Igor Costa

Modelos lineares mistos aplicados à psicolinguística experimental

24 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.

Palavras A	Palavras B
175.56	253.84
183.46	210.67
193.83	215.57
209.54	237.16
211.8	214.41
192.31	261.46
233.17	190.66
202.85	224.17
165.61	234.11
232	258.71
220.38	256.35
183.03	202.75
154.3	269.16
168.33	212.38
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,

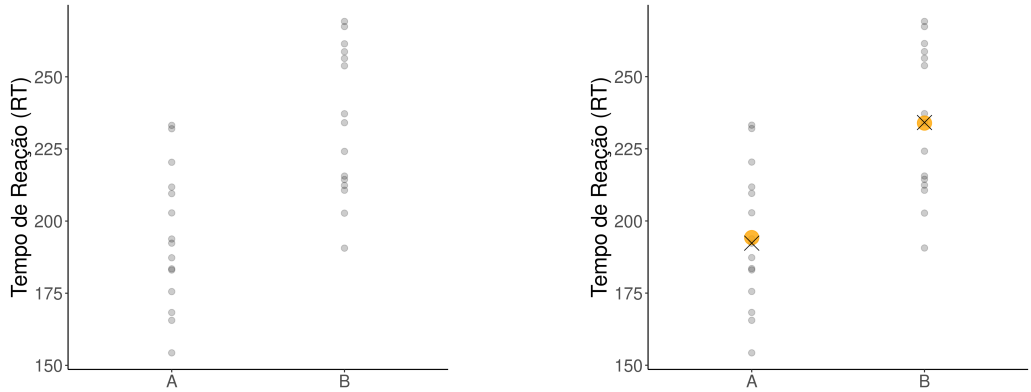

Figura 1.1 – **Painel 1:** Gráfico de dispersão (*scatterplot*) para amostras A e B. A dispersão dos dados nos parece indicar que palavras do tipo A são lidas mais rapidamente do que palavras do tipo B, mas há vários pontos que se sobrepõem. **Painel 2:** Médias (pontos laranjas) e medianas (marcacção 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 $x = \{2, 3, 5, 6, 8, 9, 2, 4, 7, 4\}$, 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[1]:

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 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:

[1] 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.

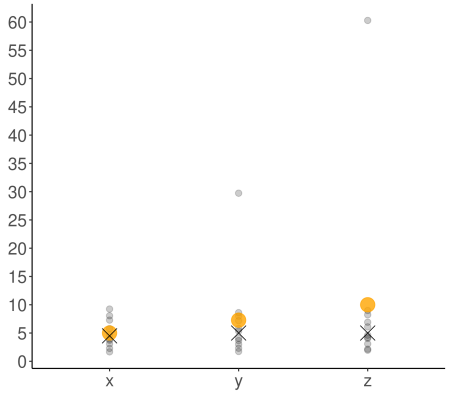

Figura 1.2 – Comparação entre a média (pontos laranjas) e a medida (marcação em "X") para os conjuntos de dados apresentados, mostrando que a mediana é uma medida robusta, estável, pouco influenciada pelos valores extremos enquanto a média, por sua vez, é uma medida pouco robusta, muito influenciada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

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:

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

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.

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.

	x	y	z
Média	5	7.27	10
Mediana	4.5	5	5

Tabela 1.3 – Relação entre média e mediana

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.

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

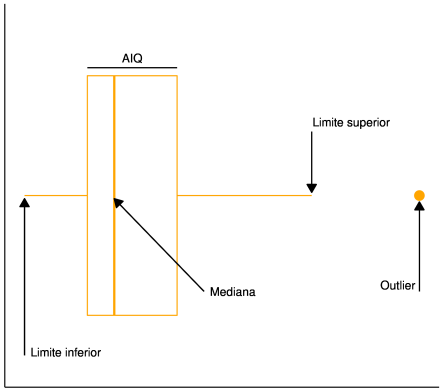

Figura 1.3 – Elementos componentes do gráfico de caixas ou *boxplot*. As caixas são delimitadas pelo primeiro quartil e pelo terceiro quartil, compreendendo, portanto, os dados mais comuns.

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 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

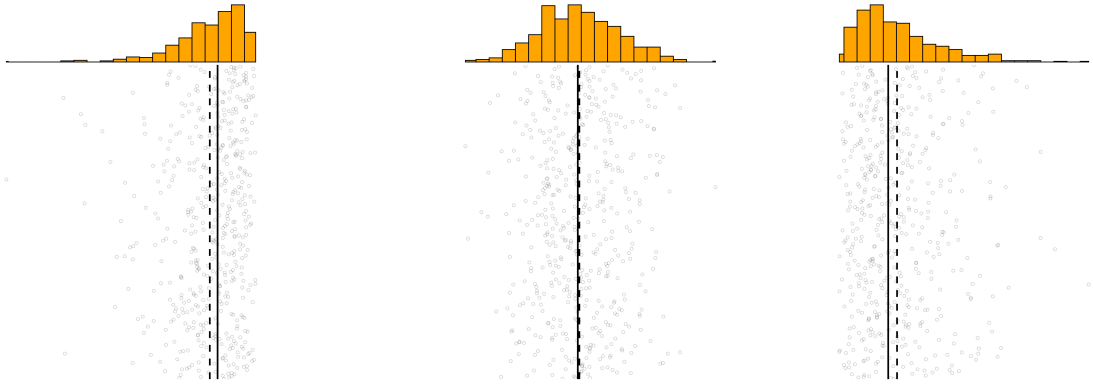

Figura 1.4 – Os três paineis mostram a relação entre as medidas da tendência central (média e mediana) com diferentes tipos de distribuição de frequência: em distribuições simétricas (como a normal - **painel central**), essas medidas tendem a coincidir; em distribuições assimétricas (como a beta - **paineis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

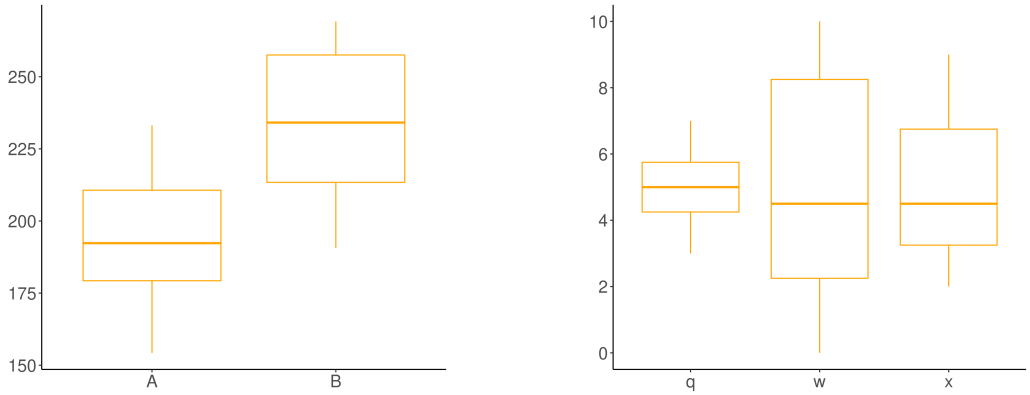

Figura 1.5 – **Painel 1:** Boxplots para as amostras A e B, mostrando a distribuição de ambos os conjuntos de dados. As caixas estão nos meios dos bigodes, demonstrando que as distribuições são simétricas e B parece ser levemente mais disperso do que A, mas uma diferença provavelmente irrelevante. **Painel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

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.

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 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

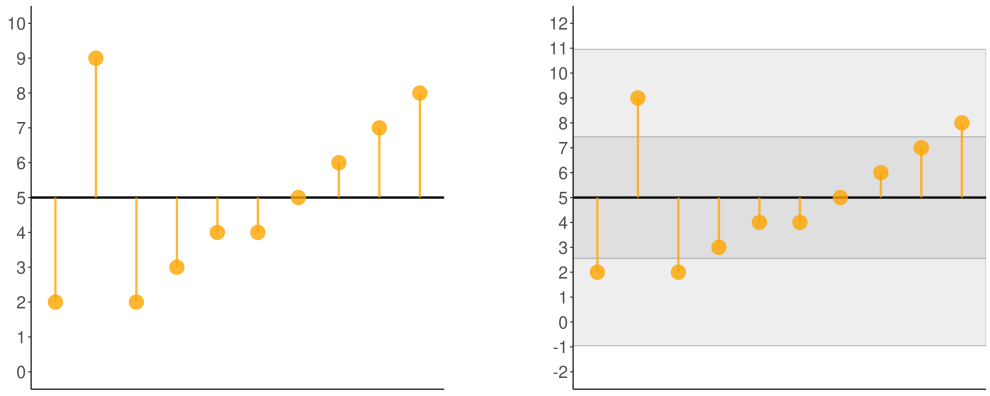

Figura 1.6 – **Painel 1:** Desvios de x em relação à média de x, ou seja, os desvios mostram o quanto cada ponto se afasta do centro dos dados, mensurando a variabilidade desses dados. **Painel 2:** A mancha mais escura mostra o desvio padrão de $x = 2.44$, ou seja, uma espécie de média da variação desses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

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:

$$\begin{aligned}
\sum (x_i - \bar{x})^2 &= 9 + 9 + 4 + 1 + 1 + 0 + 1 + 4 + 9 + 16 \\
&= 54 \\
\sum (w_i - \bar{w})^2 &= 25 + 16 + 9 + 1 + 4 + 0 + 1 + 16 + 25 + 25 \\
&= 122 \\
\sum (q_i - \bar{q})^2 &= 4 + 4 + 0 + 4 + 1 + 0 + 0 + 0 + 1 + 4 \\
&= 18
\end{aligned}$$

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 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.

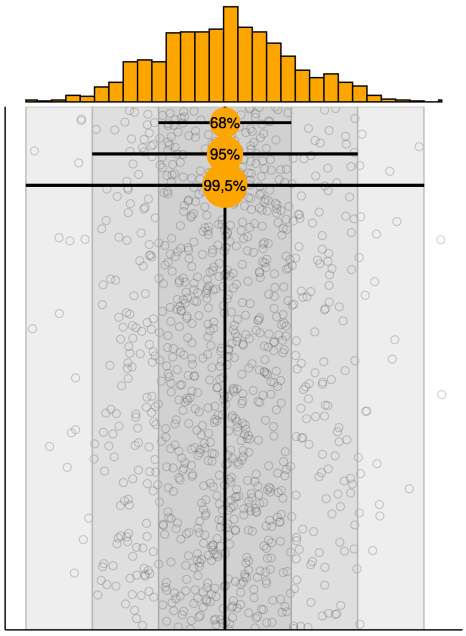

Figura 1.7 – Relação entre o desvio padrão e a curva normal para uma amostra normalmente distribuída com 1000 observações: cerca de 68% dos valores observados estão a até 1 desvio padrão da média; cerca de 95% estão a até dois desvios padrão da média; e cerca de 99.5% estão a até 3 desvios padrão da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

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.

2 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].

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

[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.

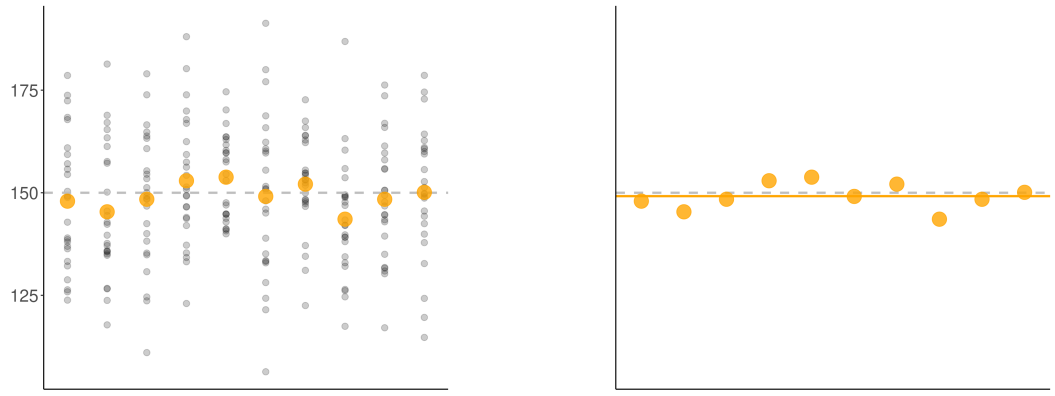

Figura 2.1 – **Painel 1:** Amostras aleatórias retiradas de uma população normalmente distribuída com média 150 e desvio padrão 15. Pontos laranjas indicam as médias de cada amostra e linha pontilhada a média da população. **Painel 2:** A média das amostras anteriores, denominada *distribuição amostral de médias*, é indicada pelos pontos laranjas. A média desses pontos (149.17), pela linha laranja.

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.

2.1 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 2.1.

Agora vamos imaginar que calculemos uma média de todas essas médias ($\bar{x}$), como mostra a linha laranja no Painel 2 da Figura 2.1. 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 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.

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

Do mesmo modo como se pode demonstrar o que foi dito acima empiricamente (a Figura 2.1 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:

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

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.

[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.

2.1.1 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 2.1. Na Figura 2.2, 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 2.1 – 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.

2.2 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 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

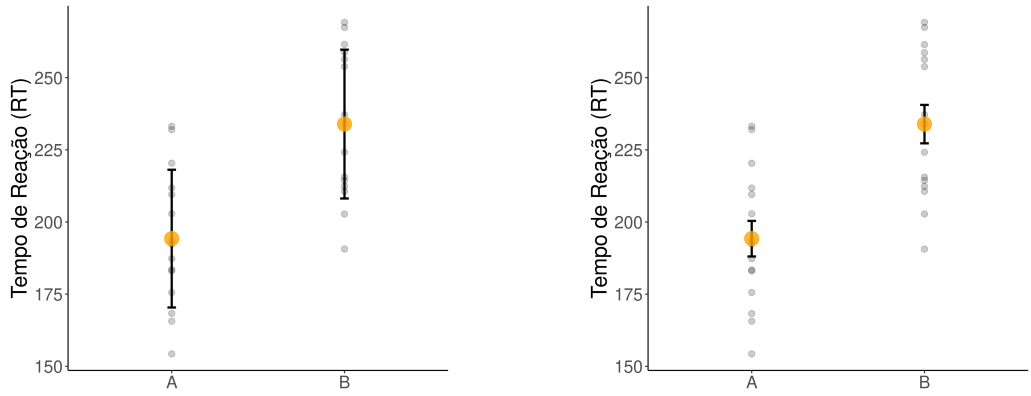

Figura 2.2 – Comparação entre o desvio padrão, estatística que mensura a variabilidade da amostra (**Painel 1**) e o erro padrão da média (**Painel 2**), que mensura o quão errado estamos, com base nessa amostra que colhemos, quanto às médias populacionais. As médias amostrais são representadas pelos pontos laranjas.

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:

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

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

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.

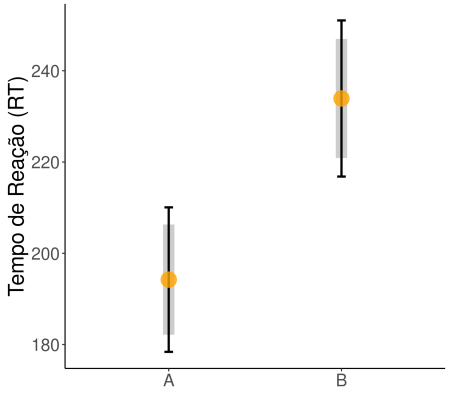

Figura 2.3 – Comparação entre os intervalos de confiança com grau de confiança de 95% (barras cinzas) e 99% (linhas pretas), mostrando que, quanto mais confiança temos, maior o nosso intervalo de valores possíveis para a média.

Observe na Figura 2.3 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).

2.2.1 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.

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 2.4 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**

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

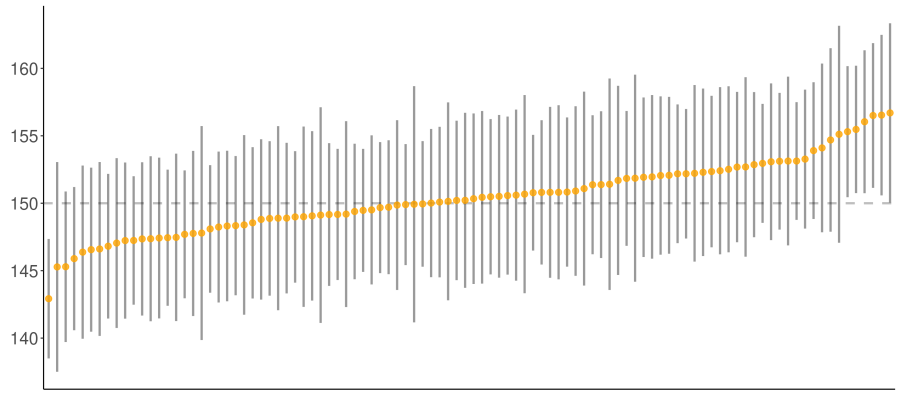

Figura 2.4 – Médias e intervalos de confiança para 100 amostras aleatórias de tamanho $n = 25$ retiradas de uma população normalmente distribuída com média $\mu = 150$ e desvio padrão $\sigma = 15$. Dos 100 intervalos calculados, apenas 7 não contêm a média populacional. Observe que cada amostra apresenta um desvio padrão distinto, logo, um intervalo distinto, mas a grande maioria deles contém o verdadeiro parâmetro populacional.

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[5]. 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.

2.3 Análise de Variância

Ao final da seção anterior, calculamos o erro padrão e o intervalo de confiança para os nossos dados. Contudo, até agora não falamos uma linha sequer sobre nossos sujeitos experimentais. Simplesmente olhamos para os dois conjuntos de dados como se fossem independentes, não relacionados. A partir de agora, portanto, vamos mudar isso e dizer que na verdade esses dados *não* são independentes, informando que tanto Palavras A quanto Palavras B foram lidas pelos mesmos sujeitos, como mostra a Tabela 2.2. Se, portanto, calcularmos o nosso intervalo de confiança do modo como feito acima, estamos tratando as nossas duas amostras como independentes quando elas na verdade não são. **Aquele método, portanto, não vai funcionar para os nossos dados.**[6] Por isso, agora, vamos partir para um método mais complexo. Esse método será a chamada Análise de Variância (ANOVA), que busca explicar a variabilidade dos dados após a aplicação do tratamento (no nosso caso, o tipo de palavra, se A ou B) e sem qualquer tratamento.

Para isso, vamos *ajustar alguns modelos matemáticos a nossos dados*. Não vamos detalhar agora o que estamos querendo dizer com essa afirmação, pois esperamos que até o final deste capítulo as coisas estejam mais claras para você, mas podemos adiantar que iremos em busca de um equação matemática que faça uma boa descrição dos nossos dados. Certamente, essa descrição não será perfeita (a realidade é sempre mais complexa do que nossas descrições, obviamente), mas poderemos achar uma que nos seja útil[7].

2.3.1 Ajustando uma ANOVA "na mão"

Calcular uma ANOVA ("na mão") não é muito difícil, apenas um pouco trabalhoso. Basicamente, vamos calcular desvios em relação a médias (coisa que você já sabe o que é e como se faz) e vamos calcular quadrados dos desvios (que você também já sabe como se faz) e vamos fazer a soma dos quadrados dos desvios (que você, adivinha, já sabe como se faz). O mais importante, no entanto, não é entender as contas, mas os princípios que estão

[5] Se você quiser uma abordagem didática sobre o tema, recomendamos a leitura de Howell (2009) e para uma defesa técnica rigorosa – e fervorosa – dos intervalos bayesianos, a referência clássica é Jaynes (1976).

[6] Ainda assim, é possível calcular um interval de confiança para dados como esses. Masson e Loftus (2003) propõem um método. Contudo, o método que propõem depende de entender a mecânica por trás das Análises de Variância.

[7] *"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."* (BOX; DRAPER, 1986).

Partic.	A	B	Média/Partic.
1	175.56	253.84	214.7
2	183.46	210.67	197.065
3	193.83	215.57	204.7
4	209.54	237.16	223.35
5	211.8	214.41	213.105
6	192.31	261.46	226.885
7	233.17	190.66	211.915
8	202.85	224.17	213.51
9	165.61	234.11	199.86
10	232	258.71	245.355
11	220.38	256.35	238.365
12	183.03	202.75	192.89
13	154.3	269.16	211.73
14	168.33	212.38	190.355
15	187.29	267.38	227.335
Médias	**194.23**	**233.91**	**214.07**

Tabela 2.2 – Médias para as amostras A e B, por condição e por sujeitos

por trás dessas contas. Mas, para facilitar as coisas, vamos repetir a tabela dos nossos dados abaixo (no início da página seguinte), com algumas modificações.

Primeiro, vamos ajustar um *modelo nulo* a nossos dados e compará-lo a um *modelo "cheio"*. Isso significa que vamos considerar a variação total dos nossos dados sem nenhum dos nossos tratamentos e compará-la com a variação com os tratamentos de modo a verificar se estes ajudaram a reduzir essa variação.

Isso é fácil de se fazer. Primeiro, vamos calcular uma média para os dados sem considerar o tipo de palavra, ou seja, tomar todos os valores de tempo obtidos para A e para B, somá-los e dividir pelo número de observações que fizemos (30). Essa média é 214.07 e muita vezes é chamada de grande média (*grand mean*). Ela está apresentada na Tabela 2.2 na última célula do canto inferior direito. Agora que temos uma média global, podemos calcular os desvios em relação a ela (cada um dos valores observados menos a grande média) e o quadrado dos resíduos. Por fim, fazemos a soma de quadrados desses resíduos, que é: 29088.81.

$$
\begin{aligned}
\mathrm{SQ}_{\mathrm{Total}} &= \sum (x_i - \bar{x}) \\
&= (175.56 - 214.07)^2 + (183.46 - 214.07)^2 + \ldots + (227.34 - 214.07)^2 \\
&= 29088.81 \\
SQ_{Trat} &= n \sum (\bar{x}_j - \bar{x}) \\
&= 15[(194.23 - 214.07)^2 + (233.92 - 214.07)^2] \\
&= 11813.53 \\
SQ_{Res} &= \sum (x_{ij} - \bar{x})
\end{aligned}
$$

Antes de prosseguir, pense sobre o que fizemos: nós simplesmente calculamos a variabilidade total dos dados, desconsiderando qualquer tratamento aplicado a eles. Por isso,

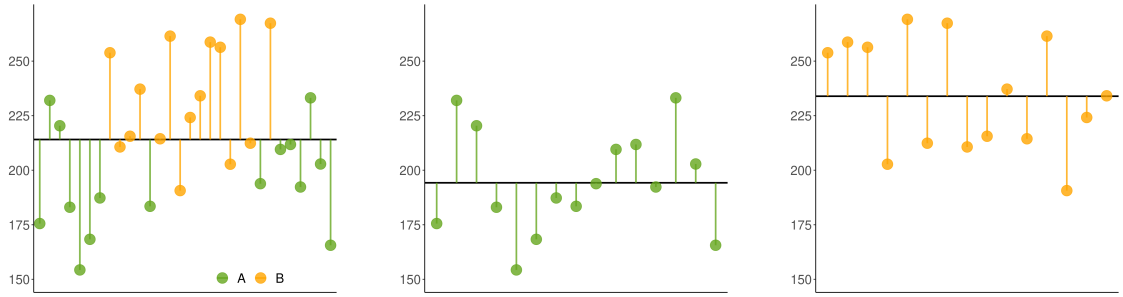

Figura 2.5 – Os três paineis mostram os desvios dos dados em três cenários. **Painel 1:** Desvios de todos os dados em relação à grande média, usados para se calcular a variância total ou Soma de Quadrados Total. **Paineis 2 e 3:** Mostram os mesmos dados do *Painel 1*, mas agora ilustrando, respectivamente, os desvios de A em relação à média de A (verde); e os desvios de B em relação à média de B (laranja). A soma dos quadrados desses desvios nos dá o quanto de informação de erro temos ao considerar os tratamentos, ou seja, a Soma de Quadrados dos Resíduos.

vamos chamar esse valor de *soma de quadrados total*. No Painel 1 da Figura 2.5 estão os desvios em relação a essa média global, ou seja, foi a partir desses desvios que calculamos a soma de quadrados total.

Agora que já temos a variabilidade total, podemos calcular a variabilidade dos resíduos **após o tratamento**. O nosso tratamento tem dois níveis (A e B), então basta calcularmos a média para A e para B e os respectivos quadrados dos resíduos, ou seja, aquilo que já fizemos no final do capítulo sobre estatística descritiva – volte até a Tabela 2.1 e confira aqueles números. A soma dos quadrados de cada um desses grupos de resíduos é: $SQ_{ResA} = 7970.44$; $SQ_{ResB} = 9304.83$. A soma de quadrados dos resíduos, então, é a soma desses dois valores: 17275.28. Os paineis 2 e 3 da Figura 2.5 ilustram essa variação.

Nessa fórmula, x_{ij} representa cada uma das observações i da condição j, e $\bar{x}_j$ a média de cada condição j. No nosso caso, cada uma das observações de 1 a 15 para a amostra A menos a média de A; e cada uma das observações de 1 a 15 para a amostra B menos a média de B, como mostram os painéis acima.

Mais uma vez, pense sobre isso: essa variabilidade é a variabilidade dos resíduos *dado o nosso tratamento*, ou seja, é o quanto de informação de erro temos, o quanto de informação não explicada pelos tratamentos. Se esse valor for muito pequeno, o tratamento explicou muito bem os dados (a maior parte da variação é explicada pelo tratamento). Se esse valor for muito grande, o tratamento não foi muito útil para explicar os dados (a variabilidade dos resíduos continuará próxima da variabilidade total). Em outras palavras, na situação hipotética de todos os pontos nos painéis 2 e 3 da Figura 2.5 estarem exatamente sobre as linhas das respectivas médias, isso significa que SQ_{res} é zero e que não há variabilidade nos dados. Eles seriam integralmente explicados pelos tratamentos (o que obviamente é um sonho... Como dissemos antes, não existem modelos perfeitos).

Repare que, se no primeiro caso consideramos a variabilidade total, sem considerar os tratamentos; e agora consideramos a variabilidade perdida, mesmo com o tratamento, o que sobrou é a variabilidade explicada pelo tratamento, ou seja, o quanto de redução de variação

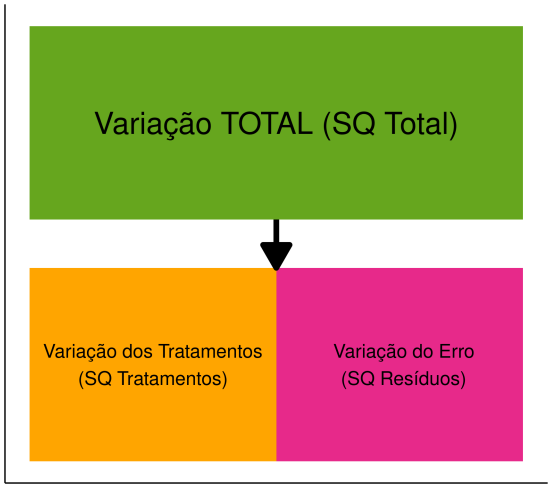

Figura 2.6 – Na Análise de Variância, a variabilidade total dos dados é dividida em seus componentes a fim de verificarmos o modelo que melhor se ajusta aos dados que temos. Daí compararmos a variabilidade explicada pelos tratamentos com a variabilidade não explicada por ele, ou seja, o quanto de variação devida ao erro temos.

tivemos depois que aplicamos o tratamento. Logo, a *soma de quadrados dos tratamentos* é simplesmente a *soma de quadrados total* menos a *soma de quadrados dos resíduos*, ou seja: $29088.81 - 17275.28 = 11813.53$. Mas, se você quiser calcular manualmente, essa soma é dada pela fórmula abaixo, onde n é o número de observações em cada tratamento e $\bar{x}_j$ é a média de cada tratamento.

Vamos resumir isso na Tabela 2.3 (que por enquanto está incompleta, mas logo iremos preenchê-la):

	G.L.	SQ	QM	F	p-valor
Tratamentos		11813.53			
Resíduos		17275.28			
Total		29088.81			

Tabela 2.3 – ANOVA 01 – somas quadráticas

Se você ficou um pouco perdido tentando acompanhar o que estivemos fazendo, basicamente foi o seguinte: calcular a variabilidade total dos dados e dividi-la em seus componentes. Parte dessa variabilidade é fruto do efeito do tratamento e parte é fruto de um valor não explicado, que chamamos de erro ou resíduos. Alguns parágrafos abaixo vamos explicar por que essa relação é importante.

Como dissemos antes, sempre que estamos estimando parâmetros populacionais, uma boa medida de quão precisos somos é o tamanho da nossa amostra. Isso será representado na coluna *graus de liberdade* (G.L.). Sabemos que o número total de observações é 30. Então vamos dizer que os *G.L. totais* são iguais a esse valor menos 1 $(n - 1)$, ou seja, 29. O dos tratamentos será igual ao número de tratamentos menos 1 $(k - 1)$. No caso, o nosso

fator palavra tem dois tratamentos, ou seja, dois níveis. Logo, nosso G.L. é 1. E o dos resíduos pode ser obtido por subtração, ou seja, *G.L. Total* menos *G.L. dos tratamentos*, ou seja, $29 - 1$, que é 28. Agora que temos isso, podemos calcular os *quadrados médios*, ou seja, a média dos quadrados, que é simplesmente a soma dos quadrados dividida pelos respectivos graus de liberdade. Pense sobre isso: estamos apenas dividindo o quanto de variação temos por uma métrica quanto ao tamanho da nossa amostra. Se tivermos muitas observações, teremos altos graus de liberdade e poderemos ter mais confiança nos nossos dados, logo a variabilidade mensurada pelos quadrados médios será menor. Logo:

	G.L.	SQ	QM	F	p-valor
Tratamentos	1	11813.53	11813.53	19.14	
Resíduos	28	17275.28	616.97		
Total	29	29088.81			

Tabela 2.4 – ANOVA 01 - graus de liberdade, quadrados médios e valor de F

Dado que chegamos até aqui, vamos pensar um pouco sobre a Tabela 2.4 a fim de entendermos o significado daquele valor de F, que ainda não sabemos calcular. A média de variação dos resíduos é bem baixa (616.97) se comparada à média de variação do tratamento (11813.53). Se dividirmos a variação dos tratamentos pela variação dos resíduos, teremos uma estimativa de quão útil esse tratamento é para explicar os dados obtidos, ou seja, qual a "proporção" na variação dos dados é explicada pelos tratamentos em relação aos erros ou resíduos.

É dessa relação que tiramos o valor de F, que é simplesmente o quadrado médio do tratamento (11813.53) dividido pelo quadrado médio dos resíduos (616.97). O que nos dá $F = 19.14$.

Para entender o que esse F significa, precisamos falar um pouco sobre distribuições de frequência, como a distribuição normal. Do mesmo modo que a área sob a curva normal pode ser dividida em probabilidades (porcentagens) dada a quantidade de desvios padrão que estamos distantes da média, existe uma distribuição de probabilidade chamada *distribuição de Fisher-Snedecor* ou distribuição-F. O valor de F dado pela ANOVA é basicamente um valor relacionado a essa distribuição, dados os graus de liberdade dos nossos tratamentos em relação aos resíduos. Sabendo que temos 1 grau de liberdade no numerador e 28 no denominador, podemos procurar em uma tabela da distribuição F qual a probabilidade de encontrarmos um valor de F igual a 19.14. Na tabela que tenho aqui em mãos (BUSSAB; MORETTIN, 2012), existe uma probabilidade menor do que 5% de encontrarmos um valor igual ou maior do que 4.20. É daí que vem o famigerado *p-valor* menor que 0.05 mostrado na Tabela 2.5 (Em geral, essa ANOVA seria reportada da seguinte maneira: $F(1, 28) = 19.14; p < 0.05$).

Já que falamos nele, vamos pensar um pouco sobre essa probabilidade. O que de fato ela quer dizer? Pense da seguinte maneira, como propõe Howell (2009): Sabemos que,

	G.L.	SQ	QM	F	p-valor
Tratamentos	1	11813.53	11813.53	19.14	0.000152
Resíduos	28	17275.28	616.97		
Total	29	29088.81			

Tabela 2.5 – ANOVA 01 - tabela completa. Para cada uma das ANOVAs a partir de agora, disponibilizaremos, também, o código do R para realizá-la. Neste caso, a tabela deve estar no formato longo, nossa variável resposta na coluna "RT", a variável preditora na coluna "palavras" e os sujeitos na coluna "sujeitos". O código nesse caso é: `aov(RT~palavras, data = dados)`.

sob a hipótese nula (H_0), as médias são iguais; e que, consequentemente, sob a hipótese alternativa (H_1), as médias são diferentes. Logo, podemos pensar que o QM_{res} é uma estimativa da variabilidade populacional; e que o QM_{trat} é uma estimativa da variabilidade populacional **se H_0 é verdadeira**, ou seja, se as médias são iguais, então não deve haver diferença entre os modelos com e sem tratamento, já que QM_{res} e QM_{trat} estão estimando a mesma coisa. Por isso, ao dividirmos QM_{trat} por QM_{res}, esperamos um valor igual a 1, **se H_0 é verdadeira**; e um valor maior do que 1, se H_0 é falsa.

Observe que, pela descrição acima, o *p-valor* é uma probabilidade condicionada pelo modelo que estamos usando. Basicamente, estamos dizendo que, assumindo um modelo para descrever os dados que diz serem nossas médias iguais, temos uma probabilidade menor do que 5% de encontrar dados como esses que encontramos. Matematicamente, podemos dizer que o *p-valor* nos dá $p(D \mid H_0)$, ou seja, a probabilidade dos dados dada a verdade da hipótese nula. Em geral, o *p-valor* é interpretado, **erroneamente**, como nos dando $p(H_0 \mid D)$, ou seja, a probabilidade da hipótese nula ser verdadeira (ou falsa, por oposição) dadas as observações que fizemos[8]. Mas o nosso modelo não pode dizer nada sobre a falsidade da hipótese nula, cuja verdade é uma assunção que nós, pesquisadores, fazemos de antemão à aplicação do modelo. Essa confusão é tão comum que gera um debate acalorado entre os estatísticos que remonta pelo menos aos anos 1930 – ver, por exemplo, Wasserstein e Lazar (2016), Greenland et al. (2016), Gigerenzer, Krauss e Vitouch (2004) e Levine et al. (2008).

Voltando ao nosso exemplo, como nosso valor de F (19.14) foi maior do que 4.20, e assumindo um modelo que nos diga que as médias são iguais, temos uma probabilidade menor do que 5% de obter dados como os que obtivemos neste experimento. Na verdade, se pedirmos para um computador calcular esse valor, ele será igual a 0.000152, justamente o valor que colocamos na Tabela 2.5.

Agora pare para pensar um minuto sobre o que estamos fazendo em termos de soma de quadrados, ou seja, em termos da variabilidade dos dados. A *soma de quadrados dos resíduos* representa a variabilidade dado o modelo completo (sem considerar os tratamentos); e a *soma de quadrados dos tratamentos* representa a variabilidade dado o modelo reduzido

[8] Se você quiser um modelo que informe esse tipo de probabilidade, então você terá que buscar a estatística bayesiana, onde esse tipo de pergunta é feita e respondida.

(considerando os tratamentos). O que essa tabela está nos dizendo é que, para os dados em questão, ao passar de um modelo que não considera os tratamentos para um modelo que considera os tratamentos, reduzimos a soma de quadrados de 17275.28 para 11813.53. Ou seja, tivemos uma redução de aproximadamente 68% na soma de quadrados. Essa, portanto, é a proporção de variação explicada pelo modelo. Em outras palavras, o modelo reduzido parece ser bem melhor porque ele se ajusta melhor aos dados: se o consideramos, podemos explicar melhor a variação encontrada. Uma vez que buscamos uma redução na soma de quadrados, esse método de estimação se chama *Método dos Mínimos Quadrados*.

$$\frac{SQ_{trat}}{SQ_{res}} = \frac{11813.53}{17275.28} = 0.6838 = 68.38\%$$

Para o experimento em questão, porém, esse modelo apresenta um grande problema: ele não considera a variabilidade devida aos sujeitos. De fato, o modelo que ajustamos seria válido apenas para o caso de amostras obtidas de *populações independentes* ou não relacionadas. Todavia, nossas amostras foram obtidas dos mesmos sujeitos, logo, elas não podem ser independentes. Vamos melhorá-lo, então, considerando esse aspecto.

2.3.2 Fatores fixos e fatores aleatórios

Para iniciar o debate sobre o problema acima, vamos começar assumindo que o modelo que acabamos de ajustar aos dados seja adequado. Se isso é verdade, agora temos confiança de que as nossas médias são diferentes não apenas na amostra que obtivemos, mas que esse efeito é um efeito real do tempo de leitura do tipo de palavra na população investigada: palavras do tipo A são lidas mais rapidamente do que palavras do tipo B. Isso significa que, se fizermos outro experimento, provavelmente obteremos uma diferença nessa direção. Por isso dizemos que nosso resultado é – ou pelo menos deveria ser – *replicável*.

Contudo, se olharmos bem para o nosso *design experimental*, vamos descobrir que talvez isso não seja totalmente verdade. Observe que nosso experimento tinha 15 sujeitos, que viram tanto palavras do tipo A quanto palavras do tipo B (tomamos medidas repetidas de cada sujeito). No entanto, esses 15 sujeitos não são toda a população de falantes de português, mas apenas uma amostra aleatória (e, supostamente, representativa dessa população). Vamos supor que nós decidamos então aplicar esse experimento mais uma vez, com 15 sujeitos distintos. Aqui há algumas possibilidades: esses 15 novos participantes são muito mais rápidos do que os primeiros; ou são muito mais lentos; ou se comportam de um modo totalmente novo e inesperado; etc. Se alguma dessas coisas acontece, não podemos ter mais confiança de que nossos resultados serão replicáveis. De fato, talvez o efeito que obtivemos seja devido aos 15 participantes específicos do meu experimento, que, por puro acaso do destino, ou alguma característica individual desse grupo, leram mais rapidamente palavras do tipo A do que palavras do tipo B.

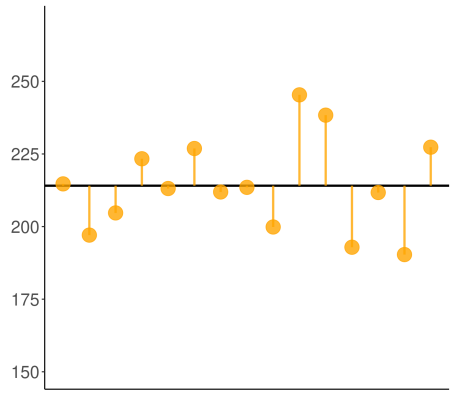

Figura 2.7 – A imagem ilustra a variação devida aos participantes do experimento, ou seja, o quanto eles variam em relação à grande média. Na ANOVA anterior, a variabilidade do erro estava inflada porque não considerava a variação dos sujeitos. Nesta, nosso modelo está controlando esse fator.

É por esse motivo que dizemos que os sujeitos, no tipo de experimento que fizemos, são chamados de um *efeito aleatório*: porque eles são uma amostra aleatória da população de interesse. Esse tipo de efeito é muito diferente do tipo de palavra (A ou B), que é um fator que nós, como cientistas, a cada vez que replicarmos o experimento, podemos controlar. Tipo de palavra, portanto, é um *efeito fixo*. Na ANOVA que ajustamos a nossos dados na seção anterior só usamos um efeito fixo (tipo de palavra), cujos níveis chamamos de *tratamento*. Todavia, ignoramos completamente a variabilidade advinda do grupo específico de sujeitos que fizeram o experimento.

Se você voltar à Tabela 2.2 no início dessa seção, verá que incluímos lá uma coluna para as médias de cada sujeito. Essas médias estão mostradas no painel acima. Como podemos ver, alguns sujeitos são rápidos em média (o 12 e o 14, por exemplo) e alguns são mais lentos em média (o 10 e o 11, por exemplo). Precisamos, portanto, ajustar uma ANOVA que considere a variabilidade dos sujeitos.

Essa ANOVA, no entanto, é um pouco mais complexa, visto que ela é o que se chama de *ANOVA para medidas repetidas*. Vamos pensar nela da forma mostrada pela Figura 2.8. Observe que, agora, estamos dividindo a variação total em dois grandes componentes. O primeiro é a variação entre os sujeitos (o quanto os sujeitos variam independente do tipo de palavras que estão lendo), ou seja, o quanto de erro temos graças às diferenças entre os sujeitos - a essa chamaremos de variação *between*. Uma vez que tivermos calculado essa variabilidade dos sujeitos, podemos simplesmente subtraí-la da variação total, obtendo a variação *within*. Logo, a variação *within* é uma variação "limpa" das diferenças entre sujeitos. Com essa variação, podemos, então, calcular normalmente se os tratamentos têm ou não efeito.[9]

[9] Um adendo aqui: é bem provável que as coisas não fiquem perfeitamente claras para você por enquanto. Pense um pouco sobre elas, mas não fique muito preso nos detalhes por agora. Mais à frente vamos dar outro exemplo e, com o tempo, as coisas vão fazendo mais sentido. Contudo, se você quiser se aprofundar no tema, recomendamos o *Capítulo 14 – Repeated-Measures Designs*, de Howell (2009).

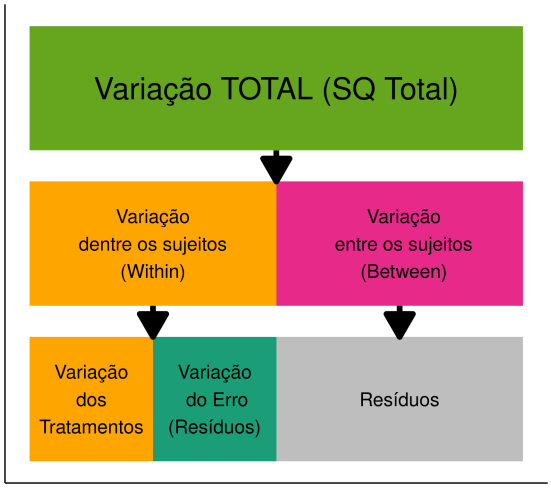

Figura 2.8 – Nesse modelo mais complexo, a variabilidade total dos dados é dividida primeiro em dois grandes componentes: a variação *between*, ou seja, quanto os sujeitos variam em média em relação à grande média; e a variação *within*, que é dividida entre a variação explicada pelos tratamentos e a variação não explicada por eles. A variação *within* é uma variação "limpa" das diferenças entre os sujeitos.

$$
\begin{aligned}
SQ_{Suj} &= k \sum (\bar{x}_{suj} - \bar{x}) \\
&= 2[(214.70 - 214.07)^2 + (197.07 - 214.07)^2 + \ldots + \quad (227.34 - 214.07)^2] \\
&= 7193.89 \\
SQ_{Res} &= SQ_{Tot} - SQ_{Trat} - SQ_{Suj} \\
&= 29088.81 - 11813.53 - 7193.89 \\
&= 10081.39
\end{aligned}
$$

Vamos, então, aos cálculos, seguindo os seguintes passos:

1. Calcular a soma de quadrados totais, como fizemos antes, que já sabemos ser 29088.81;

2. Em seguida, calcular a soma de quadrados dos tratamentos, que já sabemos ser 11813.53;

3. Então, calcular a soma de quadrados *between* sujeitos, ou seja, calcular a média de cada sujeito (elas estão na tabela, como já dissemos). E, então, calcular quantos esses sujeitos se distanciam da grande média, ou seja, calcular seus desvios. Daí calcular a soma dos quadrados dos desvios e multiplicar pelo número de tratamentos (k). No nosso caso, temos 2 tratamentos (A e B), logo $k = 2$. Esse resultado, para esse caso em particular, é 7193.89.

4. Com isso, podemos então calcular a soma de quadrados dos resíduos, que é simplesmente a soma de quadrados totais, menos a soma de quadrados dos tratamentos, menos a soma de quadrados dos sujeitos, ou seja: 10081.39:

Os dados que calculamos estão na Tabela 2.6, que já inclui os graus de liberdade e os quadrados médios (a soma de quadrados dividida pelos respectivos graus de liberdade). Esses, por sua vez, foram calculados da seguinte forma:

1. *between* sujeitos: número de sujeitos menos 1, ou seja, $15 - 1 = 14$;
2. dos tratamentos: número de tratamentos (k) menos 1, ou seja, $2 - 1 = 1$;
3. total: número total de observações (n) menos 1, ou seja, $30 - 1 = 29$.

Os graus de liberdade dos resíduos foram calculados por subtração, do mesmo modo como as somas quadráticas.

	G.L.	SQ	QM	F	p-valor
Between sujeitos:					
Resíduos	14	7193.89	513.84		
Within sujeitos:					
Tratamentos	1	11813.53	11813.53	16.40	0.00119
Resíduos	14	10081.39	720.08		
Total	29	29088.81			

Tabela 2.6 – ANOVA 02 - medidas repetidas para sujeitos. O código do R para esta ANOVA é: `aov(RT~palavras + Error(sujeitos), data = dados)`.

Observe que agora temos uma ANOVA que controla não só a variância dos tratamentos (QM_{trat}), mas também a variância dos sujeitos (QM_{suj}). Assim sendo, dado que a fórmula para o cálculo de F se mantém a mesma, ou seja, dividir o quadrado médio do tratamento pelo dos resíduos, fica a pergunta: o que mudou na nossa análise?

	G.L.	SQ	QM	F	p-valor
Tratamentos	1	11813.53	11813.53	19.14	0.000152
Resíduos	28	17275.28	616.97		
Total	29	29088.81			

Tabela 2.7 – Repetição da Tabela 2.5 para comparação.

Retorne a comparar as duas tabelas de ANOVA que ajustamos (a primeira foi repetida acima para facilitar o cotejamento). Lembre-se de que na primeira ANOVA que calculamos, sem controlar a variância dos sujeitos, F era igual a 19.14. Na ANOVA que acabamos de calcular, porém, F foi igual a 16.40, quase 3 unidades menor. Apesar da redução dos graus de liberdade (de 28 para 14), essa redução de F gerou um aumento no *p-valor*, de 0.000152 para 0.00119 (uma casa decimal). Isso é um indicativo de que nosso valor de F estava inflado no primeiro caso, sendo maior do que de fato deveria ser.

Pense um pouco sobre por que isso ocorreu. Na primeira ANOVA, o quadrado médio dos resíduos era 616.97, mas agora ele é 720.09, um valor maior. Logo, ao dividir o

quadrado médio do tratamento por esse denominador maior, teremos um resultado menor de F, já que o quadrado médio do tratamento não mudou. Isso aconteceu porque agora consideramos a variação dos sujeitos como parte do "erro experimental".

Isso explica muito a ideia por trás de um fator aleatório. Ele é um fator que, apesar de não ser de interesse do experimentador, de fato precisa ser controlado, pois pode enviesar a análise. No fundo, isso não fez grande diferença para esse experimento em particular, mostrando que o tipo de palavra parece efetivamente levar a maiores ou menores tempos de reação. Contudo, em outro experimento, poderíamos ter um valor significativo sem que ele realmente fosse – mostraremos um caso desse tipo daqui a pouco.

Aqui cabe, ainda, um breve comentário: quando controlamos a variação dos sujeitos, controlados em média o quanto eles são mais rápidos ou mais lentos em relação à grande média. Mas ignoramos completamente o fato de os sujeitos poderem interagir com o tipo de palavra. Dizendo que outro modo: um sujeito qualquer pode ser mais rápido ou mais lento em média, mas ele poderia ser afetado de maneiras distintas por palavras do tipo A ou por palavras do tipo B. O ideal, portanto, é que, para esse modelo, considerássemos também essa variabilidade. Contudo, se fizéssemos isso, teríamos uma ANOVA idêntica à que acabamos de calcular[10], isso porque temos apenas uma observação por condição por sujeito: cada sujeito vê apenas uma palavra do tipo A na condição A e uma palavra do tipo B na condição B. Normalmente não é assim que se elaboram experimentos em psicolinguística. O que se faz é termos múltiplas observações por condição por sujeito, por exemplo, 4 palavras do tipo A e 4 do tipo B, sendo que cada sujeito vê as 4 palavras em cada condição. Esse é um *design* mais complexo e as contas para ele são um pouco diferentes. Para abordar esse tipo de experimento, vamos nos despedir desses dados com os quais até agora viemos trabalhando e passemos a olhar para um novo, em que colocaremos o problema dos itens experimentais.

2.3.3 Itens como efeito aleatório

Para o experimento que propusemos inicialmente, um *design within sujeitos*, ou seja, tomando medidas repetidas de cada participante – cada participante tinha seu tempo medido em palavras do tipo A e também em palavras do tipo B –, a ANOVA que ajustamos está quase adequada (estaria perfeita se tivéssemos calculado a interação entre sujeitos e o tipo de palavra). No entanto, observe que até agora deixamos – propositalmente – uma informação de lado. Que palavras são essas para quais os tempos estão sendo mensurados?

Dado que não falamos nada sobre isso, vamos supor que, para palavras do tipo A, tenhamos colhido aleatoriamente 15 palavras num dicionário; e para palavras do tipo B, a mesma coisa, ou seja, temos 30 itens experimentais. Observe que, se esse for o caso, temos

[10] Se você quiser confirmar, compare o resultado dessa ANOVA (`aov(RT~palavras + Error(sujeitos), data = dados)` com o resultado dessa `aov(RT~palavras + Error(sujeitos/palavras), data = dados)`.

um *design within sujeitos* (o mesmo sujeito viu todas as condições), mas *between itens* (o mesmo item só aparecia em uma condição). Isso é o mesmo que dizer que sujeito e tipo de palavra eram fatores cruzados (*crossed factors*) e itens é um fator aninhado (*nested factor*) em tipo de palavra, ou seja, não temos medidas repetidas para os itens: cada participante viu um item diferente.

Como o problema dos efeitos aleatórios está diretamente ligado à replicabilidade do experimento, como vimos anteriormente, então vamos imaginar que decidimos reaplicar esse experimento em três cenários distintos:

1. *Cenário 1: vamos usar o mesmo design, os mesmos sujeitos e os mesmos itens.* Para esse caso, mesmo com tudo exatamente igual, parece óbvio que o resultado não será idêntico ao obtido na primeira realização do experimento. Isso vai acontecer por causa da variabilidade inerente a qualquer situação, a variação incontrolada, ou seja, o fator de erro.

2. *Cenário 2: vamos usar o mesmo design, os mesmos itens, mas sujeitos diferentes.* Para esse caso, além do fator de erro, temos ainda o fato de os sujeitos que fizeram esse experimento serem mais rápidos ou mais lentos do que os do primeiro experimento. Logo, nosso resultado provavelmente não será idêntico devido a dois fatores: a variabilidade incontrolada (o erro) e a variabilidade dos sujeitos.

3. *Cenário 3: vamos usar o mesmo design, os mesmos sujeitos, mas itens diferentes.* Do mesmo modo como anteriormente, os novos itens podem ser lidos mais rapidamente ou mais demoradamente. Logo, nosso resultado pode não ser o mesmo por dois motivos distintos: a variabilidade incontrolada (o erro) e a variabilidade dos itens.

Dizendo de outro modo, é preciso controlar a variância não só dos sujeitos a fim de evitar um valor de F inflado, mas também a variância dos itens. Se não fizermos isso, corremos o risco de obter um valor de F significativo que na verdade não adveio do efeito do tratamento, mas simplesmente da variabilidade dos itens amostrados. Ajustar um modelo a esses dados que não considere a variabilidade dos itens é cair naquilo que se chama a *falácia da língua como um efeito fixo*, um problema estatístico que foi descrito pela primeira vez por Clark (1973) e que tem suas raízes em Coleman (1964). Recomendamos, também, a leitura de Raaijmakers (2003) para um apanhado geral sobre o tema e de Carver (1978) para um debate sobre o assunto.

2.3.4 Mais um exemplo fictício

Vamos começar com um exemplo simples a fim de verificarmos como a variabilidade dos itens pode influenciar dramaticamente na nossa análise[11]. Imagine que recrutamos

[11] Aqui me cabe uma *mea culpa* importante: esses dados são claramente não normais e excessivamente simplórios. Obviamente que uma ANOVA não seria um modelo adequado para eles. O ponto, porém,

cinco sujeitos e os submetemos à leitura de palavras no singular e no plural. Para tanto, selecionamos, aleatoriamente, três palavras no dicionário e mensuramos o comportamento desses sujeitos lendo essas palavras (no singular e no plural). Para o caso em questão, não importa que medidas tomamos – esse é apenas um exemplo didático para explicar o problema. Queremos saber, portanto, se o número da palavra (singular ou plural) afeta o comportamento do sujeito. Os dados amostrados estão na tabela abaixo.

Número	Palavra		Número	Palavra		
singular	p_1	1	plural	p_1	15	
singular	p_2	10	plural	p_2	6	
singular	p_3	5	plural	p_3	12	
singular	p_1	8	plural	p_1	13	
singular	p_2	12	plural	p_2	7	
singular	p_3	4	plural	p_3	16	
Médias		6.66			11.5	**9.08**

Tabela 2.8 – Dados para os itens

Observe que, olhando para as médias de cada grupo, podemos ficar felizes: há uma grande diferença entre as médias obtidas na condição singular (6.66) e na condição plural (11.5). O valor em negrito (9.08) é a grande média, ou seja, a média de todos os valores. Mas será que a diferença entre as médias de singular e de plural é significativa? Primeiro, vamos calcular uma ANOVA simples, considerando número como fator fixo e ignorando os itens – para esse exemplo específico, vamos ignorar os sujeitos também. Como você já sabe, essa ANOVA não é adequada, já que ela considera que as amostras são independentes, mas, no nosso caso, elas não são, visto que os mesmos sujeitos viram tanto palavras no singular quanto no plural.

Esses dados estão resumidos na Tabela 2.9 da ANOVA.

	G.L.	SQ	QM	F	p-valor
Tratamentos	1	70.27	70.27	4.16	0.068
Resíduos	10	168.64	16.26		
Total	11	238.91			

Tabela 2.9 – ANOVA 03 - Desconsiderando a variabilidade dos itens. O código para esta ANOVA é: `aov(valor~numero, data = dados)`.

não é esse, mas sim o entendimento da maquinaria por trás da ANOVA e os motivos pelos quais elas não são adequadas para analisar, na maioria das vezes, dados linguísticos. Logo, peço encarecidamente que o leitor releve essa questão por enquanto. A partir do próximo capítulo trabalharemos com dados reais e não faremos mais esse tipo de coisa.

$$
\begin{aligned}
SQ_{Total} &= \sum (x_i - \bar{x}) \\
&= (1 - 9.08)^2 + (10 - 9.08)^2 + \ldots + (7 - 9.08)^2 + (16 - 9.08)^2 \\
&= 238.91 \\
SQ_{Trat} &= \sum (\bar{x}_j - \bar{x}) \\
&= 6[(6.66 - 9.08)^2 + (11.50 - 9.08)^2] \\
&= 70.27 \\
SQ_{Res} &= SQ_{Total} - SQ_{Trat} \\
&= 238.91 - 70.27 \\
&= 168.64
\end{aligned}
$$

Observe que obtivemos um p-valor comumente dito *marginalmente* significativo (com 1 e 10 graus de liberdade, minha tabela diz que deveríamos ter um valor de F maior do que 4.96 para ser significativo a 5%). Ora, muitos artigos reportam valores marginais. Se replicássemos, portanto, esse experimento, poderíamos crer que é bem provável que o efeito de número se mostre um efeito real.

A questão é que, nessa próxima replicação, teríamos que selecionar um novo conjunto de itens. Precisamos, portanto, incluir os itens como um fator aleatório cruzado (*crossed* ou *within*) no fator Número e calcular outra ANOVA. Para isso, vamos calcular a média de cada um dos itens:

p1	p2	p3
9.25	8.75	9.25

Tabela 2.10 – Médias para os itens experimentais.

Agora, vamos subtrair cada um desses valores da grande média e multiplicá-lo pelo *número de observações por item*, ou seja, 4. Vamos chamá-la de *soma de quadrados between itens*, já que ela nos mostra a variabilidade total dos itens independente da condição em que aparecem.

$$
\begin{aligned}
SQ_{Bet} &= w \sum (\bar{x}_{item} - \bar{x})^2 \\
&= 4[(9.25 - 9.08)^2 + (8.75 - 9.08)^2 + (9.25 - 9.08)^2] \\
&= 0.666
\end{aligned}
$$

Pense um pouco sobre essa soma. Ela nos deu um valor bem baixo se comparado ao restante, mas o que ela está estimando? A conta que acabamos de fazer mostra o quanto esses itens variam em média em relação à grande média (9.08). E isso de fato é muito

pouco. A média de todos eles é bem próxima daquela média global. No entanto, mesmo essa pequena variação pode afetar nossos resultados, como veremos em seguida. Até aqui, nada diferente do que fizemos para os sujeitos.

Vamos, então, construir nossa tabela de ANOVA, que está abaixo. Observe que ela é idêntica à tabela 2.6 de ANOVA que usamos quando incluímos os sujeitos na nossa análise. Isso porque nós dividimos a variabilidade em dois grupos: a variabilidade *between itens* (variação total dos itens) e a variabilidade *within itens* (variação dos itens dentro de cada grupo). Lembre-se, também, que duas das somas quadráticas aí presentes nós já calculamos quando fizemos a ANOVA simples: a dos *tratamentos* e a *total*.

	G.L.	SQ	QM	F	p-valor
Between itens:					
Resíduos	2	0.666	0.333		
Within itens:					
Tratamentos	1	70.27	70.27	3.34	0.105
Resíduos	8	167.97	20.99		
Total	11	238.91			

Tabela 2.11 – ANOVA 04 – medidas repetidas para os itens. O código para esta ANOVA é `aov(valor~numero + Error(palavras), data = dados)`, idêntica à ANOVA 02 (Tabela 2.6), que ajustamos considerando os sujeitos.

Antes de entrarmos no p-valor, cabe ainda um comentário sobre os graus de liberdade, que foram calculados da seguinte forma:

1. *between itens*: número de itens menos 1, ou seja, $3 - 1 = 2$;
2. dos tratamentos: número de tratamentos (k) menos 1, ou seja, $2 - 1 = 1$;
3. total: número total de observações (n) menos 1, ou seja, $12 - 1 = 11$.

Mais uma vez, os graus de liberdade dos resíduos foram calculados por subtração, do mesmo modo como as somas quadráticas.

Então, vamos ao p-valor: como você deve ter notado, nosso p-valor não é mais sequer marginalmente significativo. Ele foi de 0.06 na primeira ANOVA para 0.1 na segunda (a minha tabela diz que o valor de F teria que ser maior do que 5.32, com 1 e 8 graus de liberdade). Isso ocorreu porque o quadrado médio dos resíduos aumentou de 16.86 para 20.99, enquanto os graus de liberdade caíram de 10 para 8. Mas o que isso significa em termos teóricos? Por que isso ocorreu?

Como você deve se lembrar, as médias de cada uma das condições eram bem diferentes (6.66 e 11.5). O problema é que, apesar de as médias serem bem distintas entre as condições, havia uma pequena variabilidade entre os itens que estava a comprometer nossa análise.

Como vimos na ANOVA acima, o efeito significativo desapareceu quando controlamos a variabilidade dos itens experimentais. Observe, no entanto, que esse não é o modelo

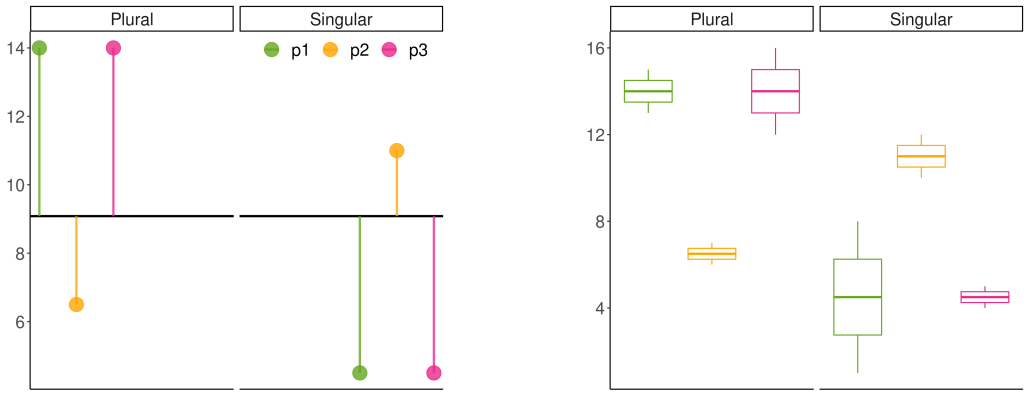

Figura 2.9 – **Painel 1:** Desvios dos itens experimentais por célula, ou seja, o quanto a média de cada item por condição se afasta da grande média. **Painel 2:** Boxplot com o comportamento de cada item por condição. Ambos os paineis mostram que o item $p2$ tem um comportamento discrepante dos demais.

mais complexo permitido para esse *design*, em que itens são um fator cruzado (*crossed* ou *within* tratamentos) com Número. O modelo que acabamos de ajustar é um modelo que simplesmente considera os itens como um fator (aleatório) e calcula um efeito para ele. Nesse caso, o modelo considera a variação total de cada item em relação à grande média[12]. O problema é que os itens parecem interagir com o Número. Enquanto alguns itens (p_1 e p_3) têm valores menores para singular, outros (p_2) têm valores maiores, como mostra a Tabela 2.12.

	p_1	p_2	p_3
plural	14	6.5	14
singular	4.5	11	4.5

Tabela 2.12 – Médias por itens por condição

Como você pode notar, o plural de fato tem uma média maior do que o singular. Todavia, há uma enorme variabilidade entre os itens, já que p_1 e p_3 são muito consistentes (têm maiores valores no plural e menores no singular), mas p_2 não. Esse item em particular tem menores tempos no plural e maiores no singular.

Imagine que você vai replicar esse experimento. Vamos supor, então, que você tenha amostrado aleatoriamente vários itens. Todavia, ao contrário do que ocorreu neste experimento, na replicação você selecionou, por acaso, itens muito parecidos com p_2. Se isso de fato ocorrer, o que vai acontecer com a diferença entre as médias? Vai se inverter: agora o singular terá médias maiores do que o plural. Por outro lado, se você amostrar, nessa replicação, vários itens parecidos com p_1 e/ou p_3, então a diferença entre as médias tende a se manter. Em outras palavras, se você não controla a variância dos itens, como poderá

[12] Quando estivermos tratando de modelos mistos, diremos que estamos calculando "interceptos" para os itens.

saber que o efeito que obteve é devido ao tratamento e não ao conjunto particular de itens que foi amostrado?

Devido a esse fato, para o *design* particular com que estamos trabalhando, precisamos considerar a interação entre o fator fixo (Número) e o fator aleatório (Item)[13]. Logo, temos mais contas a fazer. A nossa sorte é que já sabemos:

1. a soma quadrática total: 238.91;
2. a soma quadrática dos tratamentos: 70.27;
3. a soma quadrática *between* itens: 0.666.

Vamos, então, calcular a soma quadrática da interação *itens* × *tratamentos*. Para fazer isso, vamos calcular uma coisa chamada *soma de quadrados das células*. Observe a Tabela 2.13, simplesmente uma reorganização da Tabela 2.8 com algumas médias a mais:

	singular	plural
p_1	1	15
	8	13
Médias	4.5	14
p_2	10	6
	12	7
Médias	11	6.5
p_3	5	12
	4	16
Médias	4.5	14

Tabela 2.13 – Cálculo das médias por células (cruzamento itens × tratamentos)

Cada cruzamento de um nível do fator tratamento com um nível do fator item é chamado de uma célula. Para cada uma dessas células, calculamos uma média. Observe as médias aí presentes e veja como elas se relacionam com o Painel 2 da Figura 2.9, acima. Essa tabela, portanto, é uma descrição daquela variação que vimos no gráfico. A *soma de quadrados das células* será a soma de quadrados da média de cada célula menos a grande média (Painel 1 da Figura 2.9)multiplicada pelo número de observações em cada célula (n).

[13] No âmbito do modelo misto, diremos que estamos calculando "inclinações" (*slopes*) para os itens em função dos tratamentos.)

$$
\begin{aligned}
SQ_{Cell} &= n\sum(\bar{x}_{Cell} - \bar{x})^2 \\
&= 2[(4.5 - 9.08)^2 + (14 - 9.08)^2 + \ldots + (14 - 9.08)^2] \\
&= 201.41 \\
SQ_{Trat\times Itens} &= SQ_{Cell} - SQ_{Tratamentos} - SQ_{Itens} \\
&= 201.41 - 70.27 - 0.666 \\
&= 130.66 \\
SQ_{Res} &= SQ_{Total} - SQ_{Cell} \\
&= 238.91 - 201.41 \\
&= 37.5
\end{aligned}
$$

Vamos pensar um pouco sobre isso. Nós calculamos a soma quadrática dos tratamentos (digamos, o efeito dos tratamentos), a soma quadrática dos itens (o efeito dos itens, ou seja, quanto em média cada item varia em relação à grande média) e a soma quadrática das células (o quanto cada item varia em função da condição – a média de cada item específico dentro de cada condição em relação à grande média). A soma quadrática da interação, portanto, será dada pela fórmula:

E os resíduos finais serão dados pela fórmula:

Os graus de liberdade, por sua vez, serão dados por:

1. Itens: número de itens menos 1, ou seja, $3 - 1 = 2$;
2. Tratamentos: número de tratamentos menos 1, ou seja, $2 - 1 = 1$;
3. *Between*: soma dos graus de liberdade dos itens mais dos tratamentos, ou seja, $3 + 2 = 5$;
4. Resíduos da interação entre Tratamentos ×Itens é obtida por subtração: *between* - itens - tratamentos $= 5 - 2 - 1 = 2$
5. Resíduos totais (*Within*): Total (11)–between (5) $= 11 - 5 = 6$.

Observe que a SQ das células não aparece na tabela. Ela é usada apenas para fazer os cálculos.

Se buscarmos em uma tabela o valor de F (com 1 e 2 graus de liberdade) a fim de obtermos um p-valor significativo a 5%, ela nos dará 18.51, ou seja, teríamos que ter um valor de F maior do que isso, mas estamos muito longe desse número, indicando que a variabilidade dos itens é tão grande, para esses dados, que não podemos rejeitar a hipótese nula.

No caso específico desse experimento, é impossível separar o efeito do tratamento do efeito dos itens na resposta obtida. Por isso, apesar de a primeira ANOVA ter apresentado

	G.L.	SQ	QM	F	p-valor
Between itens:	5				
Resíduos	2	0.666	0.333		
Tratamentos	1	70.27	70.27	1.075	0.408
Resíduos $Trat \times Itens$	2	130.66	65.33		
Within itens:					
Resíduos	6	37.5	6.25		
Total	11	238.91			

Tabela 2.14 – ANOVA 05 – medidas repetidas para os itens com interação, ou seja, itens (palavras) aninhados (*nested*) em número. O código para esta ANOVA é: `aov(valor~numero + Error(palavras/numero), data = dados)`.

o p-valor marginalmente significativo, quando ajustamos o modelo correto, isso não ocorreu, ou seja, a ANOVA avaliou a variabilidade dos itens, dos tratamentos e disse para gente: "dada essa variação toda, não é possível rejeitar a hipótese nula".

2.3.5 O problema final: sujeitos e itens na mesma análise

Se você chegou até aqui, deve ter percebido que a ANOVA é uma poderosa ferramenta de análise de dados, mas ela tem um problema fatal para o caso da maioria dos experimentos da área de psicolinguística: ela não consegue incorporar, em sua análise, mais de um efeito aleatório e os dados de psicolinguística geralmente têm dois: sujeitos, escolhidos aleatoriamente da população; e itens linguísticos, que são uma amostra aleatória de todos os itens possíveis na língua.

A solução estatística clássica para esse problema foi dada por Clark (1973). Em vez de calcular o p-valor com base na distribuição F, ele deveria ser calculado com base no que ele chamou de *mínimo quasi F* ou *mín-F'*, para os íntimos. Esse valor nos daria, resumidamente, um p-valor que compensaria a variabilidade dos itens e dos sujeitos e forneceria um p-valor confiável, não enviesado. Para calculá-lo, basta realizar duas ANOVAs, uma que considera os sujeitos como fator aleatório (F_1) e uma que considera os itens como fator aleatório (F_2). Com os valores de F desses modelos, Clark (1973) mostra que o *mín-F'* é dado por:

$$\text{Min-F'} = \frac{F_1 \times F_2}{(F_1 + F_2)}$$

Essa solução, no entanto, foi considerada, posteriormente, como muito conservadora (ver referências em Raaijmakers (2003) e o que se tornou padrão na literatura da área foi simplesmente calcular F_1 e F_2 e considerar um resultado significativo **se ambas as análises** fossem significativas.

Deve-se destacar, ainda, que, além da solução estatística, existe uma solução teórica, ou seja, construir um *design experimental* que controle a variabilidade dos itens, deixando apenas os sujeitos como um fator aleatório. Esse tipo de *design* é chamado de *design*

contrabalanceado e pode ser visto em Raaijmakers (2003). Como esse autor mostra, porém, há uma questão interessante de *sociologia da ciência* no caso dos usos desse tipo de experimento, visto que alguns pareceristas tendem a não aceitar as análises estatísticas que não apresentam uma ANOVA para os itens (F_2), mesmo quando ela é desnecessária.

Isso se manteve mais ou menos assim, com muita discussão e polêmica ao longo dos anos, até 2008, quando dois artigos introduziram um novo paradigma na área: Baayen, Davidson e Bates (2008) e Jaeger (2008). Esses autores argumentaram em favor de as análises que consideram sujeitos e itens como fatores aleatórios serem realizadas com os chamados *modelos mistos* ou *modelos hierárquicos* (ver, também, Godoy e Nunes (2020)). O interessante é que, apesar de esses modelos serem novos na psicolinguística, eles já vinham sendo utilizados há longo tempo na sociolinguística, pelo menos os modelos para variáveis dependentes categóricas, os chamados *modelos logísticos* (SANKOFF; LABOV, 1979), que discutiremos mais à frente.

Na próxima seção discutiremos os modelos mistos em geral – agora com dados de experimentos reais –, como ajustá-los, quais vantagens eles têm em relação à Análise de Variância, além de verificarmos os pressupostos desses modelos, assunto no qual sequer tocamos até agora.

Referências

BAAYEN, R. H.; DAVIDSON, D. J.; BATES, D. M. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, Elsevier, v. 59, n. 4, p. 390–412, 2008.

BOX, G.; DRAPER, N. *Empirical Model-Building and Response Surfaces*. [S.l.]: Wiley–Blackwell, 1986.

BUSSAB, W. O.; MORETTIN, P. A. *Estatística Básica*. [S.l.]: Editora Saraiva, 2012.

CARVER, R. P. Sense and nonsense about generalizing to a language population. *Journal of Reading Behavior*, Sage Publications, v. 10, n. 1, p. 25–33, 1978.

CLARK, H. H. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of verbal learning and verbal behavior*, Elsevier, v. 12, n. 4, p. 335–359, 1973.

COLEMAN, E. B. Generalizing to a language population. *Psychological Reports*, SAGE Publications Sage CA: Los Angeles, CA, v. 14, n. 1, p. 219–226, 1964.

GIGERENZER, G.; KRAUSS, S.; VITOUCH, O. The null ritual: What you always wanted to know about significance testing but were afraid to ask. In: KAPLAN, D. (Ed.). *The Sage handbook of quantitative methodology for the social sciences*. Oaks: Sage Publications, 2004. cap. 21, p. 391–408.

GODOY, M. C.; NUNES, M. A. Uma comparação entre anova e modelos lineares mistos para análise de dados de tempo de resposta. *Revista da ABRALIN*, v. 19, n. 1, p. 1–23, 2020.

GREENLAND, S. et al. Statistical tests, p-values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, Springer, v. 31, p. 337–350, 2016.

HOWELL, D. C. *Statistical methods for psychology*. [S.l.]: Cengage Learning, 2009.

JAEGER, T. F. Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of memory and language*, Elsevier, v. 59, n. 4, p. 434–446, 2008.

JAYNES, E. T. Confidence intervals vs bayesian intervals. In: HARPER; HOOKER (Ed.). *Foundations of probability theory, statistical inference, and statistical theorys of science*. Dordrecht-Holland: D. Reidel Publishing Company, 1976. v. 2, p. 175–257.

LEVINE, T. R. et al. A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research*, Oxford Academic, v. 34, n. 2, p. 171–187, 2008.

MASSON, M. E. J.; LOFTUS, G. R. Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, v. 57, n. 3, p. 203–220, 2003.

RAAIJMAKERS, J. G. W. A further look at the"language-as-fixed-effect fallacy". *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, Canadian Psychological Association, v. 57, n. 3, p. 141, 2003.

SANKOFF, D.; LABOV, W. On the uses of variable rules. *Language in society*, Cambridge University Press, v. 8, n. 2-3, p. 189–222, 1979.

WASSERSTEIN, R. L.; LAZAR, N. A. The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, Taylor & Francis, v. 70, n. 2, p. 129–133, 2016.