Estatística aplicada à psicolinguística experimental

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

Howell, 2010: 462

Justificativa prévia:

O material que você agora tem em mãos propõe-se a ser uma introdução a conceitos básicos de estatística descritiva e estatística inferencial. O público a que se destina são estudantes da área de psicolinguística completamente leigos em estatística. Nosso objetivo, aqui, não é prover uma explicação exaustiva desses conceitos, mas apenas introduzi-los de maneira mais ou menos intuitiva, de modo que esses alunos possam, a partir dessa introdução, se aprofundar em textos mais densos da área que, muitas vezes, assumem um conhecimento prévio que os alunos não têm.

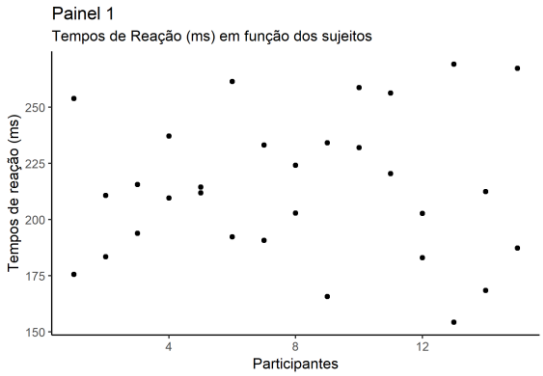
Nossa abordagem partirá dos conceitos mais básicos em estatística descritiva (*medidas de posição* e *de dispersão*) e então entrará brevemente em inferência estatística a fim de introduzir aquilo que ficou conhecido na literatura como *falácia da língua como efeito fixo* (Clarke, 1974; Coleman, 1964). Nesse processo, em vários momentos daremos grandes saltos conceituais. Por isso, consideramos adequado que o leitor interessado busque se aprofundar mais no tema por conta própria, tanto lendo livros básicos de estatística quanto lendo os artigos que tratam dos problemas estatísticos em psicolinguística (que recomendamos na bibliografia anexa). Recomendamos a todos, portanto, que esse material *não seja usado* como guia final para suas análises.

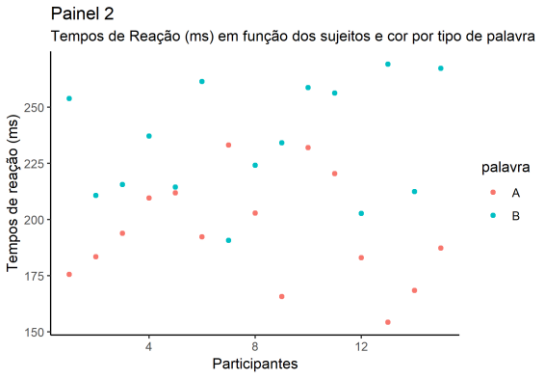
Alguns conceitos básicos em estatística descritiva

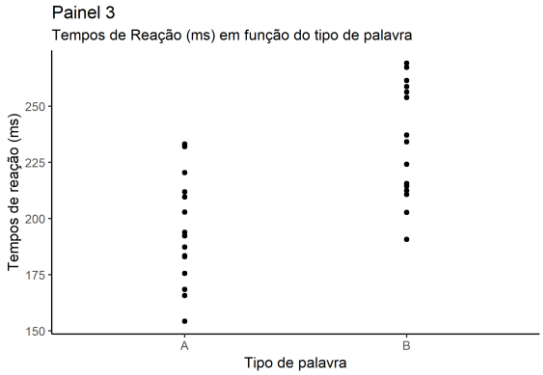
Imaginemos dois conjuntos de dados, digamos, as medidas do tempo de reação (RT, do inglês *reaction time*) para duas amostras (A e B) retiradas de populações distintas. Os dados amostrais estão abaixo.

Participantes	A	B
1	175.56	253.84
2	183.46	210.67
3	193.83	215.57
4	209.54	237.16
5	211.80	214.41
6	192.31	261.46
7	233.17	190.66
8	202.85	224.17
9	165.61	234.11
10	232.00	258.71
11	220.38	256.35
12	183.03	202.75
13	154.30	269.16
14	168.33	212.38
15	187.29	267.38

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







O painel 1 é pouco informativo sobre os dados, já que não nos permite visualizar as condições experimentais. O painel 2 nos mostra com mais clareza a distribuição dos dados para a amostra A (laranja) e para a amostra B (azul), sugerindo que os tempos de reação para B parecem ser maiores do que os tempos para A. O painel 3, então, agrupa os pontos para cada condição, ignorando os sujeitos, o que deixa ainda mais claro que, apesar de muitos valores sobrepostos, há a sugestão de uma leve tendência de B ser maior do que A.

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

1. Medidas da tendência central

1.1. A 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:

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

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

Por exemplo, se x é o conjunto de dados abaixo:

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

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

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

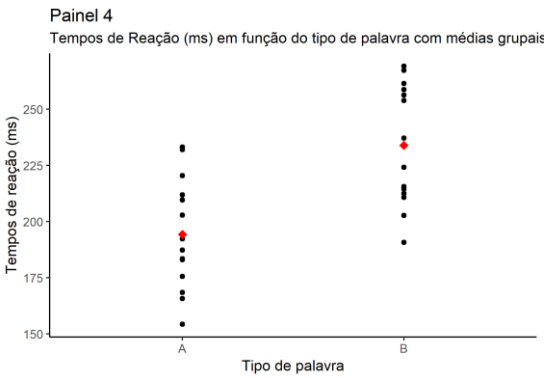
Se a variável x estivesse disposta em uma tabela vertical (a mais comum para a análise de dados), poderíamos dizer que o índice subscrito a cada *x* seria cada uma das linhas da tabela, como abaixo:

Linha	x	
1	2	x_1=2
2	3	x_2=3
3	5	x_3=5
4	6	...
5	8	
6	9	
7	2	
8	4	
9	7	
10	4	x_{10}=4

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, essa notação não seja motivo para complicação no entendimento.

Isso tendo sido dito, passemos a uma análise inicial dos valores contidos nas amostras A e B.

Para as amostras A e B, as médias dos tempos de reação são, respectivamente: 194,2 e 233,9 milisegundos – uma confirmação, até certo ponto, de que o valor do RT para B realmente é mais alto do que o valor para A. Essa diferença pode ser representada no gráfico de pontos. Losângulos vermelhos indicam as médias de cada grupo:



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.

1.2. A 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:

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

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

	x	y	z
Média	5	7,27	10
Mediana	4,5	5	5

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

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

1.3. Voltando ao exemplo

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

	A	B
Média	194,2	233,9
Mediana	192,3	234,1

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.

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.

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, 2, 3, 4, 4, 5, 6, 7, 8, 9}
y={2, 2, 3, 4, 4, 5, 6, 7, 8, 9, 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.

	Mínimo	1º quartil (q_1) (25%)	2º quartil (q_2) (50%)	3º quartil (q_3) (75%)	Máximo
x	2		4,5		9
y	2		5		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. Assim:

x = {2, 2, **3**, 4, 4,} (4,5) {5, 6, **7**, 8, 9}

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

	Mínimo	1º quartil (q_1) (25%)	2º quartil (q_2) (50%)	3º quartil (q_3) (75%)	Máximo
x	2	3	4,5	7	9
y	2	3	5	8	30

Tendo descoberto os quartis, temos uma visão global dos dados. Assim, conhecemos os valores centrais da amostra ou população que estamos estudando, ou seja, os valores que se encontram entre o 1º e o 3º quartis, excluindo-se, portanto, os extremos. Essa diferença é chamada de *Amplitude Interquartil* (AIQ = $q_3 - q_1$). Essa sumarização dos dados nos dá uma visão mais global dos valores com que estamos trabalhando, mostrando como os valores encontrados se distribuem. Nos casos acima, temos que AIQ (X) = 7 – 3 = 4; e AIQ (Y) = 8 – 3 = 5. Assim, y parece ter uma dispersão levemente maior do que x.

2.1.1. 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:

$$lim.inferior = q_1 - (1{,}5)\text{AIQ}$$
$$lim.superior = q_3 + (1{,}5)\text{AIQ}$$

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:

$$lim.inferior(x) = 3 - (1{,}5) \text{ x } 4 = -3$$
$$lim.superior = 7 + (1{,}5) \text{ x } 4 = 13$$

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.

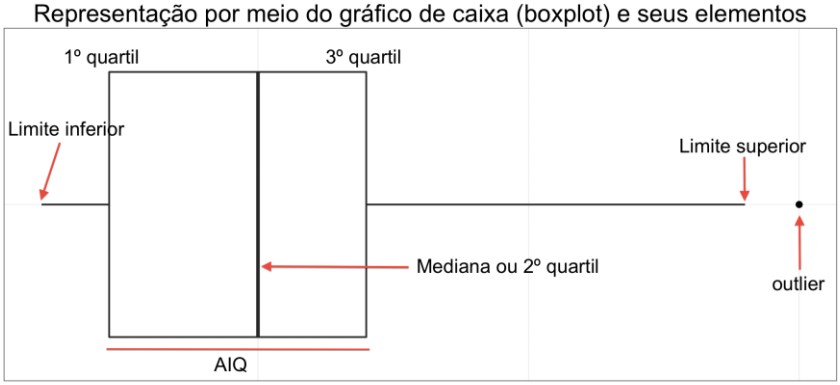
Por exemplo: imaginemos que um nutricionista mediu a massa (em kg) de uma população qualquer de adultos e encontrou q_1 = 50 kg, e q_3 = 100 kg. Nesse caso, a AIQ é 50 kg (100 kg – 50 kg) e o limite superior é 100 + 1,5 (50) = 175 kg. Assim, qualquer valor acima desse seria considerado um *outlier*. 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.

2.1.2. Boxplot e a representação dos quartis

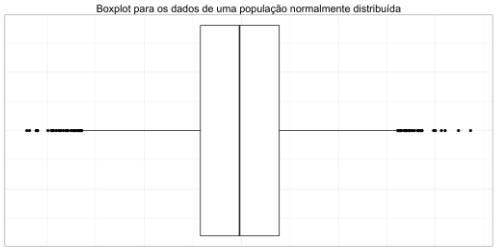
Isso tendo sido colocado, podemos passar à próxima etapa, que é a apresentação e o entendimento da importância das medidas até agora apresentadas. A sumarização dos dados como proposta acima é normalmente feita com um tipo de gráfico próprio, chamado de *gráfico de caixas*, ou *gráfico de caixa e bigodes*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

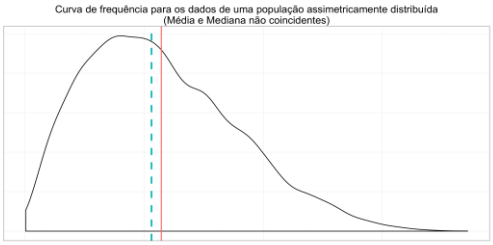
Representação por meio do gráfico de caixa (boxplot) e seus elementos

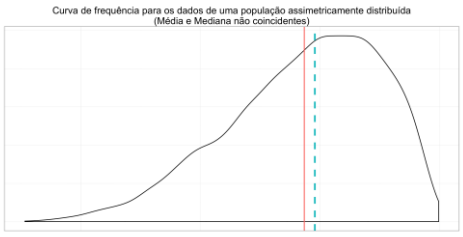


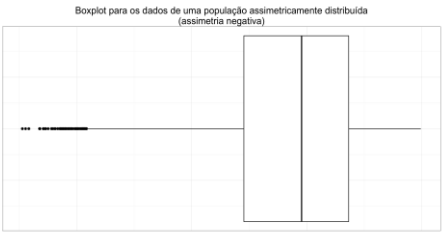
O *boxplot*, porém, não é apenas uma apresentação visual das medidas que até agora vislumbramos, mas também uma representação gráfica da curva de frequência dos dados coletados, indicando se os dados se distribuem simetricamente em torno da média e da mediana ou se os dados estão distribuídos assimetricamente (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 vermelha) e a mediana (linha azul). Nas distribuições simétricas, média e mediana coincidem. No entanto, como a média é uma medida pouco robusta, nas distribuições assimétricas, ela é "puxada" em direção à cauda mais longa, ou seja, os valores extremos na amostra ou população influenciam no valor da média. Foi por esse motivo que, quando calculamos média e mediana, algumas páginas antes, dissemos que, como elas eram próximas, já tínhamos uma noção de que nossos dados eram simétricos.





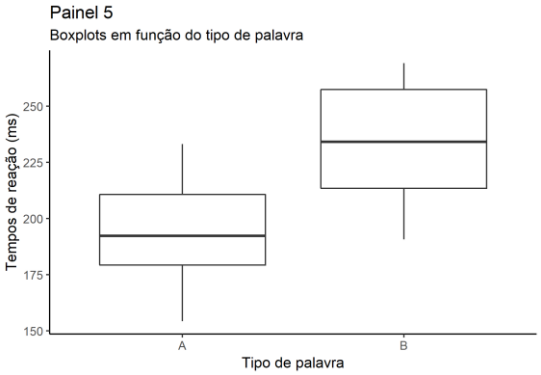




Com esses conceitos em mãos, podemos fazer uma análise mais precisa das nossas amostras A e B, que, até agora, tinham sido descritas apenas pela média e pela mediana. Assim:

	Mínimo	1º quartil (q_1) (25%)	2º quartil (q_2) (50%)	3º quartil (q_3) (75%)	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

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:

$$AIQ\ (A) = q_3 - q_1 = 210{,}7 - 179{,}3 = 31{,}4 \text{ milisegundos}$$

$$AIQ\ (B) = q_3 - q_1 = 257{,}5 - 213{,}4 = 44{,}1 \text{ milisegundos}$$

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

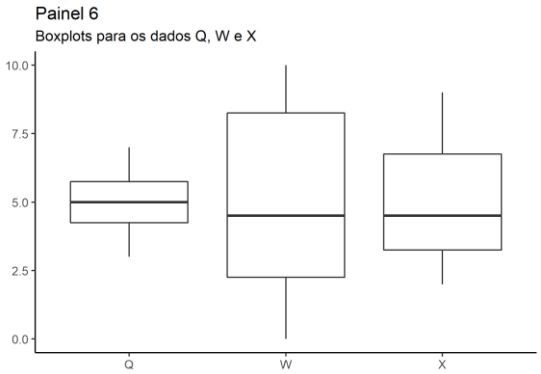
2.2. Desvios em relação à média

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}

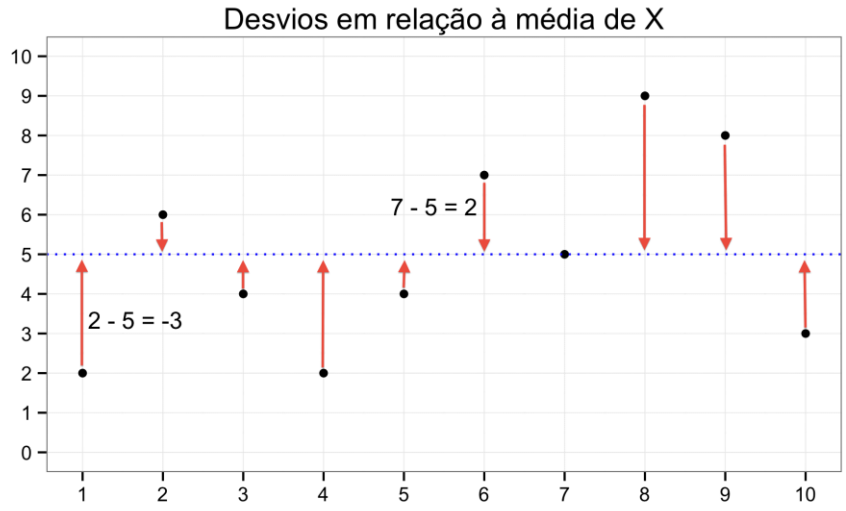
Apesar de a média ser idêntica, os dados não têm a mesma distribuição. Isso porque w cobre quase toda a gama de inteiros de 0 a 10, exceto 7, enquanto q fica restrito ao intervalo entre 3 e 7. x, por sua vez, fica numa espécie de meio termo entre ambos, cobrindo uma gama maior de valores do que q, mas 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, no gráfico abaixo, o conjunto x *plotado* aleatoriamente em torno da média de x. As setas vermelhas indicam os desvios de cada valor de xdessa média, ou seja, a distância que estão de $\bar{x}$. Com isso, podemos ter um vislumbre da dispersão de x em torno da média.



(Não se assuste com as várias colunas dessa tabela. Vamos continuar voltando a ela e explicar tudo o que está aí. Por enquanto, porém, ficaremos nas três primeiras colunas.)

	x	$\bar{x}$	$x_i - \bar{x}$	$\lvert x_i - \bar{x}\rvert$	$(x_i - \bar{x})^2$	w	$\bar{w}$	$w_i - \bar{w}$	$\lvert w_i - \bar{w}\rvert$	$(w_i - \bar{w})^2$	w	$\bar{q}$	$q_i - \bar{q}$	$\lvert q_i - \bar{q}\rvert$	$(q_i - \bar{q})^2$
	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

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.

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

Você pode, se quiser e não confiar na tabela acima, 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.

2.3. 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 dada. 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.

2.4. Variância

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

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

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

Da mesma forma que fizemos para os desvios originais, podemos, também para o quadrado dos desvios, calcular uma espécie de média dessa dispersão. Basta, portanto, dividir essa soma por *n*. A essa espécie de "média dos quadrados dos desvios" damos o

nome de *variância*. Na verdade, para pequenas amostras, o ideal é que a variância seja calculada dividindo-se aquela soma por *n – 1*. Para grandes amostras, não há diferença entre os valores dos dois métodos. Se fizéssemos isso para os dados acima, teríamos que as variâncias seriam:

$$var(x) = \frac{54}{10-1} = \frac{54}{9} = 6{,}0$$

$$var(w) = \frac{122}{10-1} = \frac{122}{9} = 13{,}55$$

$$var(q) = \frac{18}{10-1} = \frac{18}{9} = 2{,}0$$

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*.

2.5. 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:

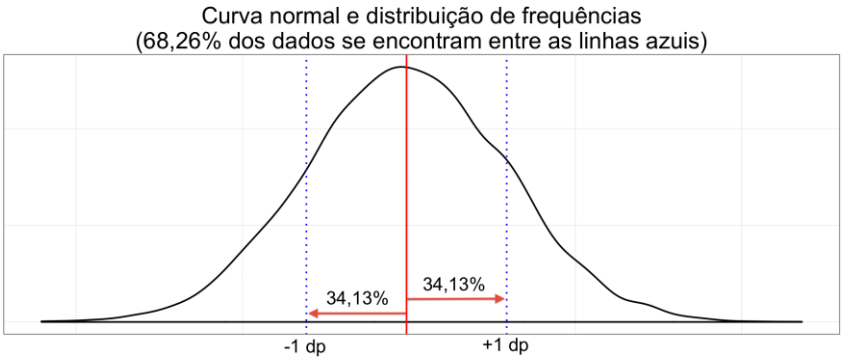
	x	w	q
Variância	6	13,55	2
Desvio padrão	2,44	3,68	1,41

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*.

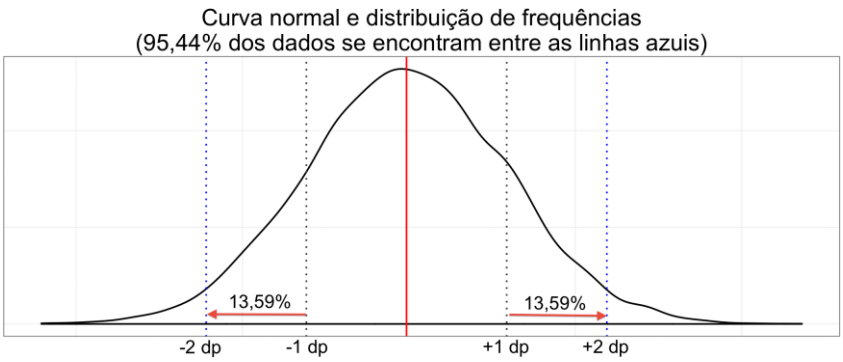
Um comentário importante deve ser feito aqui. Você não deve confundir *desvio padrão* com *erro padrão*. O desvio padrão, como vimos, é uma estatística descritiva. Ele busca mensurar o quanto os dados se dispersam em relação à média. O erro padrão, por outro lado, é uma estatística inferencial – sobre 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.

2.5.1. Desvio padrão e curva normal

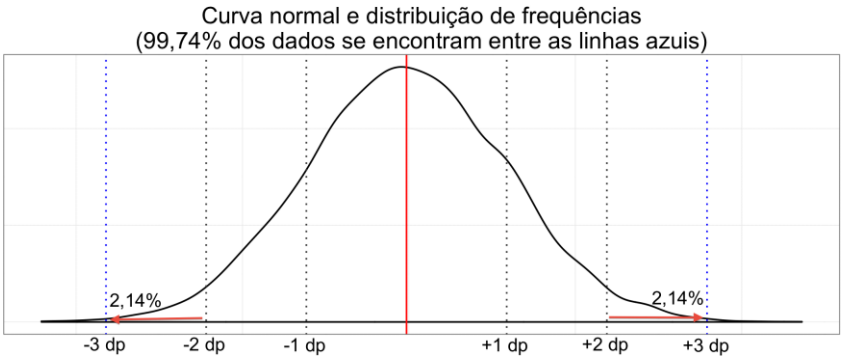
Mais importante ainda é a relação que se pode estabelecer entre o desvio padrão e os dados normalmente distribuídos. Isso porque os dados podem ser divididos em unidades de desvio padrão em relação à média, para mais ou para menos, sendo provado que a área sob a curva normal é constante dentro dos limites 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, como ilustra a figura. 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), como mostra a figura:



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), como mostra a figura:



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

2.6. 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 as seguintes estatísticas para aqueles dados, lembrando que q_2 = mediana:

	Mínimo	1º quartil (q_1) (25%)	2º quartil (q_2) (50%)	3º quartil (q_3) (75%)	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

Calculemos, então, as demais medidas de dispersão, que estão resumidas abaixo. 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.

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	233.92	19.92	19.92	396.86
183.46	194.23	-10.77	10.77	116.01	210.67	233.92	-23.25	23.25	540.5
193.83	194.23	-0.4	0.4	0.16	215.57	233.92	-18.35	18.35	336.67
209.54	194.23	15.31	15.31	234.38	237.16	233.92	3.24	3.24	10.51
211.8	194.23	17.57	17.57	308.68	214.41	233.92	-19.51	19.51	380.59
192.31	194.23	-1.92	1.92	3.69	261.46	233.92	27.54	27.54	758.53
233.17	194.23	38.94	38.94	1516.27	190.66	233.92	-43.26	43.26	1871.31
202.85	194.23	8.62	8.62	74.29	224.17	233.92	-9.75	9.75	95.04
165.61	194.23	-28.62	28.62	819.14	234.11	233.92	0.19	0.19	0.04
232	194.23	37.77	37.77	1426.52	258.71	233.92	24.79	24.79	614.61
220.38	194.23	26.15	26.15	683.79	256.35	233.92	22.43	22.43	503.16
183.03	194.23	-11.2	11.2	125.45	202.75	233.92	-31.17	31.17	971.49
154.3	194.23	-39.93	39.93	1594.46	269.16	233.92	35.24	35.24	1241.95
168.33	194.23	-25.9	25.9	670.84	212.38	233.92	-21.54	21.54	463.91
187.29	194.23	-6.94	6.94	48.17	267.38	233.92	33.46	33.46	1119.66
Soma dos desvios		0					0		
Soma dos módulos			288,71					333,64	
Desvio médio			19,25					22,24	
Soma dos quadrados				7970,44					9304,83
Variância				569,32					664,63
Desvio padrão				23,86					25,78

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

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

ferramenta para fazer inferências sobre aquilo que desconhecemos. Esse tópico (a *inferência estatística*) será abordado na próxima seção.

3. 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.

Um outro modo de dizer isso é falando que nós calculamos as *estatísticas* das amostras a fim de estimar os *parâmetros* populacionais. A grande questão aqui, porém, é que nós não podemos realizar um monte de experimentos e comparar os valores obtidos com cada um deles a fim de verificar se são consistentes ou não. Experimentos

são caros, trabalhosos, demandam grande preparação, equipamentos, tempo, participantes, horas de laboratório, etc. Para fazer essa estimativa, contamos apenas com o nosso experimento, o único que conseguimos realizar. A partir dele, temos que *adivinhar* se estamos próximos dos parâmetros populacionais ou não. Mas a nossa adivinhação é rigorosa: ela vai se valer das estatísticas que até agora calculamos, ou seja, vai se valer do fato de que sabemos os pontos em torno dos quais os dados se organizam e, talvez mais importante, como eles se dispersam em torno desses pontos.

3.1. Erro padrão

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) conhecemos. Para cada uma dessas amostras, calculamos uma média. Parece óbvio que cada uma dessas médias não é exatamente igual à média populacional (μ). Algumas estarão mais próximas, outras mais distantes.

Agora vamos imaginar que calculemos uma média de todas essas médias ($\bar{x}$). 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).

Do mesmo modo como se pode demonstrar o que foi dito acima empiricamente, 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:

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

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*. Esse é dado pela fórmula, em que *n* é o tamanho da nossa amostra:

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

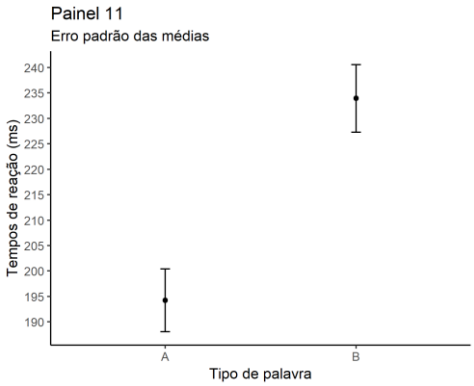
Como *n* (o tamanho da amostra) está no denominador, então, quanto maior for nossa amostra, menor será o erro padrão e, portanto, mais confiança podemos ter na nossa média amostral ($\bar{x}$) como estimativa da média populacional (μ).

3.1.1 Voltando ao nosso exemplo

Para os dados do nosso experimento, os desvios padrões e erros padrões estão resumidos na tabela abaixo:

	Média	Desvio padrão	Erro padrão
A	194,23	23,86	4,35
B	233,91	25,78	4,70

No painel 11, abaixo, o ponto central representa as médias de A e de B e as barras o erro padrão de cada uma delas (a média $\pm$ o erro padrão).



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 confiantes de estarmos "acertando" as médias, então, parece de fato que nossas médias são diferentes 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 confiante podemos estar quanto a essa distância. Para saber isso, vamos introduzir o conceito de intervalo de confiança, diretamente relacionado ao erro padrão.

3.2. 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 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 algum montante seriam raros. Esse montante é dado pela fórmula:

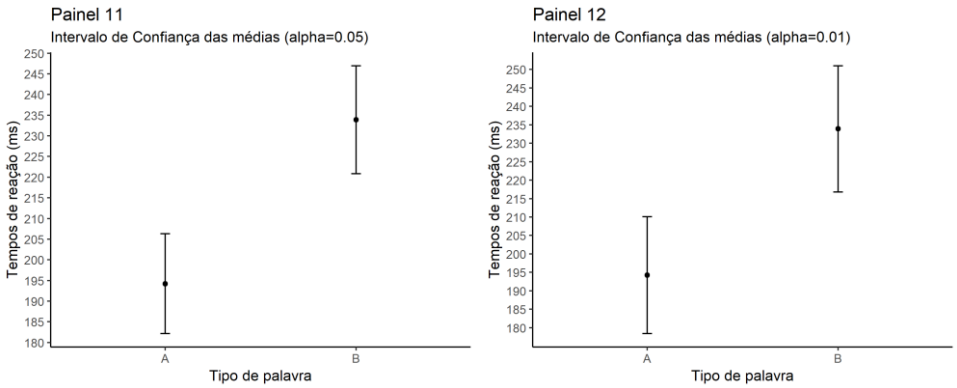
$$\frac{1{,}96\ x\ desvio\ padrão}{\sqrt{n}} = \frac{1{,}96\ x\ 23{,}86}{\sqrt{15}} = \frac{46{,}76}{3{,}87} = 12{,}08 \text{ ms.}$$

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.) em que possivelmente nossa média se encontra.

Em outras palavras, valores que fossem menores do que a média menos 12,08ms: 194,23 – 12,07 = 182,15ms; e maiores do que a média mais 12,08ms: 194,23 + 12,08 = 206,30ms são valores que têm apenas 5% de chance de ocorrer, dada a variabilidade dos dados amostrados. Com isso, poderíamos dizer que o intervalo de confiança para a média, com um *coeficiente de confiança* de 0,95 é: IC(0,95)=182,15 – 206,30.

Isso significa que nós temos 95% de confiança de que a média populacional está nesse intervalo? Em geral, é desse modo que os dados são reportados. No entanto, há um grande debate sobre como, de fato, os intervalos de confiança devem ser interpretados. Esse debate passa ao largo do escopo deste material, sobretudo porque envolve duas visões filosóficas distintas da estatística. Se você quiser uma abordagem didática sobre o tema, recomendamos a leitura de Howell (2010: 192).

Observe nos painéis abaixo 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).



3.3. Análise de Variância

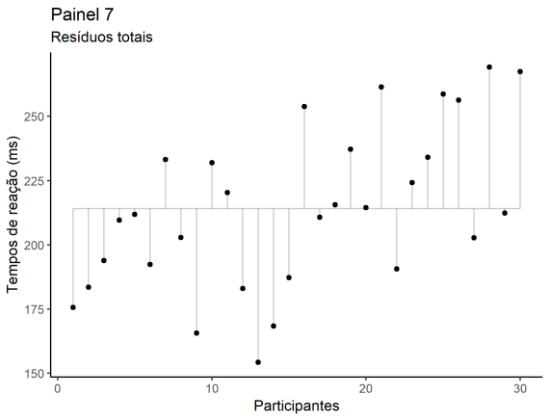
Dado o que temos até agora, vamos tentar estimar se a diferença entre as médias obtidas para palavras do tipo A e palavras do tipo B é de fato significativa, ou seja, se a diferença de fato representa uma diferença na população da qual foram extraídas. O erro padrão e os intervalos de confiança já nos deram esse indicativo, mas 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.

Calcular uma ANOVA 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 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 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 ms e muita vezes é chamada de grande média (*great mean*). Agora que temos uma média global, podemos calcular os desvios em relação a ela, que vamos chamar de *resíduos* ou *erro* (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.

Participantes	A	B	Médias dos participantes
1	175,56	253,84	214,70
2	183,46	210,67	197,07
3	193,83	215,57	204,70
4	209,54	237,16	223,35
5	211,80	214,41	213,11
6	192,31	261,46	226,89
7	233,17	190,66	211,92
8	202,85	224,17	213,51
9	165,61	234,11	199,86
10	232,00	258,71	245,36
11	220,38	256,35	238,37
12	183,03	202,75	192,89
13	154,30	269,16	211,73
14	168,33	212,38	190,36
15	187,29	267,38	227,34
Médias	194,23	233,92	**214,17**

Antes de prosseguir, pense sobre o que fizemos: nós simplesmente calculamos a **variabilidade total** dos dados, desconsiderando qualquer tratamento aplicado a eles. Por isso, vamos chamar esse valor de *soma de quadrados total*. No gráfico abaixo estão os desvios em relação a essa média global, ou seja, foi a partir desses desvios que calculamos a soma de quadrados total.

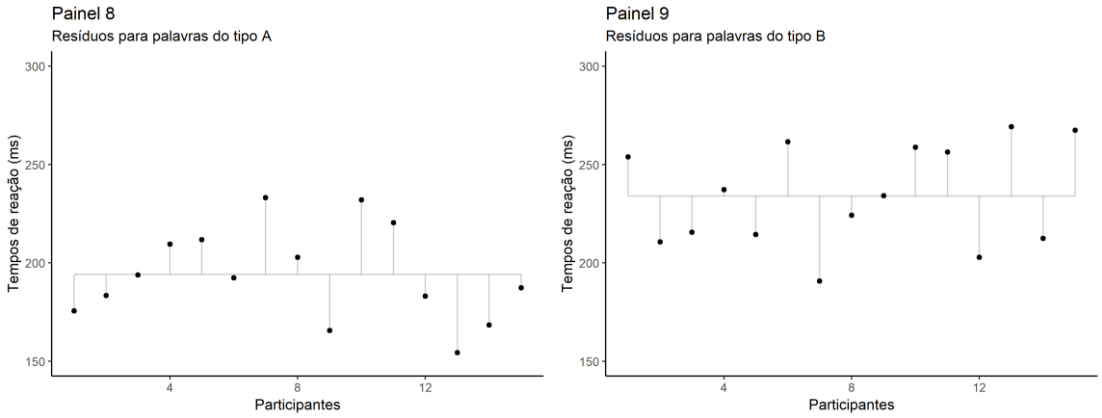


$$SQ_{Total} = \sum (x_i - \bar{x})^2$$

$$= (175{,}56 - 214{,}17)^2 + (183{,}46 - 214{,}17)^2 + \ldots$$

$$+ (227{,}34 - 214{,}14)^2 = 29088{,}81$$

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 aquela tabela e confira aqueles números. A soma dos quadrados de cada um desses grupos de resíduos é: $SQRes_A$: 7970,44; $SQRes_B$: 9304,83. A *soma de quadrados dos resíduos*, então, é a soma desses dois valores: 17275,28.



$$SQ_{Resíduos} = \sum (x_{ij} - \bar{x}_j)^2$$

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 8 e 9 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.

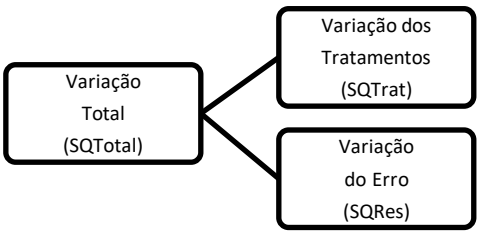
Repare que, se no primeiro caso consideramos a variabilidade total, sem considerar os tratamentos; e agora consideramos a variabilidade perdida, mesmo com o tratamento, o que sobrou é a variabilidade explicada pelo tratamento, ou seja, o quanto de redução de variação tivemos *depois* que aplicamos o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

$$SQ_{Tratamento} = n\sum(\bar{x}_j - \bar{x})^2 = 15[(194{,}23 - 214{,}17)^2 + (233{,}92 - 214{,}17)^2] = 11813{,}53$$

Vamos resumir isso na tabela abaixo (que por enquanto está incompleta, mas já já iremos preenchê-la):

	G.L.	S. Q.	Q.M.	F
Tratamento		11813,53		
Resíduos		17275,28		
Total		29088,81		

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 de 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 Tratamento, 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. Logo:

	G.L.	S. Q.	Q.M.	F
Tratamento	1	11813,53	11813,53	19,14
Resíduos	28	17275,28	616,97	
Total	29	29088,81	1003,06	

Dado que chegamos até aqui, vamos pensar um pouco sobre essa tabela 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 (11.813,53). 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 encontramos um valor de F igual a 19,14. Na tabela que tenho aqui em mãos (Bussab & Morettin, 2012: 514), existe

apenas 5% de chance de encontrarmos um valor igual ou maior do que 4,20. Como achamos muito mais do que isso (19,14), temos confiança, a 95%, de que nossas médias são de fato diferentes e que não ocorreram por acaso. Esse é o chamado p-valor da nossa ANOVA, que é menor do que 0,05, representando uma *diferença significativa*. Na verdade, se pedirmos para um computador calcular esse valor, ele será igual a 0,000152. Não queremos entrar na discussão sobre o que de fato esse número significa[1].

Completando a nossa tabela da ANOVA:

	G.L.	S. Q.	Q.M.	F	p-valor
Tratamento	1	11813,53	11813,53	19,14	0,000152
Resíduos	28	17275,28	616,97		
Total	29	29088,81	1003,06		

Um outro modo de pensar no valor de F é partir do seguinte, como propõe Howell (2010). Sabemos que, sob a hipótese nula (H_0), as médias são iguais; e que, consequentemente, sob a hipótese alternativa (H_1), as médias são diferentes. Podemos também 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. Logo, 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.

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 total* 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* (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 29088,81 para 11813,53. Ou seja, tivemos uma redução de aproximadamente 40% na soma de quadrados. Essa, portanto, é a proporção de variação explicada pelo modelo. Em outras palavras, o modelo reduzido parece ser bem melhor

[1] Como bem informa Winter (2020: 157+), o teste de significância é método de trabalho vinculado a uma corrente da estatística chamada de *frequentismo* e não é unanimidade entre os especialistas, tendo já recebido muitas críticas. Recomendamos a leitura do autor para uma introdução ao tema.

porque ele *se ajusta melhor aos dados*: se o consideramos, podemos explicar melhor a variação encontrada.

$$\frac{11813{,}53}{29088{,}81} = 0{,}4061 = 40{,}61\%$$

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.

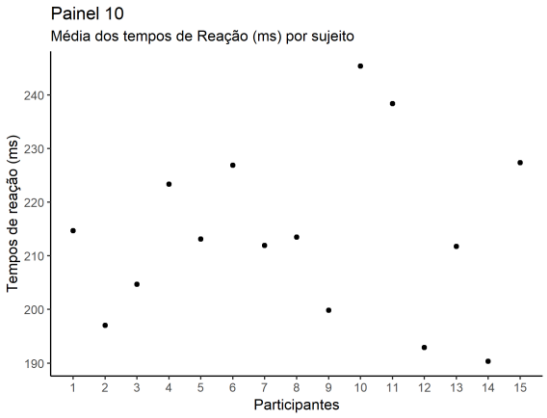
3.3.1. Fatores fixos e fatores aleatórios

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

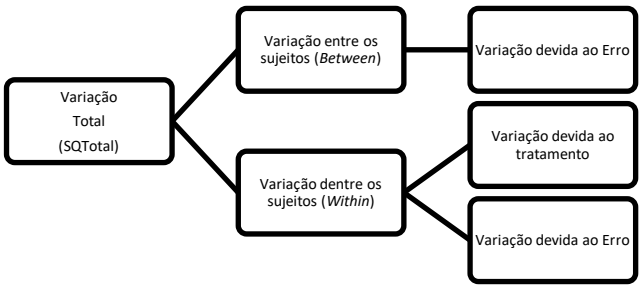
alguma característica individual desse grupo, leram mais rapidamente palavras do tipo A do que palavras do tipo B.

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), que chamamos de tratamento. Todavia, ignoramos completamente a variabilidade advinda do grupo específico de sujeitos que fizeram o experimento.



Se você voltar á tabela 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 seguinte forma:



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. 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. (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 aqui. Mais à frente vamos dar outro exemplo e, com o tempo, as coisas vão fazendo mais sentido. Agora, se você quiser se aprofundar no tema, recomendamos o *Capítulo 14 – Repeated-Measures Designs*, de Howell, 2010).

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 média global, 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.

$$SQ_{Sujeitos} = k\sum(\bar{x}_{suj} - \bar{x})^2$$
$$= 2[(214{,}70 - 214{,}17)^2 + (197{,}07 - 214{,}17)^2$$
$$+ \cdots (227{,}34 - 214{,}17)^2] = 7193{,}89$$

(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:

29088,81-11813,53-7193, 894=10081,39

Os dados que calculamos estão na tabela de ANOVA abaixo, 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: (i) *between sujeitos*: número de sujeitos menos 1, ou seja, 15-1 = 14; (ii) dos *tratamentos*: número de tratamentos (*k*) menos 1, ou seja, 2-1=1; (iii) *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.	S. Q.	Q.M.	F	p-valor
Between sujeitos					
Resíduos*	14	7193,89	513,84		
Within sujeitos					
Tratamento	1	11813,53	11813,53	16,40	0,00119
Resíduos*	14	10081,39	720,09		
Total	29	29088,81			

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?

Tabela para o modelo 1: sem considerar sujeitos					
	G.L.	S. Q.	Q.M.	F	p-valor
Tratamento	1	11813,53	11813,53	19,14	0,000152
Resíduos	28	17275,28	616,97		
Total	29	29088,81			

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.

Aqui cabe, ainda, um breve comentário: apesar de esse *design* ser um com medidas repetidas, ele é muito simples, porque temos apenas uma observação por condição por sujeito: cada sujeito vê apenas uma palavra do tipo A na condição A e uma palavra na condição 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. Não vamos abordá-lo agora, portanto. Vamos deixá-lo um pouco mais para frente e passar a um outro problema fundamental: o caso dos itens experimentais.

3.3.2. Itens como efeito aleatório

Para o experimento que propusemos inicialmente, um design *within subjects*, 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 adequada. 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 um *design within subjects* (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* e aninhado em *sujeitos*, 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:

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.

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.

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 (1974) e que tem suas raízes em Coleman (1964).

Recomendamos, também, a leitura de Raijmakers (2003) para um apanhado geral sobre o tema.

3.3.3. Um exemplo fictício

Vamos começar com um exemplo simples a fim de verificarmos como a variabilidade dos itens pode influenciar dramaticamente na nossa análise. Imagine que recrutamos 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 afeta o comportamento do sujeito. Os dados amostrados estão na tabela abaixo.

Sujeito	Número	Palavra	Valor	Número	Palavra	Valor	
1	Singular	p1	1	Plural	p1	15	
2	Singular	p2	10	Plural	p2	6	
3	Singular	p3	5	Plural	p3	12	
4	Singular	p1	8	Plural	p1	13	
5	Singular	p2	12	Plural	p2	7	
6	Singular	p3	4	Plural	p3	16	
Médias			6.66			11.5	**9.08**

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.

$$SQ_{Total} = \sum (x - \bar{x})^2$$
$$= (1 - 9.08)^2 + (10 - 9.08)^2 + \ldots + (7 - 9.08)^2 + (16 - 9.08)^2$$
$$= 238.91$$

$$SQ_{Tratamento} = n\sum(\bar{x}_j - \bar{x})^2 = 6[(6.66 - 9.08)^2 + (11.50 - 9.08)^2] = 70.27$$

$$SQ_{Resíduos} = SQ_{Total} - SQ_{Tratamento} = 238.91 - 70.27 = 168.64$$

Resumindo esses dados na tabela da ANOVA:

	G.L.	S. Q.	Q.M.	F	p-valor
Tratamento	1	70.27	70.27	4.16	0.068
Resíduos	10	168.64	16.86		
Total	11	238.91			

Observe que obtivemos um p-valor 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. Talvez nós precisemos apenas aumentar nossa amostra ou algo assim a fim de obter um valor significativo. Se replicássemos, portanto, esse experimento, é bem provável que o efeito de número seja 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 condição e calcular outra ANOVA. Para isso, vamos calcular a média de cada um dos itens:

p1	p2	p3
9.25	8.75	9.25

Agora, vamos subtrair cada um desses valores da grande média e multiplicá-lo por 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.

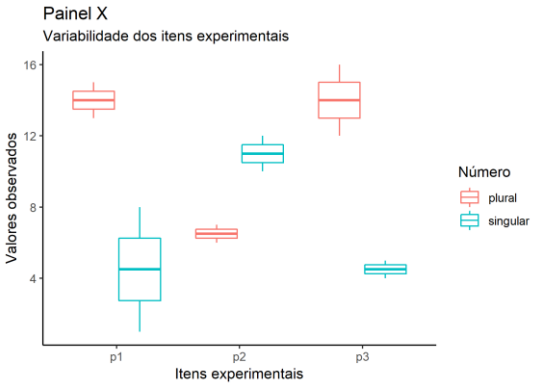
$$SQ_{Between\ Itens} = w\sum(\bar{x}_{item} - \bar{x})^2$$
$$= 4[(9.25 - 9.08)^2 + (8.75 - 9.08)^2 + (9.25 - 9.08)^2] = 0.666$$

Agora que temos esse valor, vamos construir nossa tabela de ANOVA, que está abaixo. Observe que ela é idêntica à tabela 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 três das somas quadráticas aí presentes nós já calculamos quando fizemos a ANOVA simples: a *between itens*, a dos *tratamentos* e a *total*.

	G.L.	S. Q.	Q.M.	F	p-valor
Between itens					
Resíduos*	2	0.666	0.333		
Within Itens					
Tratamento	1	70.27	70.12	3.32	0.105
Resíduos*	8	168.64	21.08		
Total	11	238.91			

Antes de entrarmos no p-valor, cabe ainda um comentário sobre os graus de liberdade, que foram calculados da seguinte forma: (i) *between itens*: número de itens menos 1, ou seja, 3-1 = 2; (ii) dos *tratamentos*: número de tratamentos (k) menos 1, ou seja, 2-1=1; (iii) *total*: número total de observações (*n*) menos 1, ou seja, 12-1=11. 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 21.08, enquanto os graus de liberdade caíram de 10 para 8. Mas o que isso significa em termos teóricos? Por que isso ocorreu? Observe o painel abaixo, em que apresentamos os *boxplots* para cada um dos itens em cada condição:



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 enorme variabilidade entre os itens. Enquanto p1ural aumenta a variável resposta para p1 e p3, ele diminui para p2.

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 p2. 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 p1 e/ou p3, então a diferença entre as médias tende a se manter. Em outras palavras, se você não controla a variância dos itens, como poderá saber que o efeito que obteve é devido ao tratamento e não ao conjunto particular de itens que foi amostrado?

No caso específico desse experimento, não tem como saber. É impossível separar o efeito do tratamento do efeito dos itens na resposta obtida. Por isso, apesar de a primeira ANOVA ter apresentado 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 pra gente: “dada essa variação toda, querido, não tenho firmeza para rejeitar a hipótese nula”.

Nesse ponto, é preciso deixar algo claro. Não há nenhum problema em usar ANOVA, ela não é um modelo obsoleto ou ruim ou problemático, mas é preciso usar a ANOVA adequada ao seu *design* experimental e, para o *design* em questão, o modelo correto é aquele que controla a variabilidade dos itens.

4. Modelos lineares mistos

Nosso material acabou por aqui. Não vamos fazer uma discussão sobre modelos mistos nele. Como dissemos na justificativa prévia, esse material tem por objetivo apenas fazer uma breve introdução a alguns conceitos fundamentais. Sobre modelos mistos, discutiremos ao longo do curso.

Apesar disso, vamos deixar abaixo duas coisas, primeiro, uma discussão sobre um design experimental específico, a fim de situar o leitor no debate sobre estruturas de fatores aleatórios e como elas estão vinculadas ao design experimental. Em segundo lugar, quatro modelos de design experimental básicos e a estrutura de fatores aleatórios adequada a cada um deles.

Se você está lendo isso antes do curso, é provável que não irá entender muito, mas pode ser que te ajude se você ler depois.

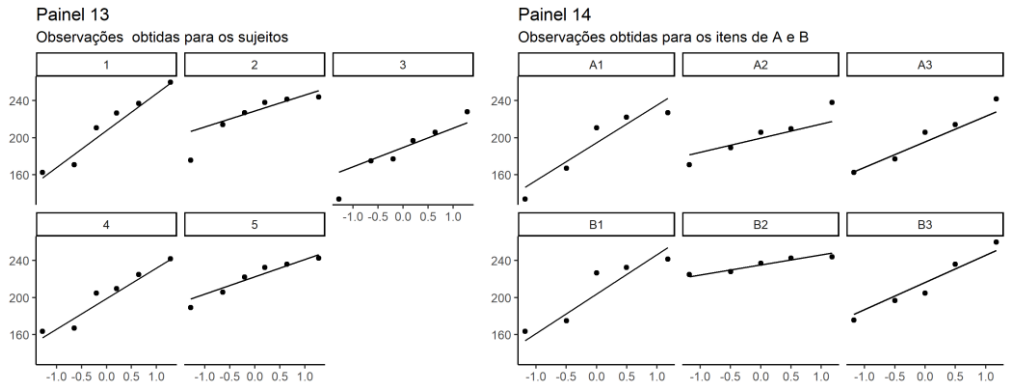
4.1. Investigando outros designs experimentais

Imagine que realizamos o seguinte experimento: coletamos aleatoriamente 3 palavras do tipo A (A1, A2 e A3) e 3 palavras do tipo B (B1, B2 e B3). Selecionamos aleatoriamente 5 sujeitos e mensuramos seus tempos de reação a cada uma dessas palavras. Temos, mais uma vez, 30 observações (15 para o tipo A e 15 para o tipo B). Os dados obtidos com esse suposto experimento estão abaixo:

Participantes	Itens A	Tempos de A	Itens B	Tempos de B
1	A1	210.64	B1	226.67
1	A2	170.66	B2	237.16
1	A3	162.56	B3	260.03
2	A1	226.83	B1	241.41
2	A2	238.03	B2	243.91
2	A3	214.24	B3	175.56
3	A1	133.68	B1	175
3	A2	205.92	B2	227.87
3	A3	177.12	B3	196.67
4	A1	167.05	B1	163.64
4	A2	209.75	B2	224.99
4	A3	241.78	B3	204.92
5	A1	222	B1	232.54
5	A2	189.14	B2	242.6
5	A3	205.98	B3	236.11

Observe que esse é um design *within subjects*, ou seja, os mesmos 5 sujeitos veem ambas as condições experimentais (A e B); e *between itens*, ou seja, itens diferentes são usados nas condições A (A1, A2 e A3) e B (B1, B2 e B3). Isso é o mesmo que dizer que *sujeito* e *tipo de palavra* são fatores cruzados (*crossed factors*) e *itens* é um fator aninhado (*nested factor*) em *tipo de palavra*. No entanto, tomamos medidas repetidas para os sujeitos e também para os itens, visto que cada item (A1, B1, etc.) era visto por sujeitos distintos, mais de uma vez. Na verdade, para cada item foram mensurados 5 tempos diferentes (um para cada sujeito).

Observe os painéis abaixo, que resumem esses dados: há seis observações para cada um dos 5 sujeitos (painel 13); e 5 observações para cada um dos 6 itens (painel 14).



Observe que o caso dos itens, nesse experimento, é similar ao caso dos sujeitos no experimento anterior. Os três itens escolhidos para cada condição não são todos os itens possíveis para aquela condição. Eles são três itens aleatórios. Eles são uma representação da população de palavras do tipo A e do tipo B. Do mesmo modo como esperamos que sujeitos diferentes tivessem tempos de reação diferentes, é plausível pensar que itens diferentes podem provocar tempos de reação diferentes. Se, da próxima vez que fizermos esse experimento, escolhermos um conjunto diferente de itens (hipoteticamente, A4, A5 e A6; B4, B5 e B6), então, não temos garantias de que teremos um resultado semelhante, porque esse conjunto particular pode provocar tempos de reação mais rápidos ou mais lentos. Essa informação, portanto, nos diz que precisamos incorporar, no nosso modelo misto, *interceptos* para sujeitos e *interceptos* para itens como efeitos aleatórios.

Paremos, então, para pensar sobre a inclinação (ou *slope*). Como mostra o painel 13, cada sujeito foi submetido aos itens da condição A e aos da condição B. Logo, a condição pode afetar o resultado dos sujeitos (um sujeito pode ser mais rápido em A do que em B; outro pode ser mais rápido em B do que em A; outro pode ser igualmente rápido em ambas; etc.). Desse modo, precisamos considerar essa variabilidade e incluir *slopes* aleatórios para sujeitos. No entanto, como mostra o painel 15, os itens de A e de B são distintos. Logo, não pode ser que um item seja afetado pela interação com o tratamento, ele não pode ser lido mais rápido ou devagar dado o tratamento A ou B. Logo, para esse design, não é preciso *slopes* para itens.

4.2. Um pouco sobre designs e fatores aleatórios

Seguindo a proposta de Barr et al. (2013), ou seja, ajustar o modelo misto segundo um critério orientado pelo design experimental (*design driven*) e não orientado pelos dados (*data driven*), apresentamos abaixo quatro designs experimentais e a sua estrutura de fatores aleatórios ideal. Como discutido no artigo dos autores, isso não significa que essa estrutura é a que deve ser mantida. Muitas vezes o modelo pode estar sobreajustado, alcançando a singularidade (variância igual a zero), de modo que o slope ou intercepto deve ser retirado do modelo.

Design 1

Sujeitos e itens aleatórios;

Tratamento fixo;

Sujeitos aninhados em tratamentos (*between subjects*);

Itens aninhados em tratamentos (*between itens*).

Esse design precisa apenas de interceptos aleatórios para sujeitos e itens.

Design 2

Sujeitos e itens aleatórios;

Tratamento fixo;

Sujeitos cruzados com tratamentos (*within subjects*);

Itens aninhados em tratamentos (*between itens*).

Esse design precisa de interceptos para sujeitos e itens e de slopes para sujeitos em função dos tratamentos.

Design 3

Sujeitos e itens aleatórios;

Tratamento fixo;

Sujeitos aninhados em tratamentos (*between subjects*);

Itens cruzados com tratamentos (*within itens*).

Esse design precisa de interceptos para sujeitos e itens e slopes para itens em função dos tratamentos.

Design 4

Sujeitos e itens aleatórios;

Tratamento fixo;

Sujeitos cruzados com tratamentos (*within subjects*);

Itens cruzados com tratamentos (*within itens*).

Esse design precisa de interceptos para sujeitos e itens e slopes para sujeitos e itens dados os tratamentos.