

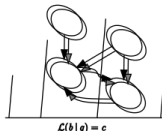
# Estatística Univariada

## Portal Metabolômica Brasil

R. R. da Silva<sup>1</sup>

<sup>1</sup>Departamento de Ciências BioMoleculares  
Faculdade de Ciências Farmacêuticas

23 de novembro de 2023



# Sumário

- 1 Testes de hipóteses
- 2 Comparação entre 2 populações
  - Amostras independentes
- 3 Análise de Variância
- 4 Comparações múltiplas de hipótese
- 5 Volcano plot

# Inferência estatística<sup>1</sup>

Na **inferência estatística** os dois principais objetivos são:

- **Estimar** um parâmetro populacional
  - Estimativa pontual
  - Estimativa intervalar
- **Testar** uma hipótese ou afirmativa sobre um parâmetro populacional







## Exemplo 8.1

Para averiguar se um tratamento é eficaz contra a doença, selecionamos uma amostra de 30 indivíduos submetidos ao tratamento.

Assumimos que todos os elementos da amostra  $X_1, \dots, X_{30}$  possuem a mesma distribuição:  $X_i \sim N(\mu, 36)$ , onde:

- $\mu = 14$  se o tratamento for eficiente
- $\mu = 18$  se o tratamento não for eficiente

Se a média da amostra for próxima de 14, temos **evidências** de que o tratamento é eficaz. Se for mais próxima de 18, as **evidências** são **contrárias** ao tratamento.

Então a pergunta é: o quão próximo é “próximo”?







# Tipos de hipóteses

Quando fazemos um teste de hipótese, chegamos a um dos dois possíveis resultados:

- **Rejeitar**  $H_0$ : em favor da hipótese alternativa  $H_a$
- **Não rejeitar**  $H_0$ : e conclui-se que não existem diferenças

## Atenção!

- O termo **aceitar** a hipótese nula é filosoficamente incorreto, pois não se pode aceitar uma hipótese baseada apenas em evidências amostrais (mesmo em um teste de hipótese formal).
- E ainda existe um **erro** associado a todo teste de hipótese

# Hipóteses

## Hipótese simples:

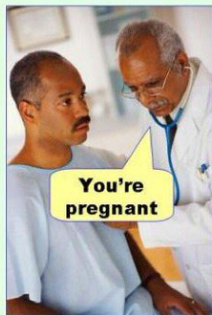
- $H_0$  : O tratamento não é eficaz ( $\mu = 18$ )
- $H_a$  : O tratamento é eficaz ( $\mu = 14$ )

## Hipóteses compostas:

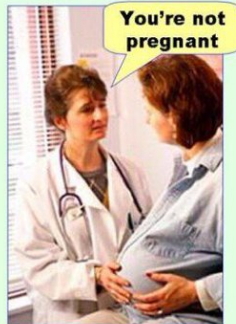
- Hipótese unilateral à esquerda
  - $H_0$  : O tratamento não é eficaz ( $\mu = 18$ );
  - $H_1$  : O tratamento é eficaz ( $\mu < 18$ ).
- Hipótese bilateral:
  - $H_0$  : O tratamento não é eficaz ( $\mu = 18$ );
  - $H_1$  : O tratamento é eficaz ( $\mu \neq 18$ ).

# Erros ao realizar um teste de hipótese

**Type I error**  
(false positive)



**Type II error**  
(false negative)



# Erros ao realizar um teste de hipótese

- **Erro Tipo I:** rejeitar  $H_0$ , quando  $H_0$  é verdadeira.
- **Erro Tipo II:** não rejeitar  $H_0$  quando  $H_0$  é falsa.

	$H_0$ verdadeira	$H_0$ falsa
Não rejeitar $H_0$	Decisão correta	Erro tipo II
Rejeitar $H_0$	Erro tipo I	Decisão correta

# Erros ao realizar um teste de hipótese

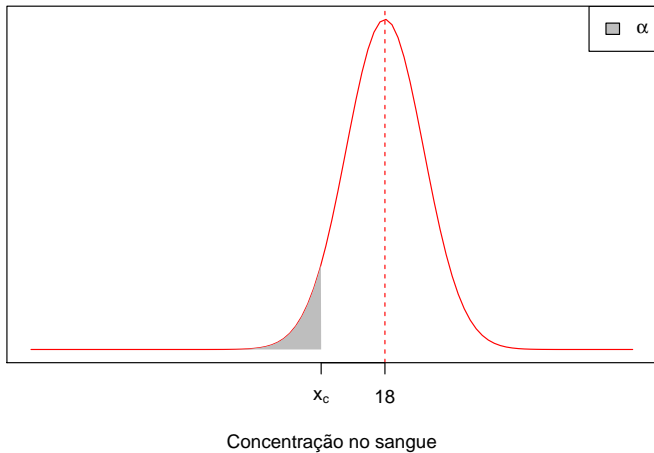
Definimos por  $\alpha$  e  $\beta$  as probabilidades de cometer os erros do tipo I e II:

- $\alpha = P(\text{erro tipo I}) = P(\text{rejeitar } H_0 \mid H_0 \text{ verdadeira})$
- $\beta = P(\text{erro tipo II}) = P(\text{não rejeitar } H_0 \mid H_0 \text{ falsa})$

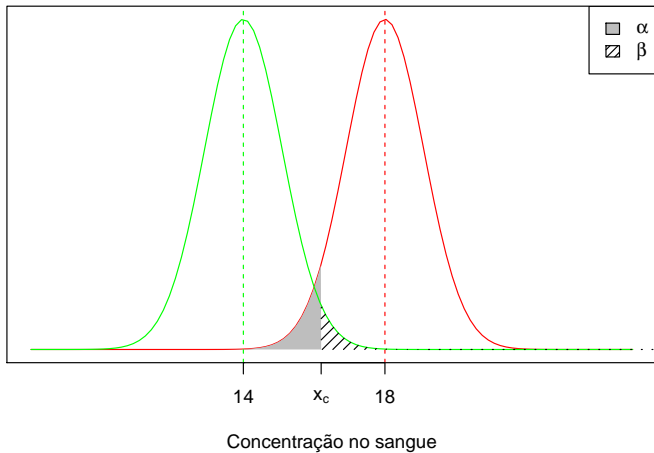
No exemplo 8.1, se  $H_0 : \mu = 18$  e  $H_a : \mu < 18$ , então:

- $\alpha =$   
 $P(\text{concluir que o tratamento é eficaz quando na verdade não é})$
- $\beta =$   
 $P(\text{concluir que o tratamento não é eficaz quando na verdade é})$

# Erros ao realizar um teste de hipótese



# Erros ao realizar um teste de hipótese





# Erros ao realizar um teste de hipótese

A situação ideal é aquela em que ambas as probabilidades,  $\alpha$  e  $\beta$ , são próximas de zero.

No entanto, à medida que diminuimos  $\alpha$ , a probabilidade  $\beta$  tende a aumentar.

Levando isso em conta, ao formular as hipóteses, **devemos cuidar para que o erro mais importante a ser evitado seja o erro do tipo I.**

Por isso, a probabilidade  $\alpha$  recebe o nome de **nível de significância** do teste, e é esse erro que devemos controlar.

# Valor crítico

Supondo  $\alpha$  conhecido podemos determinar o valor crítico  $x_c$ .

$$\begin{aligned}\alpha &= P(\text{erro tipo I}) = P(\text{rejeitar } H_0 \mid H_0 \text{ verdadeira}) \\ &= P(\bar{X} < x_c \mid \mu = 18) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{x_c - 18}{6/\sqrt{30}}\right) \\ &= P(Z < z_c)\end{aligned}$$

com  $Z \sim N(0, 1)$ .

## Obtendo o valor crítico

Dado  $\alpha$  encontramos  $z_c$  na tabela normal padrão.

Obtemos  $x_c$

$$z_c = \frac{x_c - 18}{6/\sqrt{30}} \Rightarrow x_c = 18 + z_c \frac{6}{\sqrt{30}}$$

Supondo  $\alpha = 0.05$  temos

$$0.05 = P(Z < z_c) \Rightarrow z_c = -1.64$$

logo

$$x_c = 18 - 1.64 \frac{6}{\sqrt{30}} = 16.2$$

# Região Crítica

Dada uma amostra, se  $\bar{x}_{obs} < 16.2$ , **rejeitamos**  $H_0$ , concluindo que o tratamento é eficaz.

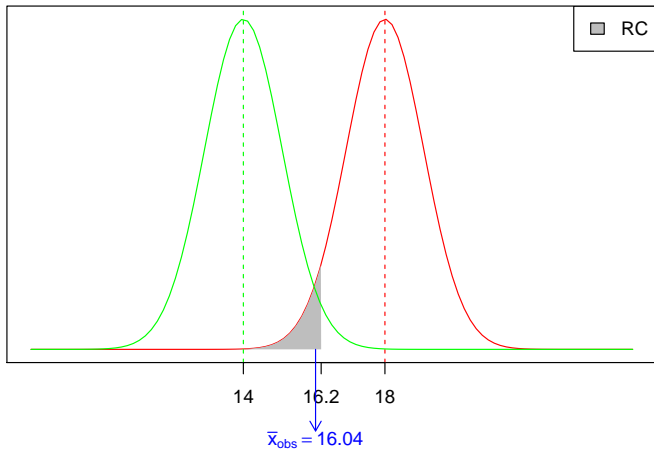
O conjunto dos números reais menores que 16.2 é denominado de **Região de Rejeição** ou **Região Crítica** (RC), isto é:

$$RC = \{x \in \mathbb{R} : x < 16.2\}.$$

No exemplo 8.1, se a média amostral dos 30 indivíduos foi  $\bar{x}_{obs} = 16.04$ , então **rejeitamos**  $H_0$ , ao nível de significância  $\alpha = 0.05$ .

Nesse caso,  $\bar{x}_{obs} < x_c$  está dentro da RC.

# Região Crítica



# Teste de hipótese bilateral

Definindo as hipóteses

$$H_0 : \mu = \mu_0 \quad \text{e} \quad H_a : \mu \neq \mu_0$$

A Região Crítica será dada por

$$RC = \{x \in \mathbb{R} \mid x < x_{c1} \quad \text{ou} \quad x > x_{c2}\}$$

Para um valor de  $\alpha$  fixado, determinamos  $x_{c1}$  e  $x_{c2}$  de modo que

$$P(\bar{X} < x_{c1} \cup \bar{X} > x_{c2}) = \alpha$$

Assim, distribuimos a área  $\alpha$  igualmente entre as duas partes da RC

$$P(\bar{X} < x_{c1}) = \frac{\alpha}{2} \quad \text{e} \quad P(\bar{X} > x_{c2}) = \frac{\alpha}{2}$$



# Etapas de um teste de hipótese

- 1 Estabelecer as hipóteses nula e alternativa.
- 2 Definir a forma da região crítica, com base na hipótese alternativa.
- 3 Identificar a distribuição do estimador e obter sua estimativa.
- 4 Fixar  $\alpha$  e obter a região crítica.
- 5 Concluir o teste com base na estimativa e na região crítica.



# $P$ -valor

Em geral,  $\alpha$  é pré-fixado para construir a regra de decisão.

Uma alternativa é deixar em aberto a escolha de  $\alpha$  para quem for tomar a decisão.

A ideia é calcular, **supondo que a hipótese nula é verdadeira**, a probabilidade de se obter estimativas mais extremas do que aquela fornecida pela amostra.

Essa probabilidade é chamada de **nível descritivo**, denotada por  $\alpha^*$  (ou  $P$ -valor).

Valores pequenos de  $\alpha^*$  evidenciam que a hipótese nula é falsa.

O conceito de “pequeno” fica para quem decide qual  $\alpha$  deve usar para comparar com  $\alpha^*$ .

## P-valor

Para **testes unilaterais**, sendo  $H_0 : \mu = \mu_0$ , a expressão de  $\alpha^*$  depende da hipótese alternativa:

$$\alpha^* = P(\bar{X} < \bar{x}_{obs} \mid H_0 \text{ verdadeira}) \quad \text{para } H_a : \mu < \mu_0$$

$$\alpha^* = P(\bar{X} > \bar{x}_{obs} \mid H_0 \text{ verdadeira}) \quad \text{para } H_a : \mu > \mu_0$$

Para **testes bilaterais**, temos  $H_0 : \mu = \mu_0$  contra  $H_0 : \mu \neq \mu_0$ , a definição do nível descritivo depende da relação entre  $\bar{x}_{obs}$  e  $\mu_0$ :

$$\alpha^* = 2 \times P(\bar{X} < \bar{x}_{obs} \mid H_0 \text{ verdadeira}) \quad \text{se } \bar{x}_{obs} < \mu_0$$

$$\alpha^* = 2 \times P(\bar{X} > \bar{x}_{obs} \mid H_0 \text{ verdadeira}) \quad \text{se } \bar{x}_{obs} > \mu_0$$

Como estamos calculando a probabilidade para apenas uma das caudas, então esse valor é multiplicado por 2.

# Comparação de Duas Médias

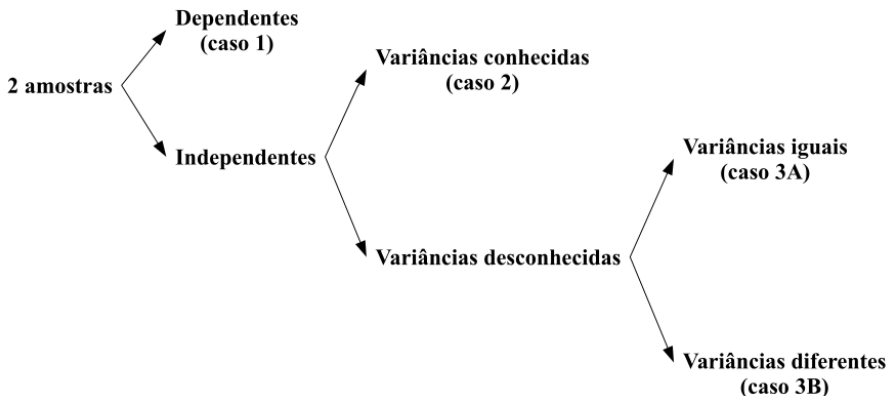
Duas das principais suposições feitas no desenvolvimento dos testes de hipóteses foram:

- Independência entre os componentes da amostra;
- Variabilidade associada aos valores populacionais e amostrais.

## Testes Paramétricos

Os testes paramétricos discutidos aqui, assumem variáveis que se comportam segundo modelo Normal, ou que as amostras são suficientemente grandes para obter uma aproximação.

# Comparação de Duas Médias



# Amostras independentes com variâncias conhecidas

Consideramos agora o teste relacionado com a situação em que queremos comparar médias de duas populações independentes, quando as correspondentes variâncias são conhecidas. A obtenção de informação a respeito do valor de variância populacional pode ser obtido de estudos anteriores ou experimentos similares.

Supondo duas populações com variâncias iguais a um valor conhecido  $\sigma_0^2$ . Além disso, admitamos que as populações seguem uma distribuição Normal, com médias  $\mu_1$  e  $\mu_2$ . Obtendo amostras aleatórias independentes  $(X_1, \dots, X_{n1})$  e  $(Y_1, \dots, Y_{n2})$  de cada população com, tamanhos  $n_1$  e  $n_2$  para as duas amostras, podemos testar as seguintes hipóteses:

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

# Amostras independentes com variâncias conhecidas

Como estamos interessados em determinar se a diferença é estatisticamente significativa, podemos ainda reescrever as hipóteses em termos de  $\mu_D = \mu_1 - \mu_2$ , isto é:

$$H_0 : \mu_D = 0 \quad (\text{As médias populacionais são iguais;})$$

$$H_a : \mu_D \neq 0 \quad (\text{As médias populacionais não são iguais;},$$

o que sugere trabalharmos com o *estimador* de  $\mu_D$ :

$$\bar{D} = \bar{X} - \bar{Y}$$

Com as suposições feitas, temos

$$X_i \sim N(\mu_1, \sigma_0^2), i = 1, 2, \dots, n_1;$$

$$Y_i \sim N(\mu_2, \sigma_0^2), i = 1, 2, \dots, n_2;$$

# Amostras independentes com variâncias conhecidas

Pela independência dessas variáveis,  $\bar{D}$  terá distribuição Normal com média  $E(\bar{D}) = \mu_D$  e quanto à variância, temos:

$$\begin{aligned} Var(\bar{D}) &= Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) \\ &= \frac{\sigma_0^2}{n_1} + \frac{\sigma_0^2}{n_2} = \sigma_0^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \end{aligned}$$

Note que a independência entre as amostras foi necessária para obter essa variância.

# Resumo

**Tabela 9.1: Comparação de médias para duas populações**

Situação	Estimadores
Amostras Pareadas (Caso 1)	$\bar{D} = \frac{\sum_{i=1}^n D_i}{n}$ $S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2$ $T = \frac{\bar{D} - \mu_D}{S_D / \sqrt{n}} \sim t_{(n-1)}$
Amostras Independentes (Caso 2) Variâncias conhecidas	$\bar{D} = \bar{X} - \bar{Y}$ $\text{Var}(\bar{D}) = \sigma_X^2 / n_1 + \sigma_Y^2 / n_2$ $Z = \frac{\bar{D} - \mu_D}{\sqrt{\sigma_X^2 / n_1 + \sigma_Y^2 / n_2}}$



# Resumo

**Tabela 9.1: Comparação de médias para duas populações**

Situação	Estimadores
Amostras Independentes (Caso 3A) Variâncias desconhecidas e iguais	$\bar{D} = \bar{X} - \bar{Y}$ $S_c^2 = \frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{(n_1 - 1) + (n_2 - 1)}$ $T = \frac{\bar{D} - \mu_D}{\sqrt{S_c^2(1/n_1 + 1/n_2)}}$
Amostras Independentes (Caso 3B) Variâncias desconhecidas e diferentes	$\bar{D} = \bar{X} - \bar{Y}$ $\hat{\sigma}_D^2 = S_X^2/n_1 + S_Y^2/n_2$ $T = \frac{\bar{D} - \mu_D}{\sqrt{S_X^2/n_1 + S_Y^2/n_2}}$

# Análise de Variância

Consideramos nesta seção o caso de comparação de três ou mais populações, definidas por uma variável qualitativa (fator) através de testes com as correspondentes médias. Iniciamos com o caso em que as amostras de cada população têm o mesmo tamanho.

Consideraremos um modelo estatístico, em que cada observação  $Y$  pode ser decomposta em duas componentes: *sistemática* e *aleatória*, esta última representando variações individuais e todos os fatores que não são explicados pela parte sistemática.

Matematicamente, podemos escrever

$$Y = \mu + e.$$

# Análise de Variância

Se  $Y$  representa a observação associada a uma unidade experimental, a parte sistemática  $\mu$  pode ser vista como média populacional, que é fixa, e a parte aleatória  $e$  como a informação referente a outros fatores que podem influir nas observações mas não são incorporadas em  $\mu$ .

Suponha que estamos interessados em compara as médias de  $K$  populações, isto é, queremos testar

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_K;$$

$H_a$  : pelo menos uma das médias  $\mu_i$  é diferente das demais.

# Análise de Variância

Definimos as quantidades *Soma de Quadrados Dentro* (SQD)

$$SQD = \sum_{i=1}^K \sum_{j=1}^m (Y_{ij} - \hat{\mu}_i)^2 = \sum_{i=1}^K \sum_{j=1}^m (Y_{ij} - \bar{Y}_i)^2 = \sum_{i=1}^K \sum_{j=1}^m Y_{ij}^2 - m \sum_{i=1}^K \bar{Y}_i^2$$

e *Soma de Quadrados Total* (SQT)

$$SQT = \sum_{i=1}^K \sum_{j=1}^m (Y_{ij} - \hat{\mu})^2 = \sum_{i=1}^K \sum_{j=1}^m (Y_{ij} - \bar{Y})^2 = \sum_{i=1}^K \sum_{j=1}^m Y_{ij}^2 - mK\bar{Y}^2.$$

# Análise de Variância

A diferença entre SQT e SQD representa a *soma de quadrados entre* será denotada por SQE, isto é,

$$SQE = SQT - SQD.$$

Das expressões para soma de quadrados total e de dentro, segue que:

$$SQE = m \sum_{i=1}^K (\bar{Y}_i - \bar{Y})^2 = m \left( \sum_{i=1}^K \bar{Y}_i^2 - K \bar{Y}^2 \right).$$

# Análise de Variância

Cada uma das somas de quadrados envolve um certo número de quantidades que estão sendo estimadas. Por exemplo, SQT contém  $\bar{Y}$  e SQD contém  $\bar{Y}_i, i = 1, \dots, K$ . Levando este fato em consideração e o número de observações nas amostras, definimos os correspondentes quadrados médios:

$$QMT = \frac{SQT}{Km - 1} : \text{quadrado médio total;}$$

$$QMD = \frac{SQD}{Km - K} = \frac{SQD}{K(m - 1)} : \text{quadrado médio dentro;}$$

$$QME = \frac{SQE}{K - 1} : \text{quadrado médio entre.}$$

Note que, nesse caso, é preciso calcular as três quantidades anteriores pois QMT não é igual à soma de QMD com QME.

# Análise de Variância

O teste estatístico para hipótese  $H_0$  envolve os quadrados médios. Se QME for grande comparado à QMD, a parte sistemática do modelo estará captando grande parte da informação dos dados e a hipótese  $H_0$  deverá ser rejeitada. Definimos, então, a quantidade

$$F = \frac{QME}{QMD}.$$

# Análise de Variância

Temos, agora condições de encontrar o valor crítico de  $f_c$  e determinar a região crítico do teste, que será da forma

$$RC = \{f \in \mathbb{R}^+ : f > f_c\}$$

Das três suposições feitas, a mais importante é a segunda,  $Var(Y_{ij}) = \sigma^2$ , para  $i = 1, \dots, K$  e  $j = 1, \dots, m$ , que tem o nome técnico de *homocedasticidade*.



# Análise de Variância

A discussão sobre o comportamento dos erros e das somas de quadrados é resumida na **Tabela de Análise de Variância (ANOVA)** Tabela 9.2:

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	F
Entre	K-1	SQE	QME	QME/QMD
Dentro	K(m-1)	SQD	QMD	
Total	Km-1	SQT		

# Métodos de comparações múltiplas de hipótese

- São aplicadas após a rejeição de  $H_0$  pela estatística F da ANOVA;
- São métodos para corrigir a inflação do nível de significância global decorrente do teste de um grande número de hipótese;
- Isso é feito principalmente de duas formas:
  - 1 Corrigir-se o p-valor após os testes de hipótese individuais para ter nível de significância global  $\alpha_k$  desejado.
  - 2 Emprega-se uma estatística de teste de hipótese que incorpore o número de hipóteses para ter nível de significância global  $\alpha_k$  desejado.

## Correção do p-valor pelo método de Bonferroni

- Agora serão testadas separadamente um conjunto de hipóteses

$$H_0 : m_i = m_j \quad \forall \quad i \neq j$$

- Se as hipóteses forem para todos os pares possíveis de  $i, j \in \{1; \dots; k\}$ , já foi visto que totalizam  $u = \binom{k}{2}$ .
- Dado um nível de significância global, para  $p$  hipóteses independentes, o nível de significância individual  $\alpha$  corrigido é

$$\alpha_p = 1 - (1 - \alpha)^p \quad \text{logo}$$

$$\alpha = 1 - (1 - \alpha_p)^{1/p} \approx \alpha_p/p$$

- Dessa forma, o p-valor do teste t individual é multiplicado por  $p$  para corrigir pela quantidade de hipóteses.

# Método de Benjamini–Hochberg

- Testando  $m$  hipóteses

$$H_1, H_2, \dots, H_m$$

- Obter os *Valores-P*

$$p_1, p_2, \dots, p_m$$

- Ordenar os *Valores-P*

$$p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$$

- Seja  $\alpha$  a taxa de falsas descobertas (FDR) que se quer controlar, encontre

$$p_{(i)} \leq \frac{i}{m} \alpha$$

- Faça o Valor-P correspondente o ponto de corte, a taxa FDR está controlada em  $\alpha$ .

## Comparações múltiplas pelo teste de Tukey

- Outra opção é trocar a estatística de teste  $\Rightarrow$  outra distribuição amostral.
- Pelo teste de Tukey, rejeita-se a hipótese de igualdade de duas médias quando

$$q_0 = \frac{abs(\bar{y}_i - \bar{y}_j)}{ep(\bar{y}_i - \bar{y}_j)} = \frac{abs(\bar{y}_i - \bar{y}_j)}{\sqrt{2s^2/r}} > q_{\alpha, \nu, k},$$

em que  $q_{\alpha, \nu, k}$  é o quantil superior da distribuição da amplitude total studentizada e  $ep(.)$  denota o erro padrão.

- Ou seja, não se usa mais a distribuição t mas sim esta que incorpora o número de tratamentos ( $k$ ) como parâmetro.
- Pelo uso desta estatística de teste se faz o controle para manter o nível de significância global no valor desejado.

# Pressuposições

As pressuposições gerais de testes paramétricos são:

- As populações sendo comparadas são normalmente distribuídas.
- A amostra é representativa da população.
- Os dados estão em uma escala intervalar ou proporcional.

Teste não paramétricos podem ser usados quando:

- Os dados não se ajustam a uma distribuição especificada e pressuposições não são feitas.
- Dados são medidos em qualquer escala.

# Pressuposições

- Pode-se inspecionar os pressupostos pela análise de resíduos.
- Também é possível aplicar testes de hipótese para os pressupostos.
- Os pressupostos não devem ser violados para a validade das inferências.

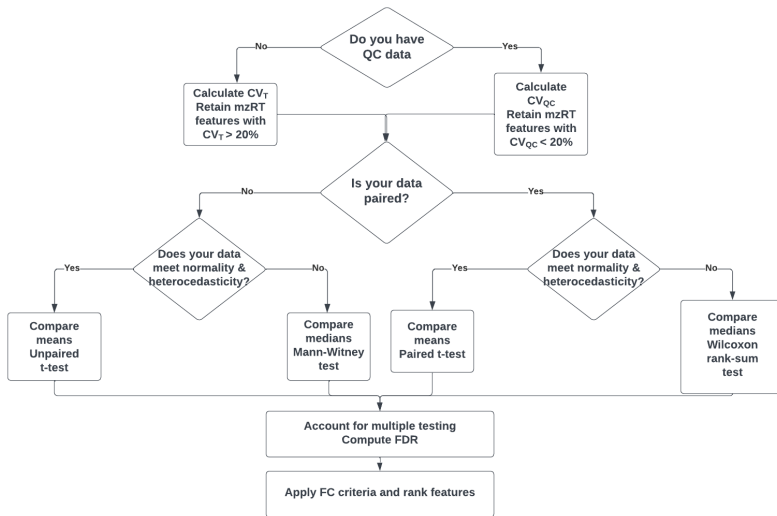
# Qual teste aplicar?<sup>2</sup>

Desenho Experimental	Distribuição Normal	Distribuição Não Normal
	Compare médias	Compare medianas
Dados não pareados	Teste-t	Mann-Whitney
Dados pareados	Teste-t pareado	Wilcoxon
>2 grupos não pareados	Anova	Kruskal-Wallis
>2 grupos pareados	Anova - medidas repetidas	Friedman

<sup>2</sup>Vinaixa et al. 2012.



# Fluxo de análises completo

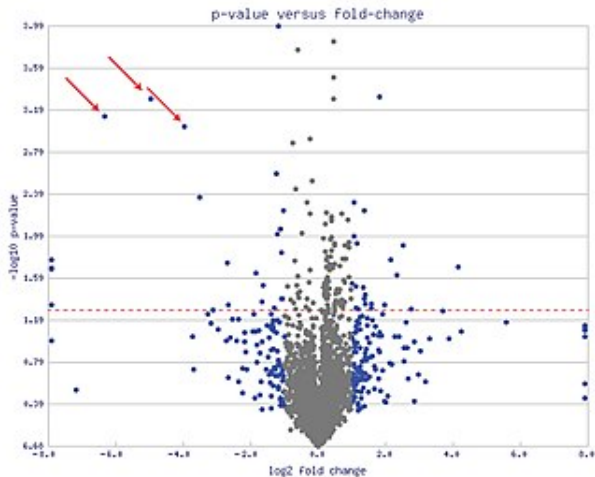


# Volcano plot

Um gráfico de *Volcano* é um tipo de gráfico de dispersão usado para identificar rapidamente mudanças em grandes conjuntos de dados. Suas principais características são

- Representa graficamente a significância versus a magnitude da diferença em y e x, respectivamente.
- Usa o negativo do logaritmo do valor p no eixo y (geralmente na base 10).
- O logaritmo da razão entre as médias dos grupos comparados (geralmente na base 2) é utilizado no eixo x.

# Volcano plot



# Referências bibliográficas



Marcos Nascimento Magalhães e  
Antonio Carlos Pedroso de Lima. *Noções de  
probabilidade e estatística*. Vol. 5. Editora da  
Universidade de São Paulo, 2002.



Maria Vinaixa et al. "A guideline to univariate statistical  
analysis for LC/MS-based untargeted  
metabolomics-derived data". Em: *Metabolites* 2.4  
(2012), pp. 775–795.