# 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

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):

- Delineamento com uma Fator Aleatório

# 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

resposta da i-ésima unidade experimental exposta ao j-ésimo tratamento

$n_j$

nj réplicas no tratamento j

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

# Motivação

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

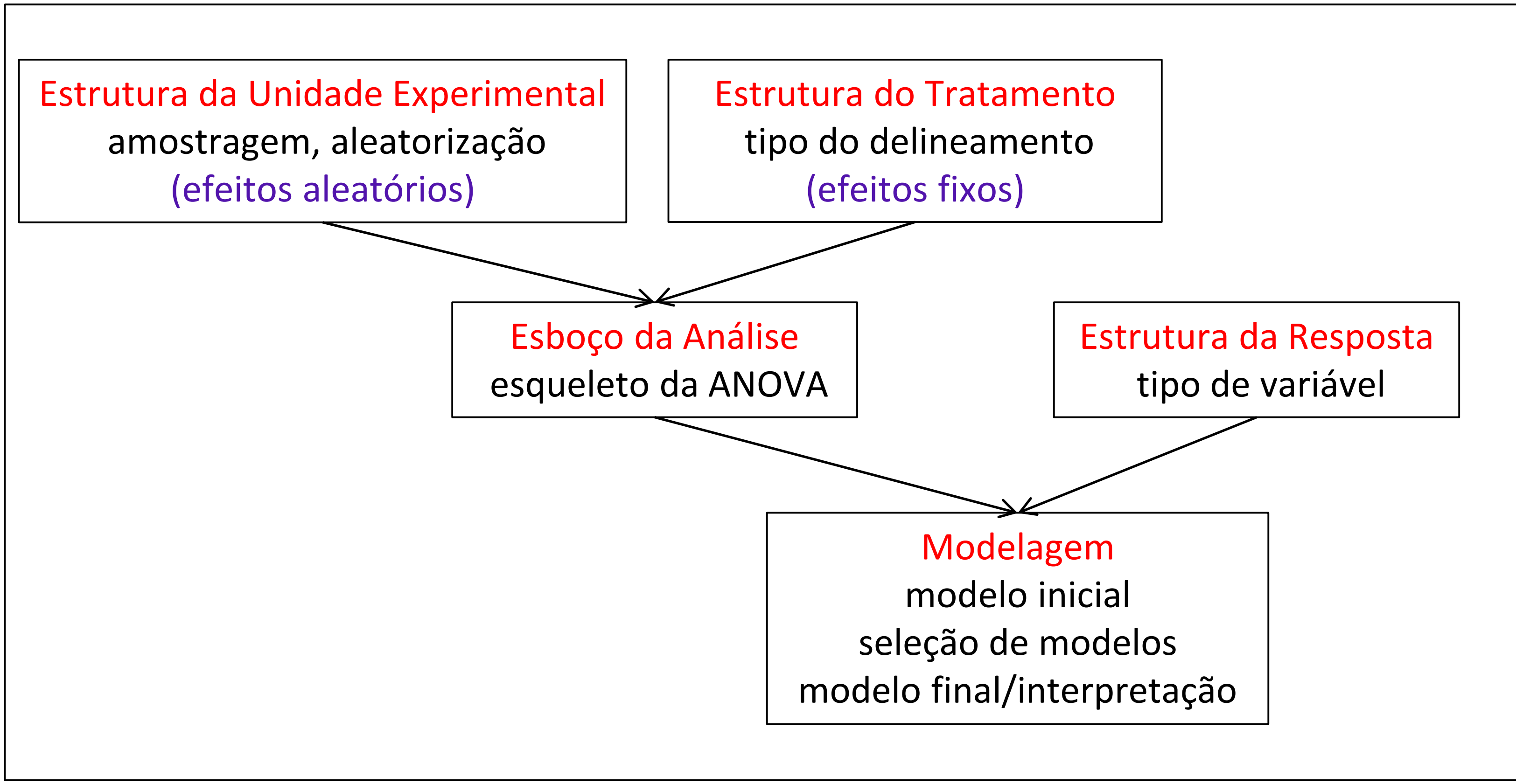
**Dados: Avaliação de uma Resposta sob 4 tratamentos**

| 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

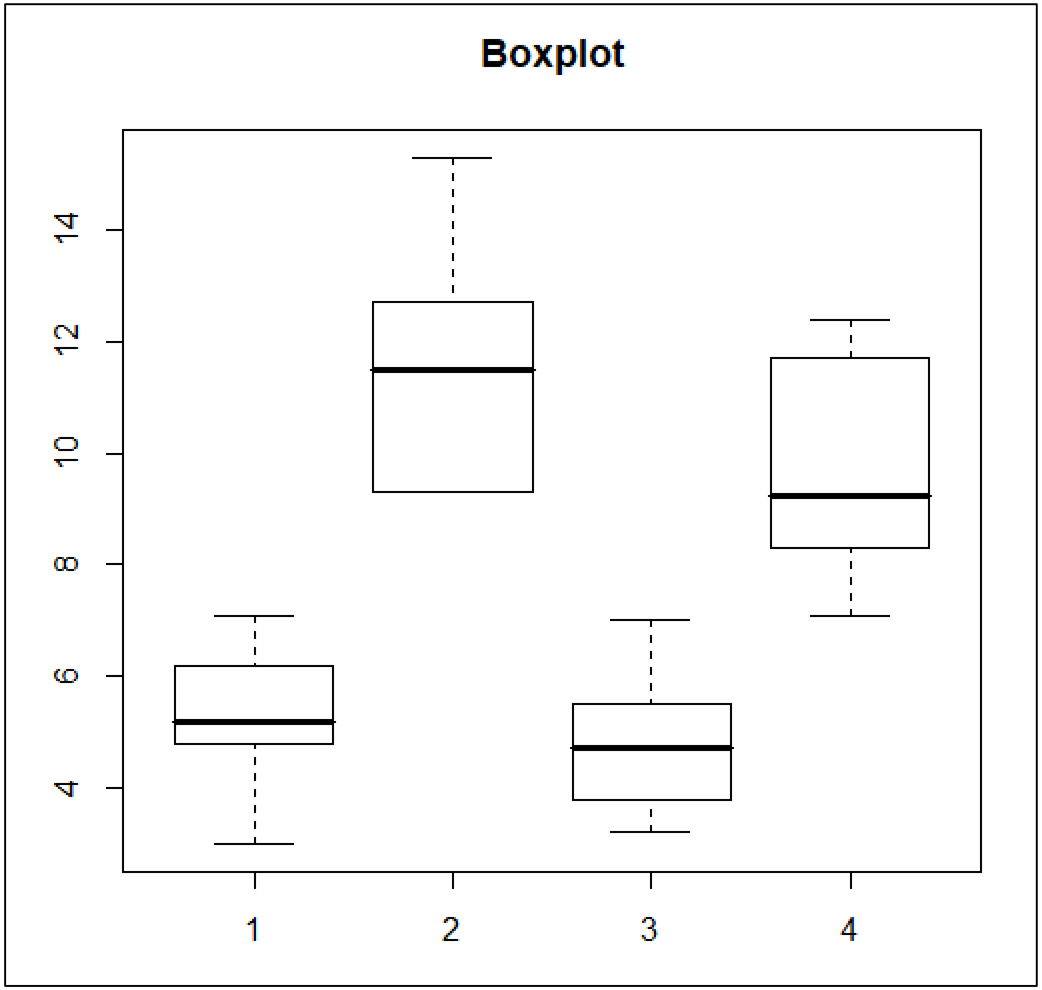
Variabilidade DENTRO dos Tratamentos

| 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 |

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

Há evidência amostral para a existência de efeito do tratamento? Discutir.




*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!*

# Modelos ANOVA (Análise de Variância)

- Considere a seguinte identidade:

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

Variável resposta | Média geral | Efeito de tratamento | Erro

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

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

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 | $\sum_{ij} \left(y_{ij} - \bar{y}\right)^2$ | | | |

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

QMResidual estima a variância de Y

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

# 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: $H_0 : \mu_j = \mu,\ j = 1,\ldots,a; \quad H_1 :$ 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:**

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

**Modelo de Médias**

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

**Modelo de Desvios de Média**

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

**Modelo de Casela de Referência**

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

$$E\left(y_{ij}\right) = \mu_j; \qquad Var\left(y_{ij}\right) = Var\left(e_{ij}\right) = \sigma^2$$

# 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" pode garantir isso!

$$y_{ij} = \mu_j + e_{ij}; \qquad \Rightarrow \quad \underbrace{Y_{n\times 1}}_{\text{Vetor de resposta}} = \underbrace{X_{n\times a}}_{\text{Matriz de Panejamento}} \; \underbrace{\beta_{a\times 1}}_{\text{Vetor de parâmteros}} + \underbrace{e_{n\times 1}}_{\text{Vetor de resíduos}}$$

# ANOVA – Análises de Diagnóstico

$$y_{ij} = \underbrace{\bar{y}_j}_{\hat{y}_{ij}} + \underbrace{\boxed{\left(y_{ij} - \bar{y}_j\right)}}_{\hat{e}_{ij}}$$

y ajustado resíduo

- Análises Descritivas dos Resíduos

$\Rightarrow$ Histogramas e box-plots dos resíduos

$\Rightarrow$ Quantis dos resíduos contra quantis da normal

} Distribuição simétrica? Normalidade?

$\Rightarrow$ $\underbrace{\hat{e}_j \times \text{ordem(obs)}}$ erros independentes

$\underbrace{\hat{e}_j \times \hat{y}_j}$ homocedasticidade

tendências não modeladas

# ANOVA – Análises de Diagnóstico

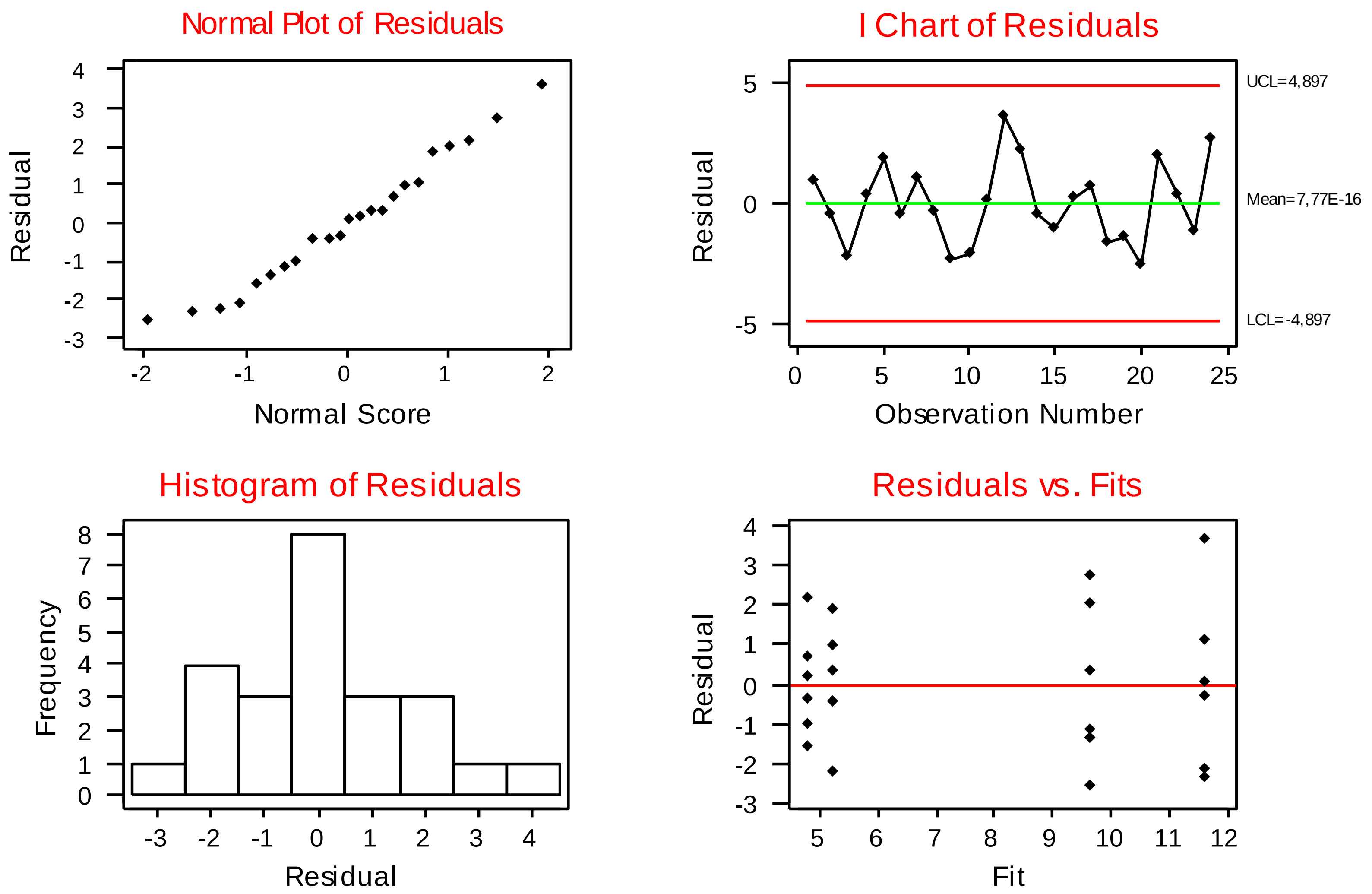
$$\hat{e}_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \overline{y}_j = y_{ij} - h_{ii} y_{ij} = (1 - h_{ii}) y_{ij}$$ Resíduo

$$h_{ii} = X_i (X'X)^{-1} X_i$$

$$\tilde{e}_{ij} = \frac{\hat{e}_{ij}}{\sqrt{QM \operatorname{Re} s}}$$ Resíduo semi-studentizado

$$r_{ij} = \frac{\hat{e}_{ij}}{\sqrt{QM \operatorname{Re} s (1 - h_{ii})}}$$ Resíduo studentizado

# Residual Model Diagnostics




**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)

# Hipótese de Interesse

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

$$\begin{cases} \mathbf{H_0:}\ \mu_j = \mu \Leftrightarrow \tau_j = 0; \quad j = 1, \ldots, a \\ \mathbf{H_1:}\ \text{existe pelo menos uma diferença} \end{cases}$$

**Existe evidência de diferenças entre as médias ?**

$$\Updownarrow$$

**As diferenças ENTRE médias são maiores que a variação DENTRO dos tratamentos ?**

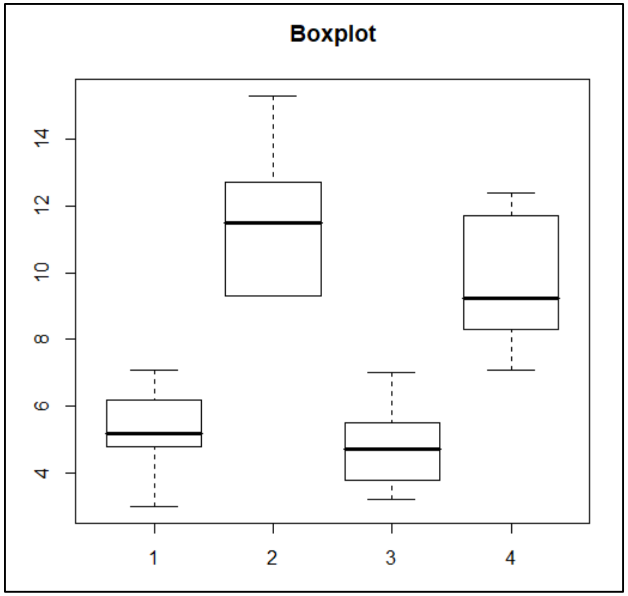
No caso da rejeição de $H_0$ ($p \leq \alpha$) e da análise de diagnóstico indicar que as premissas estão satisfeitas, o próximo passo da análise é a realização de "comparações múltiplas" entre as médias.

# Comparações Múltiplas

**Considerando os dados do exemplo:**

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

**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




Discuta as limitações envolvidas na realização de testes "t" individuais para todas as possíveis comparações entre duas médias!

# Comparações Múltiplas

- **Intervalo de Confiança para $\mu_j$** (j=1,...,a)

$$\bar{y}_j - t_{(\alpha/2;r-1)}\sqrt{\frac{QM\,\mathrm{Res}}{n_j}} \leq \mu_j \leq \bar{y}_j + t_{(\alpha/2;r-1)}\sqrt{\frac{QM\,\mathrm{Res}}{n_j}}$$

- **Intervalo de Confiança para a Diferença entre Médias**

$$\left(\bar{y}_j - \bar{y}_k\right) \pm t_{(\alpha/2;n_j+n_k-2)}\sqrt{QM\,\mathrm{Res}\left(\frac{1}{n_j}+\frac{1}{n_k}\right)}$$

**Qual é o nível de confiança global** (ou, equivalentemente, o nível de significância global) quando são realizadas muitos testes simultaneamente (muitas comparações entre pares de médias)?

# Tabela de ANOVA

**One-way ANOVA: Resp versus Trat**

```
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
                                   Individual 95% CIs For Mean
                                   Based on Pooled StDev
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                     6,0       9,0      12,0
```

**Defina H a hipótese a ser testada via a estatística F**

**Suposições envolvidas**

**Conclusão**

# Comparações Múltiplas

**Nível de Significância Único para a Família de Comparações**

**4 Tratamentos** $\mu_1 \quad \mu_2 \quad \mu_3 \quad \mu_4$

**Comparações de interesse:**

$$\begin{cases} H_{01}: \mu_1 = \mu_2 & \Rightarrow \alpha_1 \\ H_{02}: \mu_1 = \mu_3 & \Rightarrow \alpha_2 \\ H_{03}: \mu_1 = \mu_4 & \Rightarrow \alpha_3 \\ H_{04}: \mu_2 = \mu_3 & \Rightarrow \alpha_4 \\ H_{05}: \mu_2 = \mu_4 & \Rightarrow \alpha_5 \\ H_{06}: \mu_3 = \mu_4 & \Rightarrow \alpha_6 \end{cases}$$

$\alpha$?

**Para testar as 6 hipóteses, qual é o nível de significância global para esta família de comparações entre pares de medias?**

# Comparações Múltiplas

- **Comparações aos Pares (pares de médias sendo comparadas)**

$$H: \mu_j = \mu_k \quad \Leftrightarrow \quad \mu_j - \mu_k = 0$$

- **Contrastes entre Médias**

$$H: \ \mu_1 = \frac{\mu_2 + \mu_3}{2} \quad \Leftrightarrow \quad \mu_1 - \frac{\mu_2 + \mu_3}{2} = 0$$

- **Combinações Lineares de Médias**

$$H: \ 0,30\ \mu_1 + 0,20\ \mu_2 + 0,50\ \mu_3 = 0$$

# Comparações Múltiplas

- **Método de Tukey**

$$H: \mu_j = \mu_k \quad ; \quad j \neq k \qquad \Rightarrow \binom{a}{2} = \frac{a(a-1)}{2} \text{ comparações}$$

**Intervalo de Confiança Simultâneo de Tukey**

$$\left(\bar{Y}_j - \bar{Y}_k\right) \pm q_{a;\nu;\alpha/2}\; s_c \sqrt{\frac{1}{n}}$$

**n-a**

**Delineamento Balanceado**

$q_{a;\nu;\alpha/2}$ : range studentizado (valor tabelado)

Critério HSD (Honest Significant Difference)

$$\left(\bar{Y}_j - \bar{Y}_k\right) \pm \frac{q_{a;\nu;\alpha/2}}{\sqrt{2}}\; s_c \sqrt{\frac{1}{n_j} + \frac{1}{n_k}}$$

**Não Balanceado**

**Table D.8:** Percent points for the Studentized range.

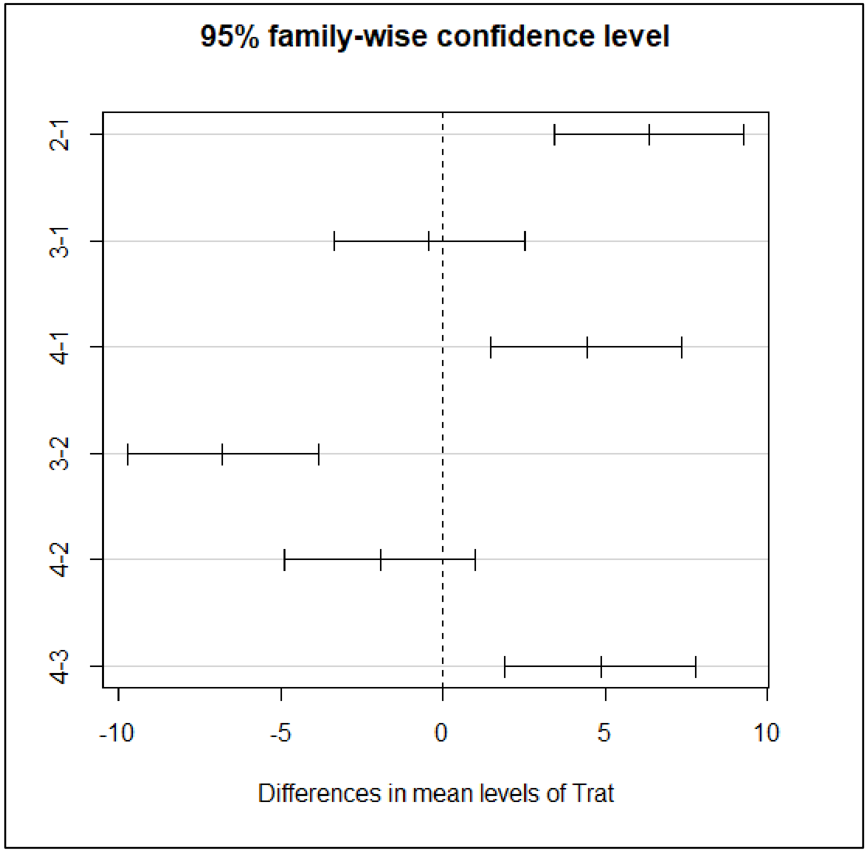
Table entries are $q_{.05}(K, \nu)$.

| $\nu$ | K 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 15 | 20 | 30 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 18.0 | 27.0 | 32.8 | 37.1 | 40.4 | 43.1 | 45.4 | 47.4 | 49.1 | 55.4 | 59.6 | 65.1 | 71.7 |
| 2 | 6.09 | 8.33 | 9.80 | 10.9 | 11.7 | 12.4 | 13.0 | 13.5 | 14.0 | 15.7 | 16.8 | 18.3 | 20.0 |
| 3 | 4.50 | 5.91 | 6.82 | 7.50 | 8.04 | 8.48 | 8.85 | 9.18 | 9.46 | 10.5 | 11.2 | 12.2 | 13.4 |
| 4 | 3.93 | 5.04 | 5.76 | 6.29 | 6.71 | 7.05 | 7.35 | 7.60 | 7.83 | 8.66 | 9.23 | 10.0 | 10.9 |
| 5 | 3.64 | 4.60 | 5.22 | 5.67 | 6.03 | 6.33 | 6.58 | 6.80 | 6.99 | 7.72 | 8.21 | 8.87 | 9.67 |
| 6 | 3.46 | 4.34 | 4.90 | 5.30 | 5.63 | 5.90 | 6.12 | 6.32 | 6.49 | 7.14 | 7.59 | 8.19 | 8.91 |
| 7 | 3.34 | 4.16 | 4.68 | 5.06 | 5.36 | 5.61 | 5.82 | 6.00 | 6.16 | 6.76 | 7.17 | 7.73 | 8.40 |
| 8 | 3.26 | 4.04 | 4.53 | 4.89 | 5.17 | 5.40 | 5.60 | 5.77 | 5.92 | 6.48 | 6.87 | 7.40 | 8.03 |
| 9 | 3.20 | 3.95 | 4.41 | 4.76 | 5.02 | 5.24 | 5.43 | 5.59 | 5.74 | 6.28 | 6.64 | 7.14 | 7.75 |
| 10 | 3.15 | 3.88 | 4.33 | 4.65 | 4.91 | 5.12 | 5.30 | 5.46 | 5.60 | 6.11 | 6.47 | 6.95 | 7.53 |
| 11 | 3.11 | 3.82 | 4.26 | 4.57 | 4.82 | 5.03 | 5.20 | 5.35 | 5.49 | 5.98 | 6.33 | 6.79 | 7.35 |
| 12 | 3.08 | 3.77 | 4.20 | 4.51 | 4.75 | 4.95 | 5.12 | 5.27 | 5.39 | 5.88 | 6.21 | 6.66 | 7.21 |
| 13 | 3.06 | 3.73 | 4.15 | 4.45 | 4.69 | 4.88 | 5.05 | 5.19 | 5.32 | 5.79 | 6.11 | 6.55 | 7.08 |
| 14 | 3.03 | 3.70 | 4.11 | 4.41 | 4.64 | 4.83 | 4.99 | 5.13 | 5.25 | 5.71 | 6.03 | 6.46 | 6.98 |
| 15 | 3.01 | 3.67 | 4.08 | 4.37 | 4.59 | 4.78 | 4.94 | 5.08 | 5.20 | 5.65 | 5.96 | 6.38 | 6.89 |
| 16 | 3.00 | 3.65 | 4.05 | 4.33 | 4.56 | 4.74 | 4.90 | 5.03 | 5.15 | 5.59 | 5.90 | 6.31 | 6.81 |
| 17 | 2.98 | 3.63 | 4.02 | 4.30 | 4.52 | 4.70 | 4.86 | 4.99 | 5.11 | 5.54 | 5.84 | 6.25 | 6.74 |
| 18 | 2.97 | 3.61 | 4.00 | 4.28 | 4.49 | 4.67 | 4.82 | 4.96 | 5.07 | 5.50 | 5.79 | 6.20 | 6.68 |
| 19 | 2.96 | 3.59 | 3.98 | 4.25 | 4.47 | 4.65 | 4.79 | 4.92 | 5.04 | 5.46 | 5.75 | 6.15 | 6.63 |
| 20 | 2.95 | 3.58 | 3.96 | 4.23 | 4.45 | 4.62 | 4.77 | 4.90 | 5.01 | 5.43 | 5.71 | 6.10 | 6.58 |
| 21 | 2.94 | 3.56 | 3.94 | 4.21 | 4.42 | 4.60 | 4.74 | 4.87 | 4.98 | 5.40 | 5.68 | 6.07 | 6.53 |
| 22 | 2.93 | 3.55 | 3.93 | 4.20 | 4.41 | 4.58 | 4.72 | 4.85 | 4.96 | 5.37 | 5.65 | 6.03 | 6.49 |
| 23 | 2.93 | 3.54 | 3.91 | 4.18 | 4.39 | 4.56 | 4.70 | 4.83 | 4.94 | 5.34 | 5.62 | 6.00 | 6.45 |
| 24 | 2.92 | 3.53 | 3.90 | 4.17 | 4.37 | 4.54 | 4.68 | 4.81 | 4.92 | 5.32 | 5.59 | 5.97 | 6.42 |
| 25 | 2.91 | 3.52 | 3.89 | 4.15 | 4.36 | 4.53 | 4.67 | 4.79 | 4.90 | 5.30 | 5.57 | 5.94 | 6.39 |
| 26 | 2.91 | 3.51 | 3.88 | 4.14 | 4.35 | 4.51 | 4.65 | 4.77 | 4.88 | 5.28 | 5.55 | 5.92 | 6.36 |
| 27 | 2.90 | 3.51 | 3.87 | 4.13 | 4.33 | 4.50 | 4.64 | 4.76 | 4.86 | 5.26 | 5.53 | 5.89 | 6.34 |
| 28 | 2.90 | 3.50 | 3.86 | 4.12 | 4.32 | 4.49 | 4.62 | 4.74 | 4.85 | 5.24 | 5.51 | 5.87 | 6.31 |
| 29 | 2.89 | 3.49 | 3.85 | 4.11 | 4.31 | 4.47 | 4.61 | 4.73 | 4.84 | 5.23 | 5.49 | 5.85 | 6.29 |
| 30 | 2.89 | 3.49 | 3.85 | 4.10 | 4.30 | 4.46 | 4.60 | 4.72 | 4.82 | 5.21 | 5.47 | 5.83 | 6.27 |
| 35 | 2.87 | 3.46 | 3.81 | 4.07 | 4.26 | 4.42 | 4.56 | 4.67 | 4.77 | 5.15 | 5.41 | 5.76 | 6.18 |
| 40 | 2.86 | 3.44 | 3.79 | 4.04 | 4.23 | 4.39 | 4.52 | 4.63 | 4.73 | 5.11 | 5.36 | 5.70 | 6.11 |
| 45 | 2.85 | 3.43 | 3.77 | 4.02 | 4.21 | 4.36 | 4.49 | 4.61 | 4.70 | 5.07 | 5.32 | 5.66 | 6.06 |
| 50 | 2.84 | 3.42 | 3.76 | 4.00 | 4.19 | 4.34 | 4.47 | 4.58 | 4.68 | 5.04 | 5.29 | 5.62 | 6.02 |
| 100 | 2.81 | 3.36 | 3.70 | 3.93 | 4.11 | 4.26 | 4.38 | 4.48 | 4.58 | 4.92 | 5.15 | 5.46 | 5.83 |
| $\infty$ | 2.77 | 3.31 | 3.63 | 3.86 | 4.03 | 4.17 | 4.29 | 4.39 | 4.47 | 4.80 | 5.01 | 5.30 | 5.65 |

# Comparações Múltiplas

## Intervalos de Confiança Simultâneos de Tukey a 95%

| | 1 | 2 | 3 |
|---|---|---|---|
| 2 | -9.303<br>-3.464<br>$\mu_1 < \mu_2$ | | |
| 3 | -2.486<br>3.353<br>$\mu_1 = \mu_3$ | 3.897<br>9.736<br>$\mu_2 > \mu_3$ | |
| 4 | -7.336<br>-1.497<br>$\mu_1 < \mu_4$ | -0.953<br>4.886<br>$\mu_2 = \mu_4$ | -7.770<br>-1.930<br>$\mu_3 < \mu_4$ |




**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
```

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

# Comparações Múltiplas

- **Método de Contrastes Ortogonais**
- **Método de Dunnet**
- **Método de Scheffé**
- **Método de Bonferroni**
- **Método FDR**

# Comparações Múltiplas

- **Método de Bonferroni**

$$H : \sum_{j=1}^{a} c_j \, \mu_j = 0 \quad ; \quad \sum c_j = 0$$

$$\sum_{j=1}^{a} c_j \, \bar{Y}_j \; \pm \; B \sqrt{\hat{V}\left(\sum c_j \bar{Y}_j\right)}$$

**Intervalo de Confiança Simultâneo de Bonferroni (é mais conservador que o de Tukey)**

$$B = t_{(n-a);(1-\alpha/2K)}$$

**Correção de Bonferroni para "K" múltiplos testes**

# Comparações Múltiplas

## Correções para Múltiplos Testes

| Rejeitar $H_{0j}$ se: | Método |
|---|---|
| $p_{(k)} < \alpha / K$ | Correção de Bonferroni para múltiplos testes |
| $p_{(j)} < \alpha /(K - j + 1)$ $para \ \ todo \ \ j = 1, \ldots, k$ | Correção de Holm (Controle “forte” da taxa de erro para os múltiplos testes) |
| $p_{(j)} < j\alpha / K$ $para \ \ algum \ \ j \geq k$ | Correção FDR (Taxa de Falsa Descoberta): Benjamini-Hochberg Controle menos conservador da taxa de erro para os múltiplos testes) |

K: número total de testes $\alpha$: nível de significância global fixado

$p_{(j)}$: nível descritivo (p-valor) ordenado, $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(K)}$

```
     pt.valor     adjustb      adjusth    adjustfdr
[1,] 8.792659e-05 0.0005275596 0.0005275596 0.0004928938
[2,] 1.642979e-04 0.0009857876 0.0008214896 0.0004928938
[3,] 7.230700e-04 0.0043384199 0.0028922799 0.0014461400
[4,] 1.528013e-03 0.0091680765 0.0045840383 0.0022920191
[5,] 1.539205e-01 0.9235227555 0.3078409185 0.1847045511
[6,] 5.980538e-01 1.0000000000 0.5980537708 0.5980537708
```

# Comparações Múltiplas

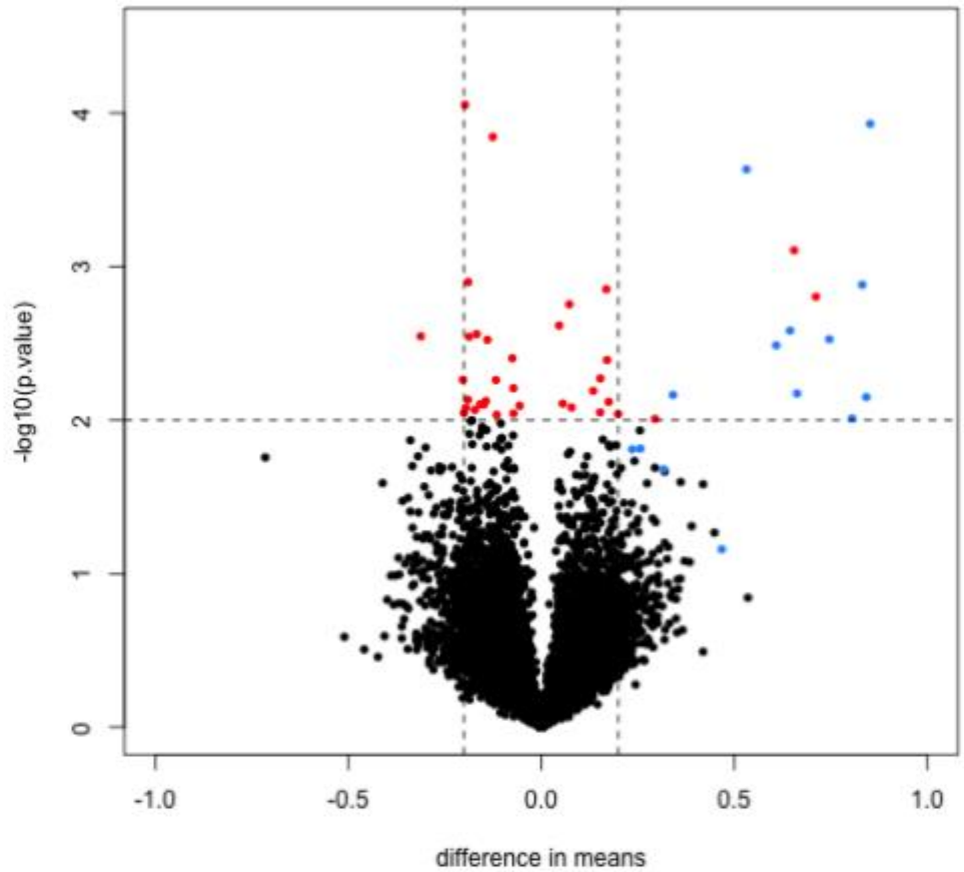
## Correções para Múltiplos Testes

**Valores-p dos Testes "t" bicaudal, Bolferroni, Holm e FDR (False Discover Rate)**

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

# Gráfico Vulcão

Distribuição dos valores-p de acordo com a expressão gênica diferencial entre dois tecidos.




**Volcano plot** for t-test comparing two groups. Spiked-in genes are denoted with blue. Among the rest of the genes, those with p-value < 0.01 are denoted with red.

Necessidade de adotar uma correção para os múltiplos testes.

Irizarry and Love, 2015

# Referência Bibliográfica

Box, G.E.; Hunter, W.G. and Hunter, J.S. (1978). *Statistics for Experimenters*. An Introduction to Designs, Data Analysis and Model Building. John Wiley & Sons.

→ Neter, J.K.; Kutner, M; Wasserman, W. and Nachtsheim, C.J. (2016). *Applied Linear Statistical Models*. 4th ed., Times Mirror Higher Education Group.

Oehlert, GW. *A First Course in Design and Analysis of Experiments*. Univ. of Minnesota, Licensed by Creative Commons, 2010.