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

3.3. Análise de Variância

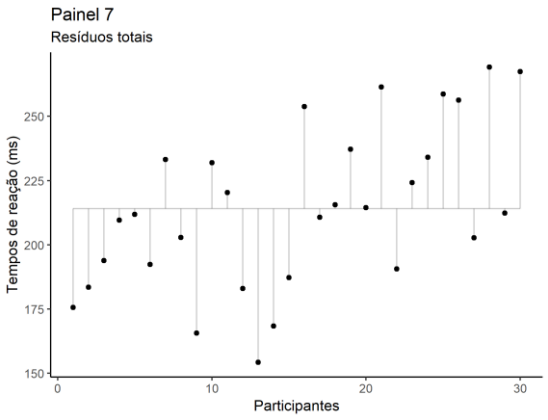
Dado o que temos até agora, vamos tentar estimar se a diferença entre as médias obtidas para palavras do tipo A e palavras do tipo B é de fato significativa, ou seja, se a diferença de fato representa uma diferença na população da qual foram extraídas. O erro padrão e os intervalos de confiança já nos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

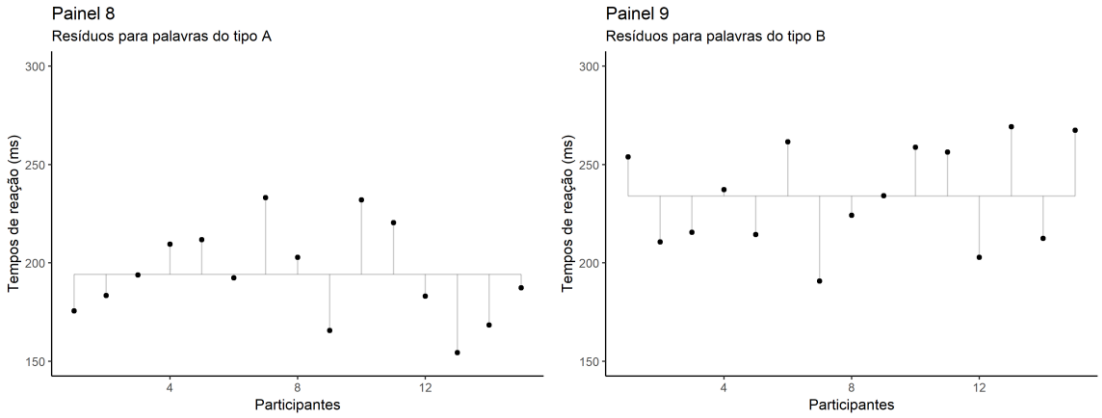
Participantes	A	B	Médias dos participantes
1	175,56	253,84	214,70
2	183,46	210,67	197,07
3	193,83	215,57	204,70
4	209,54	237,16	223,35
5	211,80	214,41	213,11
6	192,31	261,46	226,89
7	233,17	190,66	211,92
8	202,85	224,17	213,51
9	165,61	234,11	199,86
10	232,00	258,71	245,36
11	220,38	256,35	238,37
12	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.

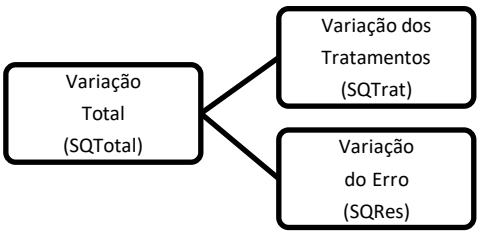
Repare que, se no primeiro caso consideramos a variabilidade total, sem considerar os tratamentos; e agora consideramos a variabilidade perdida, mesmo com o tratamento, o que sobrou é a variabilidade explicada pelo tratamento, ou seja, o quanto de redução de variação tivemos *depois* que aplicamos o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

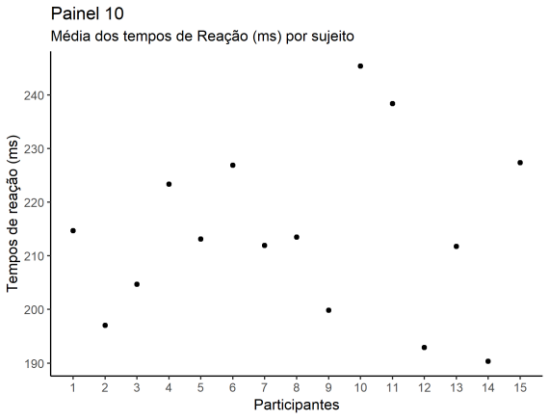
3.3.1. Fatores fixos e fatores aleatórios

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

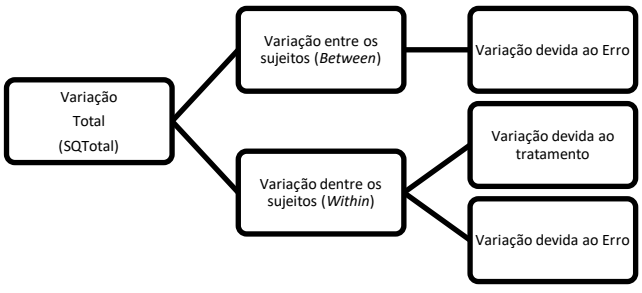
alguma característica individual desse grupo, leram mais rapidamente palavras do tipo A do que palavras do tipo B.

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.

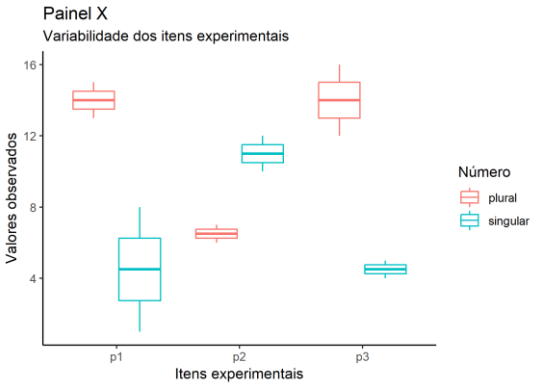
$$SQ_{Between\ Itens} = w\sum(\bar{x}_{item} - \bar{x})^2$$
$$= 4[(9.25 - 9.08)^2 + (8.75 - 9.08)^2 + (9.25 - 9.08)^2] = 0.666$$

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.

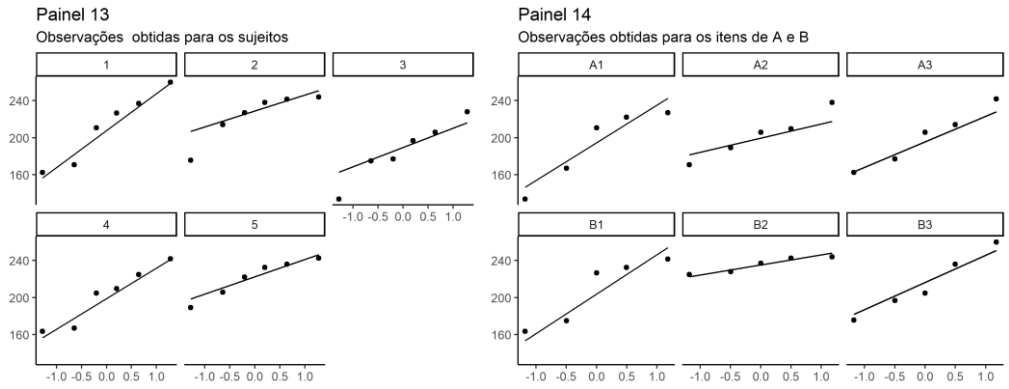
4.1. Investigando outros designs experimentais

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.