IBI5086
Introdução a Métodos Estatísticos para a Bioinformática

Profa. Júlia Maria Pavan Soler

pavan@ime.usp.br

IME/USP – 2º Semestre/2020

Programa

- Álgebra linear básica: cálculo matricial, determinantes, sistemas lineares, produto interno, norma, ortogonalidade, autovalores e autovetores
- Estrutura de Dados: variáveis (resposta, explicativa), unidades amostrais e experimentais

1.1. Comparação de Grupos (2 ou mais): Testes Clássicos (teste t, Wilcoxon, modelos ANOVA) e Testes de Aleatorização, Comparações Múltiplas

1.2. Análise de Tabelas de Contingência: Testes Qui-Quadrado, Regressão Logística.

2. Análise Multivariada de Dados: Componentes Principais, Análise Discriminante e Classificação, Correlação Canônica, modelos MANOVA

3. Simulação de Monte Carlo, Intervalos de Confiança Bootstrap

Planejamento de Experimentos e Modelos ANOVA (Análise de Variância)

$$Y = f(X) + e$$

Variável resposta quantitativa | Fatores (preditores) | Erro aleatório

- Estrutura dos Fatores (Tratamentos – variável X):

Já vimos:

✓ **Delineamento com Um único Fator e seus níveis**

- Delineamento Fatorial Cruzado

- Delineamento Fatorial Hierárquico (aninhado, *nested*)

- Estrutura das unidades amostrais (Aleatorização dos Tratamentos)

✓ **Delineamento Completamente Aleatorizado (DCA)**

- Delineamento Aleatorizado em Blocos Completos (DABC):

- Delineamento com uma Fator Aleatório

DCA Fatorial

Estrutura de aleatorização

Estrutura de Tratamentos

A_1			A_2			. . .	A_a		
B_1	B_2	... B_b	B_1	B_2	... B_b	. . .	B_1	B_2	... B_b
-	-	-	-	Y_{ijk}	-		-	-	-
-	-	-					-	-	-
-	-	-					-	-	-

Trat A em a níveis
Trat B em b níveis

- Y_{ijk}: resposta do i-ésimo indivíduo submetido ao j-ésimo nível do tratamento A e k-ésimo nível do tratamento B

- Estrutura de Tratamento $\Rightarrow$ 2 ou + Fatores Cruzados (Fatorial a x b)

- Delineamento Completamente Aleatorizado (DCA) com réplicas em cada combinação dos níveis dos fatores

Exemplo

Dados: Uma resposta de interesse é avaliada em unidades experimentais (n=24) submetidas a dois tratamentos (A e B), cada um em dois níveis (Baixo e Alto)

T1	T2	T3	T4	← Um Fator em 4 níveis
Baixo		Alto		← Trat A (Fator1)
Baixo	Alto	Baixo	Alto	← Trat B (Fator2)
6,2	12,7	7,0	8,3	
4,8	11,3	4,4	7,1	
3,0	9,3	3,8	11,7	
5,6	9,5	5,0	10,0	
7,1	11,7	5,5	8,5	
4,8	15,3	3,2	12,4	

A atribuição aleatória dos Tratamentos (4) às unidades experimentais (n=24).
Delineamento balanceado com 6 réplicas

Exemplo: DCA com 1 Fator em 4 níveis

Passo1: Entender o estudo

Dados de um DCA com 1 Fator em 4 níveis

T1	T2	T3	T4
6,2	12,7	7,0	8,3
4,8	11,3	4,4	7,1
3,0	9,3	3,8	11,7
5,6	9,5	5,0	10,0
7,1	11,7	5,5	8,5
4,8	15,3	3,2	12,4

$H_0 : \mu_j = \mu$

H_1 : Existe pelo menos uma diferença

Passo2: Adotar um modelo teórico de análise

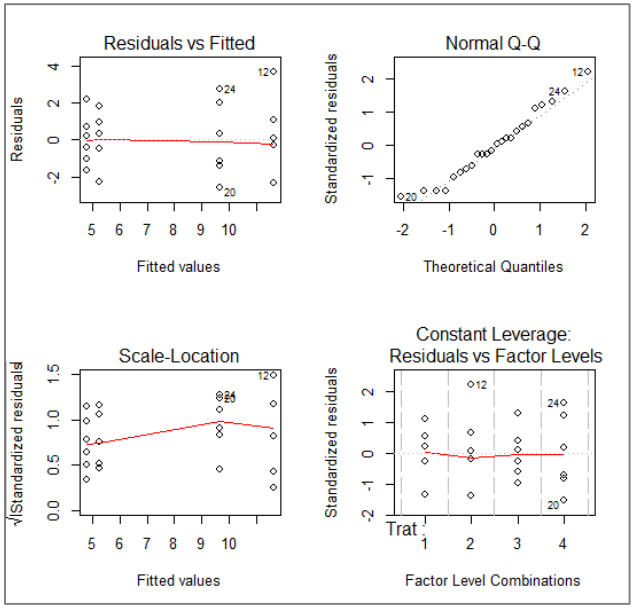
$$y_{ij} = \mu_j + e_{ij} = \mu_1 + \tau_j + e_{ij};$$

$$j = 2,3,4;\ i = 1,...,6$$

$$e_{ij} \overset{iid}{\sim} N\left(0;\sigma^2\right);$$

$$y_{ij} \sim N\left(\mu_j;\sigma^2\right)$$

Passo3: Ajuste do modelo e nálise de diagnóstico



Passo4: Análise dos resultados e Conclusão

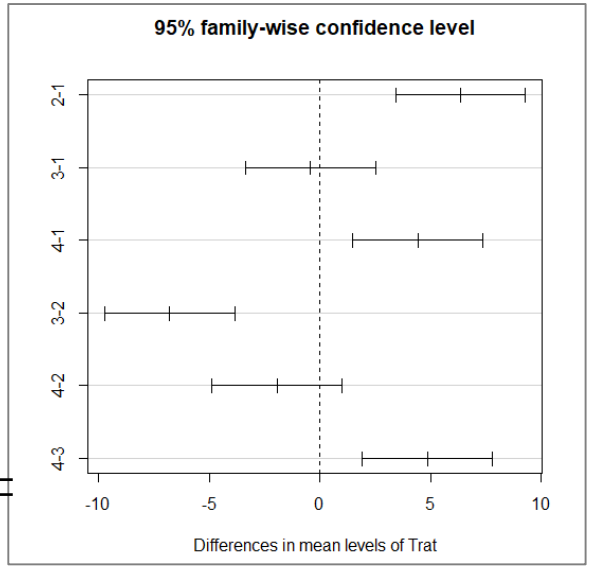
Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Trat	3	199.937	66.646	20.16	2.924e-06
Residuals	20	66.117	3.306		

Coefficients

(Intercept)	Trat2	Trat3	Trat4
5.250	6.350	-0.433	4.417

$$\left(\mu_2 = \mu_4\right) > \left(\mu_1 = \mu_3\right) \Leftarrow$$



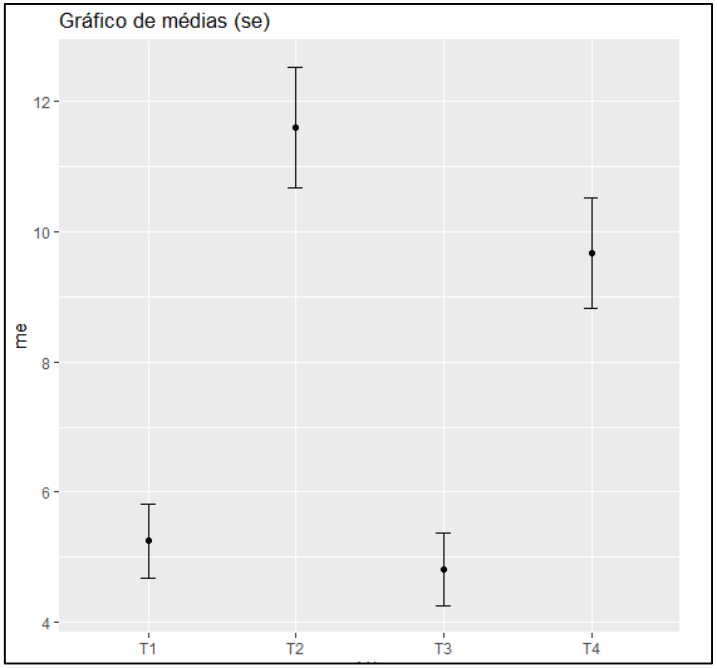
DCA – Estrutura de Tratamento

Dados de um DCA com 1 Fator em 4 níveis

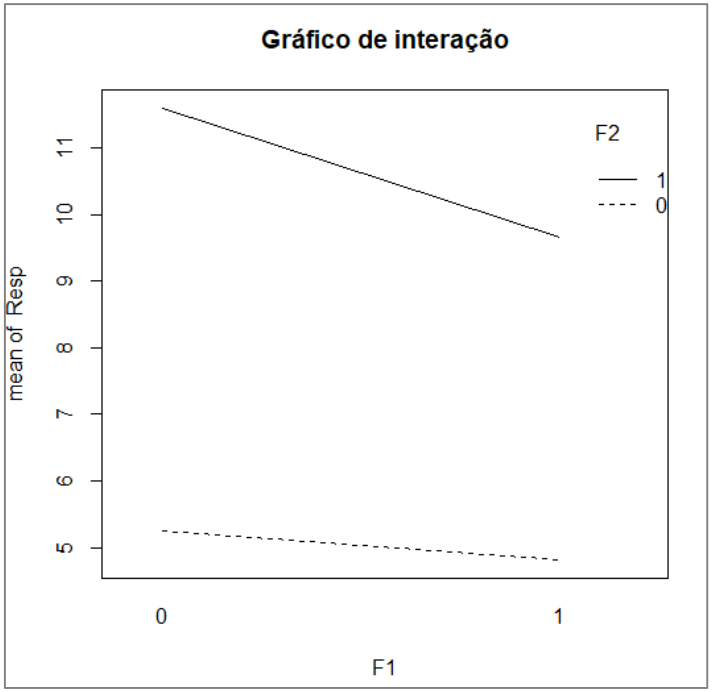
T1	T2	T3	T4
6,2	12,7	7,0	8,3
4,8	11,3	4,4	7,1
3,0	9,3	3,8	11,7
5,6	9,5	5,0	10,0
7,1	11,7	5,5	8,5
4,8	15,3	3,2	12,4

1 Fator em 4 níveis ⇒ Fatorial 2x2

T1	T2	T3	T4	
Baixo	Baixo	Alto	Alto	Trat A (F1)
Baixo	Alto	Baixo	Alto	Trat B (F2)



⇒



DCA – Estrutura de Tratamento

DCA com 1 Fator em J níveis:

$$y_{ij} = \mu_j + e_{ij} = \mu_1 + \tau_j + e_{ij};$$

$$e_{ij} \overset{iid}{\sim} N\left(0;\sigma^2\right); \quad y_{ij} \overset{ind}{\sim} N\left(\mu_j;\sigma^2\right); \quad j = 1,...,J; \quad i = 1,..,n_j = r$$

DCA Fatorial axb: $y_{ijk} = \mu_{jk} + e_{ijk} = \mu_{00} + \tau_j + \beta_k + \gamma_{jk} + e_{ijk};$

J = axb

$$e_{ijk} \overset{iid}{\sim} N\left(0;\sigma^2\right); \quad y_{ijk} \overset{ind}{\sim} N\left(\mu_{jk};\sigma^2\right); \quad j = 1,..,a; k = 1,...,b; \; i = 1,..,n_{jk} = r$$

y_{ijk} : resposta da unidade *i* submetida aos níveis *j* do Fator A e *k* do Fator B

μ_{00} : valor esperado da resposta para o Trat de referência (A e B nos níveis baixos)

τ_j : desvio em relação a μ_{00} devido ao efeito (principal) do Fator A no nível j

β_k : desvio em relação a μ_{00} devido ao efeito (principal) do Fator B no nível k

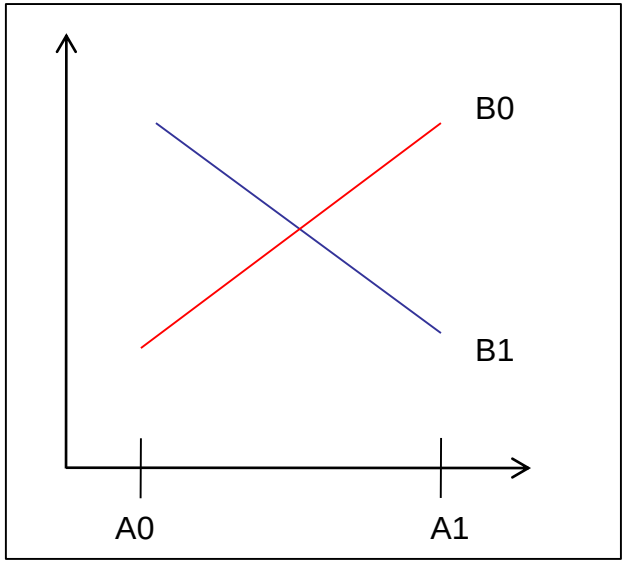
γ_{jk} : efeito de interação entre os fatores A e B. É o desvio do efeito aditivo dos fatores

e_{ijk} : efeito aleatório, suposto normal, independente e homocedástico.

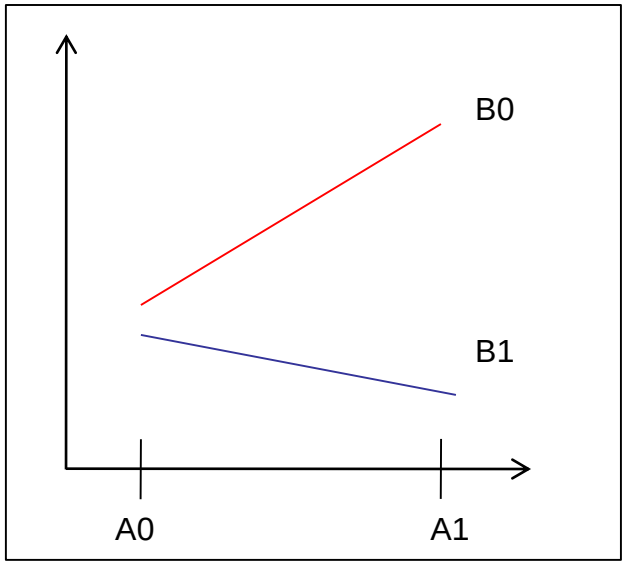
Perfis de Médias – Efeito de Interação
Fatorial 2x2

Existência de Interação

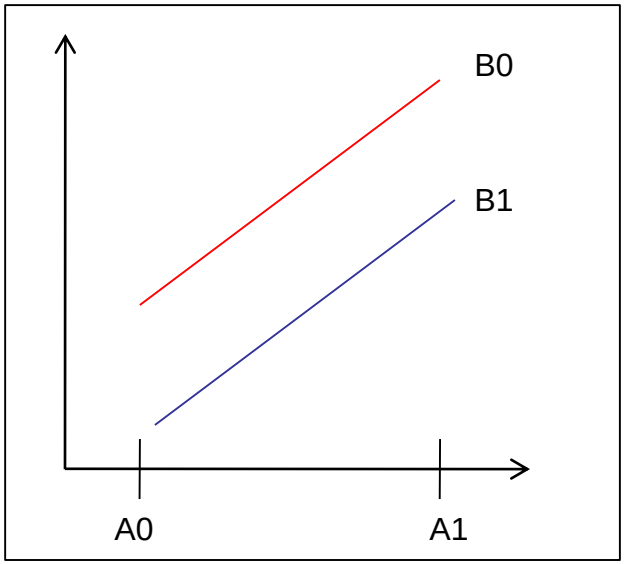
A direção do Efeito de B muda de acordo com o nível de A



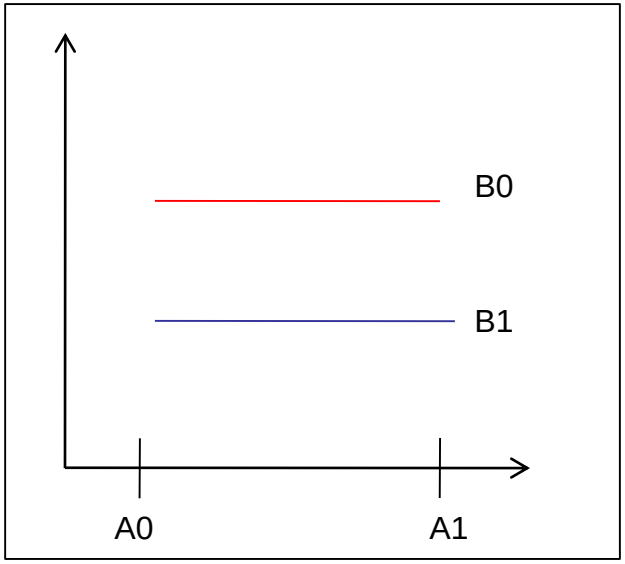
A magnitude do Efeito de B muda de acordo com o nível de A



Não Existência de Interação

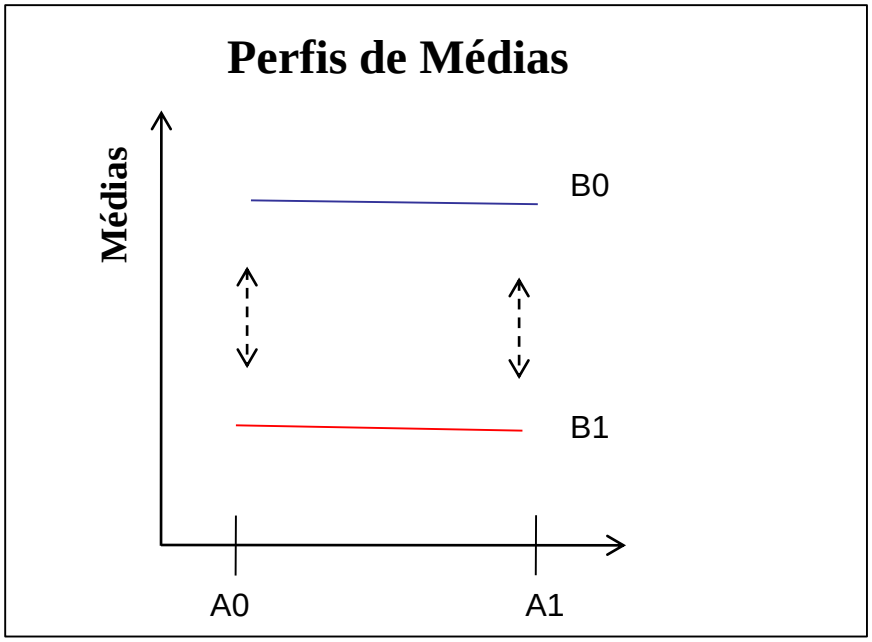


Efeito principal tanto de A como de B

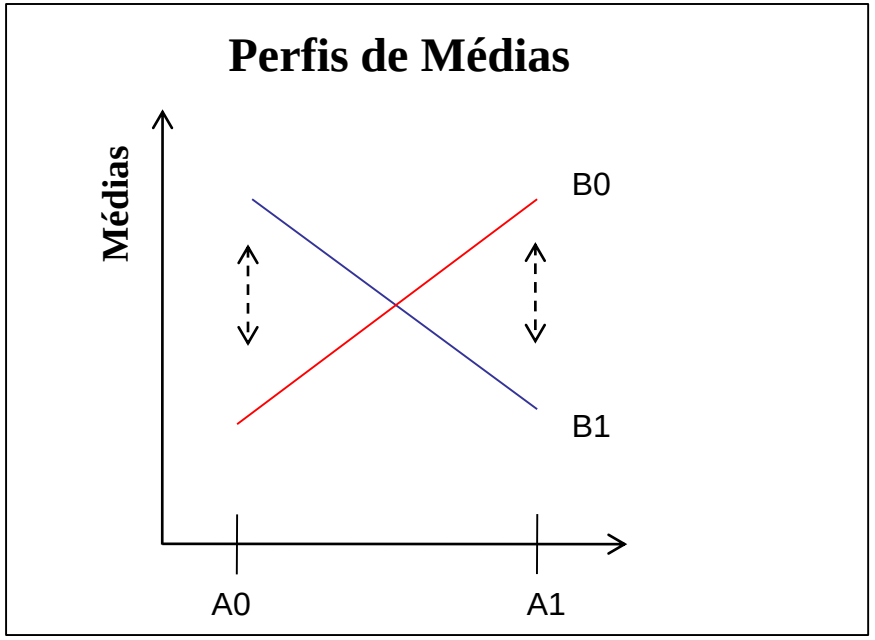


Efeito principal somente de B

Perfis de Médias – Efeito de Interação
Fatorial 2x2



Inexistência de interação entre os fatores (paralelismo)



Existência de interação entre os fatores (desvio do paralelismo)

$$H_0 : \mu_{00} - \mu_{01} = \mu_{10} - \mu_{11}$$

$$H_1 : \mu_{00} - \mu_{01} \neq \mu_{10} - \mu_{11}$$

Efeito de interação é definido como um contraste entre médias

DCA com um Fator em J níveis
Decomposição de Fontes de Variação

F.V.	nºg.l.	SQ
Entre Trat.	**J-1**	$\sum_j n_j(\bar{y}_{.j} - \bar{y})^2$
Resíduo	**n-J**	$\sum_{ij}(y_{ij} - \bar{y}_{.j})^2$
Total	**n-1**	$\sum_{ij}(y_{ij} - \bar{y})^2$

Fator A **a-1**

Fator B **b-1**

Interação **(a-1)(b-1)**

(a-1)+(b-1)+(a-1(b-1)=ab-1=J-1

J=ab n=abr n_j=r

Decomposição da soma de quadrados Entre Tratamentos e do correspondente número de graus de liberdade de acordo com Efeitos Principais e de Interação entre os fatores

Delineamento Fatorial
Modelo e Fontes de Variação

$$y_{ijk} = \mu_{jk} + \varepsilon_{ijk} = \mu + \tau_j + \beta_k + \gamma_{jk} + e_{ijk}; \quad e_{ijk} \overset{iid}{\sim} N\left(0; \sigma^2\right)$$

Identidade útil:

$$y_{ijk} - \bar{y} = (\bar{y}_{j.} - \bar{y}) + (\bar{y}_{.k} - \bar{y}) + (\bar{y}_{jk} - \bar{y}_{j.} - \bar{y}_{.k} + \bar{y}) + (y_{ijk} - \bar{y}_{jk})$$

$\sum_{ijk}(y_{ijk} - \bar{y})^2$	$\sum_{j} br(\bar{y}_{j.} - \bar{y})^2$	$\sum_{k} ar(\bar{y}_{.k} - \bar{y})^2$	$\sum_{jk} r(\bar{y}_{jk} - \bar{y}_{j.} - \bar{y}_{.k} + \bar{y})^2$	$\sum_{ijk}(y_{ijk} - \bar{y})^2$
SQTotal	**SQA**	**SQB**	**SQA*B**	**SQRes**
	associadas aos Efeitos principais dos fatores A e B		associada ao Efeito de Interação	

Tabela de ANOVA
DCA – Fatorial axb

F.V.	g.l.	SQ	QM	F
Fator A	**a-1**	$\sum_{j} br(\bar{y}_{j.} - \bar{y})^2$	**QM(A)**	**QM(A)/QMRes**
Fator B	**b-1**	$\sum_{k} ar(\bar{y}_{.k} - \bar{y})^2$	**QM(B)**	**QM(B)/QMRes**
Interação A*B	**(a-1)(b-1)**	$\sum_{jk} r(\bar{y}_{jk} - \bar{y}_{j.} - \bar{y}_{.k} + \bar{y})^2$	**QM(A*B)**	**QM(A*B)/QMRes**
Resíduo	**n-(ab-1)**	$\sum_{ijk} (y_{ijk} - \bar{y}_{jk})^2$	**QMRes**	
Total	**n-1**	$\sum_{ijk} (y_{ijk} - \bar{y})^2$		

Testar primeiro o efeito de interação (H_{01}: $\gamma_{ij}=0$)
Se H_{01} for rejeitada, realizar comparações múltiplas para estudar o efeito de interação
Se H_{01} não for rejeitada, testar os efeitos principais dos fatores (modelos reduzidos, aditivos)

Tabelas ANOVA

Dados de um DCA com 1 Fator em 4 níveis

T1	T2	T3	T4
6,2	12,7	7,0	8,3
4,8	11,3	4,4	7,1
3,0	9,3	3,8	11,7
5,6	9,5	5,0	10,0
7,1	11,7	5,5	8,5
4,8	15,3	3,2	12,4

Tabela ANOVA: Modelo de 1 Fator em 4 níveis

```
          Df  Sum Sq Mean Sq F value    Pr(>F)
Trat       3 199.937  66.646   20.16 2.924e-06
Residuals 20  66.117   3.306
```

Coefficients (codificação: Trat=1,2,3,4)

```
(Intercept)  Trat2   Trat3   Trat4
      5.250  6.350  -0.433   4.417
```

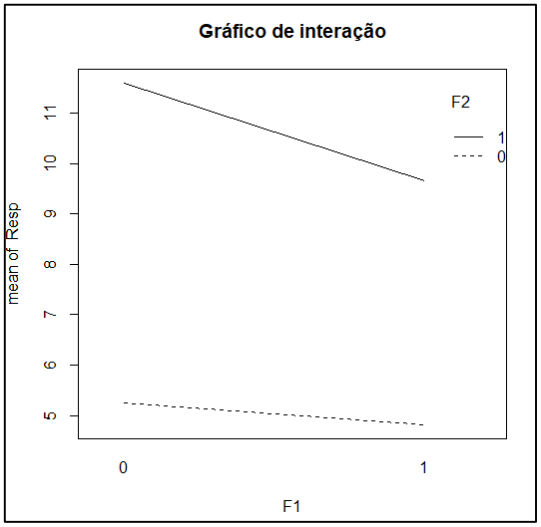


Tabela Anova: Modelo Fatorial 2x2 (Interação)

```
          Df  Sum Sq Mean Sq F value    Pr(>F)
F1         1   8.402   8.402  2.5415    0.1266
F2         1 188.160 188.160 56.9176 2.849e-07
F1:F2      1   3.375   3.375  1.0209    0.3244
Residuals 20  66.117   3.306
```

Coefficients (codificação:F1(0=30% 1=100%) F2(0=SN 1=N)

```
(Intercept)      F1      F2   F1:F2
      5.250  -0.433   6.350   -1.50
```

Obter Modelos reduzidos (sem interação → só F2)

ANÁLISE DE VARIÂNCIA – MODELO REDUZIDO

EFEITOS PRINCIPAIS

Source	DF	SS	MS	F	P
F1	1	8,640	8,640	2,64	0,119
F2	1	189,282	189,282	57,81	0,000
Error	21	68,757	3,274		
Total	23	266,678			

Concl. ?

ANÁLISE DE VARIÂNCIA – MODELO "MAIS" REDUZIDO

EFEITO PRINCIPAL DE F2

Source	DF	SS	MS	F	P
F2	1	189,28	189,28	53,80	0,000
Error	22	77,40	3,52		
Total	23	266,68			

Concl. ?

Não é necessário aplicar o teste de Tukey neste caso.

O teste F (=53,80) é equivalente ao teste t para amostras independentes (sob homocedasticidade) $t^2_{22} = F_{1,22}$

Delineamento Aleatorizado em Blocos Completos

	Tratamentos			
Bloco	**T_1**	**T_2**	**. . .**	**T_a**
B_1	**Y_{11}**	**Y_{21}**	**. . .**	**Y_{a1}**
B_2	**Y_{12}**	**Y_{22}**	**. . .**	**Y_{a2}**
	. . .	**. . .**	**Y_{ij}**	**. . .**
B_r	**Y_{1r}**	**Y_{2r}**	**. . .**	**Y_{ar}**

"aleatorização restrita" dentro de blocos

a unidades experimentais dentro de cada bloco são aleatoriamente atribuídas aos tratamentos

r replicações em cada tratamento

Delineamento Aleatorizado em Blocos Completos

Estrutura dos Dados

- **Fatores**
  - **Tratamento:** ***fator de interesse***
  - **Bloco:** ***fator de controle***
- **Aleatorização Restrita (Dentro do Bloco)**
- **Variável Resposta**

Vantagem da Blocagem

eliminar o efeito de uma fonte de variação conhecida (bloco)

⬇

redução do efeito residual

Exemplo

<u>Dados</u>: Medidas realizadas de acordo com um tratamento sob Blocagem (fator de Controle)

	Tratamento			
Bloco	**T1**	**T2**	**T3**	**T4**
B1	**6,2**	**12,7**	**7,0**	**8,3**
B2	**4,8**	**11,3**	**4,4**	**7,1**
B3	**3,0**	**9,3**	**3,8**	**11,7**
B4	**5,6**	**9,5**	**5,0**	**10,0**
B5	**7,1**	**11,7**	**5,5**	**8,5**
B6	**4,8**	**15,3**	**3,2**	**12,4**

Estrutura de Blocos (hipotético): coluna de água de um rio.

Delineamento Completamente Aleatorizado
Delineamento Aleatorizado em Blocos Completos

DCA

T1	T2	…	Ta	
⋮	⋮		⋮	
n_1	n_2	…	n_a	n

Aleatorização irrestrita das *n* unidades experimentais aos *a* tratamentos

DAB

	T1	T2	…	Ta
B1	⋮	⋮		⋮
B2				
…				
Br				
	r	r	…	r

Aleatorização restrita das *a* unidades experimentais dentro de cada um dos *r* blocos

Modelo Teórico - Aditividade

$$y_{ij} = \mu_{ij} + e_{ij}$$

$$= \underbrace{\mu_j + \beta_i}_{\text{componente fixo}} + e_{ij}$$

$$e_{ij} \overset{iid}{\sim} N\left(0; \sigma^2\right)$$

componente aleatório

$$y_{ij} = \mu + \tau_j + \beta_i + e_{ij}$$

efeito do tratamento

Efeito do bloco

$$\sum_{j=1}^{J} \tau_j = \sum_{i=1}^{r} \beta_i = 0$$

Suposição: Modelo Aditivo (não há efeito de interação entre os fatores)

Modelo Teórico / Estimadores

$$y_{ij} = \mu_{ij} + e_{ij} = \mu + \tau_j + \beta_i + e_{ij}; \quad e_{ij} \overset{iid}{\sim} N\left(0; \sigma^2\right)$$

identidade útil na obtenção das estatísticas

Resíduo é o efeito de interação entre os fatores

$$y_{ij} = \bar{y} + (\bar{y}_{.j} - \bar{y}) + (\bar{y}_{i.} - \bar{y}) + (y_{ij} - \bar{y}_{.j} - \bar{y}_{i.} + \bar{y})$$

Valor observado

$\hat{\tau}_j$ $\hat{\beta}_i$ $\hat{e}_{ij}$

$\hat{y}_{ij}$ **Valor predito**

Hipótese de Interesse

$$Y_{ij} \sim N(\mu_j + \beta_i ; \sigma^2)$$

$$H_0 : \mu_1 = \mu_2 = ... = \mu_J = \mu$$

H_1: **existe pelo menos uma diferença**

Em geral, não há interesse em testar o efeito do fator Bloco (apesar de, teoricamente, ser possível testar!)

$$H_0 : \beta_1 = \beta_2 = ... = \beta_r = \beta$$

Qual é a estatística deste teste ?

Porém, note que não é possível testar o efeito de interação entre os fatores tratamento e bloco! Este efeito define o resíduo do modelo.

Tabela de ANOVA

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_J = \mu$$

F.V.	g l	SQ	QM	F	p
TRAT	**J-1**	$\sum r\,(\bar{y}_{.j} - \bar{y})^2$	**SQTrat/(J-1)**	$\frac{\text{QMTrat}}{\text{QMRes}}$	
BLOCO	**r-1**	$\sum J\,(\bar{y}_{i.} - \bar{y})^2$			
RESÍDUO	**(J-1)(r-1)**	$\sum_{ij} (y_{ij} - \bar{y}_{.j} - \bar{y}_{i.} + \bar{y})^2$	**SQRes/(J-1)(r-1)**		
TOTAL	**rJ-1**	$\sum_{ij} (y_{ij} - \bar{y})^2$			

(r-1)+(J-1)(r-1) = rJ-J

g.l. da interação

$$F = \frac{QMTrat}{QM\,\text{Res}} \overset{H_0}{\sim} F_{(J-1),(J-1)(r-1)}$$

Como tomar decisão sobre H_0 ?

Em geral, não há interesse em testar o efeito do fator Bloco e, além das premissas clássicas, a validade do modelo aditivo deve ser investigada.

Tabela de ANOVA - DABC

Compare com a tabela de ANOVA do correspondente DCA.

H: $\mu_1 = \mu_2 = \ldots = \mu_a = \mu$

F.V.	g.l.	SQ	QM	F	p
TRAT	**3**	**201.448**	**67.149**	**19.79**	**0.000**
BLOCO	**5**	**14.343**	**2.869**	**0.85**	**0.538**
RESÍDUO	**15**	**50.346**	**3.392**		
TOTAL	**23**	**266.678**			

Com a inclusão do fator bloco, houve ganho em precisão na estimação do efeito do tratamento?

Compare o número de graus de liberdade dos delineamentos DCA e DABC definidos para o mesmo conjunto de dados!

Modelos ANOVA

DCA: Delineamento Completamente Aleatorizado

Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Trat	3	199.937	66.646	20.16	2.924e-06	***
Residuals	20	66.117	3.306			

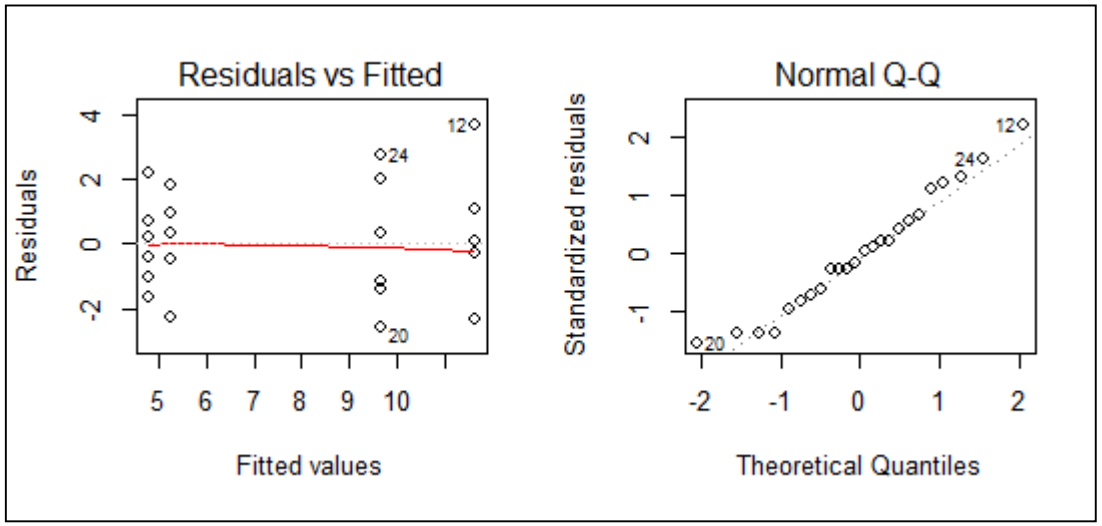
DABC: Delineamento Aleatorizado em Blocos Completos

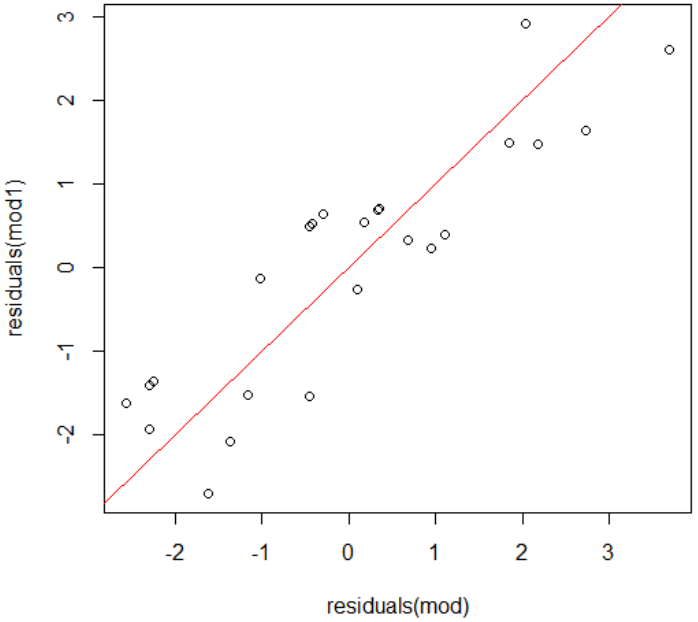
Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Trat	3	199.937	66.646	19.3593	2.04e-05	***
Bloco	5	14.478	2.896	0.8411	0.5412	
Residuals	15	51.638	3.443			

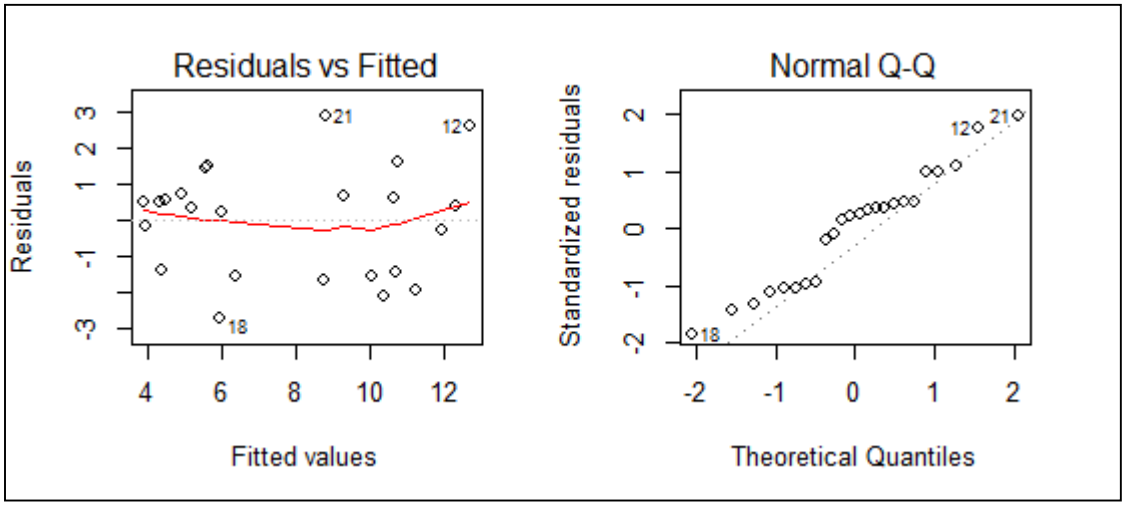
Modelos ANOVA

ANOVA-DCA





ANOVA-DABC



Não há evidência amostral de ganho em precisão devido à inclusão do fator Bloco: sob o DCA (QMRes=3,306) e sob o DABC (QMRes=3,443)

Dados: Respostas avaliadas sob os Tratamentos (T1, T2, T3 e T4)

	trat	bloco	resp
1	1	1	6.2
2	2	1	12.7
3	3	1	7.0
4	4	1	8.3
5	1	2	4.8
6	2	2	11.3
7	3	2	4.4
8	4	2	7.1
9	1	3	3.0
10	2	3	9.3
11	3	3	3.8
12	4	3	11.7
13	1	4	5.6
14	2	4	9.3
15	3	4	5.0
16	4	4	10.0
17	1	5	7.1
18	2	5	11.7
19	3	5	5.5
20	4	5	8.5
21	1	6	4.8
22	2	6	15.3
23	3	6	3.2
24	4	6	12.4

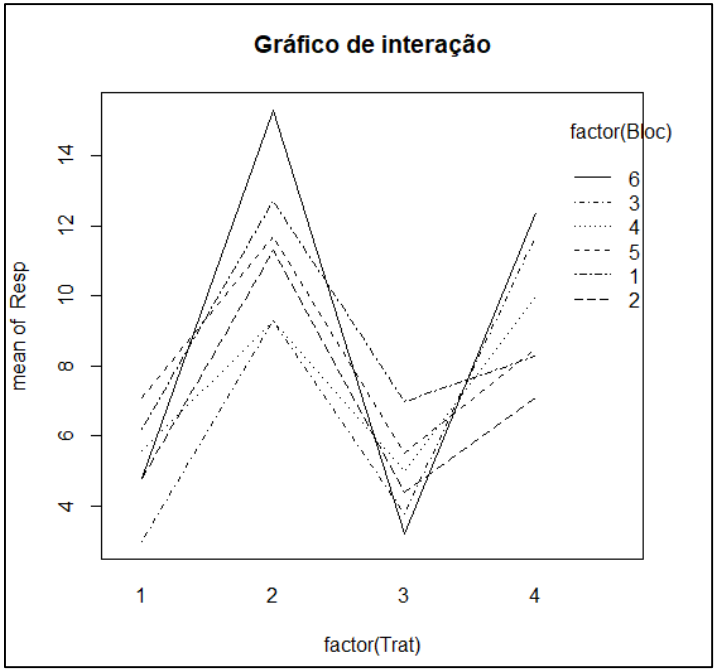
Estimativas dos parâmetros do modelo (no "R")

Intercept	trat2	trat3	trat4
5.9666667	6.3500000	-0.4333333	4.4166667

bloco2	bloco3	bloco4	bloco5	bloco6
-1.6500000	-1.6000000	-1.0750000	-0.3500000	0.3750000

Interprete as estimativas dos parâmetros do modelo estrutural adotado na análise dos dados (sob o DABC)!

Há indicação de efeito de interação entre os fatores Tratamento e Bloco?



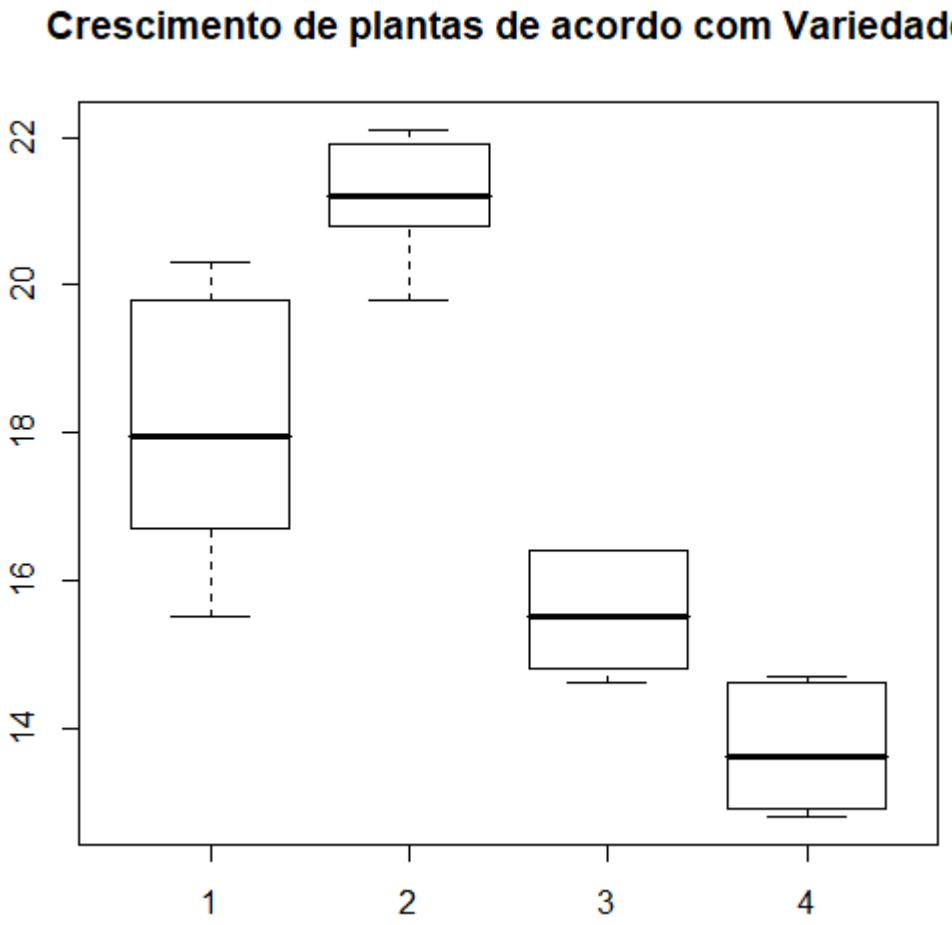
Exemplo

Dados: Crescimento de plantas (cm) de acordo com a variedade

Canteiro	Var1	Var2	Var3	Var4
1	19,8	21,9	16,4	14,7
2	16,7	19,8	15,4	13,5
3	17,7	21,0	14,8	12,8
4	18,2	21,4	15,6	13,7
5	20,3	22,1	16,4	14,6
6	15,5	20,8	14,6	12,9

Blocos: canteiros

Modelo ANOVA - DABC



	Var1	Var2	Var3	Var4
Média	18.03	21.17	15.53	13.70
Desvio Padrão	1.82	0.84	0.77	0.81

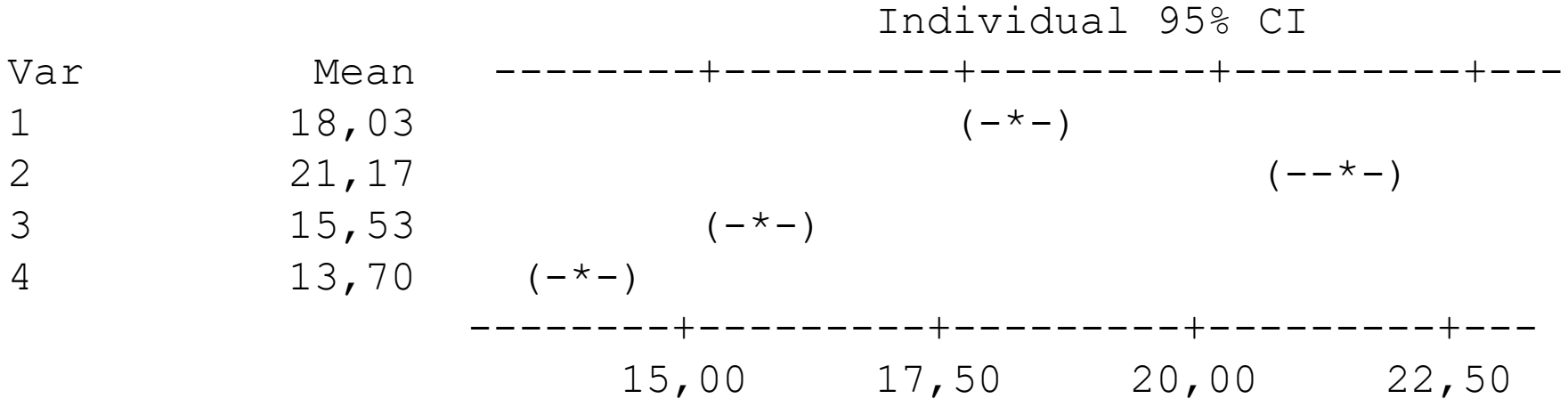
Hipóteses ?

Suposições ?

Analysis of Variance Table

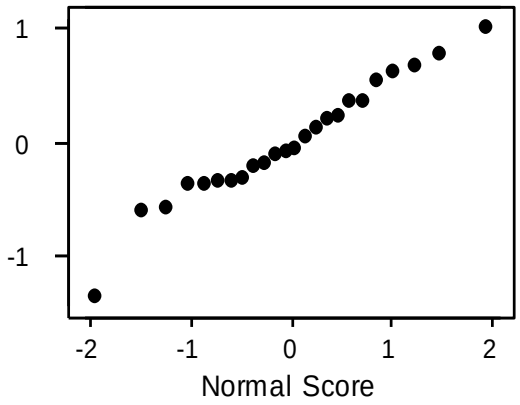
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(Var)	3	188.538	62.846	144.4369	2.742e-11
factor(Bloco)	5	19.793	3.959	9.0981	0.0003857
Residuals	15	6.527	0.435		
Total	23	214.858			

Concl. ?

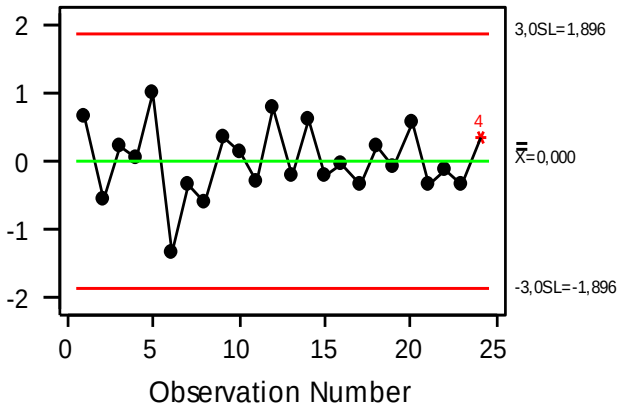


Residual Model Diagnostics

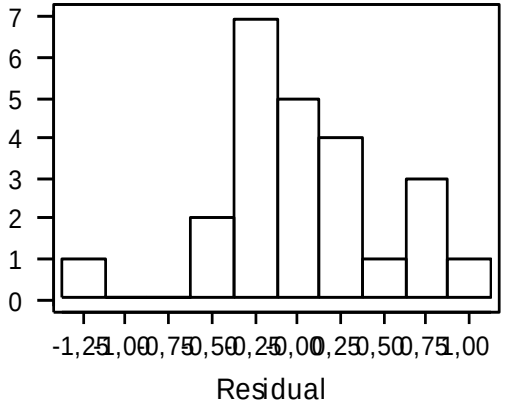
Normal Plot of Residuals



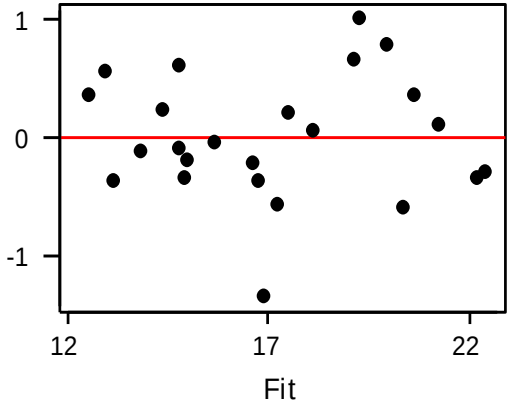
I Chart of Residuals



Histogram of Residuals



Residuals vs. Fits



As suposições do modelo estão satisfeitas ?

Modelos ANOVA – DCA e DABC

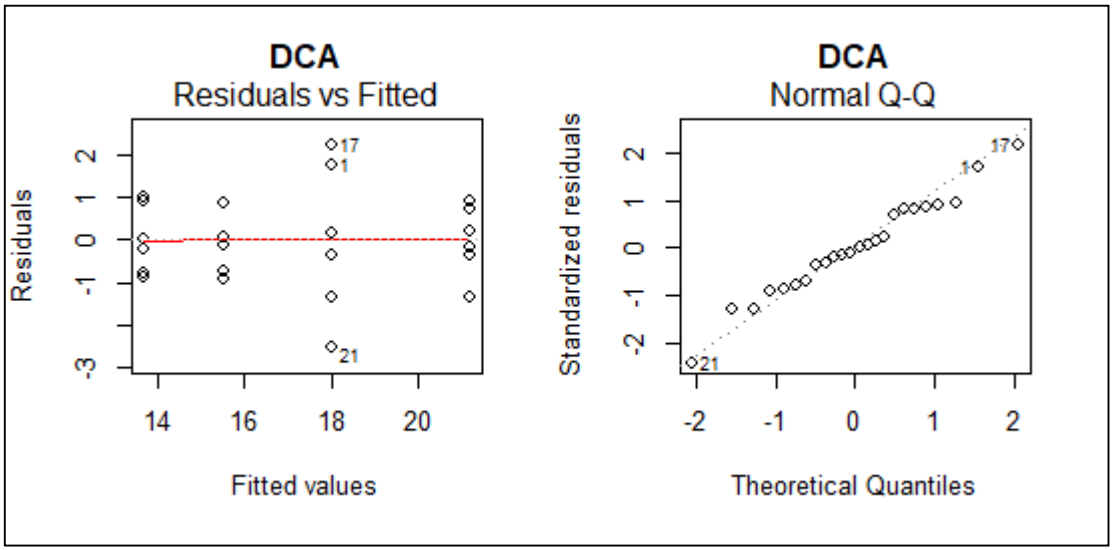
Delineamento Completamente Aleatorizado - DCA

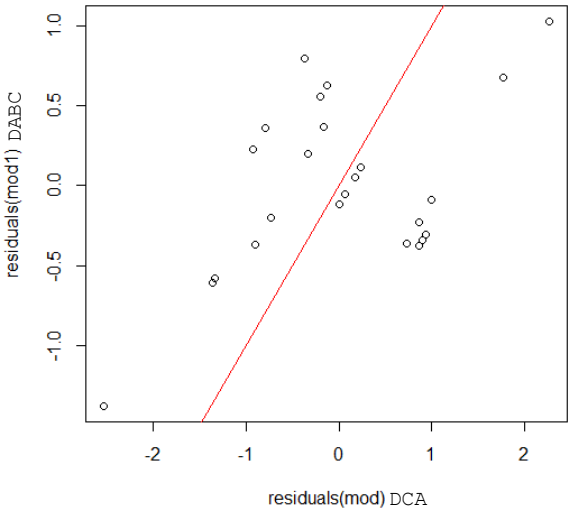
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
factor(Var)	3	188.54	62.846	47.755	2.654e-09	***
Residuals	20	26.32	1.316			

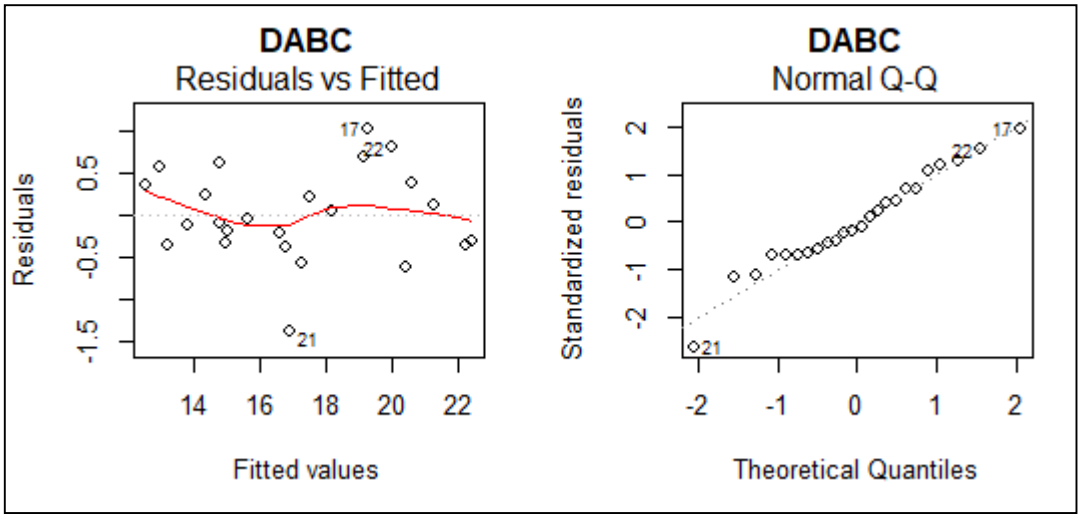
Delineamento Aleatorizado em Blocos Completos - DABC

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
factor(Var)	3	188.538	62.846	144.4369	2.742e-11	***
factor(Bloco)	5	19.793	3.959	9.0981	0.0003857	***
Residuals	15	6.527	0.435			

Modelos ANOVA – DCA e DABC







Maior variância dos resíduos sob o DCA (QMRes=1,316) relativamente ao DABC (QMRes=0,435)

DABC - Tukey Simultaneous Tests (efeito do fator Variedade)

Response Variable resp

All Pairwise Comparisons among Levels of var

Hipóteses ?

var = 1 subtracted from:

Level var	Difference of Means	SE of Difference	T-Value	Adjusted P-Value
2	3,133	0,3808	8,23	0,0000
3	-2,500	0,3808	-6,56	0,0001
4	-4,333	0,3808	-11,38	0,0000

(21,17-18,03)/(√2*0.435/6)

var = 2 subtracted from:

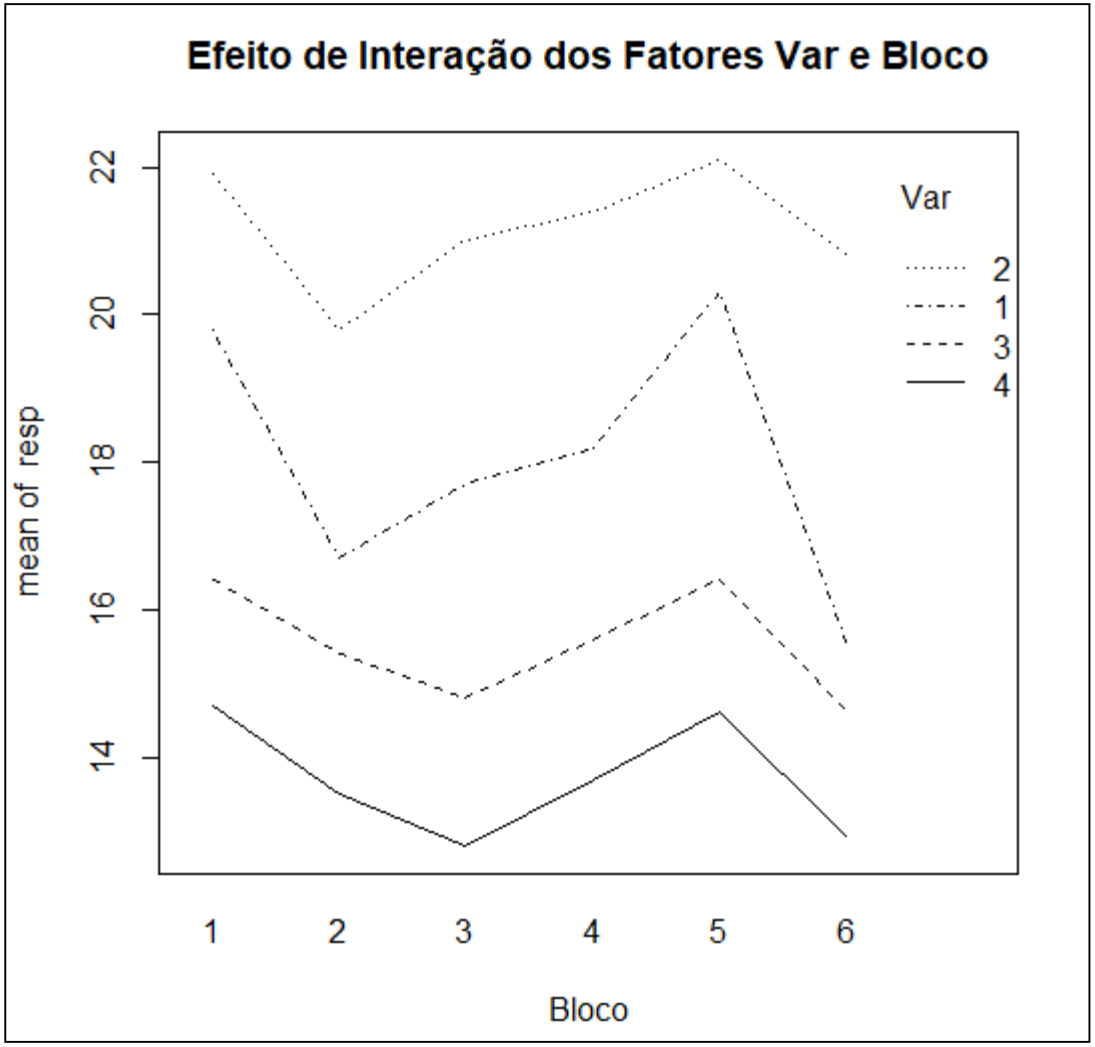
Level var	Difference of Means	SE of Difference	T-Value	Adjusted P-Value
3	-5,633	0,3808	-14,79	0,0000
4	-7,467	0,3808	-19,61	0,0000

Conclusão?

var = 3 subtracted from:

Level var	Difference of Means	SE of Difference	T-Value	Adjusted P-Value
4	-1,833	0,3808	-4,814	0,0012

Modelo ANOVA - DABC



No Delineamento Aleatorizado em Blocos Completos (DABC) o resíduo é o Efeito de Interação entre os fatores Tratamento e Bloco.

Pelo gráfico de perfis individuais há indicação de efeito de interação?

Note que o paralelismo dos perfis é uma indicação de ausência de interação!