

Mini-Curso de Estatística  
Curso de Ecologia da Floresta Amazônica  
PDBFF - Manaus, AM, Brasil - Agosto de 2008

Adriano Sanches Melo  
Dep. Ecologia, Instituto de Biociências  
Universidade Federal do Rio Grande do Sul  
C.P. 15007 - Porto Alegre - RS - Brasil

[adrimelo@ufrgs.br](mailto:adrimelo@ufrgs.br)  
[www.ecologia.ufrgs.br/~adrimelo/](http://www.ecologia.ufrgs.br/~adrimelo/)

## Objetivos

- A coisa não é tão difícil assim...
- Importância do planejamento
- Idéia de modelos lineares
- Familiarização com análises
- Resolução de problemas em computador (Systat)

## Visão Geral

- Planejamento, tipos de variáveis
- Modelos Lineares
- Inferências: teste de  $b_x$  e partição variância
- Blocos
- Interação
- Modelos Lineares - Exemplos resolvidos
  - Teste t
  - Regressão linear simples
  - Anova 1 fator
  - Anova 1 fator + bloco
  - Anova 2 fatores
  - Qui-quadrado
- Análise multivariada exploratória
  - Índices de similaridade
  - Noções de Classificação e Ordenação
- Tarde – Resolução de exercícios

0) Planejamento: Razão de ser tão importante.

Trabalhos A.S. Flecker

1) Modelos lineares.

A base para os testes de hipóteses tradicionais.

$$Y_i = b_0 + b_1 X_i + e_i$$

2) Análises são basicamente as mesmas...

O que muda é a natureza (contínua, categórica) das variáveis dependentes e independentes.

-Cada capítulo, uma análise...

Exemplos de modelos lineares

3) Variáveis dependentes e independentes.

(respostas)

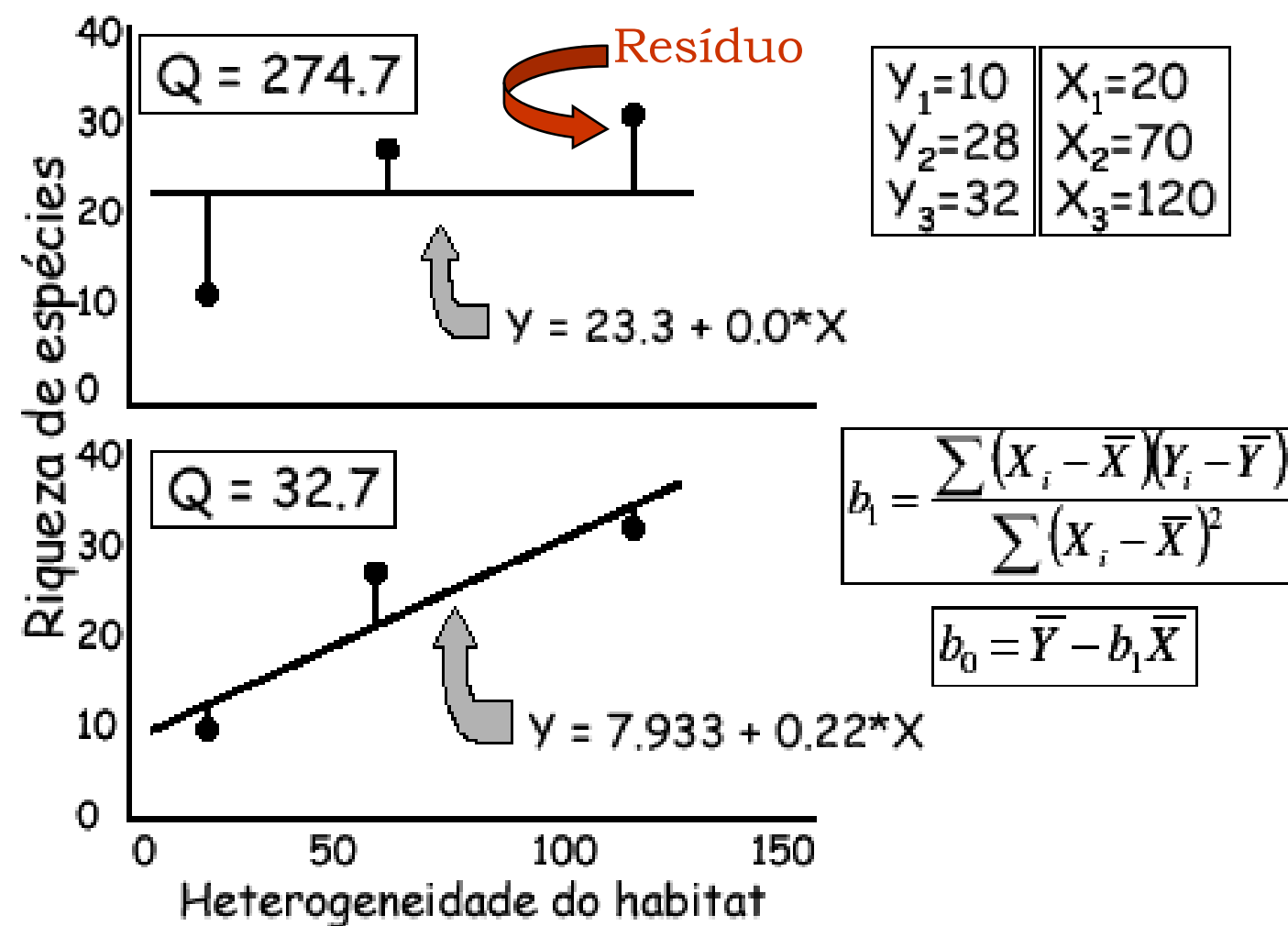
(preditoras)

Exemplos

#### 4) O modelo de regressão linear simples.

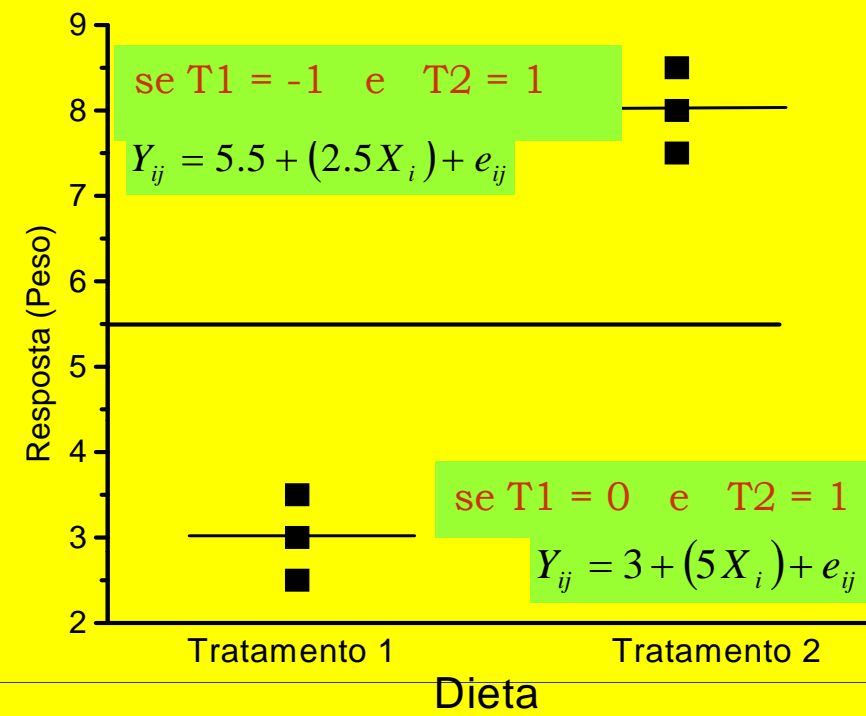
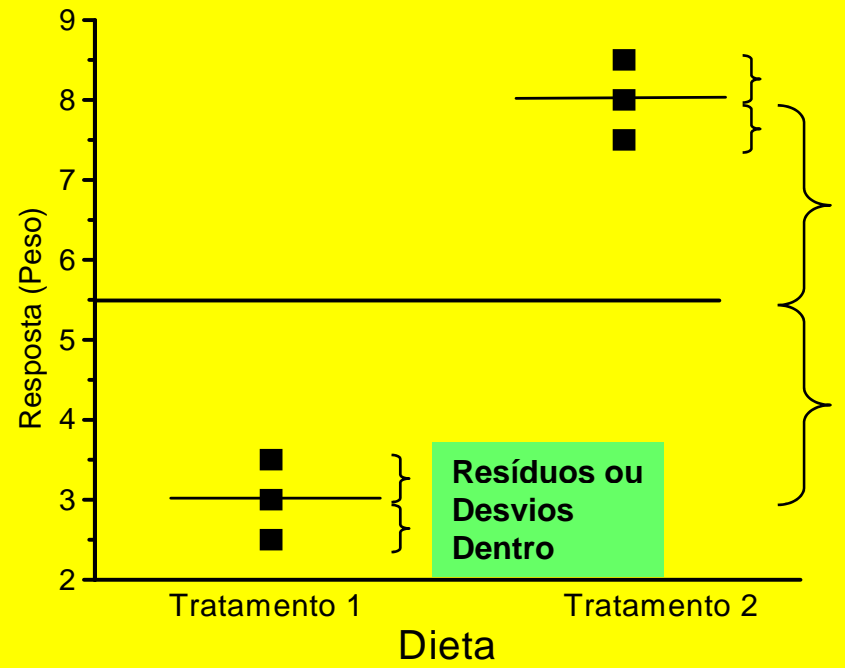
Ajuste do modelo

$$Q = \text{Soma} ( Y_i - [ b_0 + b_1 X_i ] )^2$$



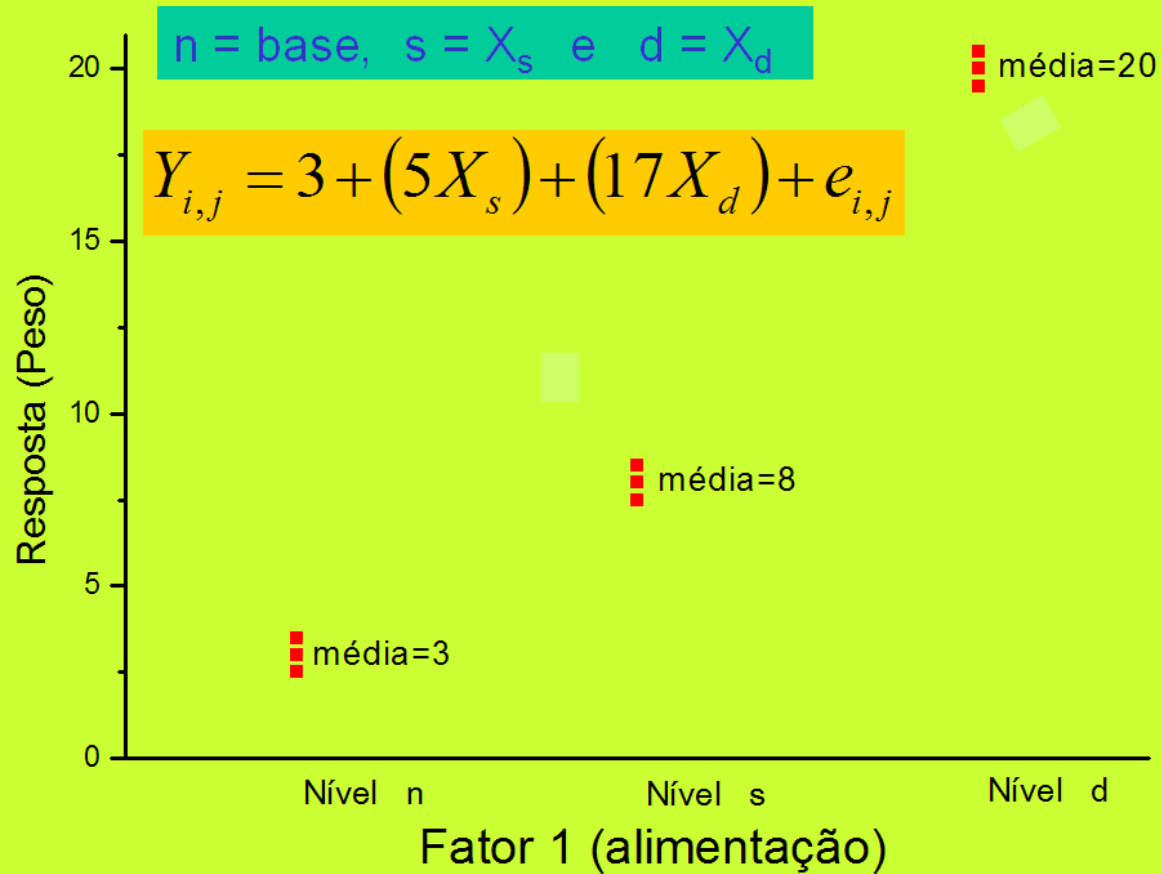
Média também minimiza Q !!!

5a) Variáveis preditoras categóricas: variáveis indicadoras  
2 níveis



## 5b) Variáveis preditoras categóricas: variáveis indicadoras > 2 níveis

X	Y
n	2.5
n	3
n	3.5
s	7.5
s	8
s	8.5
d	19.5
d	20
d	20.5



## 6) Modelos Lineares (As 'diferentes' análises....)

Variável dependente: Contínua

		Número de variáveis independentes		
		1	2	3
Tipo variável independente	contínua	regressão simples	regressão múltipla	regressão múltipla
	qualitativa	teste $t$ (1-2 níveis) 1-anova (>2 níveis)	test $t$ pareado 2-anova 1-anova + bloco	3-anova 2-anova + bloco 1-anova + 2 blocos
	mixta	----	Ancova	Ancova



## 7) Inferências

- Forma geral para testar a 'significância' de uma variável preditora.
- Pergunta que se faz: Vale a pena incluir no modelo?
- Duas formas de interpretar

### 7a) Coeficientes são diferentes de '0' ?

Numa regressão linear simples:

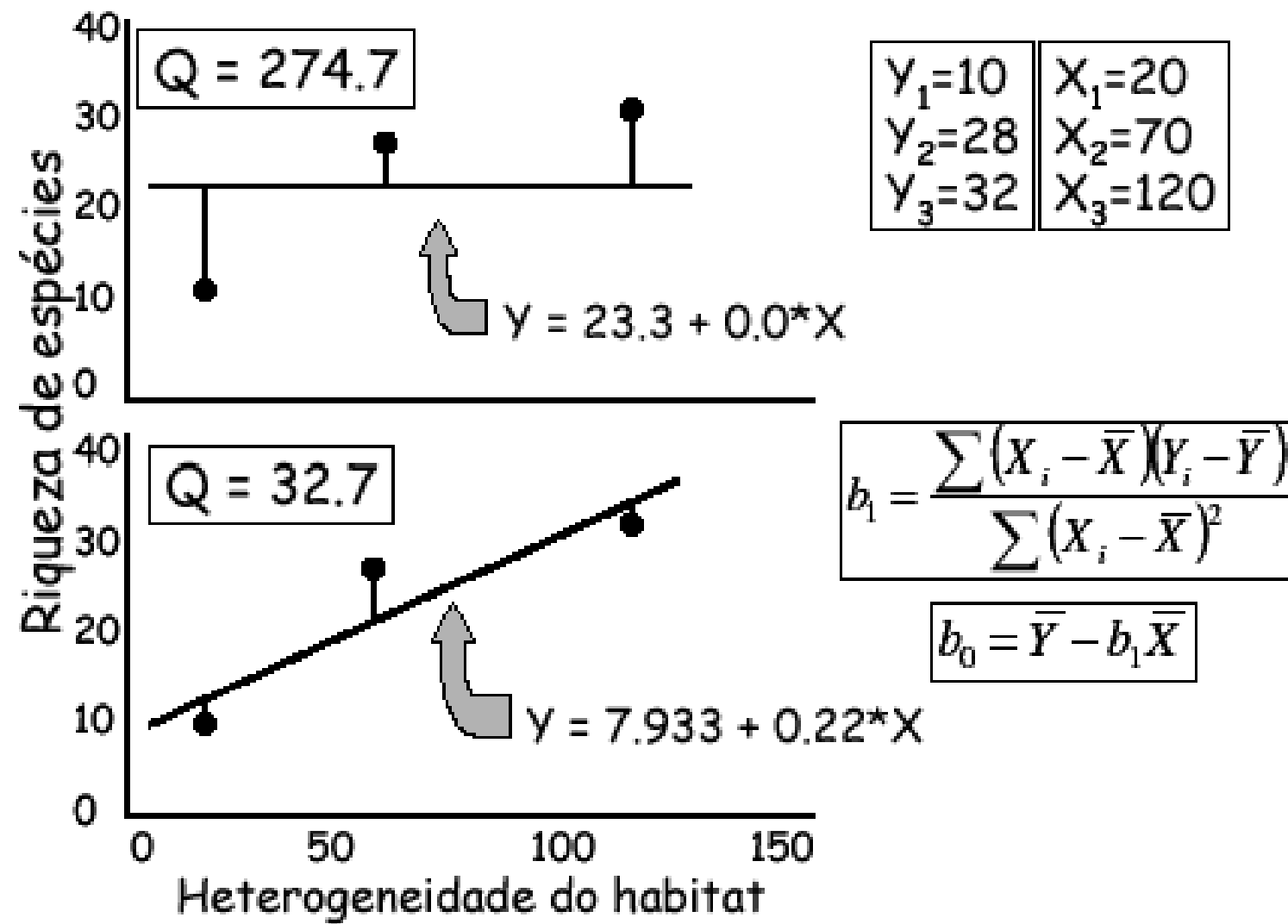
Vale a pena incluir  $b_1X$ ?

A variável  $X$  aumenta a explicação consideravelmente?

Teste:  $H_0: b_1 = 0$  ( $Y = b_0$ )  
ou  $H_1: b_1 \neq 0$  ( $Y = b_0 + b_1X$ )

7b) Partição de Variância

$$Q = \sum (Y_i - [b_0 + b_1 X_i])^2$$



## Resultado Systat Exemplo Anterior

Dep Var: VAR00005 N: 3 Multiple R: 0.939 **Squared multiple R:0.881**

Adjusted squared multiple R: 0.762 Standard error of estimate:5.715

Effect	Coefficient	Std Error	Std Coef	Tolerance	t	P(2 Tail)
CONSTANT	7.933	6.550	0.000	.	1.211	0.439
VAR00004	<b>0.220</b>	<b>0.081</b>	<b>0.939</b>	1.000	<b>2.722</b>	<b>0.224</b>

### Analysis of Variance

Source	Sum-of-Squares	df	Mean-Square	F-ratio	P
Regression	<b>242.000</b>	1	242.000	7.408	<b>0.224</b>
Residual	<b>32.667</b>	1	32.667		

$$274.667 = 242.000 + 32.667$$

$$SSTotal = SSReg + SSE$$

$$R^2 = 242 / 274.667 = 0.881$$

### SSTotal

$$(10 - 23.33)^2 = 177.78$$

$$(28 - 23.33)^2 = 21.78$$

$$(32 - 23.33)^2 = \underline{23.33}$$

$$\mathbf{274.67}$$

### Y ajustado

$$7.93 + 0.22 * 20 = 12.33$$

$$7.93 + 0.22 * 70 = 23.33$$

$$7.93 + 0.22 * 120 = 34.33$$

### SSE

$$(10 - 12.33)^2 = 5.44$$

$$(28 - 23.33)^2 = 21.78$$

$$(32 - 34.33)^2 = \underline{5.44}$$

$$\mathbf{32.67}$$

-- Por que dividir por (n-1) ?

-- A idéia de *graus de liberdade* (Crawley p.36)

$$\text{Variância} = s^2 = \frac{SS}{n-1}$$

Suponha termos 5 números e que sua média seja 4. A soma dos números, portanto, deve ser 20. Vamos ver quais números poderiam ser:

- 1) O primeiro número pode ser qualquer um; por exemplo o 2

2				
---	--	--	--	--

- 2) O segundo número pode ser qualquer um; por exemplo o 7

2	7			
---	---	--	--	--

- 3) O terceiro número pode ser qualquer um; por exemplo o 4

2	7	4		
---	---	---	--	--

- 4) O quarto número pode ser qualquer um; por exemplo o 0

2	7	4	0	
---	---	---	---	--

- 5) Não temos escolha para o quinto número; ele DEVE ser 7

2	7	4	0	7
---	---	---	---	---

## 7b) Partição de Variância

SSTotal (SST) = SSRregressão (SSR) + SSResíduo (SSE)

$$\underbrace{Y_i - \bar{Y}}_{\text{SST}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\text{SSR}} + \underbrace{Y_i - \hat{Y}_i}_{\text{SSE}}$$

$$SST = \sum (Y_i - \bar{Y})^2 \longrightarrow MST = Var = \frac{\sum (Y_i - \bar{Y})^2}{n-1}$$

$$SSR = \sum (\hat{Y}_i - \bar{Y})^2 \longrightarrow MSR = \frac{SSR}{1}$$

$$SSE = \sum (Y_i - \hat{Y}_i)^2 \longrightarrow MSE = \frac{SSE}{n-2}$$

## 7e) Graus de Liberdade (gl ou df)

**Total** =  $n - 1$

**Modelo** = número de parâmetros exceto constante ( $b_0$ )

Variável contínua = 1

Variável categórica = níveis - 1

**Resíduo** = gl Total - gl Modelo

## 8) Refinando nosso estudo: Uso de blocos

Experimento *Prochilodus*.

3 níveis do fator de estudo

Co = controle;

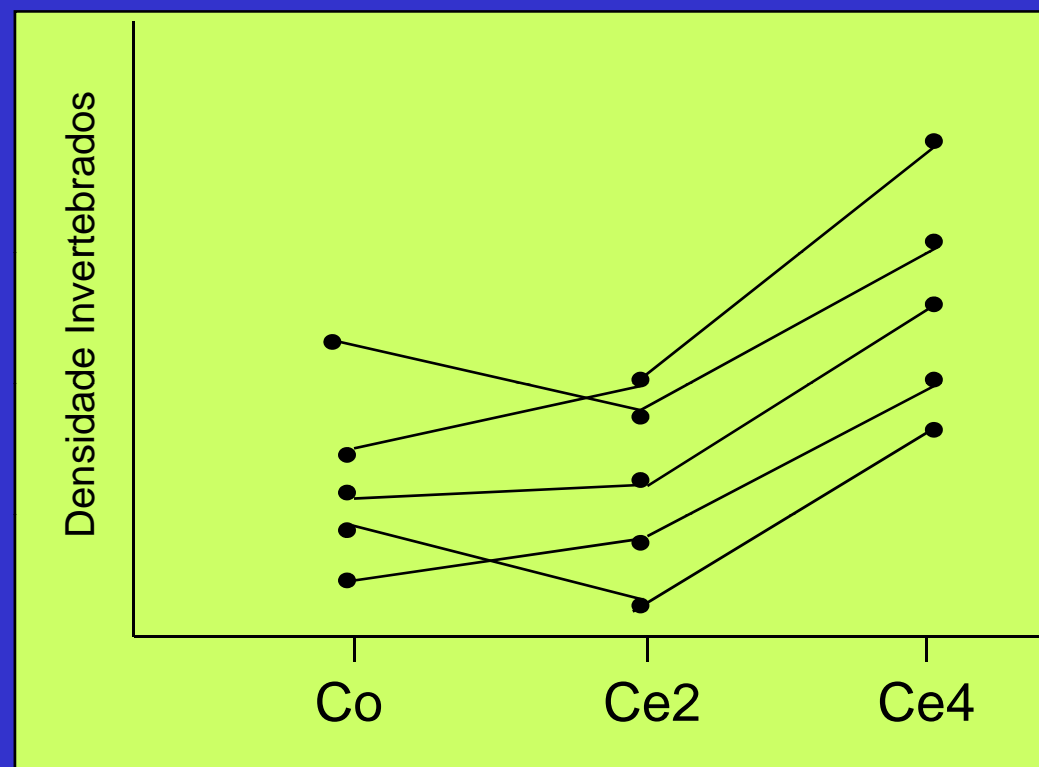
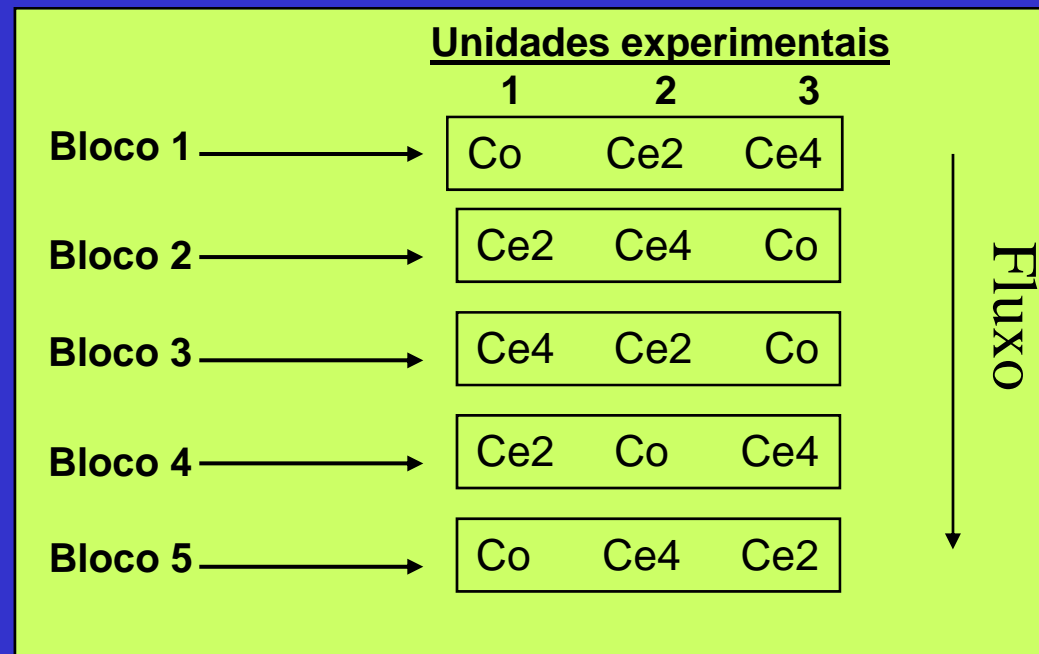
Ce2 = Controle de procedimento

Ce4 = gaiola de exclusão 4 lados

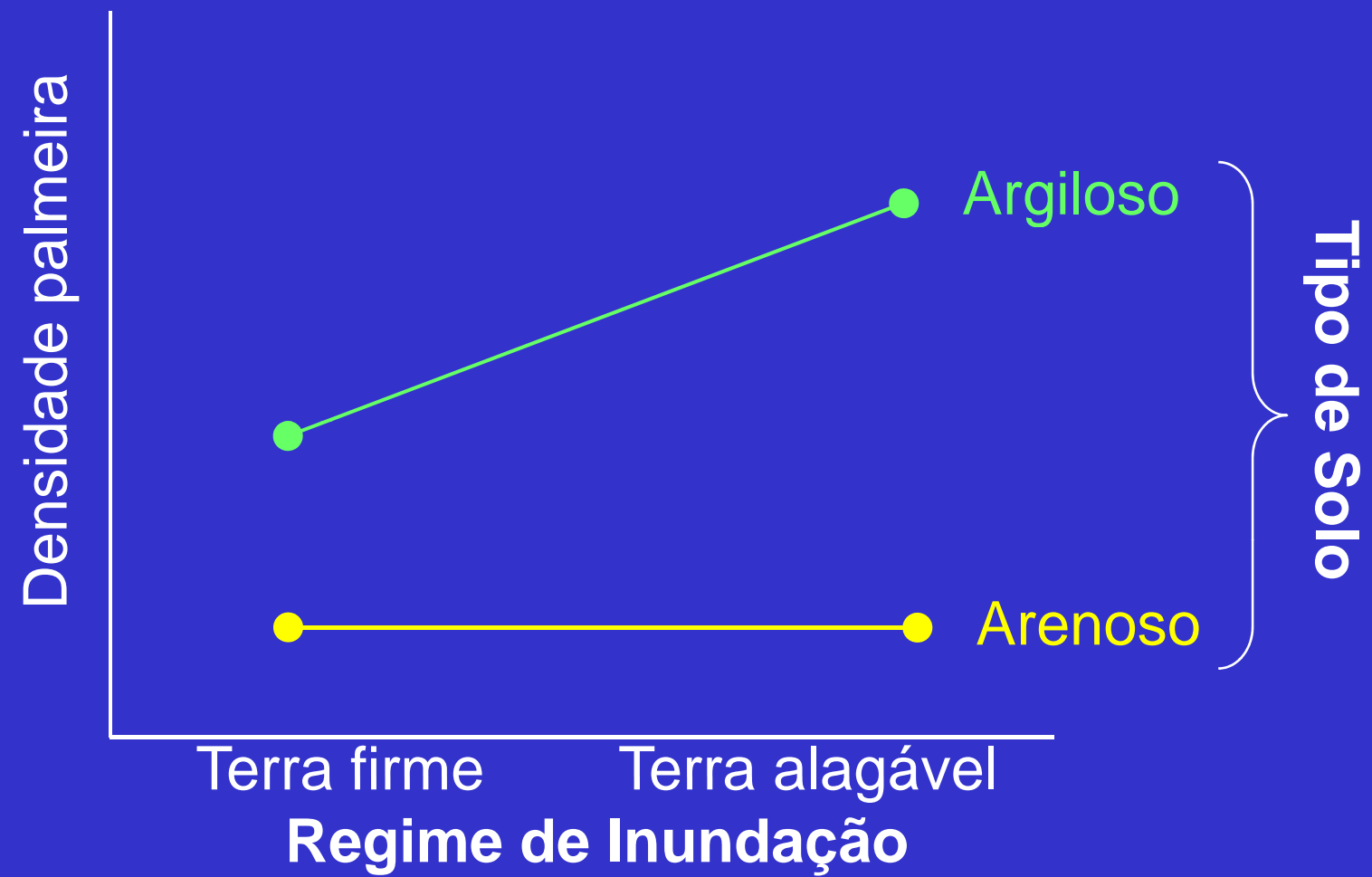
*O trabalho é feito por corredeiras, cada corredeira sendo 1 bloco.*

Note que em cada corredeira existe uma réplica do tratamento.

Poderíamos ter mais de uma réplica por corredeira. Note também que a posição do tratamento dentro da corredeira é aleatória. Este desenho experimental em blocos seria útil no caso de haver indícios de que a fauna entre corredeiras seja distinta, ou seja que as corredeiras sejam muito diferentes umas das outras.



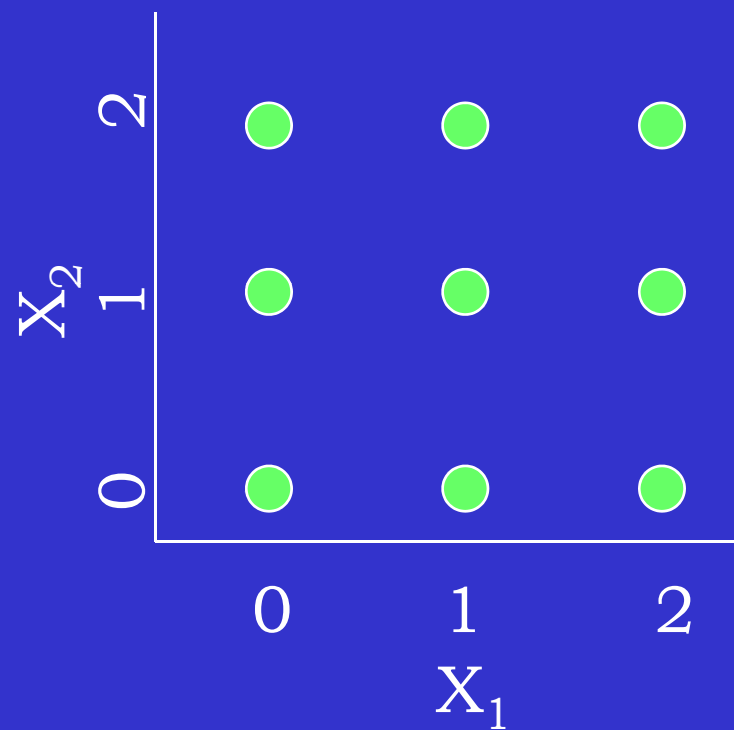
9) Quando temos mais de uma variável preditora  
Interação



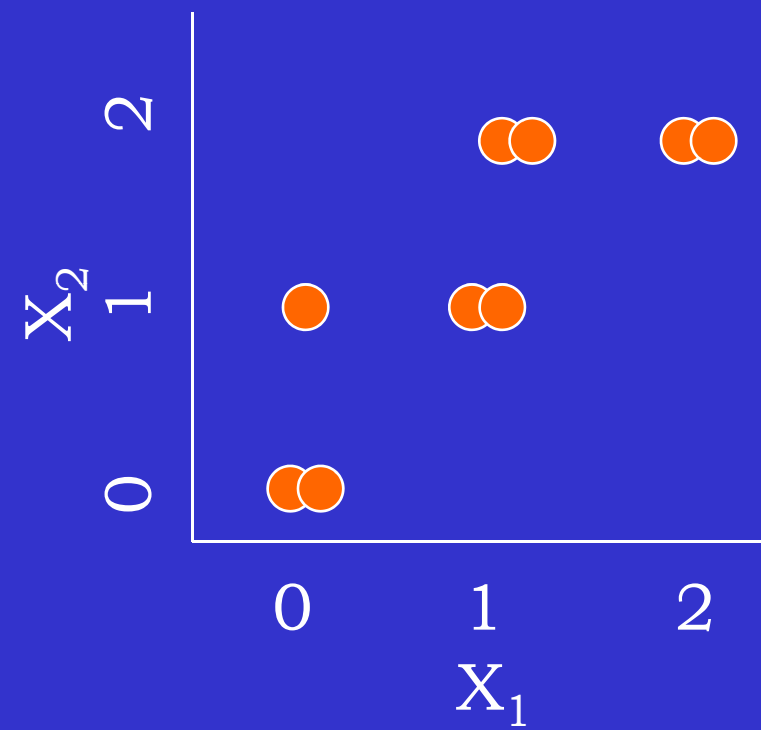


## 10) Quando temos mais de uma variável preditora

Análises fatoriais  
Correlação entre  
variáveis preditoras é '0'



Análises não fatoriais  
Correlação entre variáveis  
preditoras é diferente de '0'  
Problemas com  
multicolinearidade ....



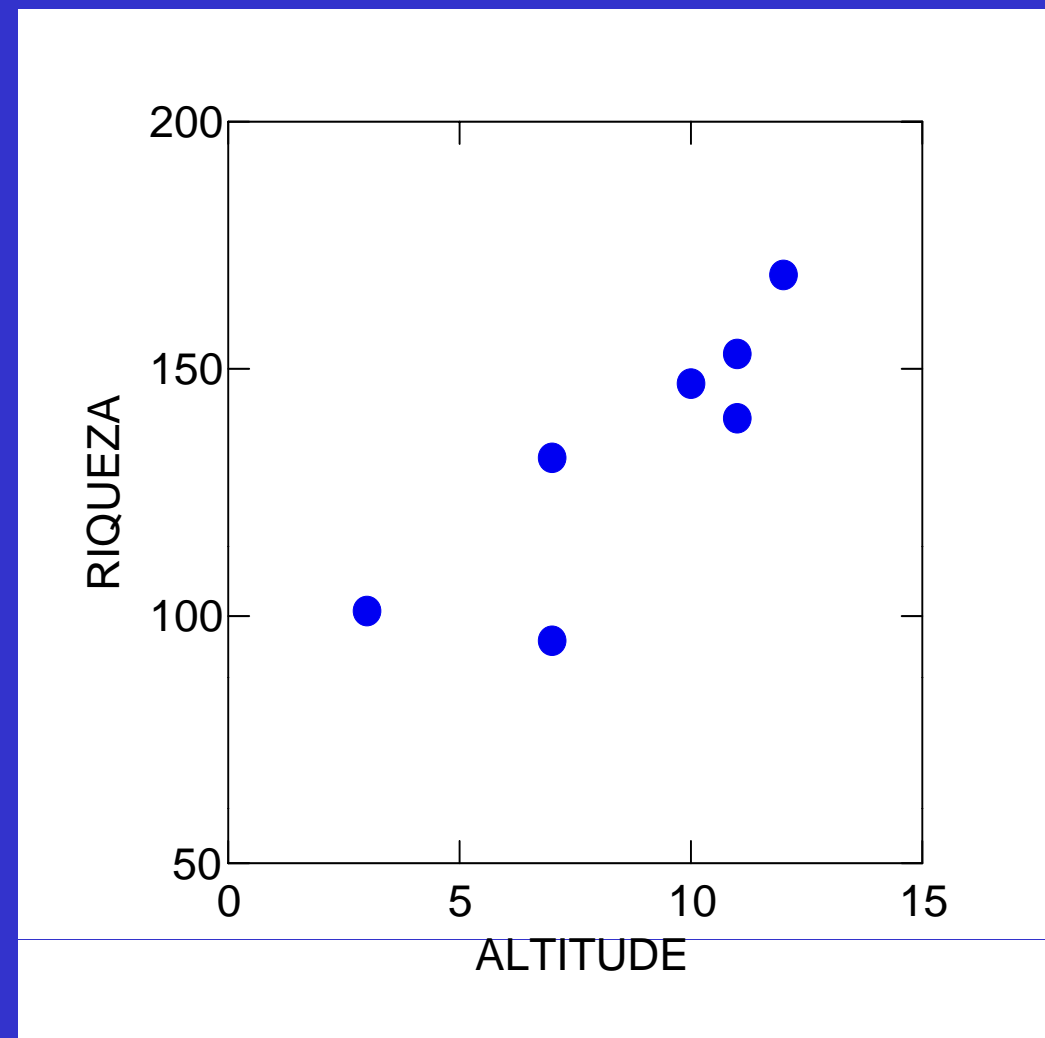
## 11) Modelos Lineares Generalizados (GLM)

Variável dependente: Qualitativa

		Número de variáveis independentes		
		1	2	3
Tipo variável independente	contínua	regressão logística simples	regressão logística múltipla	regressão logística múltipla
	qualitativa	teste G teste $\chi^2$	modelos log-lineares	modelos log-lineares
	mixta	----	regressão logística múltipla	regressão logística múltipla

## Na prática - Regressão Linear Simples

Passo 1 - Escolha do modelo



S	Altitude
101	3
95	7
132	7
147	10
140	11
153	11
169	12

## Na prática - Regressão Linear Simples

### Passo 2 - Ajuste do modelo

### Resultado Systat

Dep Var: RIQUEZA N: 7 **Multiple R: 0.861** **Squared multiple R: 0.741**  
Adjusted squared multiple R: 0.689 Standard error of estimate: 15.111

Effect	Coefficient	Std Error	Std Coef	Tolerance	t	P(2 Tail)
CONSTANT	70.344	17.746	0.000	.	3.964	0.011
<b>ALTITUDE</b>	<b>7.288</b>	<b>1.928</b>	0.861	1.000	<b>3.780</b>	<b>0.013</b>

Effect	Coefficient	Lower	< 95%>	Upper
CONSTANT	70.344	24.727		115.961
<b>ALTITUDE</b>	<b>7.288</b>	<b>2.332</b>		<b>12.245</b>

-----  
Correlation matrix of regression coefficients

	CONSTANT	ALTITUDE
CONSTANT	1.000	
ALTITUDE	-0.947	1.000

-----  
**Analysis of Variance**

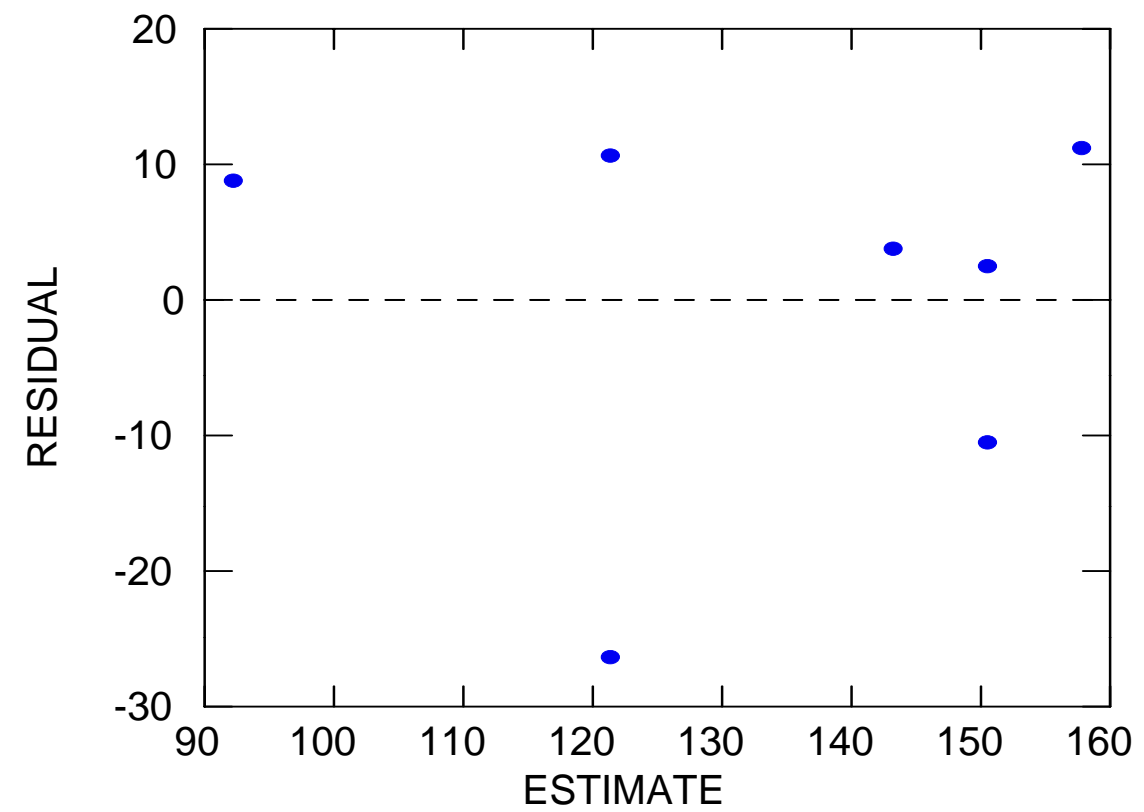
Source	Sum-of-Squares	df	Mean-Square	F-ratio	P
Regression	3263.108	1	3263.108	14.290	0.013
Residual	1141.749	5	228.350		

-----

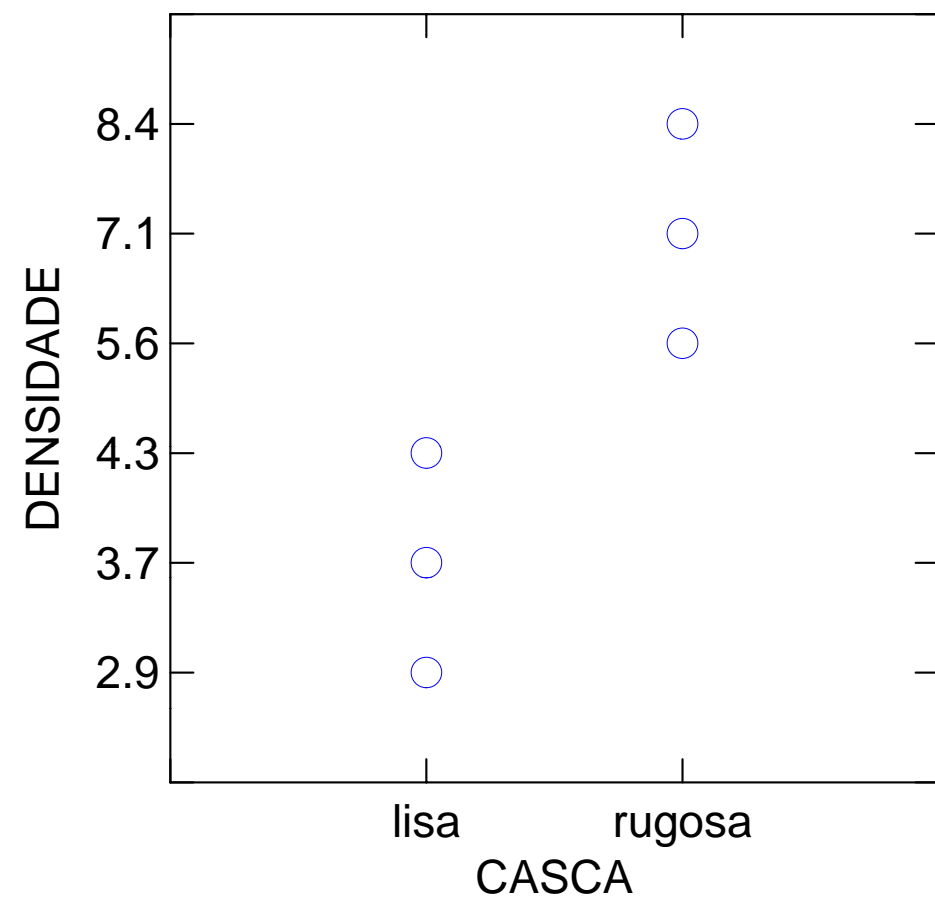
## Na prática - Regressão Linear Simples

Passo 3 - Adequação do modelo (transformação ?)

Plot of Residuals against Predicted Values



## Na prática - Teste t



Densidade	Tipo de casca
3.7	lisa
4.3	lisa
2.9	lisa
5.6	rugosa
8.4	rugosa
7.1	rugosa

Médias  
geral = 5.333  
lisa = 3.633  
rugosa = 7.033

## Na prática Teste t

$$\begin{aligned}
 (3.7 - 3.633)^2 &= 0.004489 \\
 (4.3 - 3.633)^2 &= 0.444889 \\
 (2.9 - 3.633)^2 &= 0.537289 \\
 (5.6 - 7.033)^2 &= 2.053489 \\
 (8.4 - 7.033)^2 &= 1.868689 \\
 (7.1 - 7.033)^2 &= 0.004489 \\
 \hline
 &4.913334
 \end{aligned}$$

$$\begin{aligned}
 (3.633 - 5.333)^2 &= 2.89 \\
 (7.033 - 5.333)^2 &= 2.89 \\
 \hline
 &5.78
 \end{aligned}$$

$$\begin{aligned}
 n &= 3 \\
 5.78 * 3 &= 17.34
 \end{aligned}$$

D	casca
3.7	lisa
4.3	lisa
2.9	lisa
5.6	rugosa
8.4	rugosa
7.1	rugosa

## Resultado Systat

Dep Var:DENSIDADE N: 6 Multiple R:0.883 Squared multiple R:0.779

Estimates of effects  $B = (X'X)^{-1} X'Y$

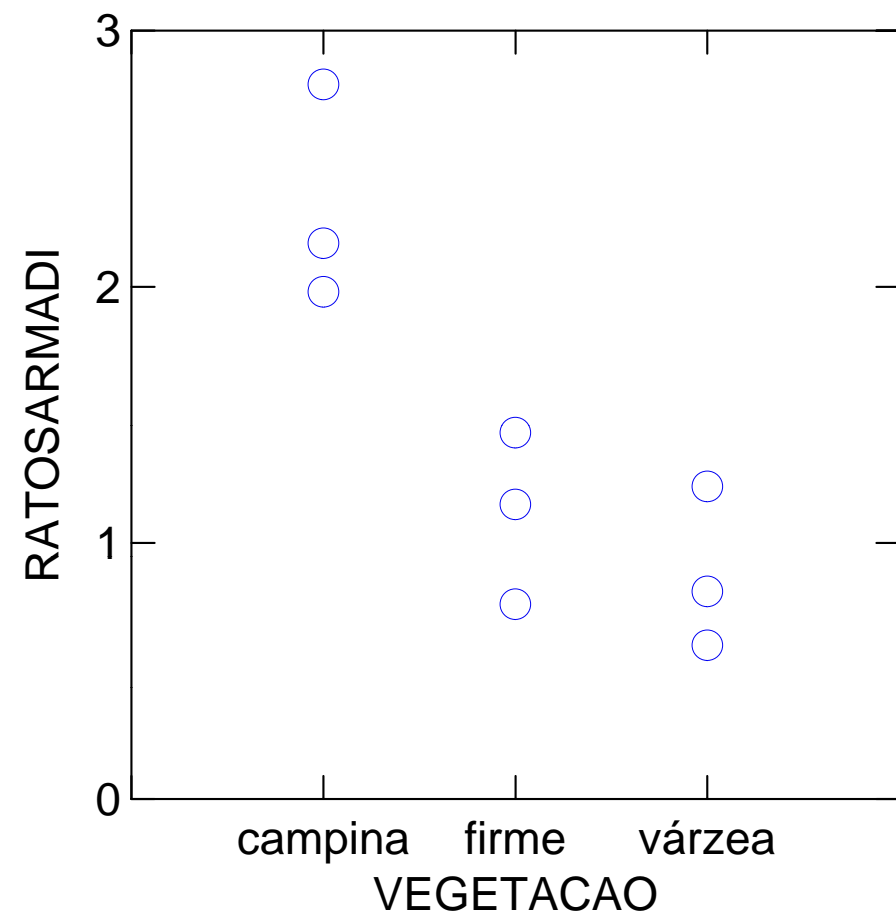
		DENSIDADE
CONSTANT		7.033
CASCA\$	lisa	-3.400

Médias  
 geral = 5.333  
 lisa = 3.633  
 rugosa = 7.033

### Analysis of Variance

Source	Sum-of-Squares	df	Mean-Square	F-ratio	P
Regression	17.340	1	17.340	14.117	0.020
Residual	4.913	4	1.228		

## Na prática - Anova 1 fator



Ratos/armadilha	Vegetação
0.60	várzea
0.81	várzea
1.22	várzea
1.43	firme
1.15	firme
0.76	firme
2.17	campina
1.98	campina
2.79	campina

Médias  
 geral = 1.434  
 várzea = 0.877  
 firme = 1.113  
 campina = 2.313



## Na prática - Anova 1 fator

### Resultado Systat

Dep Var:RATOSARMADI N:9 Multiple R:0.905 Squared multiple R:0.819  
-1

Estimates of effects  $B = (X'X)^{-1} X'Y$

	RATOSARMADI
CONSTANT	0.877
VEGETACAO\$ campina	1.437
VEGETACAO\$ firme	0.237

#### Médias

geral	= 1.434
várzea	= 0.877
firme	= 1.113
campina	= 2.313

#### Analysis of Variance

Source	Sum-of-Squares	df	Mean-Square	F-ratio	P
Regression	3.560	2	1.780	13.619	0.006
Residual	0.784	6	0.131		

## Na prática - Anova 1 fator

### Resultado Systat - teste *a posteriori*

COL/  
ROW VEGETACAO\$  
1 **campina**  
2 **firme**  
3 **várzea**

Using least squares means.  
Post Hoc test of RATOSARMADI

-----  
Using model MSE of 0.131 with 6 df.  
Matrix of pairwise mean differences:

	1	2	3
1	0.000		
2	-1.200	0.000	
3	-1.437	-0.237	0.000

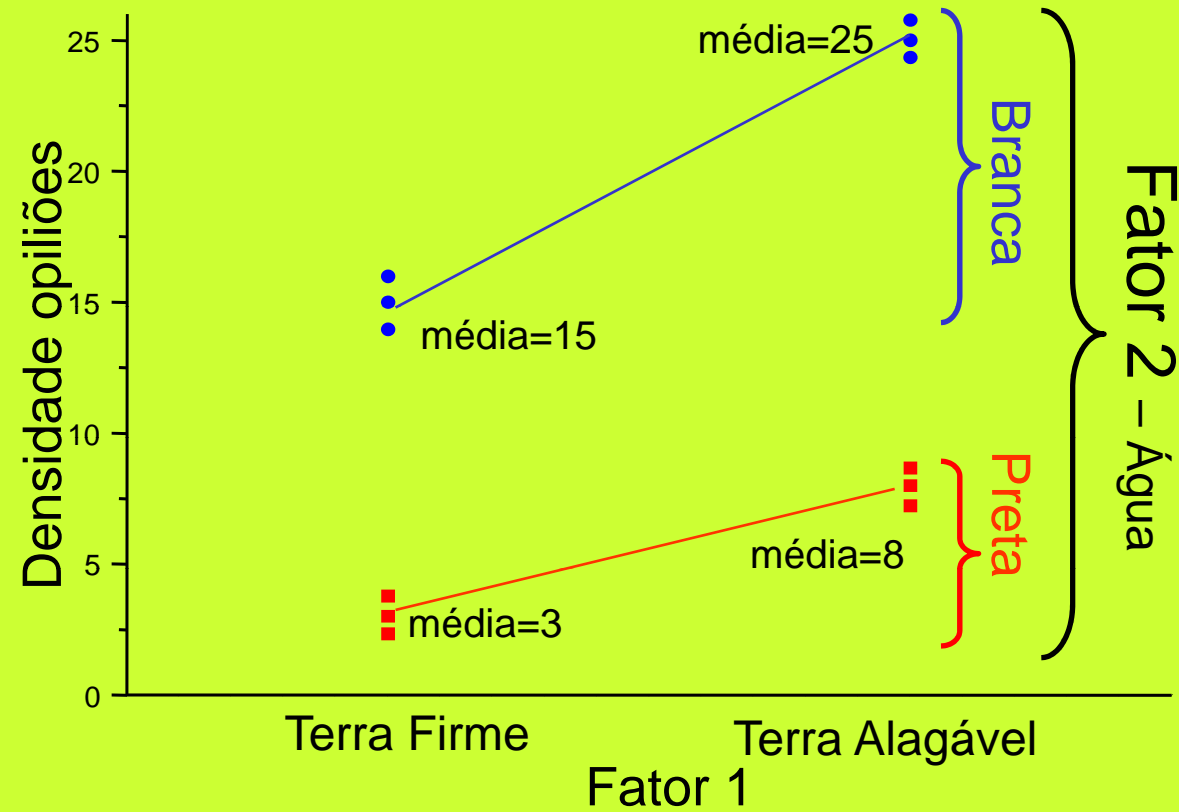
**Tukey HSD Multiple Comparisons.**

**Matrix of pairwise comparison probabilities:**

	1	2	3
1	1.000		
2	0.016	1.000	
3	0.007	0.716	1.000

-----

## Na prática - Anova 2 fatores



Dens. opções	Bacia	Alaga
2.5	Negro	firme
3	Negro	firme
3.5	Negro	firme
7.5	Negro	alaga
8	Negro	alaga
8.5	Negro	alaga
14.5	Solim	firme
15	Solim	firme
15.5	Solim	firme
24.5	Solim	alaga
25	Solim	alaga
25.5	Solim	alaga

## Na prática - Anova 2 fatores

## Resultado Systat

Dep Var:DENSOPILIO N:12 Multiple R:0.999 Squared multiple R:0.998

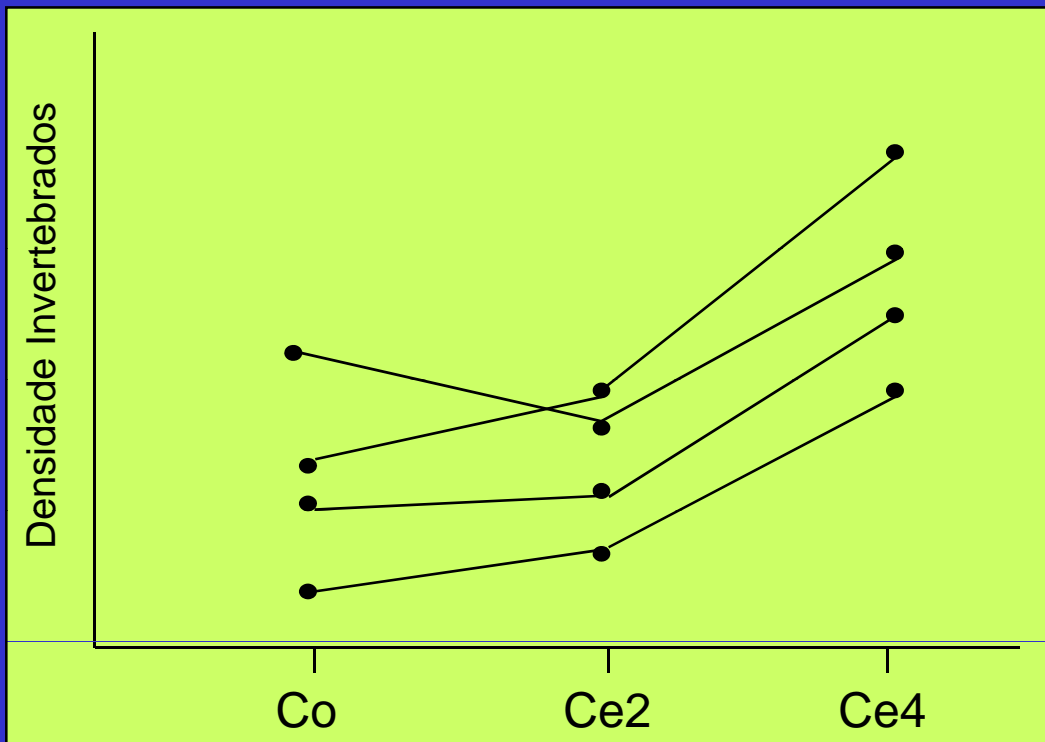
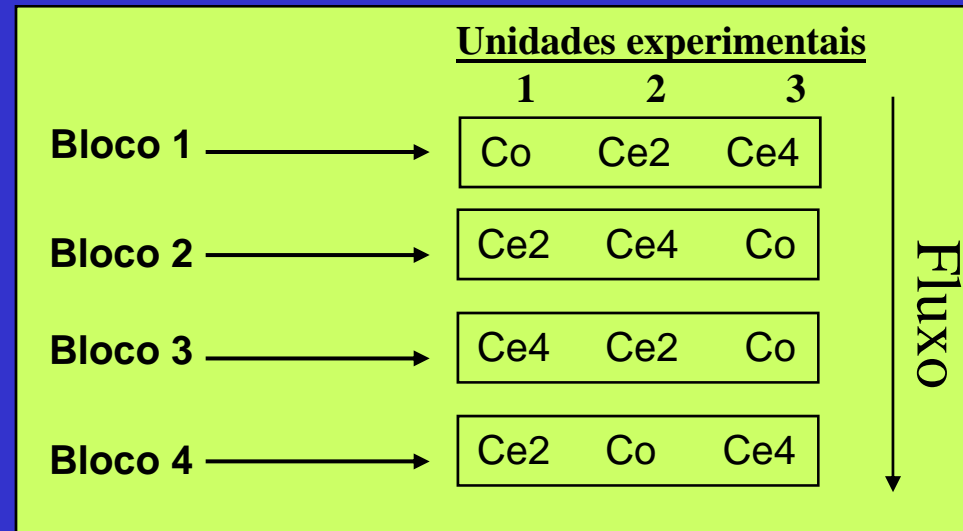
```
                DENSOPILIO
CONSTANT                12.750
RIO$      Negro        -7.250
ALAGAMENTO$ alaga      3.750
ALAGAMENTO$ alaga
RIO$      Negro        -1.250
```

### Analysis of Variance

Source	Sum-of-Squares	df	Mean-Square	F-ratio	P
RIO\$	630.750	1	630.750	2523.000	0.000
ALAGAMENTO\$	168.750	1	168.750	675.000	0.000
<b>ALAGAMENTO\$*RIO\$</b>	<b>18.750</b>	<b>1</b>	<b>18.750</b>	<b>75.000</b>	<b>0.000</b>
Error	2.000	8	0.250		

---

## Na prática Anova 1 fator + bloco



Densid Algas	Trat	Bloco
2.5	cont	1
3	ce2	1
4.5	ce4	1
17.5	cont	2
17	ce2	2
19.5	ce4	2
14.5	cont	3
15	ce2	3
15.5	ce4	3
4.5	cont	4
5	ce2	4
5.5	ce4	4

cont = controle  
ce2 = cont. proced.  
ce4 = exclusão

## Na prática Anova 1 fator + bloco

### Resultado Systat com bloco

Dep Var:VAR00001 N:12 Multiple R:0.998 Squared multiple R: 0.997

Analysis of Variance

Source	Sum-of-Squares	df	Mean-Square	F-ratio	P
<b>TRATAM\$</b>	<b>5.167</b>	<b>2</b>	<b>2.583</b>	<b>10.333</b>	<b>0.011</b>
BLOCO\$	474.000	3	158.000	632.000	0.000
Error	1.500	6	0.250		

### Resultado Systat sem bloco

Dep Var: VAR00001 N: 12 Multiple R: 0.104 Squared multiple R: 0.011

Analysis of Variance

Source	Sum-of-Squares	df	Mean-Square	F-ratio	P
<b>TRATAM\$</b>	<b>5.167</b>	<b>2</b>	<b>2.583</b>	<b>0.049</b>	<b>0.953</b>
Error	475.500	9	52.833		

## Na prática - Qui-quadrado

A presença de mosquitos depende da espécie de bromélia?

	Presença mosquito (resposta)		Totais
	Sim	Não	
Bromélia A	18	15	33
Bromélia B	8	32	40
Totais	26	47	73

## Na prática - Qui-quadrado

	Presença mosquito (resposta)		Totais
	Sim	Não	
Bromélia A	18 (11.75)	15 (21.25)	33
Bromélia B	8 (14.25)	32 (25.75)	40
Totais	26	47	73

Frequência Esperada

$$\frac{\text{total linha} * \text{total coluna}}{\text{total geral}}$$

$$(33 * 26) / 73 = 11.75$$

$$\chi^2_{Pearson} = \sum \frac{(O - E)^2}{E}$$

$$\chi^2_{Pearson} = \frac{(18 - 11.75)^2}{11.75} + \frac{(8 - 14.25)^2}{14.25} + \frac{(15 - 21.25)^2}{21.25} + \frac{(32 - 25.75)^2}{25.75} = 3.32 + 2.74 + 1.83 + 1.52 = 9.42$$

$$df = (\text{linhas} - 1) * (\text{colunas} - 1) = 1 * 1 = 1$$

Probabilidade de obter o valor 9.42 com 1 df = 0.0022



## Na prática - Qui-quadrado

### Resultado Systat

Frequência	Mosquito	Bromélia
18	sim	a
8	sim	b
15	nao	a
32	nao	b

Case frequencies determined by value of variable **FREQ.**

Frequencies

MOSQUITO\$ (rows) by BROMELIA\$ (columns)

	a	b	Total
nao	15	32	47
sim	18	8	26
Total	33	40	73

Test statistic	Value	df	Prob
<b>Pearson Chi-square</b>	<b>9.410</b>	<b>1.000</b>	<b>0.002</b>



## Análise Multivariada Exploratória

Matrix de locais (objetos) por espécies (variáveis)

	sp1	sp2	sp3	sp4	sp5	sp6	sp7
Sitio1	0	5	12	23	23	68	0
Sitio2	5	10	24	0	45	52	6
Sitio3	78	79	0	0	2	0	54
Sitio4	45	79	0	0	3	0	68
Sitio5	6	2	34	0	10	0	2

## Índices de similaridade: Qualitativos: Presença e ausência

		amostra 1	
		+	-
amostra 2	+	a	b
	-	c	d

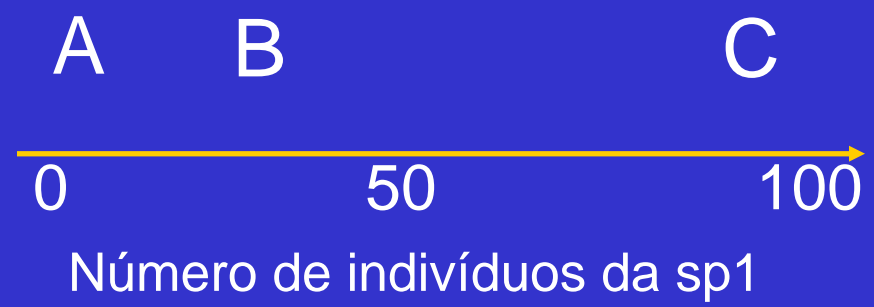
- Discordância quanto ao uso de “d”
- Geralmente variam entre 0 e 1

$$S_j = \frac{a}{a+b+c} \quad \text{Jaccard}$$

$$S_s = \frac{2a}{2a+b+c} \quad \text{Sorensen}$$

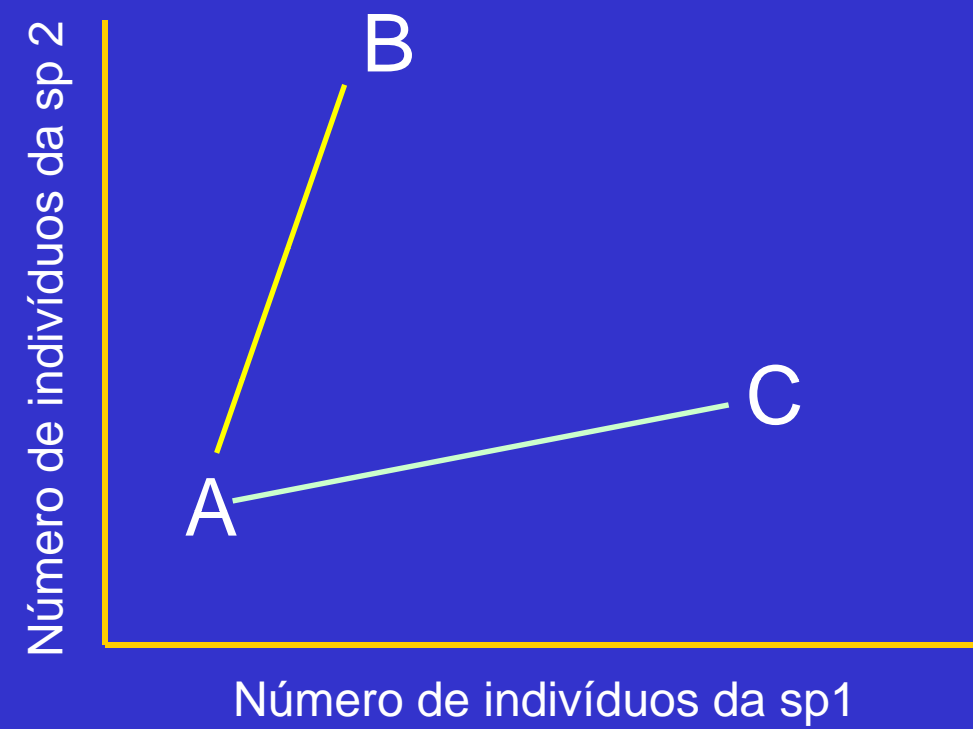
$$S_{sm} = \frac{a+d}{a+b+c+d} \quad \text{Concordância simples (“simple matching”)}$$

Índices de similaridade: Quantitativos: também abundância



## Índices de similaridade: Quantitativos: também abundância

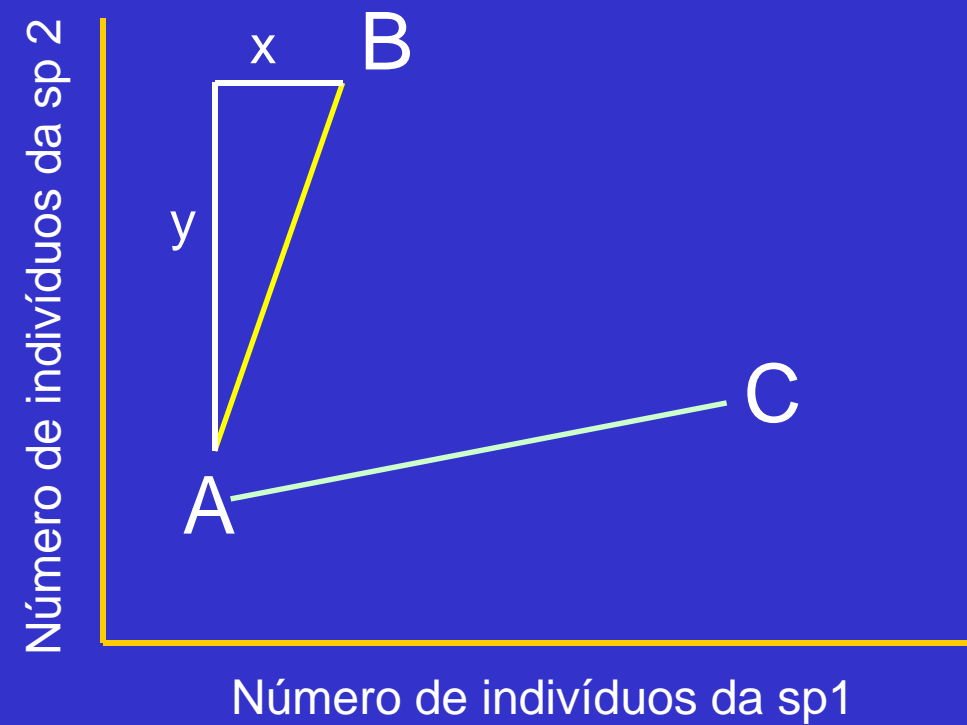
Distância Euclidiana



Índices de similaridade: Quantitativos: também abundância

Distância Euclidiana

$$D_{jk} = \sqrt{x^2 + y^2}$$

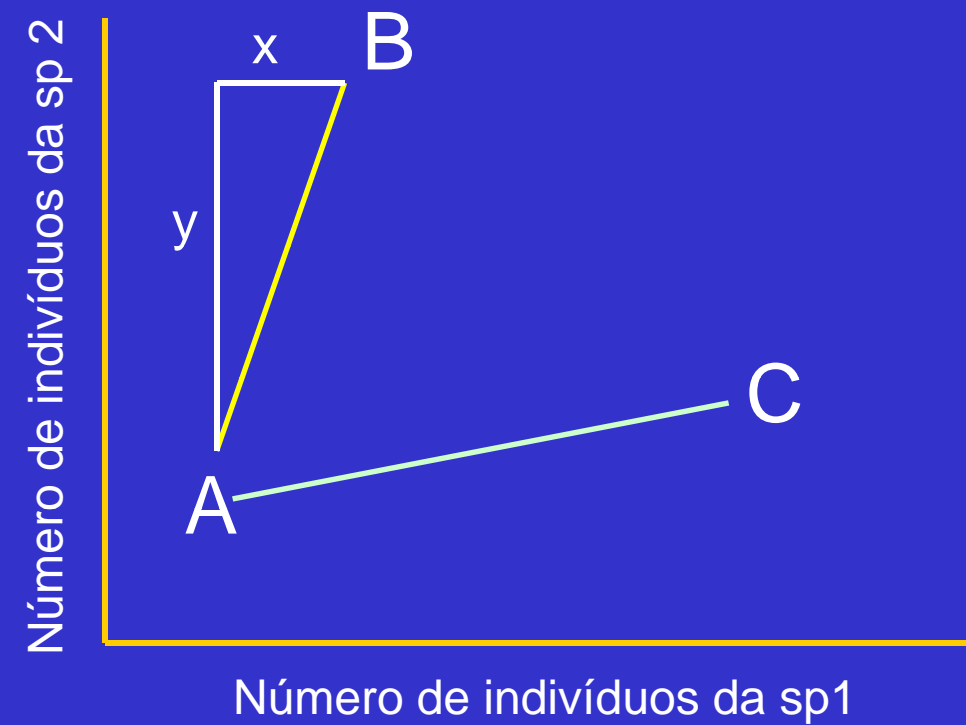


Índices de similaridade: Quantitativos: também abundância

Distância Euclidiana

$$D_{jk} = \sqrt{x^2 + y^2}$$

E para 3, 4, 5 etc espécies?

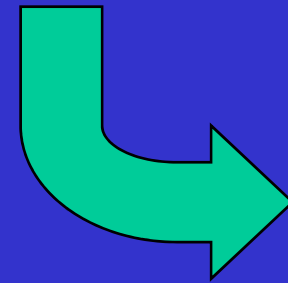




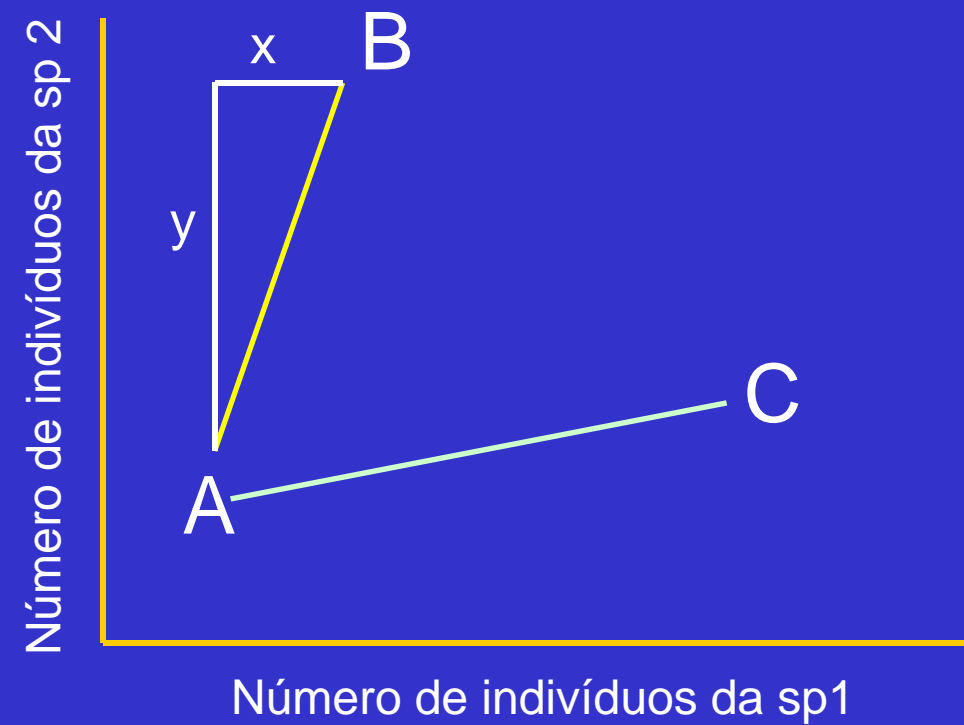
Índices de similaridade: Quantitativos: também abundância

Distância Euclidiana

$$D_{jk} = \sqrt{x^2 + y^2}$$



$$D_{jk} = \sqrt{\sum_{i=1}^s (X_{ij} - X_{ik})^2}$$



## Índices de similaridade: Quantitativos: também abundância

Distância Euclidiana

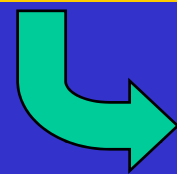
$$\Delta_{jk} = \sqrt{\sum_{i=1}^s (X_{ij} - X_{ik})^2}$$

$$d_{jk} = \sqrt{\frac{\Delta_{jk}^2}{s}}$$

varia de 0 ao  $\infty$

Distância Manhattan  
("city block")

$$\Delta_{jk} = \sum_{i=1}^s |X_{ij} - X_{ik}|$$



Distância Bray-Curtis

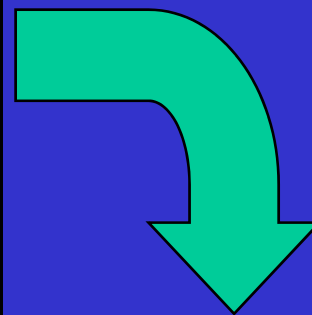
$$d_m(j, k) = \frac{\sum_{i=1}^s |X_{ij} - X_{ik}|}{\sum_{i=1}^s (X_{ij} + X_{ik})}$$

varia de 0 a 1

## Índices de similaridade

Dados originais

	sp 1	sp 2	sp 3	sp 4	sp 5
a1	5	2	5	2	1
a2	0	1	3	2	1
a3	2	1	3	2	1
b1	5	20	6	5	5
b2	12	19	4	7	11
b3	11	21	5	7	10

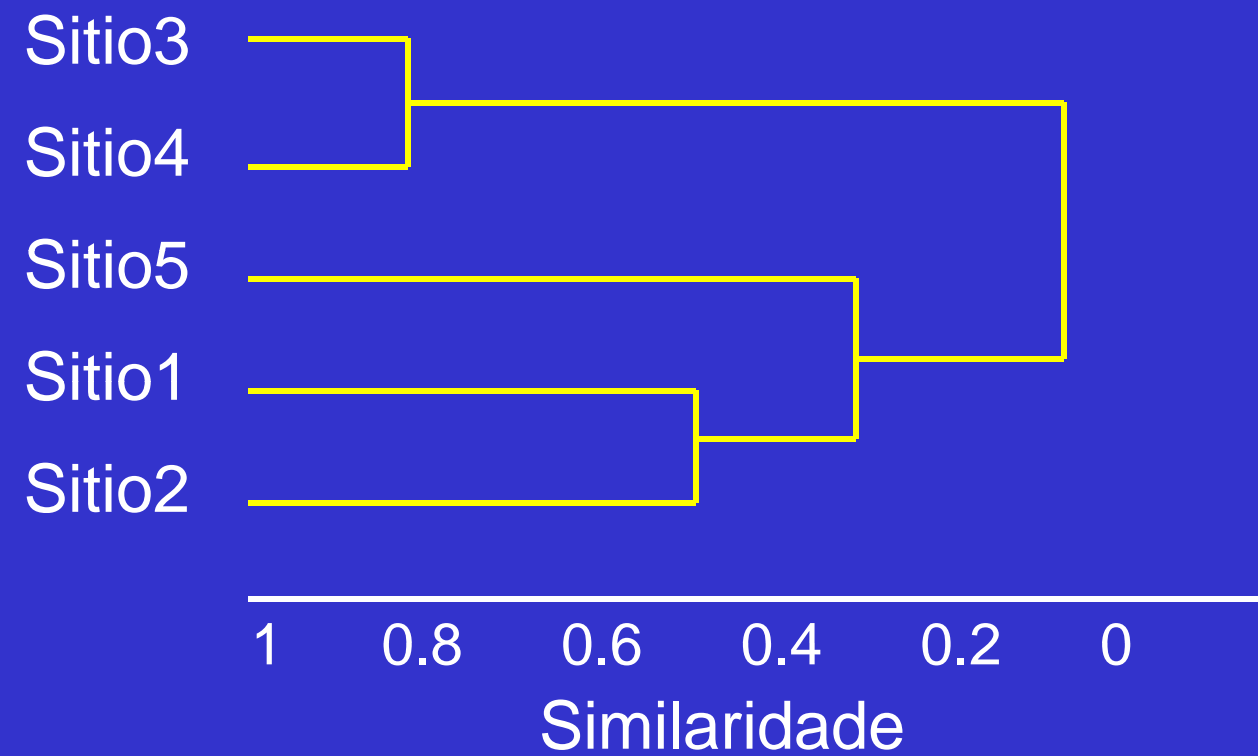


Matriz de distância  
(Bray-Curtis)

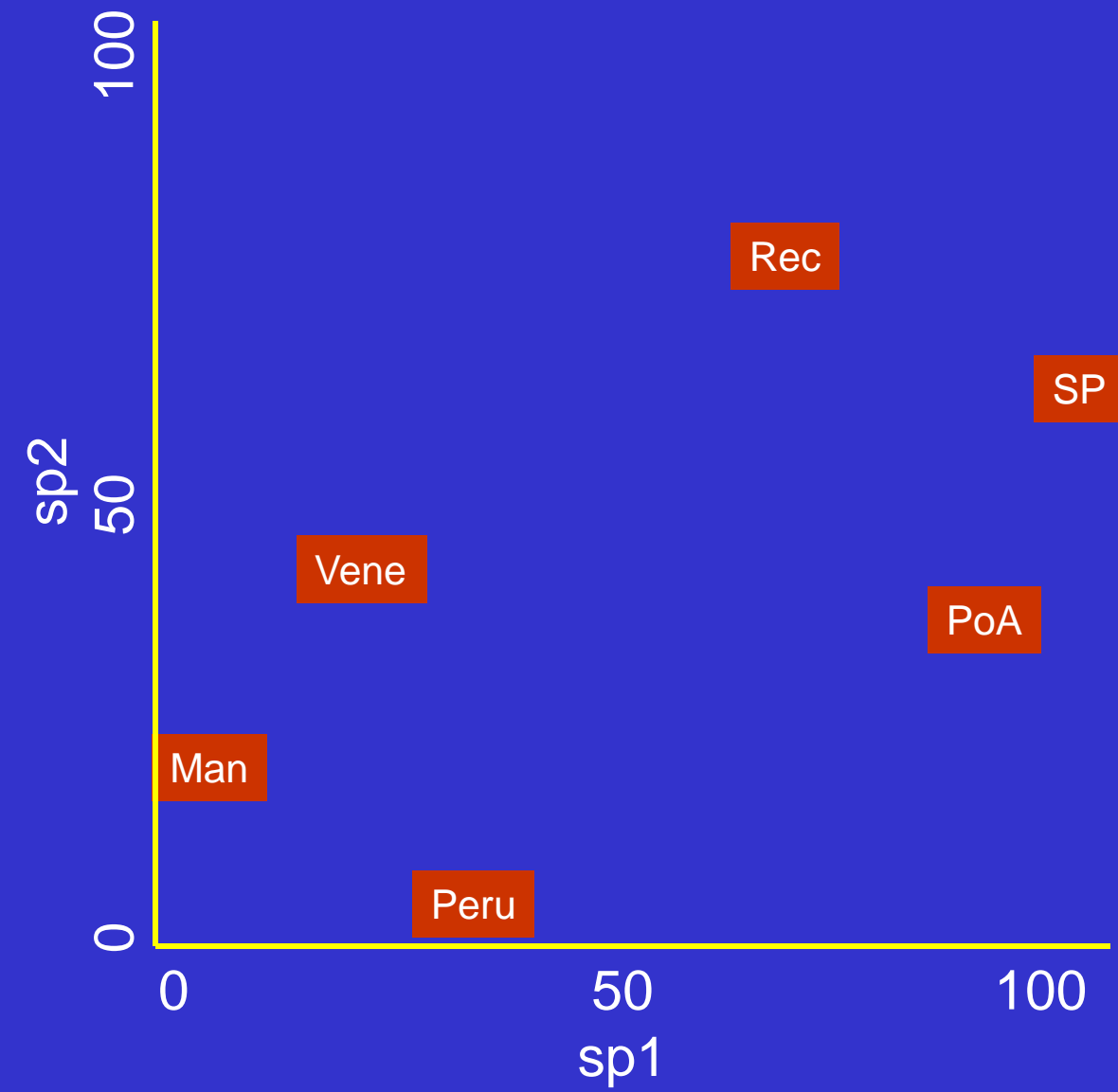
	a1	a2	a3	b1	b2	b3
a1	0					
a2	0,364	0				
a3	0,250	0,125	0			
b1	0,464	0,708	0,640	0		
b2	0,588	0,766	0,709	0,191	0	
b3	0,565	0,770	0,714	0,157	0,047	0

## Análise Multivariada Exploratória: Classificação

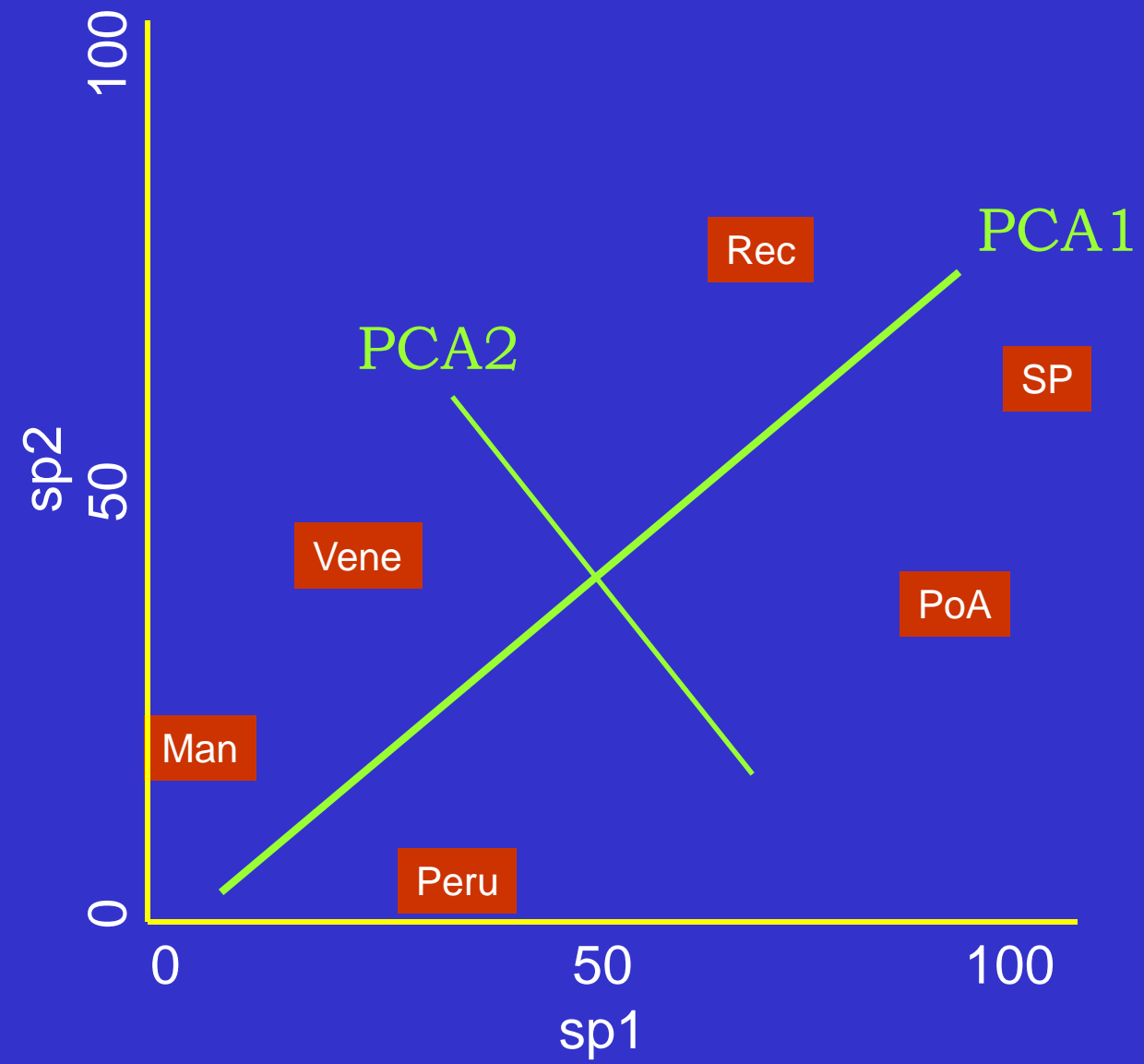
- Matriz de similaridade entre objetos (e.g. Jaccard, Bray-Curtis)
- Método de aglomeração (e.g. UPGMA)
- Resultado: Dendrograma



## Análise Multivariada Exploratória: Ordenação



## Análise Multivariada Exploratória: Ordenação



## Análise Multivariada Exploratória: Ordenação



## Análise Multivariada Exploratória: Ordenação

