

# Introdução geral aos modelos mistos

Vanderly Janeiro

Dep. de Estatística - UEM

## Conteúdo

<b>1</b>	<b>Introdução</b>	<b>2</b>
1.1	Exemplo introdutório: NIR predição de medidas em HPLC . . . . .	2
1.2	Análise simples . . . . .	2
1.3	Análise simples por uma abordagem ANOVA . . . . .	4
1.4	O problema da abordagem ANOVA . . . . .	6
1.5	O modelo misto . . . . .	7
1.6	Comparação de modelo fixo e misto . . . . .	8
1.6.1	Análise por modelo misto . . . . .	10
1.7	Exemplo com valores ausentes . . . . .	12
1.7.1	Análise por ANOVA de efeitos fixos . . . . .	12
1.7.2	Análise por modelo misto . . . . .	14
1.8	Por que usar modelos mistos? . . . . .	17
1.9	ANEXO . . . . .	19

**Obs.:** Este documento é uma tradução e alguns (poucos) incrementos do arquivo orinal encontrado no endereço <http://www2.compute.dtu.dk/courses/02429/enotepdfs/eNote-1.pdf> em 03/12/2022.

# 1 Introdução

## 1.1 Exemplo introdutório: NIR predição de medidas em HPLC

Em uma empresa farmacêutica, o uso da espectroscopia NIR (Near Infrared Reflectance) foi investigado como uma alternativa ao método mais complicado (e caro) de HPLC para determinar o conteúdo da substância ativa em pastilhas. Abaixo, as medições em 10 pastilhas são mostradas junto com as diferenças (em mg):

	pastilhas	HPLC	NIR	Dif
1	p1	10.4	10.1	0.3
2	p2	10.6	10.8	-0.2
3	p3	10.2	10.2	0.0
4	p4	10.1	9.9	0.2
5	p5	10.3	11.0	-0.7
6	p6	10.7	10.5	0.2
7	p7	10.3	10.2	0.1
8	p8	10.9	10.9	0.0
9	p9	10.1	10.4	-0.3
10	p10	9.8	9.9	-0.1
		HPLC	NIR	Dif
	Média	10.3400	10.3900	-0.0500
	DP	0.3239	0.4012	0.2953

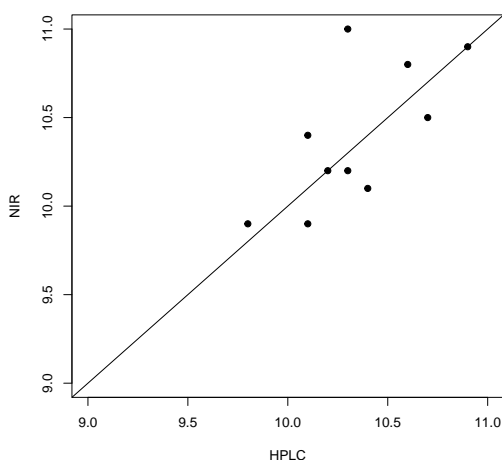


Figura 1: Dispersão espectroscopia NIR vs HPLC.

Um dos principais interesses reside na diferença média entre os dois métodos, também chamada de viés de método nesse contexto.

## 1.2 Análise simples

A abordagem mais direta para uma análise desses dados é considerar a configuração como uma configuração de duas amostras pareadas e realizar a análise de teste  $t$  pareada correspondente. A abordagem do teste  $t$  pareado corresponde a uma análise de uma amostra das diferenças, ou seja, calculando a média e o desvio padrão das 10 diferenças dadas na tabela acima:

$$\bar{d} = -0.05, \quad s_d = 0.2953$$

A incerteza da diferença estimada  $\bar{d} = -0.05$  é dado pelo erro padrão

$$EP(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{0.2953}{\sqrt{10}} = 0.0934$$

Consideremos as hipóteses:

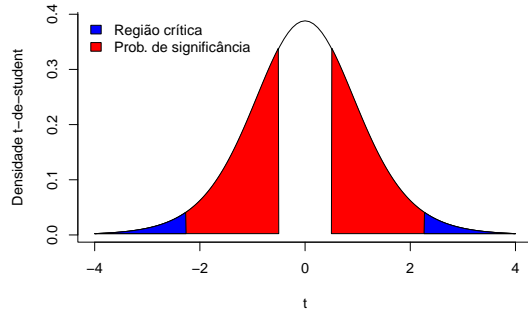
$$\begin{cases} H_0 : \ddot{d} = 0 & \text{Não há diferença entre as medidas HPLC e NIR} \\ H_1 : \ddot{d} \neq 0 & \text{NIR produz medidas diferentes de HPLC} \end{cases}$$

Um teste  $t$  para a hipótese de nenhuma diferença (sem viés de método) é então dado por

$$t_c = \frac{\bar{d} - \ddot{d}}{EP(\bar{d})} = \frac{-0.05}{0.0934} = -0.5354$$

```
(valor.p <- 2*pt(abs(t.c), df = 10-1, lower.tail = F))
```

```
[1] 0.6054
```



com um valor-p de 0.6054 não há viés de método significativo. Uma faixa de confiança de 95% para a diferença é dada por

$$\begin{aligned} IC[\ddot{d}, 95\%] &= \bar{d} \pm t_{0.975;9} EP(\ddot{d}) \\ &= -0.05 \pm 2.2622 \times 0.0934 \\ &= [-0.2613, 0.1613] \end{aligned} \tag{1}$$

mostrando que, embora não possamos reivindicar uma diferença significativa de método, a diferença pode muito bem ser tão pequena quanto -0.2613 ou tão grande quanto 0.1613.

A análise simples que acabamos de realizar é baseada em um modelo estatístico. Formalmente, o modelo estatístico é expresso por

$$d_i = \mu + \varepsilon_i, \quad \varepsilon \stackrel{iid}{\sim} N(0, \sigma^2)$$

em que:

- $\mu$  é a verdadeira média das diferenças entre os dois métodos (verdadeiro viés de mensuração devido a método );
- $\sigma$  o verdadeiro desvio padrão.

”Verdadeiro” refere-se aqui ao valor na “população”, da qual foi retirada a “amostra aleatória” de 10 pastilhas (neste caso, pastilhas com um nível nominal de 10 mg de conteúdo ativo). As diferenças  $d_i$  são definidos por  $d_i = y_{i2} - y_{i1}$ , sendo  $y_{ij}$  o valor medido pelo método  $j$  ( $j = 1$  (NIR) e  $j = 2$  (HPLC)) para pastilha,  $i$  ( $i = 1, \dots, 10$ ). Formalmente, a média e o desvio padrão das diferenças são “estimativas” dos valores populacionais e uma “notação de circunflexo (”chapéu”) é geralmente empregada:

$$\hat{\mu} = \bar{d}, \quad \hat{\sigma} = s_d$$

### 1.3 Análise simples por uma abordagem ANOVA

A configuração do teste  $t$  pareado também pode ser considerada como uma configuração de “blocos aleatórios” com 10 “blocos” (pastilhas) e dois “tratamentos” (métodos). O modelo para esta situação torna-se:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2), \quad (2)$$

em que:

- $\mu$  denota a média geral das diferenças das mensurações;
- $\alpha_i$  o efeito da  $i$ -ésima pastilha (bloco);
- $\beta_j$  o efeito do  $j$ -ésimo método.

Dados reorganizados:

	bl	met	y
1	p1	HPLC	10.4
2	p2	HPLC	10.6

```
... <NA> <NA> ...
19    p9  NIR 10.4
20   p10 NIR  9.9
```

Uma análise de variância (ANOVA) do modelo (2) resultará em:

```
fm.bl<- aov(y ~ bl + met, data=dat3)
(tb<- anova(fm.bl))
```

Analysis of Variance Table

Response: y

```
      Df Sum Sq Mean Sq F value Pr(>F)
bl      9  2.001  0.2223    5.10  0.012 *
met     1  0.013  0.0125    0.29  0.605
Residuals 9  0.392  0.0436
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
library(kableExtra)
tb4 <- kbl(tb , caption = "Análise de variância para o modelo
                        (\\ref{mod:bl})", booktabs = T)

kable_styling(tb4 ,latex_options = c( "hold_position"))
```

Tabela 1: Análise de variância para o modelo (2)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bl	9	2.0005	0.2223	5.0968	0.0118
met	1	0.0125	0.0125	0.2866	0.6054
Residuals	9	0.3925	0.0436	NA	NA

Observe que o valor-p para o efeito do método é o mesmo do teste  $t$  pareado acima. Isso é consequência do fato de que a estatística  $F$  é exatamente igual ao quadrado da estatística  $t$  pareada:

$$F_{met} = t_c^2 = (-0.5354)^2 = 0.2866$$

estimativa do desvio padrão residual  $\sigma$  dado por

$$\hat{\sigma} = \sqrt{QM_{Res}} = \sqrt{0.0436} = 0.2088$$

Observe que o valor obtido é igual ao desvio padrão das diferenças dividido por  $\sqrt{2}$ .

Lembre-se que o quadrado médio do resíduo é um estimador que denota a média ponderada das variâncias, e como temos dois grupos, podemos escrever<sup>1</sup>:

$$QM_{Res} = \frac{(n-1)s_{HPLC}^2 + (n-1)s_{NIR}^2}{(n-1) + (n-1)} = \frac{(n-1)(s_{HPLC}^2 + s_{NIR}^2)}{2(n-1)} = \frac{(s_{HPLC}^2 + s_{NIR}^2)}{2}$$

A incerteza da diferença média é dada por

$$EP(\bar{y}_2 - \bar{y}_1) = \sqrt{\hat{\sigma}^2 \left( \frac{1}{10} + \frac{1}{10} \right)} = \sqrt{0.0436 \left( \frac{1}{10} + \frac{1}{10} \right)} = 0.0934$$

exatamente como acima e, portanto, também a faixa de confiança de 95% será a mesma de (1).

## 1.4 O problema da abordagem ANOVA

A análise que acabamos de realizar é o que a maioria dos livros de estatística apresentaria como a análise dos dados do projeto de blocos completos aleatórios. E quando se trata do teste (e possível análise post hoc) das diferenças de tratamento, isso está perfeitamente correto, e qualquer *software* estatístico faria essa análise para você (em SAS: PROC GLM). O problema surge se você também pedir ao programa para lhe dar uma estimativa da incerteza das médias de tratamento individualmente, ou seja, não uma diferença de tratamento. Por exemplo, qual é a incerteza do valor médio do método NIR? Um uso comum do modelo (2) leva a:

$$EP(\bar{y}_1) = \frac{\hat{\sigma}}{\sqrt{10}} = \frac{0.2088}{\sqrt{10}} = 0.066$$

e isso é novamente o que qualquer procedimento de *software* ANOVA lhe diria. Isso NÃO está correto, no entanto! Suponha por um momento que apenas observamos o método NIR.

---

<sup>1</sup>Ver Montgomery, Design and Analysis of Experiments 8th, pág. 72

Em seguida, usaríamos esses 10 valores como uma amostra aleatória para obter o desvio padrão

$$s_1 = 0.4012 \text{ e a incerteza } \longrightarrow \text{EP}(\bar{y}_1) = \frac{s_1}{\sqrt{10}} = \frac{0.4012}{\sqrt{10}} = 0.1269$$

Observe como o modelo (2) subestima dramaticamente o que parece ser a incerteza real nos valores NIR. Isso ocorre porque a variância  $\sigma^2$  no modelo (2) mede a variabilidade residual após correção de possíveis diferenças entre as pastilhas (os efeitos das pastilhas no modelo). Na análise subsequente, e melhor análise a esse respeito, essa variabilidade entre as pastilhas que é usada. **A diferença conceitual entre as duas abordagens é se as 10 pastilhas são considerados uma amostra aleatória ou não.** Na ANOVA, cada pastilha é modelada individualmente com um efeito e os resultados da análise são válidos apenas para essas 10 pastilhas específicas. No entanto, essas 10 pastilhas específicas não são de particular interesse - estamos interessados em pastilhas em geral, e essas 10 pastilhas devem ser consideradas como representando esta população de pastilhas, ou seja, uma amostra aleatória. A desvantagem de analisar o NIR (resp. HPLC) em uma análise separada é que as informações disponíveis no conjunto de dados completo sobre a variabilidade dentro de cada pastilha não são usadas, levando a uma análise de dados ineficiente. A solução é combinar os dois em um modelo para os dados completos, que considere as pastilhas como uma amostra aleatória, ou seja: onde o efeito da pastilha é considerado um “efeito aleatório”.

## 1.5 O modelo misto

O modelo com pastilha como efeito aleatório é expresso como

$$y_{ij} = \mu + a_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2), \quad (3)$$

em que:

- $\mu$  denota a média geral das diferenças das mensurações;
- $a_i$  o efeito **aleatório**  $i$ -ésima pastilha (bloco), assumida ser independente e normalmente distribuída:  $a_i \stackrel{iid}{\sim} N(0, \sigma_T^2)$  ;
- $\beta_j$  o efeito do  $j$ -ésimo método.
- $a_i$  e  $\varepsilon_{ij}$  são independentes (importante).

Uma vez que o modelo consiste tanto em efeitos aleatórios (pastilhas) quanto em efeitos fixos (não aleatórios) (método), nos referimos a modelos desse tipo como **modelos mistos**. Cada efeito aleatório no modelo dá origem a um **componente de variância**.

O termo de erro residual  $\varepsilon_{ij}$  pode ser visto como um efeito aleatório e, portanto, a variância residual  $\sigma^2$  é um componente de variância.

## 1.6 Comparação de modelo fixo e misto

Para entender as diferenças conceituais entre os modelos (2) e (3) estudaremos três características teóricas dos modelos:

1. O valor esperado da  $ij$ -ésima observação  $\longrightarrow y_{ij}$ ;
2. A variância da  $ij$ -ésima observação  $\longrightarrow y_{ij}$ ;
3. A relação entre duas observações diferentes (covariância/correlação)

**OBS.:** Para entender a parte de covariância da tabela 2, lembre-se de que a covariância entre duas variáveis aleatórias, por ex. duas observações diferentes no conjunto de dados, expressam uma relação entre essas duas variáveis. Portanto, observações independentes têm covariância zero, que é o caso do modelo de efeitos fixos ordinários, onde todas as observações são consideradas independentes.

1.i para o modelo (2)

$$\mathbb{E}(y_{ij}) = \mathbb{E}(\mu + \alpha_i + \beta_j + \varepsilon_{ij}) = \mu + \alpha_i + \beta_j$$

1.ii para o modelo (3)

$$\mathbb{E}(y_{ij}) = \mathbb{E}(\mu + a_i + \beta_j + \varepsilon_{ij}) = \mu + \beta_j$$

2.i para o modelo (2) (lembrando que os termos do modelo são independentes)

$$\text{Var}(y_{ij}) = \text{Var}(\mu + \alpha_i + \beta_j + \varepsilon_{ij}) = \text{Var}(\varepsilon_{ij}) = \sigma^2$$

2.ii para o modelo (3)

$$\begin{aligned} \text{Var}(y_{ij}) &= \text{Var}(\mu + a_i + \beta_j + \varepsilon_{ij}) \\ &= \text{Var}(a_i + \varepsilon_{ij}) \\ &= \text{Var}(a_i) + \text{Var}(\varepsilon_{ij}) + \underbrace{2\text{Cov}(a_i, \varepsilon_{ij})}_{=0 \text{ por ind.}} \\ &= \sigma_T^2 + \sigma^2 \end{aligned}$$



3.i para o modelo (2), sendo  $y_{ij}$  e  $y_{i'j'}$  duas observações distintas

$$\begin{aligned}\text{Cov}(y_{ij}, y_{i'j'}) &= \text{Cov}(\mu + \alpha_i + \beta_j + \varepsilon_{ij}, \mu + \alpha_{i'} + \beta_{j'} + \varepsilon_{i'j'}) \\ &= \underbrace{\text{Cov}(\varepsilon_{ij}, \varepsilon_{i'j'})}_{=0 \text{ por ind.}} \\ &= 0\end{aligned}$$

3.ii para o modelo (3), sendo  $y_{ij}$  e  $y_{i'j'}$  duas observações distintas

$$\begin{aligned}\text{Cov}(y_{ij}, y_{i'j'}) &= \text{Cov}(\mu + a_i + \beta_j + \varepsilon_{ij}, \mu + a_{i'} + \beta_{j'} + \varepsilon_{i'j'}) \\ &= \text{Cov}(a_i + \varepsilon_{ij}, a_{i'} + \varepsilon_{i'j'}) \\ &= \text{Cov}(a_i, a_{i'}) + \text{Cov}(a_i, \varepsilon_{i'j'}) + \text{Cov}(\varepsilon_{ij}, a_{i'}) + \text{Cov}(\varepsilon_{ij}, \varepsilon_{i'j'})\end{aligned}$$

- se as duas observações são de pastilhas diferentes,  $i \neq i'$  as suposições de independência do modelo misto garantem que todas essas covariâncias (“relações”) entre as observações é zero;
- se duas observações provêm da mesma pastilha  $i = i'$  e  $j \neq j'$ , então  $\text{Cov}(a_i, \varepsilon_{i'j'}) = \text{Cov}(\varepsilon_{ij}, a_{i'}) = \text{Cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0$

$$\text{Cov}(y_{ij}, y_{i'j'}) = \text{Cov}(a_i, a_{i'}) = \text{Var}(a_i) = \sigma_T^2$$

Resumo das informações para comparação

Tabela 2: Resumo das informações para comparação		
Est.	Modelo fixo (2)	Modelo aleatório (3)
1. $\mathbb{E}(y_{ij})$	$\mu + \alpha_i + \beta_j$	$\mu + \beta_j$
2. $\text{Var}(y_{ij})$	$\sigma^2$	$\sigma_T^2 + \sigma^2$
3. $\text{Cov}(y_{ij}, y_{i'j'}); j \neq j'$	0	$\sigma_T^2$ (se $i = i'$ ) 0 (se $i \neq i'$ )

Os valores esperados e as variâncias mostram como os efeitos das pastilhas no modelo misto entram na parte de variância (aleatória) (como um componente de variância) em vez da

parte esperada (fixa/sistemática) do modelo. Ele também enfatiza que as expectativas sob o modelo misto não dependem de uma pastilha individual, mas é uma expectativa da média populacional das pastilhas.

Assim, as observações na mesma pastilha não são mais consideradas independentes - alguma correlação é permitida entre as observações na mesma pastilha. Isso ilustra uma característica essencial de passar de modelos de regressão/ANOVA padrão para modelos mistos.

### 1.6.1 Análise por modelo misto

Como mencionado acima, a análise dos dados com base no modelo misto é, neste caso, em grande parte, uma cópia exata da análise comum: é usada a mesma decomposição da variabilidade fornecida pela tabela ANOVA, os testes F para o método efeito (e/ou efeito de pastilha) e as comparações de métodos subsequentes são as mesmas. A incerteza do valor NIR médio no modelo misto é:

$$EP(\bar{y}_1) = \frac{\sqrt{\sigma_T^2 + \sigma^2}}{\sqrt{n}}$$

Assim, para calcular EP, precisamos estimar os dois componentes de variância. Uma maneira de fazer isso é usar as chamadas esperança dos quadrados médios. Eles fornecem a expectativa teórica dos três quadrados médios da tabela ANOVA: (não passamos por esses cálculos teóricos de esperanças aqui)

Tabela 3: Resumo das informações para comparação

FV	E(QM)	
	Modelo fixo (2)	Modelo aleatório (3)
Pastilhas	$\sigma^2 + \frac{b \sum_{i=1}^a \tau_i^2}{a-1}$	$a\sigma_T^2 + \sigma^2$
Métodos	$\sigma^2 + \frac{a \sum_{j=1}^b \beta_j^2}{b-1}$	$\sigma^2 + \frac{b \sum_{i=1}^a \beta_i^2}{a-1}$
Resíduos	$\sigma^2$	$\sigma^2$

Usando os resultados da ANOVA:

```
library(kableExtra)
tb4 <- kbl(tb , caption = "Análise de variância para o modelo
                        pastilha e método", booktabs = T)
```

```
kable_styling(tb4 ,latex_options = c( "hold_position"))
```

Tabela 4: Análise de variância para o modelo pastilha e método

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bl	9	2.0005	0.2223	5.0968	0.0118
met	1	0.0125	0.0125	0.2866	0.6054
Residuals	9	0.3925	0.0436	NA	NA

As esperanças mostram quais características dos dados entram em cada quadrado médio. Por exemplo, no quadrado médio residual entra apenas o componente de variância do erro residual, e assim esta é uma estimativa natural deste componente de variância,  $\sigma^2 = 0,0436$  (exatamente como no modelo fixo). E como a expectativa para o quadrado médio da pastilha mostra que 0,2223 é uma estimativa de  $2\sigma_T^2 + \sigma^2$ , podemos usar o valor de componente de variância do erro residual, e assim esta é uma estimativa natural deste componente de variância,  $\sigma^2$  para obter:

$$2\hat{\sigma}_T^2 + \hat{\sigma}^2 = 0.2223 \longrightarrow 2\sigma_T^2 = 0.2223 - 0.0436 \longrightarrow \sigma_T^2 = \frac{0.2223-0.0436}{2} = 0.0894$$

mostrando que a variação de pastilha para pastilha parece ter cerca de duas vezes o tamanho da variação residual. A incerteza do valor NIR médio agora se torna

$$EP(\bar{y}_1) = \frac{\sqrt{\sigma_T^2 + \sigma^2}}{\sqrt{n}} = \frac{\sqrt{0.0894 + 0.0436}}{\sqrt{10}} = 0.1153$$

agora muito mais próximo do valor encontrado na análise do dados NIR separada. E devido à especificação completa do modelo, agora conseguimos decompor a variabilidade total em seus dois componentes de variância. Isso fornece informações adicionais E tem impacto nos graus de liberdade a serem usados, quando o erro padrão for usado para testes de hipóteses e/ou cálculos de intervalo de confiança - algo ao qual retornaremos com mais detalhes posteriormente.

O exemplo mostrou como o modelo misto surge sendo a forma mais adequada de expressar um modelo estatístico apropriado à situação em questão. A mensagem principal era que, para evitar erros na direção da subestimação da incerteza, precisávamos do modelo misto. Como tal, surgiu como um mal necessário. O próximo exemplo ilustrará como o modelo misto de maneira direta pode fornecer informações sobre os principais problemas em um conjunto de dados, o que uma ANOVA fixa direta não fornece.

## 1.7 Exemplo com valores ausentes

Imagine que além das 10 pastilhas do exemplo anterior foram observadas outras 10 pastilhas, mas cada uma delas apenas para um dos dois métodos, dando a seguinte tabela de dados:

	pastilhas	HPLC	NIR	Dif
1	p1	10.4	10.1	0.3
2	p2	10.6	10.8	-0.2
3	p3	10.2	10.2	0.0
4	p4	10.1	9.9	0.2
5	p5	10.3	11.0	-0.7
6	p6	10.7	10.5	0.2
7	p7	10.3	10.2	0.1
8	p8	10.9	10.9	0.0
9	p9	10.1	10.4	-0.3
10	p10	9.8	9.9	-0.1
11	p11	NA	10.8	NA
12	p12	NA	9.8	NA
13	p13	NA	10.5	NA
14	p14	NA	10.3	NA
15	p15	NA	9.7	NA
16	p16	10.3	NA	NA
17	p17	9.6	NA	NA
18	p18	10.0	NA	NA
19	p19	10.2	NA	NA
20	p20	9.9	NA	NA

### 1.7.1 Análise por ANOVA de efeitos fixos

Uma análise de variância modelo fixo (ANOVA) (modelo (2)) resultará em:

```
fm.bl2<- aov(y ~ bl + met, data=dat4)
tb2<- anova(fm.bl2)
```

Tabela 5: Análise de variância para o modelo (2)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bl	19	3.7230	0.1959	4.4931	0.0129
met	1	0.0125	0.0125	0.2866	0.6054
Residuals	9	0.3925	0.0436	NA	NA

10 pastilhas

20 pastilhas

	Df	Sum Sq	Mean Sq		Df	Sum Sq	Mean Sq
bl	9	2.001	0.2223	bl	19	3.72	0.1959
met	1	0.013	0.0125	met	1	0.01	0.0125
Residuals	9	0.392	0.0436	Residuals	9	0.39	0.0436

Observe que apenas a linha referente a pastilhas da tabela mudou em comparação com a análise anterior. Da mesma forma, a estimativa da diferença média do método é dada como antes para ser

$$\hat{\beta}_2 - \hat{\beta}_1 = -0.05$$

usando apenas os 10 pastilhas para os quais ambas as observações estão presentes. E a incerteza também é dada apenas por esses 10 pastilhas, como antes:

$$EP(\hat{\beta}_2 - \hat{\beta}_1) = \sqrt{\hat{\sigma}^2 \left( \frac{1}{10} + \frac{1}{10} \right)} = 0.0934$$

Portanto, em resumo, a análise de efeito fixo usa apenas as informações das primeiras 10 pastilhas.

```
require(lsmeans)

# Ajuste com todos os dados
lsmeans::lsmeans(fm.bl2, pairwise ~ met)
```

```
$lsmeans
  met  lsmean      SE df lower.CL upper.CL
HPLC  10.2 0.0618   9     10.1     10.4
NIR   10.3 0.0618   9     10.1     10.4
```

Results are averaged over the levels of: bl  
 Confidence level used: 0.95

```
$contrasts
  contrast    estimate      SE df t.ratio p.value
HPLC - NIR    -0.05 0.0934   9  -0.535  0.6054
```

Results are averaged over the levels of: bl

### 1.7.2 Análise por modelo misto

Considere por um momento como uma análise das 10 pastilhas para os quais apenas um dos métodos foi observado poderia ser realizada. Este conjunto de dados pode ser considerado como duas amostras independentes - uma amostra de tamanho 5 dentro de cada método e uma configuração clássica de teste  $t$  de duas amostras:

```

 $\bar{y}_1 = 10.22, \quad s_1 = 0.4658$ 
 $\bar{y}_2 = 10.00, \quad s_2 = 0.2739$ 

tapply(dat4$y[21:30], dat4$met[21:30], mean)

## HPLC NIR
## 10.00 10.22

tapply(dat4$y[21:30], dat4$met[21:30], sd)

## HPLC NIR
## 0.2739 0.4658
```

A diferença é estimada em

$$\bar{y}_2 - \bar{y}_1 = -0.22,$$

e o erro padrão (pooled) para:

$$EP(\bar{y}_2 - \bar{y}_1) = \sqrt{s_{pool}^2 \left( \frac{1}{5} + \frac{1}{5} \right)} = \sqrt{0.146 \times \frac{2}{5}} = 0.24 \quad s_{pool}^2 = \frac{4s_1^2 + 4s_2^2}{5 + 5 - 2} = \frac{s_1^2 + s_2^2}{2} = 0.146$$

Os resultados das duas análises separadas podem ser resumidos como:

	Pastilhas 1-10	Pastilhas 11-20
Diferenças	-0.05	-0.22
EP <sup>2</sup>	0.00872	0.0584

A ANOVA de efeitos fixos usa apenas a primeira coluna de informações; seria preferível usar todas as informações. Como as duas estimativas da diferença do método têm incertezas (muito) diferentes, uma média ponderada das duas usando os erros padrão quadrados inversos como pesos pode ser calculada<sup>2</sup>:

$$\hat{\beta}_2 - \hat{\beta}_1 = \frac{-0.05 \frac{1}{0.00872} - 0.22 \frac{1}{0.0584}}{\frac{1}{0.00872} + \frac{1}{0.0584}} = -0.072$$

O uso de regras básicas de cálculo de variância fornece o erro padrão quadrado dessa média ponderada

$$\text{Var}(\hat{\beta}_2 - \hat{\beta}_1) = \left[ \text{EP}(\hat{\beta}_2 - \hat{\beta}_1) \right]^2 = \frac{1}{0.00872} + \frac{1}{0.0584} = 0.00759$$

portanto

$$\text{EP}(\hat{\beta}_2 - \hat{\beta}_1) = 0.0871$$

Observe que, além de fornecer um valor ligeiramente diferente, esse estimador também é mais preciso do que aquele baseado apenas nos tablets 1-10. Esse é o tipo de análise que o modelo misto para essa situação leva. E ao combinar os dados em uma análise (em vez de duas análises separadas), as informações sobre os dois componentes de variância são usadas de maneira otimizada. Nesse caso, os componentes de variância não são facilmente derivados da tabela ANOVA.

$$\hat{\sigma}^2 = 0.0435, \quad \hat{\sigma}_T^2 = 0.1019$$

$$\hat{\beta}_2 - \hat{\beta}_1 = -0.07211 \quad \text{EP}(\hat{\beta}_2 - \hat{\beta}_1) = 0.0871$$

Vemos como o modelo misto incorpora automaticamente todas as informações na análise da diferença do método, portanto superior a uma ANOVA de efeitos fixos puros. Este é um exemplo de como a análise pelo modelo misto automaticamente “recupera a informação entre blocos” em um delineamento de blocos incompletos, Cochran e Cox (1957).

<sup>2</sup><https://doi.org/10.5808/GI.2016.14.4.173>  
<https://doi.org/10.2307/3001666>

```
library(nlme)
fm.lme<- lme(y ~ met, random = ~1|bl, data=dat4)
```

```
summary(fm.lme)
```

Linear mixed-effects model fit by REML

Data: dat4

AIC BIC logLik

32.74 38.07 -12.37

Random effects:

Formula: ~1 | bl

(Intercept) Residual

StdDev: 0.3192 0.2085

$$\sigma_T^2 = 0.3192^2 = 0.1019$$

$$\sigma^2 = 0.2085^2 = 0.0435$$

Fixed effects: y ~ met

	Value	Std.Error	DF	t-value	p-value
(Intercept)	10.212	0.09259	19	110.29	0.0000
metNIR	0.072	0.08697	9	0.83	0.4285

Correlation:

(Intr)

metNIR -0.47

Standardized Within-Group Residuals:

	Min	Q1	Med	Q3	Max
	-1.166584	-0.438026	0.003153	0.436344	1.844778

Number of Observations: 30

Number of Groups: 20

```
require(lsmeans)
```

```
lsmeans::lsmeans(fm.lme, pairwise ~ met)
```

\$lsmeans

	met	lsmean	SE	df	lower.CL	upper.CL
HPLC	10.2	0.0926	19	10.0	10.4	
NIR	10.3	0.0926	9	10.1	10.5	



Degrees-of-freedom method: containment

Confidence level used: 0.95

\$contrasts

contrast	estimate	SE	df	t.ratio	p.value
HPLC - NIR	-0.0721	0.087	9	-0.829	0.4285

Degrees-of-freedom method: containment

## 1.8 Por que usar modelos mistos?

Acabamos de ver como o uso de modelos lineares mistos nos salvou de “cometer um erro” quando se trata da incerteza do nível NIR esperado. Mais direto ao ponto seria dizer que o modelo linear misto possibilitou ampliar a inferência estatística feita sobre o nível médio do NIR. A inferência estatística é o processo de usar dados para dizer algo sobre a(s) população(ões)/mundo real de onde os dados vieram em primeiro lugar. Estimativas de parâmetros, incertezas e intervalos de confiança e testes de hipóteses são exemplos de inferência estatística. A inferência induzida pelo modelo de efeitos fixos no exemplo introdutório é válida apenas para as 10 pastilhas específicas no experimento: a baixa incerteza é válida para a estimativa da média dos 10 valores NIR verdadeiros desconhecidos. A inferência induzida pelo modelo misto é válida para a estimativa do valor NIR médio da população de pastilhas.

Para a configuração de bloco completo randomizado, a inferência sobre as diferenças de tratamento não foi afetada pela ampliação do espaço de inferência. Em outras situações, quando os dados possuem uma estrutura hierárquica, a importância de fazer a inferência adequada é uma questão também para os testes de diferenças de tratamento. Se 20 pacientes forem alocados em dois grupos de tratamento e, posteriormente, medidos 10 vezes cada, a variabilidade essencial quando se trata de comparar os dois tratamentos provavelmente estará nas diferenças de paciente para paciente. E claramente não seria válido apenas analisar os dados como se 100 observações independentes estivessem disponíveis em cada grupo. Um modelo misto com pacientes como efeito aleatório lidaria com a situação de maneira adequada, induzindo a inferência mais provável de ser a relevante neste caso.

Também vimos acima como um modelo misto pode recuperar informações nos dados não encontrados por um modelo de efeito fixo, quando dados incompletos e/ou desbalanceados estão disponíveis. Este é um benefício importante dos modelos mistos.

A abordagem de modelo misto oferece uma maneira flexível de modelar covariância/correla-

ção nos dados. Isso é particularmente relevante para dados longitudinais ou outros tipos de dados de medidas repetidas, por ex. dados espaciais como em geoestatística. Dessa forma, a inferência adequada sobre efeitos fixos é obtida e a própria estrutura de covariância fornece informações adicionais sobre o problema em questão. A manipulação de variações não homogêneas em modelos fixos e mistos também está incluída na caixa de ferramentas.

Portanto, há muitas vantagens dos modelos mistos. E, em muitos casos, um modelo misto é realmente o único modelo razoável para os dados fornecidos. É justo admitir que também há uma potencial desvantagem. Mais suposições distributivas são feitas e aproximações são usadas na metodologia levando a resultados potencialmente enviesados. Também a alta complexidade dos (alguns) modelos torna o tratamento de dados e a comunicação dos resultados um desafio. No entanto, após este curso você deve estar pronto para enfrentar este desafio!

## 1.9 ANEXO

	Pastilhas 1-10	Pastilhas 11-20
Diferenças	$d_1$	$d_2$
EP <sup>2</sup>	EP <sub>1</sub>	EP <sub>2</sub>

Diferença conjunta:

$$\hat{\beta}_2 - \hat{\beta}_1 = \frac{d_1 \frac{1}{EP_1^2} + d_2 \frac{1}{EP_2^2}}{\frac{1}{EP_1^2} + \frac{1}{EP_2^2}}$$

Variância da diferença:

$$\begin{aligned}
\text{Var}(\hat{\beta}_2 - \hat{\beta}_1) &= \text{Var} \left[ \frac{d_1 \frac{1}{EP_1^2} + d_2 \frac{1}{EP_2^2}}{\frac{1}{EP_1^2} + \frac{1}{EP_2^2}} \right] \\
&= \frac{1}{\left( \frac{1}{EP_1^2} + \frac{1}{EP_2^2} \right)^2} \left[ \text{Var} \left( d_1 \frac{1}{EP_1^2} \right) + \text{Var} \left( d_2 \frac{1}{EP_2^2} \right) \right] \\
&= \frac{1}{\left( \frac{1}{EP_1^2} + \frac{1}{EP_2^2} \right)^2} \left[ EP_1^2 \frac{1}{(EP_1^2)^2} + EP_2^2 \frac{1}{(EP_2^2)^2} \right] \\
&= \frac{1}{\left( \frac{1}{EP_1^2} + \frac{1}{EP_2^2} \right)^2} \left[ \frac{1}{EP_1^2} + \frac{1}{EP_2^2} \right] \\
&= \frac{1}{\frac{1}{EP_1^2} + \frac{1}{EP_2^2}}
\end{aligned}$$