conta com carinho. O desvio padrão dividido pela raiz de n é simplesmente o erro padrão que calculamos na seção anterior. O intervalo de confiança, portanto, faz uso daquela estatística para calcular um grau de confiança (95%, 99%, 99,9%, etc.) na nossa estimativa amostral.

Observe na Figura 1.10 os intervalos de confiança com índice de confiança (α – a letra grega *alpha*) iguais a 0.05 (95%) e 0.01 (99%) para as médias de A e de B. Dado que os extremos das barras não se cruzam, parece razoável confiar que as nossas médias de fato são diferentes. Observe também que quanto mais confiança desejarmos (mais certeza quisermos), maior será nosso intervalo (menos precisão teremos).

1.8.2.1 Sobre a interpretação do intervalo de confiança

Neste ponto chegamos a uma questão filosófica um tanto espinhosa e que gera inúmeras polêmicas: como interpretar um intervalo de confiança? Isso significa que nós temos 95% de confiança de que a média populacional está nesse intervalo? A resposta talvez não seja tão simples.

Vamos imaginar o seguinte: nós temos uma população de média μ desconhecida, ou seja, o parâmetro populacional que desejamos estimar. Para essa população coletamos uma amostra aleatória e representativa de tamanho 25 ($n = 25$) e calculamos sua variância ($s^2 = 225$), que usamos para estimar a variância populacional ($\sigma^2 = 225$). Com isso, podemos calcular o nosso erro padrão da média, que é 3 ($\sqrt{\frac{225}{25}} = \frac{15}{5} = 3$). Se este é o erro padrão, podemos calcular um intervalo de confiança de 95%, que é de 6 unidades[4] para mais ou para menos da média da população, que não temos.

[4] Vamos chutar 2 em vez de 1.96 aqui só para facilitar as contas: $2 \times 3 = 6$.

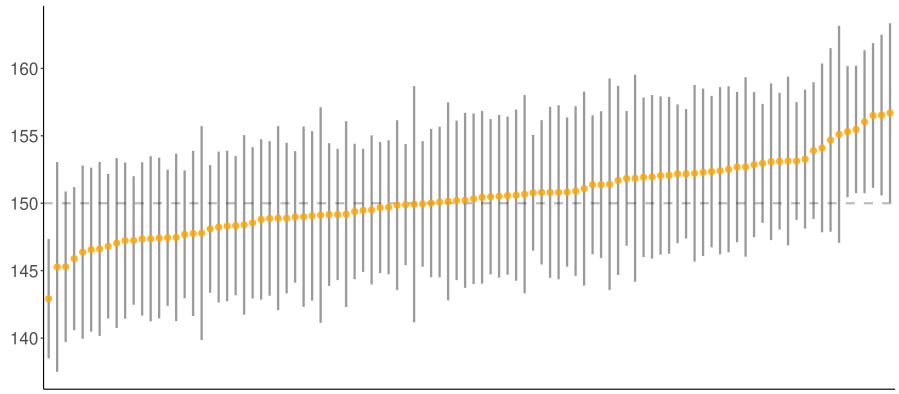

Figura 1.11 – Médias e intervalos de confiança para 100 amostras aleatórias de tamanho $n = 25$ retiradas de uma população normalmente distribuída com média $\mu = 150$ e desvio padrão $\sigma = 15$. Dos 100 intervalos calculados, apenas 7 não contêm a média populacional. Observe que cada amostra apresenta um desvio padrão distinto, logo, um intervalo distinto, mas a grande maioria deles contém o verdadeiro parâmetro populacional.

Se retirássemos uma grande quantidade de amostras dessa população e calculássemos a média e um intervalo de confiança para cada uma delas, 95% desses intervalos conteriam a verdadeira média populacional. A Figura 1.11 mostra uma centena de médias coletadas aleatoriamente de uma população com média 150 e desvio padrão 15, para as quais foram calculados intervalos de confiança. Observe que apenas 7 deles não contêm a média populacional. É isso que o intervalo de confiança nos informa: **se repetirmos o nosso experimento um número grande de vezes, 95% desses experimentos apresentarão médias cujos intervalos contêm a média populacional**. Empiricamente, portanto, o intervalo de confiança é uma excelente medida. **Mas ele não nos diz que o nosso intervalo contém o parâmetro que estamos investigando**. Se afirmamos isso, estamos dando um passo além, cujo raciocínio pode ser resumido da seguinte maneira: a teoria do intervalo de confiança me diz que 95% dos intervalos conterão o parâmetro; ora, eu coletei uma amostra, então é mais provável que o intervalo calculado para essa amostra esteja entre esses 95% do que entre os 5% restantes; logo, este intervalo específico que acabei de coletar tem 95% de chance de conter o parâmetro populacional.

Mas esse último passo é enganoso, já que não fazemos uma centena de experimentos, fazemos apenas um. Imagine que a média desse nosso único experimento (a nossa amostra) seja 142. Ora, essa é a única medida que temos, já que não podemos mensurar a população inteira. Se seguirmos o raciocínio dado acima, iríamos dizer que temos 95% de certeza de que a média populacional está entre 136 e 148 unidades e de que nossa média amostral é representativa da média da população. Mas não é! Você já sabe que nossa população tem média 150. Logo, o nosso intervalo não contém a média populacional (ele é um daqueles 5% de que falamos dois parágrafos acima). Em outras palavras, ou a minha média amostral está entre as 95% mais comuns (e tenho uma boa estimativa da população) ou ela é um valor exótico (e não tenho uma boa estimativa da população). É um jogo *tudo* ou *nada*: o

intervalo que temos em mãos ou contém o parâmetro (100% de chance de contê-lo) ou não contém o parâmetro (0% de chance de contê-lo). Não faz sentido dizer que ele tem 95% de chance de contê-lo.

Por esse motivo, muitos criticam o intervalo de confiança, sugerindo, em lugar dele, o chamado *intervalo de credibilidade*, uma estatística bayesiana cuja discussão está além do debate que realizamos por aqui. Você pode dar uma boa lida por aí em busca dessas distinções. Se você quiser uma abordagem didática sobre o tema, recomendamos a leitura de Howell (2009): 192) [1]. Mas então qual a vantagem de usar um intervalo de confiança? Parece que voltamos à estaca zero. Não é verdade. A teoria do intervalo de confiança nos garante que, no longo prazo, acertaremos o parâmetro populacional 95% das vezes. Para a ciência, no que diz respeito à replicabilidade e comparação de experimentos, é incrível, pois sabemos que até podemos errar algumas vezes, mas estaremos certos a grande maioria do tempo.

Referências

[1] David C Howell. *Statistical methods for psychology*. Cengage Learning, 2009.