Introdução a Métodos Estatísticos para a Bioinformática

Profa. Júlia Maria Pavan Soler

pavan@ime.usp.br

IBI 5086 – Bioinformática - IME/USP

2º Sem/2023

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, observações independentes ou pareadas

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.

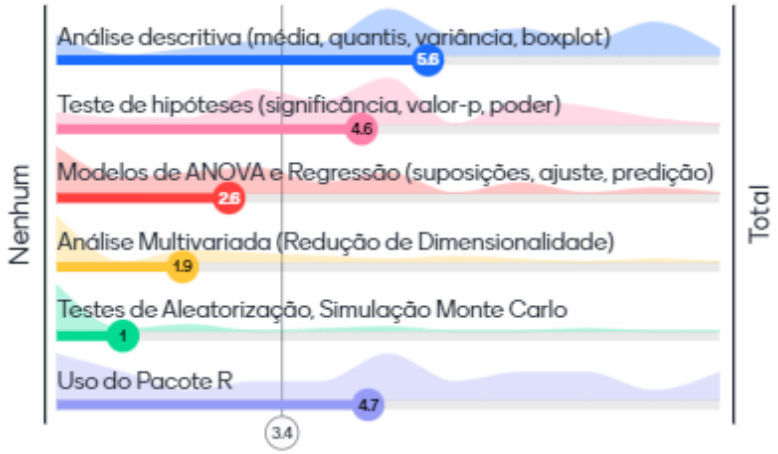
2. Análise Multivariada de Dados: Componentes Principais, Análise Discriminante e Classificação, Correlação Canônica, modelos MANOVA

3. Simulação de Monte Carlo, Intervalos de Confiança Bootstrap

Acesse mentl.com e use o código 1363 7089

Mentimeter

Pontue o seu conhecimento (domínio) sobre os seguintes conteúdos relacionados à Estatística:

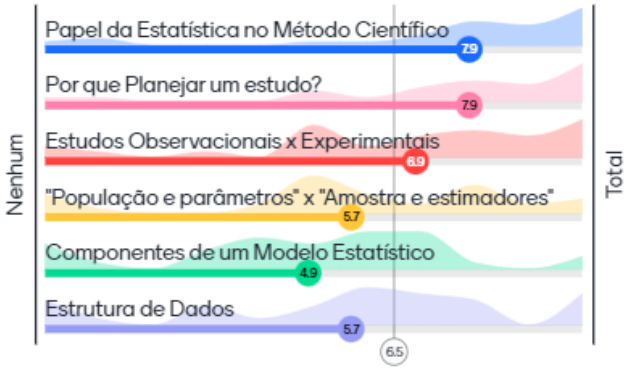


30

Acesse menti.com e use o código 1710 9522

Mentimeter

Pontue o seu entendimento sobre os seguintes conteúdos vistos na Aula100822:

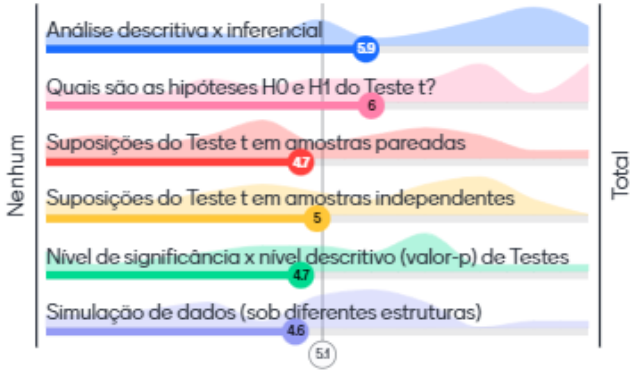


32

Acesse menti.com e use o código 5211 2482

Mentimeter

Pontue o seu entendimento sobre os seguintes conteúdos vistos na Aula170823:

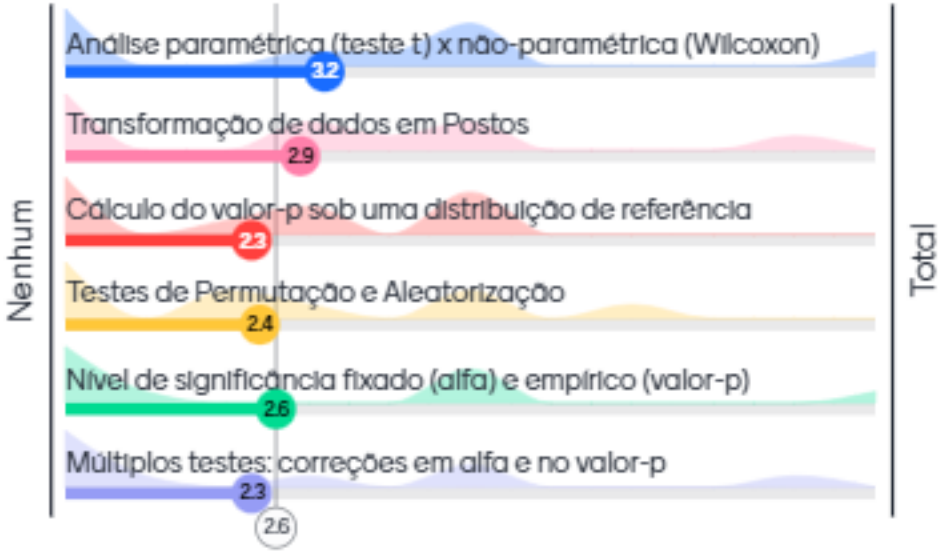


20

Acesse menti.com e use o código 1485 2289

Mentimeter

Pontue o seu entendimento sobre os seguintes conteúdos vistos na Aula240823:



10

```
> dados
   trat resp
1     A 81.4
2     A 84.5
3     A 84.8
4     A 87.3
5     A 79.7
6     A 85.1
7     A 81.7
8     A 83.7
9     A 84.3
10    B 84.7
11    B 86.1
12    B 83.2
13    B 91.9
14    B 86.3
15    B 82.6
16    B 89.1
17    B 83.8
18    B 88.5
```

Dados hipotéticos: dados da expressão gênica avaliada em tecidos hepáticos submetidos aos tratamentos A e B.

Objetivo do Estudo: Há efeito diferencial dos tratamentos na expressão gênica avaliada?

Vamos avaliar a significância desse efeito por meio de diferentes construções de testes estatísticos!

Execute os commandos do R disponibilizados para essa Aula!
Vamos discutir os resultados!
Importante se apoiar no que foi proposto no Protocolo do Estudo, o qual deve ser submetido a registro antes da coleta dos dados!

```
> expre[1:10,1:10]
         FAM IID FID MID SEX AFF   DE eQTM_DE eQTM_DE.1 eQTM_DE.2
1   DCASES20   1   0   0   1   2  758     501      3628       569
2   DCASES42   1   0   0   1   2  467    1492     20376      1621
3   DCASES62   1   0   0   1   2 1609    5802     48390     10086
4   DCASES70   1   0   0   1   2 1005    1122      9057      2235
5   DCASES90   1   0   0   1   2 1063    4260     35422      6933
6  DCASES106   1   0   0   1   2 1420     494      3010      1293
7  DCASES122   1   0   0   1   2 2047     394      5632      3664
8  DCASES126   1   0   0   1   2  644    2709      9730      1900
9  DCASES156   1   0   0   1   2  951    2746     30531      7940
10 DCASES165   1   0   0   2   2 1767     280      7860      2039

> dim(expre)
[1]  1000 12010
```

Dados do Projeto TCGA, Simulação 1, Expressão gênica.

Objetivo: Encontrar os “genes” com expressão diferencial entre as amostras Caso e Controle.

Vamos discutir como visualizar os resultados das anáilises e propor diferentes possibilidades de correção para os múltiplos testes estatísticos envolvidos!

Comparação de 2 ou mais Grupos

Planejamento de Experimentos e Modelos ANOVA (Análise de Variância)

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

Variável resposta quantitativa | Fatores (preditores) | Erro aleatório

Veremos os seguintes Planejamentos:

- Estrutura dos Fatores (Tratamentos – variável X):
  - ✓ 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):

- Delineamentos com Fatores Aleatórios

Delineamento Completamente Aleatorizado - DCA

T_1	T_2	. . .	T_a
Y_{11}	Y_{12}	. . .	Y_{1a}
Y_{21}	Y_{22}	. . .	Y_{2a}
. . .	. . .	Y_{ij}	. . .
$Y_{n_1 1}$	$Y_{n_2 2}$	. . .	$Y_{n_a a}$

Tratamentos: 1 Fator em *a* níveis
Fator de Efeito Fixo

Esquema de aleatorização:
atribuição completamente aleatória das unidades experimentais aos ***a*** tratamentos

Y_{ij}: resposta da i-ésima unidade experimental exposta ao j-ésimo tratamento

n_j: nj réplicas no tratamento j

$$\sum_{j=1}^{a} n_j = n$$

Dados balanceados ou desbalanceados

Motivação

Considere o seguinte delineamento completamente aleatorizado (DCA) com um fator fixo em 4 níveis e seis réplicas por tratamento.

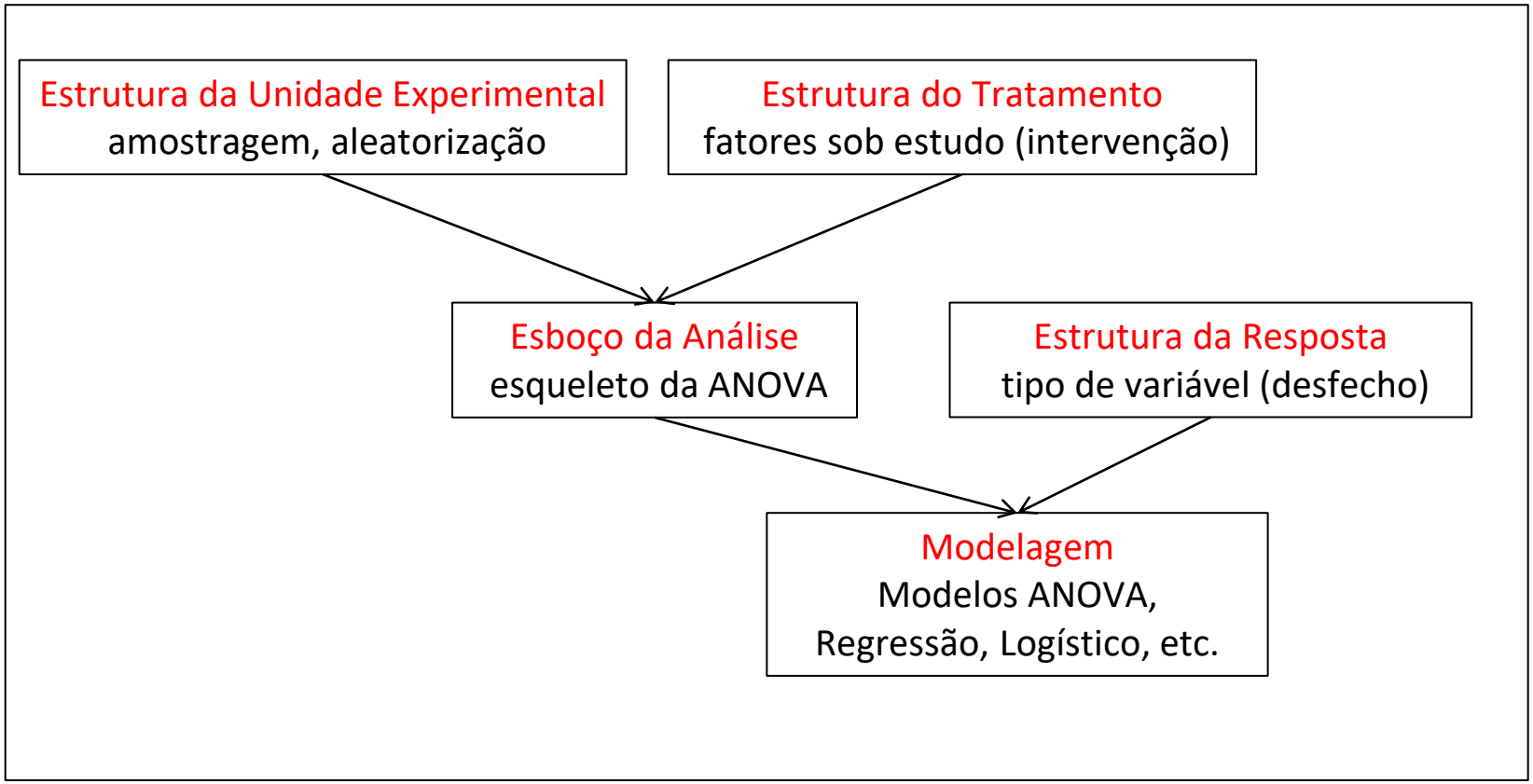
Dados: Avaliação de uma Resposta sob 4 tratamentos

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

Discuta o delineamento experimental e a estrutura dos dados.
Há evidência amostral para efeito de tratamento?

Planejamento e Análise de Dados

Estrutura Geral de Análise de Dados: (Goos and Gilmour, 2012)



Delineamento Completamente Aleatorizado - DCA

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

- ✓ **Estrutura das Unidades Experimentais**
  - ⇒ 24 unidades amostrais completamente aleatorizadas a 4 tratamentos
  - ⇒ 6 réplicas em cada tratamento (amostras balanceadas)
- ✓ **Estrutura de Tratamentos**
  - ⇒ 1 Fator (Tratamento) em 4 níveis
  - ⇒ Fator Fixo: T1, T2, T3 e T4
- ✓ **Estrutura da variável resposta**
  - ⇒ Uma única variável quantitativa de interesse

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

Estatísticas Descritivas

Variabilidade ENTRE Médias

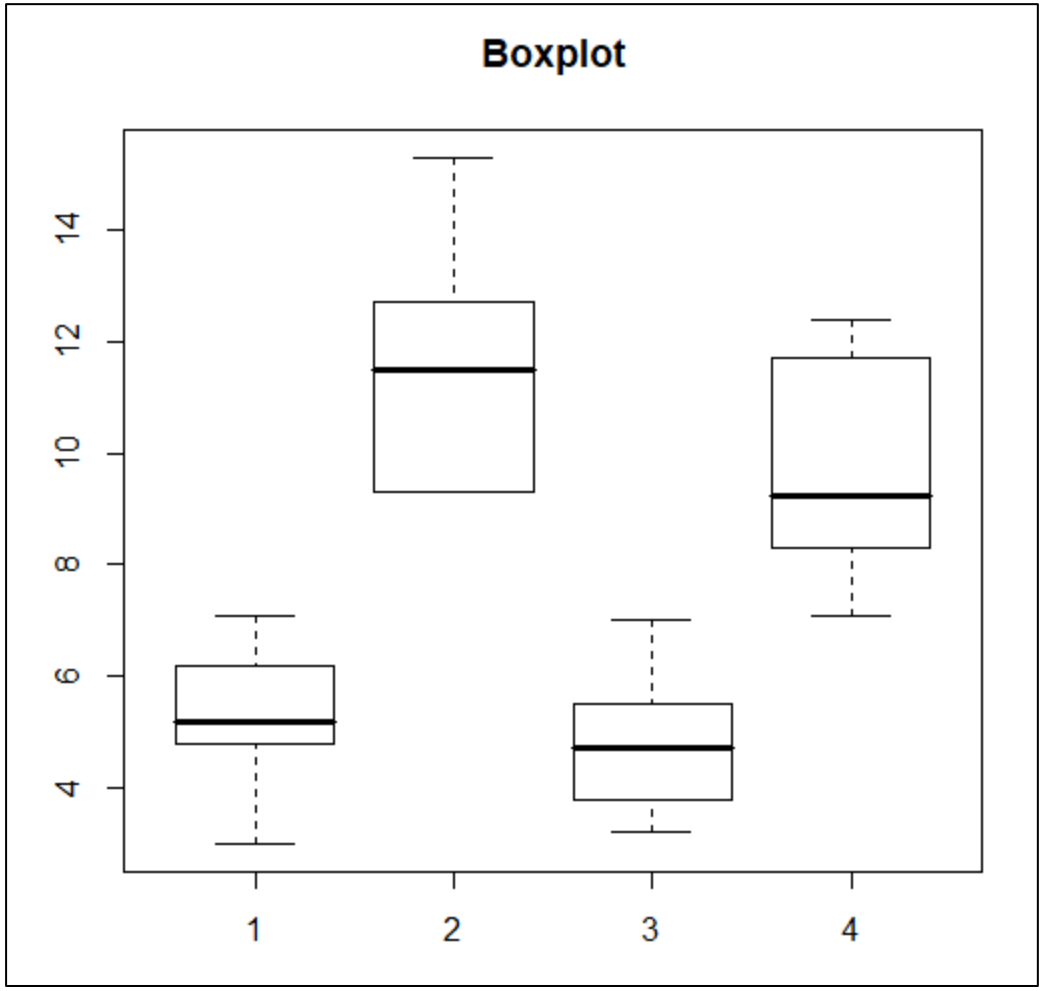
Variabilidade DENTRO dos Tratamentos

Variable	N	Mean	Median	TrMean	StDev	SE Mean
T1	6	5,250	5,200	5,250	1,408	0,575
T2	6	11,633	11,500	11,633	2,222	0,907
T3	6	4,817	4,700	4,817	1,348	0,550
T4	6	9,667	9,250	9,667	2,075	0,847

Variable	Minimum	Maximum	Q1	Q3
T1	3,000	7,100	4,350	6,425
T2	9,300	15,300	9,450	13,350
T3	3,200	7,000	3,650	5,875
T4	7,100	12,400	8,000	11,875

Há evidência amostral para a existência de efeito do tratamento?

Comparar a fonte de variação ENTRE tratamentos com a fonte de variação DENTRO de tratamentos (Análise de Variâncias – ANOVA)



n=6 é o tamanho amostral por grupo: Não é recomendável construir o boxplot com tão poucos pontos (em geral, para tamanhos amostrais maiores que 10)
Alternativa: Construir Dotplots ou Gráficos de Média (±sd ou ±se) !

Modelos ANOVA (Análise de Variância)

- Considere a seguinte identidade:

$$y_{ij} = \bar{y} + \left(\bar{y}_j - \bar{y}\right) + \left(y_{ij} - \bar{y}_j\right)$$

Variável resposta | **Média geral** | **Efeito de tratamento** | **Erro**

y_{ij}: resposta da unidade experimental *i* submetida ao *j*-ésimo tratamento

$$\sum_{ij}\left(y_{ij} - \bar{y}\right)^2 = \sum_{j} n_j\left(\bar{y}_j - \bar{y}\right)^2 + \sum_{ij}\left(y_{ij} - \bar{y}_j\right)^2$$

Soma de Quadrados Total | Soma de Quadrados de Tratamentos | Soma de Quadrados do Resíduo

Tabela de ANOVA
DCA – Um único Fator

Fonte de Variação	Número de graus de liberdade	Soma de Quadrados	Quadrado Médio	Estat. F	Valor-p
Tratamento (Entre)	a-1	$\sum_j n_j \left(\bar{y}_j - \bar{y}\right)^2$	SQTrat/(a-1)	$\frac{\text{QMTrat}}{\text{QMRes}}$	p
Resíduo (Dentro)	n-a	$\sum_{ij} \left(y_{ij} - \bar{y}_j\right)^2$	SQRes/(n-a)		
Total	n-1	$\sum_{ij} \left(y_{ij} - \bar{y}\right)^2$			

Teste F da ANOVA: teste global

$$F = \frac{QMTrat}{QM\,\text{Re}\,s} \sim F_{(a-1),(n-a)}$$

QMResidual estima a variância de Y (σ^2)

$$s_c^2 = \frac{\left(n_1 - 1\right)s_1^2 + \ldots + \left(n_a - 1\right)s_a^2}{n - a}$$

Tabela de ANOVA

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

Analysis of Variance for Resp

Source	DF	SS	MS	F	P
Trat	3	201,45	67,15	20,59	0,000
Error	20	65,23	3,26		
Total	23	266,68			

Level	N	Mean	StDev
T1	6	5,250	1,408
T2	6	11,633	2,222
T3	6	4,817	1,348
T4	6	9,667	2,075

Pooled StDev = 1,806

Hipóteses: $H_0 : \mu_j = \mu,\ j = 1,...,a;\quad H_1$: Existe pelo menos uma diferença entre as médias

Suposições: Normalidade, Independência e Homocedasticidade

Realizar análise de resíduos!

Conclusão: F=20,59 (p=0,000) Há evidência para a rejeição de H0

$\Rightarrow$ Existe pelo menos uma diferença entre as médias

Modelo ANOVA
Análise de Resíduos

Modelo Estrutural: diferentes parametrizações podem ser adotadas

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

Modelo de Médias

$$= \mu + \tau_j + e_{ij}; \quad \sum_j \tau_j = 0$$

Modelo de Desvios de Média (SAS)

$$= \begin{cases} \mu_1 + e_{i1} \\ \mu_1 + \tau_j + e_{ij}; j = 2,...,a \end{cases}$$

Modelo de Casela de Referência (R)

Modelo Distribucional: $y_{ij} = \mu_j + e_{ij}; \quad e_{ij} \overset{iid}{\sim} N\left(0;\sigma^2\right)$

$$y_{ij} \overset{ind}{\sim} N\left(\mu_j;\sigma^2\right); \quad E\left(y_{ij}\right) = \mu_j; \quad Var\left(y_{ij}\right) = Var\left(e_{ij}\right) = \sigma^2$$

Suposições:
Independência,
normalidade,
homocedasticidade

Modelo ANOVA
Análise de Resíduos

Checar as suposições do modelo:

- ✓ Normalidade
- ✓ Variância constante (homocedasticidade)
- ✓ Independência

Garantir que a estatística "F" da tabela de ANOVA tem distribuição F(a-1,n-a).

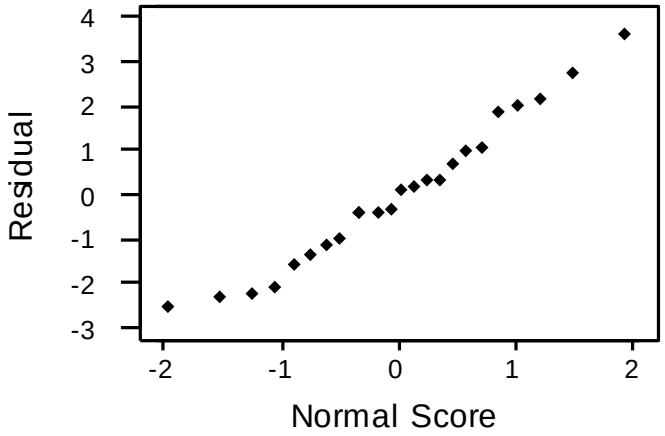
Resultado importante: "Aleatorização" dos Tratamentos às unidades experimentais pode garantir isso!

$$y_{ij} = \mu_j + e_{ij}; \qquad \Rightarrow \quad \hat{e}_{ij} = y_{ij} - \hat{\mu}_j = y_{ij} - \bar{Y}_j$$

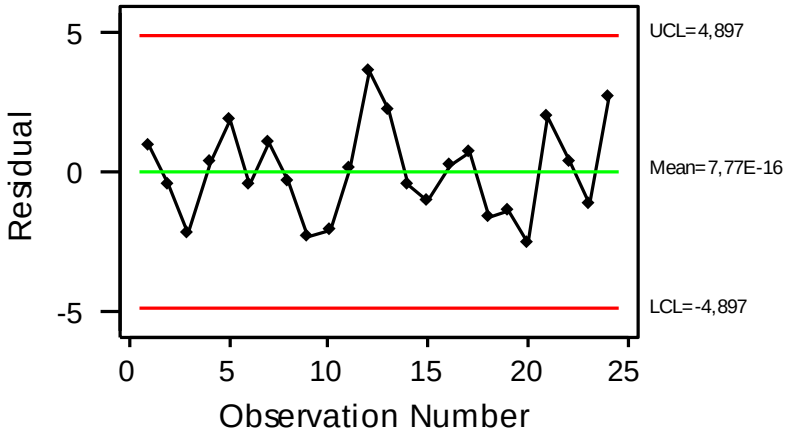
Resíduo do modelo de ANOVA ajustado
Com base nestes "n" valores devemos realizar a análise de diagnóstico das suposições adotadas

Residual Model Diagnostics

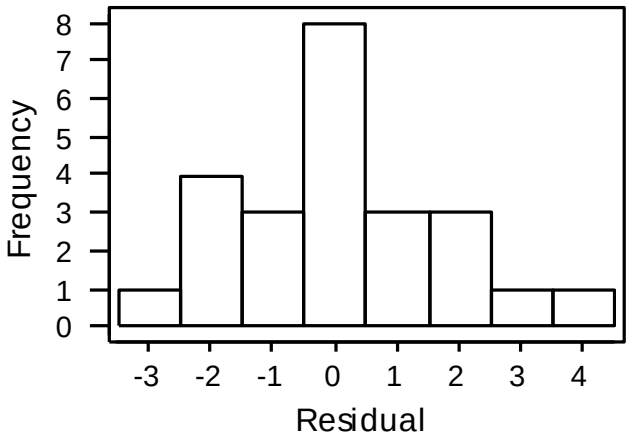
Normal Plot of Residuals



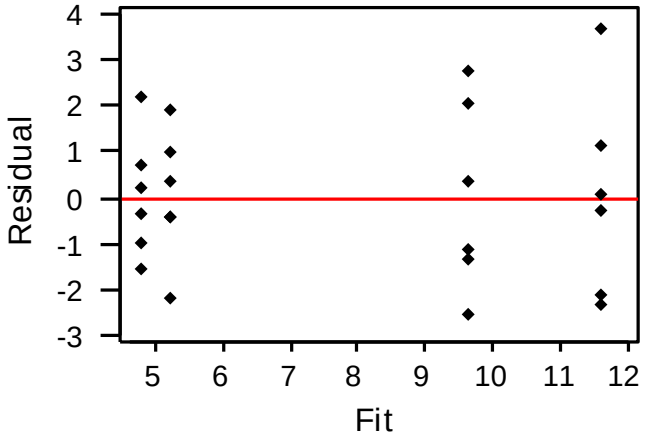
I Chart of Residuals



Histogram of Residuals



Residuals vs. Fits



E quando os dados não satisfazem as suposições impostas pelo modelo? Quais são as medidas remédio?

ANOVA – Análises de Diagnóstico

- Teste para a verificação da Normalidade

  Teste de Shapiro-Wilk, Teste de Kolmogorov-Smirnov

- Testes para a verificação da homocedasticidade (variâncias homogêneas)

  Teste de Bartlet (sensível a desvios da Normalidade)

  Teste de Levene (robusto)

Quando as suposições não estão satisfeitas:

- Transformações dos dados
- Soluções Não-Paramétricas, Testes de Aleatorização
- Adotar modelos sob premissas apropriadas (incorporar covariáveis, normalizar os dados antes da análise, uso de modelos lineares generalizados)

Hipótese de Interesse

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

H_0: $\mu_j = \mu \Leftrightarrow \tau_j = 0; \quad j = 1, ..., a$

H_1: existe pelo menos uma diferença entre as médias

Existe evidência de diferenças entre as médias?

A variação ENTRE as médias dos tratamentos é maior que a variação DENTRO dos tratamentos?

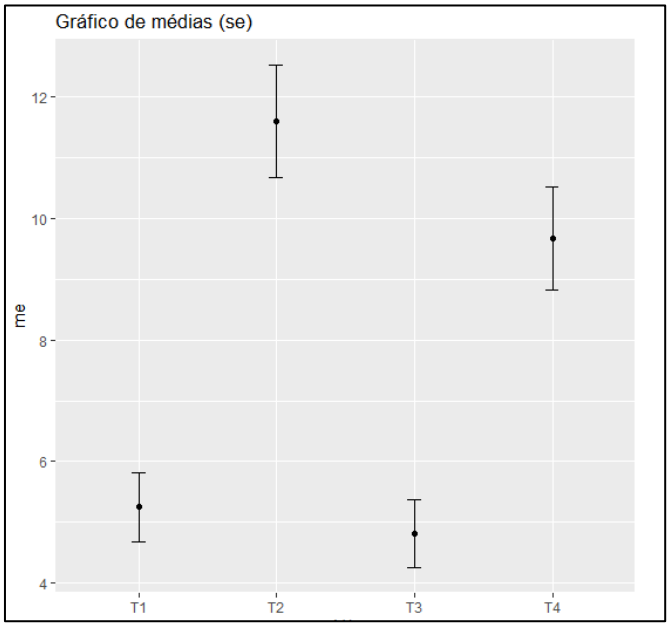
No caso da rejeição de H_0 ($p \leq \alpha$) e da análise de diagnóstico indicar que as premissas estão satisfeitas, o próximo passo da análise é a realização de "comparações múltiplas" entre as médias.

Comparações Múltiplas

Considerando os dados do exemplo:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu \quad \Leftrightarrow \quad \tau_1 = ... = \tau_4 = 0 \\ H_1 : \exists \text{ pelo menos uma diferença entre as médias} \end{cases}$$

Conclusão: F=20,59 p=0,000 $\Rightarrow$ há evidência amostral para a rejeição de H_0 $\Rightarrow$ há pelo menos uma diferença entre as médias



Realizar testes para a comparações específicas entre os grupos.

Discuta as limitações envolvidas na realização de testes “t” individuais para todas as possíveis comparações entre pares de médias!

$\Rightarrow$ Realizar Testes de Comparações múltiplas (com correção para a multiplicidade de testes)

Comparações Múltiplas

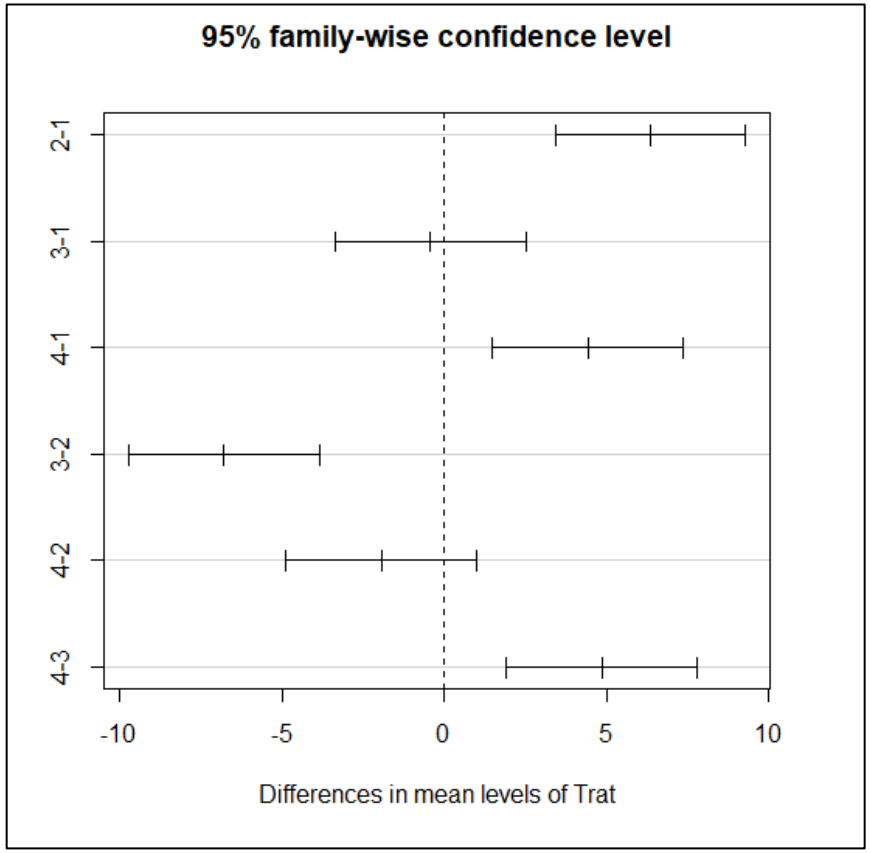
Intervalos de Confiança Simultâneos de Tukey a 95%

	1	2	3
2	-9.303 -3.464		
3	-2.486 3.353	3.897 9.736	
4	-7.336 -1.497	-0.953 4.886	-7.770 -1.930

$\mu_1 < \mu_2$

$\mu_1 = \mu_3$ $\quad \mu_2 > \mu_3$

$\mu_1 < \mu_4$ $\quad \mu_2 = \mu_4$ $\quad \mu_3 < \mu_4$



Conclusão? $(\mu_2 = \mu_4)^* > (\mu_1 = \mu_3)$ Diferenças significantes a um nível de significância global igual a 5%

Diferentes apresentações dos ICS:

```
Tukey 95,0% Simultaneous Confidence Intervals
Trat = 1  subtracted from:
Trat   Lower   Center  Upper  ------+--------+--------+--------+
2       3,464   6,3833  9,303                    (----*----)
3      -3,353  -0,4333  2,486          (----*----)
4       1,497   4,4167  7,336                (----*----)
                              ------+--------+--------+--------+
                                 -6,0       0,0       6,0      12,0
Trat = 2  subtracted from:
Trat   Lower  Center   Upper  ------+--------+--------+--------+
3      -9,736  -6,817  -3,897  (----*----)
4      -4,886  -1,967   0,953         (----*----)
                              ------+--------+--------+--------+
                                 -6,0       0,0       6,0      12,0
Trat = 3  subtracted from:
Trat  Lower  Center  Upper  ------+--------+--------+--------+
4      1,930   4,850  7,770                  (----*----)
                            ------+--------+--------+--------+
                               -6,0       0,0       6,0      12,0
```

Dados do Exemplo

```
Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Resp ~ Trat, data = dados)

$Trat
           diff       lwr       upr     p adj
2-1   6.3500000  3.411858  9.288142 0.0000360
3-1  -0.4333333 -3.371476  2.504809 0.9756358
4-1   4.4166667  1.478524  7.354809 0.0022547
3-2  -6.7833333 -9.721476 -3.845191 0.0000149
4-2  -1.9333333 -4.871476  1.004809 0.2840071
4-3   4.8500000  1.911858  7.788142 0.0008789
```

A estatística é avaliada na distribuição do range estudentizado

Conclusão? $(\mu_2 = \mu_4)^* > (\mu_1 = \mu_3)$ Diferenças significantes a um nível de significância global igual a 5%

Comparações Múltiplas

Correções para Múltiplos Testes

Valores-p dos Testes "t" bicaudais corrigidos por Bolferroni, Holm e FDR

	pt.valor	adjustb	adjusth	adjustfdr	
T1-T2	8.792659e-05	0.0005275596	0.0005275596	0.0004928938	$\mu_1 \neq \mu_2$
T1-T3	1.642979e-04	0.0009857876	0.0008214896	0.0004928938	$\mu_1 \neq \mu_3$
T1-T4	7.230700e-04	0.0043384199	0.0028922799	0.0014461400	$\mu_1 \neq \mu_4$
T2-T3	1.528013e-03	0.0091680765	0.0045840383	0.0022920191	$\mu_2 \neq \mu_3$
T2-T4	1.539205e-01	0.9235227555	0.3078409185	0.1847045511	$\mu_2 = \mu_4$
T3-T4	5.980538e-01	1.0000000000	0.5980537708	0.5980537708	$\mu_3 = \mu_4$

As 4 primeiras diferenças entre médias são significantes (para um nível de significância global igual a α=0.05, α=0.01, α=0.003)

Comparações Múltiplas

- **Método de Tukey**
- **Método de Dunnet**
- **Método de Scheffé**
- **Método de Bonferroni**
- **Método FDR**
- **Etc.**