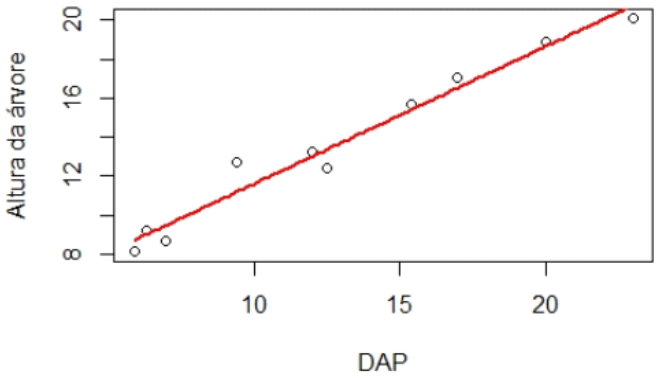
Aula 04 - Medidas de Associação

quinta-feira, 3 de setembro de 2020 16:53

ASSOCIAÇÃO ENTRE VARIÁVEIS QUANTITATIVAS

Ajuste de uma reta

- Equação matemática linear;
- Representação de um conjunto de dados;
- Relação de causa e efeito;
- Interpolação e Extrapolação.



Variáveis:

X $\Rightarrow$ Variável **Independente**

Y $\Rightarrow$ Variável **Dependente**

Equação matemática:

$$y = \alpha + \beta x,$$

---

em que α representa o intercepto e β o coeficiente angular.

Interpretação prática do parâmetro β: o quanto varia a resposta y para um acréscimo de uma unidade na variável x.

Modelo Estatístico

$$y = \alpha + \beta x + \epsilon$$

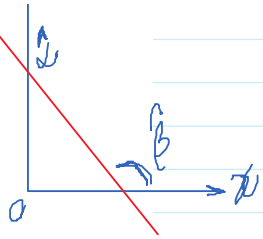
Reta ajustada:

Reta ajustada:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x,$$

ou

$$\hat{y} = a + bx$$



em que $\hat{\alpha}$ (ou a) e $\hat{\beta}$ (ou b) são as estimativas dos parâmetros α e β.

Estimativas pelo método dos mínimos quadrados:

$$\hat{\beta} = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

e

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} y_i - \hat{\beta}\sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\beta}\bar{x},$$

em que n corresponde ao tamanho da amostra.

Exemplo: Considerando-se o exemplo de altura da árvore (Y) e o diâmetro a altura do peito (X):

$n = 10$

Tabela: Etapas intermediárias

Observação	x	y	x^2	y^2	xy
1	5,9	8,1	34,8	65,9	47,9
2	6,3	9,2	39,7	84,6	57,9
3	7,0	8,7	49,0	74,9	60,6
4	9,4	12,7	88,4	161,7	119,5
5	12,0	13,2	144,0	174,8	158,6
6	12,5	12,4	156,2	154,0	155,1
7	15,4	15,7	237,2	246,2	241,6
8	17,0	17,0	289,0	290,0	289,5
9	20,0	18,9	400,0	357,4	378,1
10	23,0	20,1	529,0	402,6	461,5
Total	128,5	136,0	1967,23	2011,94	1970,51

$\sum_{i=1}^{10} x_i$

$\sum_{i=1}^{10} x_i \cdot y_i = x_1 y_1 + x_2 y_2 + \ldots + x_{10} y_{10}$

$$\hat{\beta} = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\hat{\beta} = \frac{10 \cdot 1970{,}51 - 128{,}5 \cdot 136{,}0}{10 \cdot 1967{,}23 - (128{,}5)^2}$$

$$\hat{\beta} \simeq 0{,}71$$

Na calculadora:

$(10 \times 1970{,}51 - 128{,}5 \times 136) \div (10 \times 1967{,}23 - 128{,}5^2) =$

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} y_i - \hat{\beta} \sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\beta}\bar{x},$$

$\bar{x} = \frac{128{,}5}{10}$ $\bar{x} = 12{,}85$

$\bar{y} = \frac{136}{10}$ $\bar{y} = 13{,}6$

$\hat{\alpha} = 13{,}6 - 0{,}71 \cdot 12{,}85$

$\hat{\alpha} \approx 4{,}48$

$$\hat{y} = \hat{\alpha} + \hat{\beta}x.$$

$\hat{y} = 4{,}48 + 0{,}71x$

Na calculadora:
$13{,}6 - 0{,}71 \times 12{,}85 =$

Conferindo...

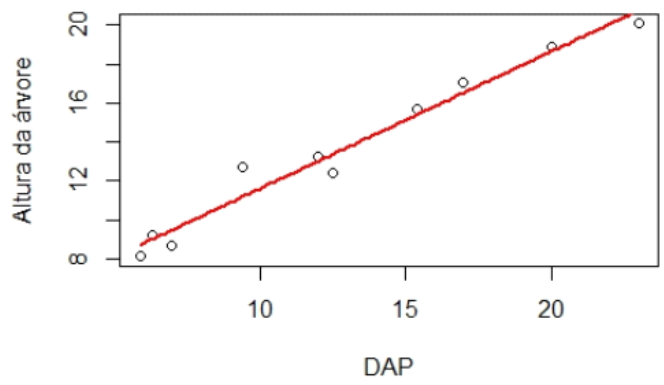
$$\hat{\beta} = \frac{10(1970{,}51) - (128{,}5)(136{,}0)}{10(1967{,}23) - (128{,}5)^2} = 0{,}7053$$

$$\hat{\alpha} = \frac{136{,}0 - 0{,}7053(128{,}5)}{10} = 4{,}5368$$

Exemplo: Considerando-se o exemplo de altura da árvore (Y) e o diâmetro a altura do peito (X):

Reta ajustada

$$\hat{y}_i = 4{,}5368 + 0{,}7053x_i.$$



Verificação da qualidade do ajuste

RESÍDUO:

Diferença entre o valor observado (y_i) e o valor predito ($\hat{y}_i$), para um determinado valor x_i:

$$e_i = y_i - \hat{y}_i.$$

$$\hat{y}_i = 4,5368 + 0,7053 x_i.$$

Tabela: Etapas intermediárias

Observação	x	y	$\hat{y}$	x^2	y^2	xy	e
1	5,9	8,1	8,7	34,8	65,9	47,9	−0,6
2	6,3	9,2	9,0	39,7	84,6	57,9	0,2
3	7,0	8,7	9,5	49,0	74,9	60,6	−0,8
4	9,4	12,7	11,2	88,4	161,7	119,5	
5	12,0	13,2	13,0	144,0	174,8	158,6	
6	12,5	12,4	13,4	156,2	154,0	155,1	
7	15,4	15,7	15,4	237,2	246,2	241,6	
8	17,0	17,0	16,5	289,0	290,0	289,5	
9	20,0	18,9	18,6	400,0	357,4	378,1	
10	23,0	20,1	18,7	529,0	402,6	461,5	
Total	128,5	136,0		1967,23	2011,94	1970,51	

Estime a altura de uma árvore cujo DAP é igual a 25 cm.

$\hat{y} = 4,5368 + 0,7053 \cdot 25$

$\hat{y} = 22,17$ m

O primeiro resíduo (simples) é dado por:

$$e_1 = 8,1 - (4,5368 + 0,7053 \times 5,9) = 8,1 - 8,7 = -0,6$$

Modelo bem ajustado:

é aquele que apresenta resíduos pequenos.

Resíduo simples $\Rightarrow$ depende das unidades de medida

$\Downarrow$

Resíduos Padronizados $\Rightarrow z_i = \frac{e_i}{\sqrt{\sum_{i=1}^{n} e_i^2/(n-2)}}$

Na prática: erro pequeno $\Rightarrow$ resíduo padronizado entre -2 e 2.

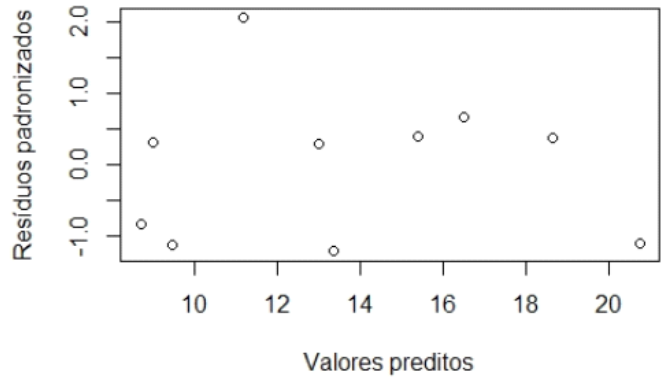


Figura: Gráfico dos valores preditos *versus* resíduos padronizados

Ideal: Gráfico sem padrão!

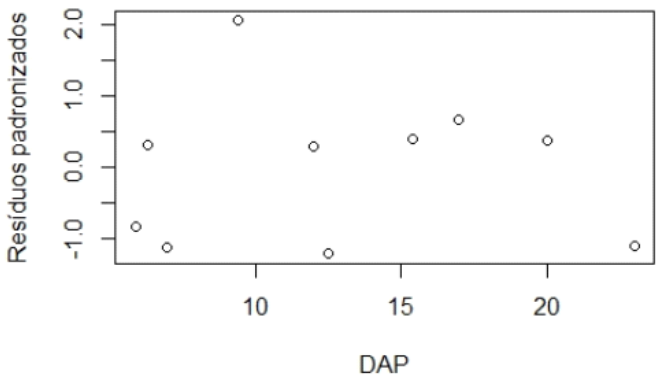


Figura: Gráfico dos valores de DAP *versus* resíduos padronizados

Ideal: Gráfico sem padrão!

ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS

Tabelas de contingência

Duas variáveis categorizadas:

- A: com s categorias $\Rightarrow A_1,\ A_2, \ldots, A_s$;
- B: com r categorias $\Rightarrow B_1,\ B_2, \ldots, B_r$.

Tabela: Distribuição conjunta de frequências das variáveis A e B, observados em n elementos

	B				
A	B_1	B_2	$\ldots$	B_r	Total
A_1	n_{11}	n_{12}	$\ldots$	n_{1r}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\ldots$	n_{2r}	$n_{2\cdot}$
$\cdot$	$\cdot$	$\cdot$	$\ldots$	$\cdot$	$\cdot$
A_s	n_{s1}	n_{s2}	$\ldots$	n_{sr}	$n_{s\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\ldots$	$n_{\cdot r}$	n

Exemplo: Na Tabela a seguir apresentamos a distribuição conjunta do comportamento de 59 capivaras com relação ao ambiente.

Tabela: Distribuição conjunta das frequências das variáveis comportamento e ambiente para 59 capivaras

	Comportamento		
Ambiente	Agressivo	Não Agressivo	Total
Restrito	22	5	27
Amplo	20	12	32
Total	42	17	59

Tabela: Distribuição conjunta das frequências das variáveis comportamento e ambiente para 59 capivaras

	Comportamento		
Ambiente	Agressivo	Não Agressivo	Total
Restrito	22	5	**27**
Amplo	20	12	**32**
Total	**42**	**17**	**59**

- **Distribuição marginal da variável Ambiente**;
- **Distribuição marginal da variável Comportamento**;

Totais marginais diferentes $\Rightarrow$ difícil visualização da associação

$\Downarrow$

Porcentagens: linhas, colunas, total

Tabela: Distribuição conjunta das frequências das variáveis comportamento e ambiente para 59 capivaras

	Comportamento		
Ambiente	Agressivo	Não Agressivo	Total
Restrito	81,5	18,5	**100**
Amplo	62,5	37,5	**100**
Total	71,2	28,8	**100**

Exemplo: Os dados da tabela a seguir têm por objetivo verificar se os caracteres ciclo (Tardio e Precoce) e virescência (Normal e Virescente), de uma progênie da espécie "X", segregam de forma independente

Tabela: Contagem de plantas segregando para dois caracteres

	Virescência		
Ciclo	Normal	Virescente	Total
Tardio	3470	910	4380
Precoce	1030	290	1320
Total	4500	1200	5700

Fonte: Andrade e Ogliari, 2007

A virescência caracteriza-se pelo desenvolvimento de cloroplastos nas pétalas, resultando no aparecimento de flores verdes

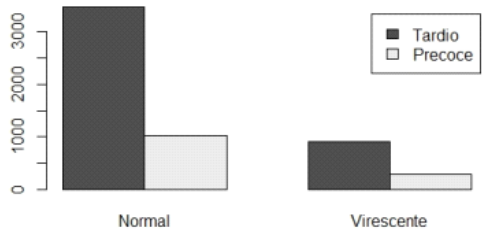


Figura: Contagem de plantas segregando para dois caracteres numa progênie da espécie "X"

Perfis-coluna

Tabela: Contagem de plantas segregando para dois caracteres

	Virescência		
Ciclo	Normal	Virescente	Total
Tardio	77,11%	75,83%	76,84%
Precoce	22,89%	24,17%	23,16%
Total	100%	100%	100%

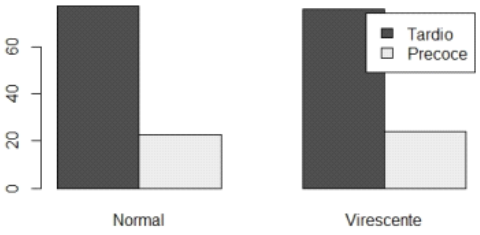


Figura: Distribuição de precocidade segundo a virescência

Exemplo: Os dados da tabela a seguir referem-se às coletas de insetos em armadilhas adesivas de duas cores, em que os indivíduos coletados de uma determinada espécie foram sexados, tendo como objetivo verificar se havia influência da cor da armadilha sobre a atração de machos e fêmeas dessa espécie.

Tabela: Números de insetos coletados em armadilhas adesivas e sexados.

Armadilha	Machos	Fêmeas	Totais
Alaranjada	246	17	263
Amarela	458	32	490
Totais	704	49	753

Fonte: Cordeiro, G.M.; Demétrio, C.G.B. Modelos Lineares Generalizados e Extensões

Tabela: Números de insetos coletados em armadilhas adesivas e sexados.

Armadilha	Machos	Fêmeas	Totais
Alaranjada	246 (32,7%)	17 (2,3%)	263 (36,0%)
Amarela	458 (60,8%)	32 (4,2%)	490 (65,0%)
Totais	704 (93,5%)	49 (6,5%)	753 (100%)

Fonte: Cordeiro, G.M.; Demétrio, C.G.B. Modelos Lineares Generalizados e Extensões

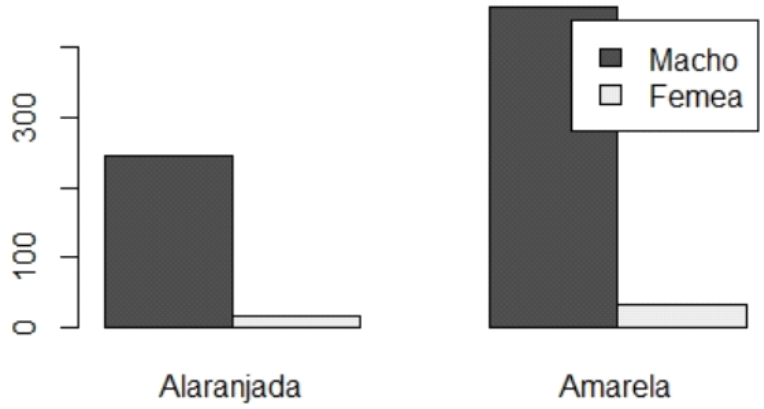


Figura: Gráfico de distribuição conjunta da cor da armadilha e sexagem dos insetos

Exemplo: A tabela a seguir refere-se ao número de pássaros de uma particular espécie, classificados de acordo com o local da floresta onde se alimentam, para duas estações do ano.

Estação do ano	Local da floresta			
	Árvores	Arbusto	Chão	Total
Primavera	30	20	9	59
Outono	13	22	26	61
Total	43	42	35	120

Fonte: Andrade e Ogliari, 2007

Perfil linha:

Estação do ano	Local da floresta			
	Árvores	Arbusto	Chão	Total
Primavera	30 (50,8%)	20 (33,9%)	9 (15,3%)	59 (100%)
Outono	13 (21,3%)	22 (36,1%)	26 (42,6%)	61 (100%)
Total	43 (35,8%)	42 (35,0%)	35 (29,2%)	120 (100%)

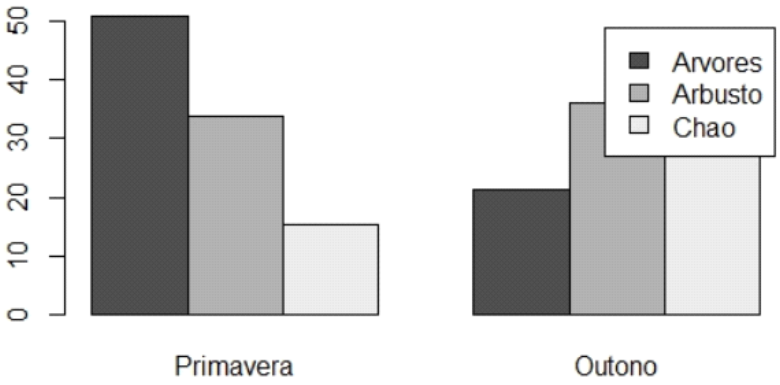


Figura: Associação entre local e estação

Estatística χ^2

Comparar a frequência observada com a frequência esperada!

Frequência esperada

Considerando-se a independência entre as variáveis,

$$fe_{ij} = \frac{n_{i.} n_{.j}}{n_{..}}.$$

$n_{i.}$ → total linha
$n_{.j}$ → total coluna
$n_{..}$ → total

Estatística χ^2

$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{(n_{ij} - fe_{ij})^2}{fe_{ij}}$$

i → linha
j → coluna

Considerando o exemplo do hábito alimentar dos pássaros e local

Estação	Local da floresta			
do ano	Árvores	Arbusto	Chão	Total

Estação do ano	Local da floresta			
	Árvores	Arbusto	Chão	Total
Primavera	30	20	9	59
Outono	13	22	26	61
Total	43	42	35	120

Fonte: Andrade e Ogliari, 2007

$$fe_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}.$$

$\frac{59 \cdot 43}{120} = 21{,}14$ $\quad \frac{59 \cdot 42}{120} = 20{,}65$ $\quad \frac{59 \cdot 35}{120} =$

... e assim por diante!

Considerando o exemplo do hábito alimentar dos pássaros e local

⇒ Frequências esperadas

Estação do ano	Local da floresta			
	Árvores	Arbusto	Chão	Total
Primavera	30 (21,14)	20 (20,65)	9 (17,21)	59
Outono	13 (21,86)	22 (21,35)	26 (17,79)	61
Total	43	42	35	120

⇒ Diferenças

Estação do ano	Local da floresta		
	Árvores	Arbusto	Chão
Primavera	8,86	-0,65	-8,21
Outono	-8,86	0,65	8,21

$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{(n_{ij} - fe_{ij})^2}{fe_{ij}}$$

$\frac{(8{,}86)^2}{21{,}14} = 3{,}7133$ $\quad \frac{(-0{,}65)^2}{20{,}65} = 0{,}0205$

Estatística:

$$\begin{aligned}
\chi^2 &= \frac{(8,86)^2}{21,14} + \frac{(-0,65)^2}{20,65} + \frac{(-8,21)^2}{17,21} + \frac{(-8,86)^2}{21,86} + \frac{(0,65)^2}{21,35} + \frac{(8,21)^2}{17,79} \\
&= 3,7133 + 0,0205 + 3,9166 + 3,5910 + 0,0198 + 3,7889 \\
&= 15,0501.
\end{aligned}$$

Coeficiente de Contingência (Pearson)

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

No exemplo:

$$C = \sqrt{\frac{15,0501}{15,0501 + 120}} = 0,3338$$

$\Downarrow$

Existe associação!

Forte ou fraca?

Depende o limite superior do coeficiente de contingência!!!

Limite superior de $C \Rightarrow \sqrt{(t-1)/t}$, em que t é o mínimo entre o número de linhas e o número de colunas.

Modificação do coeficiente de contingência

$$C^* = \frac{C}{\sqrt{(t-1)/t}}$$

No exemplo: $t = 2$,

$$C^* = \frac{0,3338}{} = 0,4721$$

$$C^* = \frac{0,3338}{\sqrt{(2-1)/2}} = 0,4721$$

Existe uma associação entre as duas variáveis mas é de fraco à moderada

ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS E QUANTITATIVAS

- Análise da variável quantitativa dentro de cada nível da variável qualitativa:
  - medidas resumo → quartis (Q_1, Q_2, Q_3)
  - histogramas
  - box plots
  - diagrama de ramos e folhas

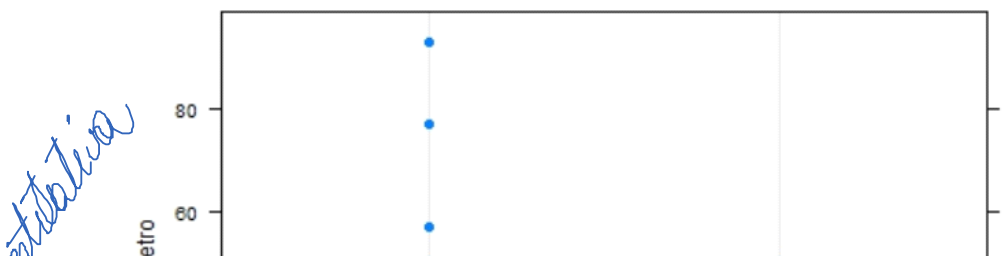
Q_2 = mediana

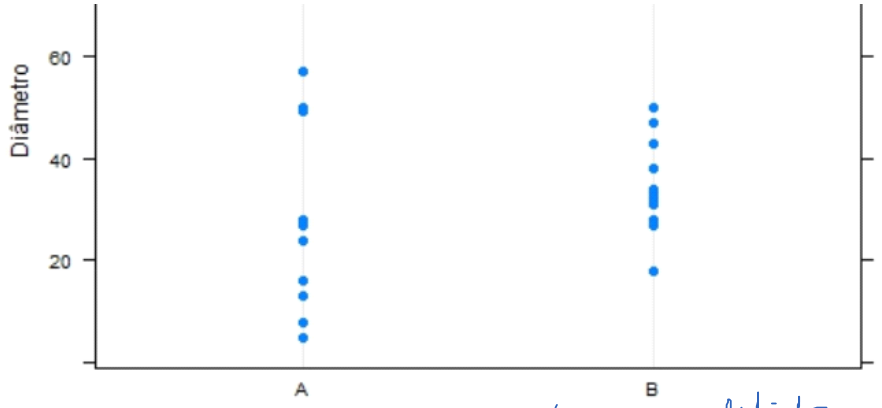
Exemplo: O dados apresentados a seguir correspondem à variável diâmetro da árvore avaliada em duas florestas (Floresta A e Floresta B).

Tabela: Diâmetro das árvores

Floresta A	16	50	13	8	5	77	93
	27	57	28	24	16	49	
Floresta B	38	43	32	18	47	33	38
	27	50	34	34	31	28	

Exemplo:O dados apresentados a seguir correspondem à variável diâmetro da árvore avaliada em duas florestas (Floresta A e Floresta B).





Exemplo: Os dados apresentados na tabela a seguir são referentes ao volume de madeira por árvore de *Eucaliptus camaldulensis*, em $m^3 10^{-4}$. São apresentados os dados de 5 progênies avaliadas.

Tabela: Volume de madeira, em $m^3 10^{-4}$

A	B	C	D	E
212	108	63	175	133
206	194	77	239	106
224	163	100	100	185
289	111	99	104	136
324	236	68	256	147
219	146	76	267	210

Exemplo: Os dados apresentados são referentes ao volume de madeira por árvore de *Eucaliptus camaldulensis*, em $m^3 10^{-4}$. São apresentados os dados de 5 progênies avaliadas.

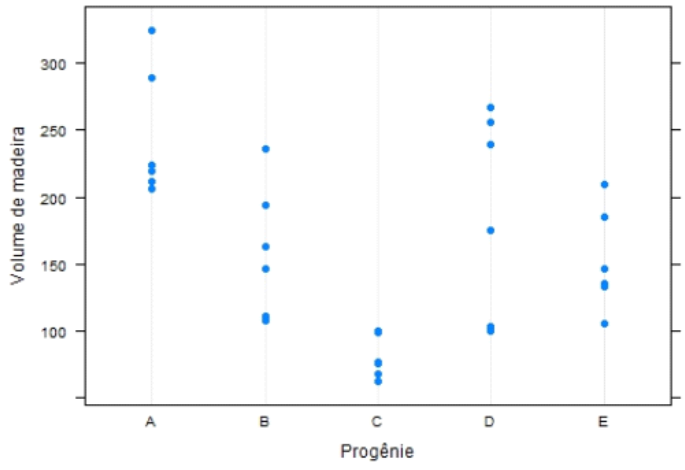


Figura: Gráfico de pontos para a variável volume de madeira por progênie

Exemplo: Os dados apresentados são referentes ao volume de madeira por árvore de *Eucaliptus camaldulensis*, em $m^3 10^{-4}$. São apresentados os dados de 5 progênies avaliadas.

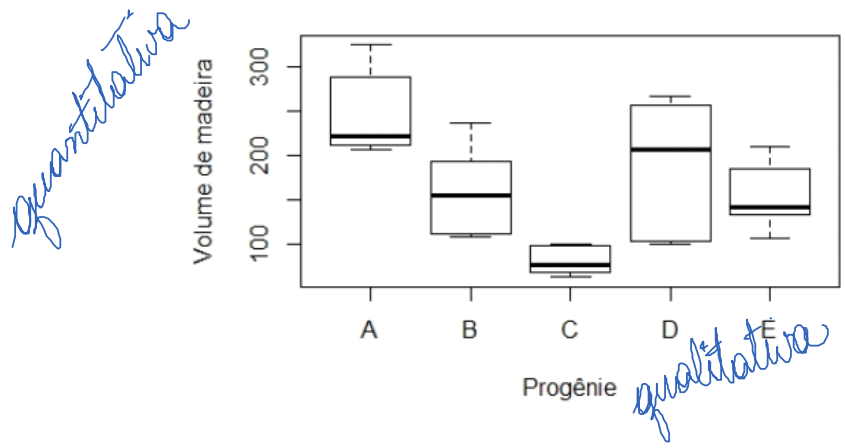


Figura: Box plot para a variável volume de madeira por progênie

Exercícios

Exercício 1:
Considerando o consumo individual diário de proteínas de origem animal, em gramas, e o coeficiente de natalidade, em 14 países.

País	Consumo diário	Coeficiente de Natalidade			
Formosa	4,7	45,6			
Malásia	7,5	39,7			
Índia	8,7	33,0			
Japão	9,7	27,0			
Iugoslávia	11,2	25,9			
Grécia	15,2	23,5			
Itália	15,2	23,4			
Bulgária	16,8	22,2			
Alemanha	37,3	20,0			
Irlanda	46,7	19,1			
Dinamarca	56,1	18,3			
Austrália	59,9	18,0			
Estados Unidos	61,4	17,9			
Suécia	62,6	15,0			

Fonte: VIEIRA, Sônia. **Introdução à Bioestatística**. Rio de Janeiro: Campus, p.80, 1989.

Calcule o coeficiente de correlação de Pearson e escreva o que se pode dizer sobre isso, encontre a equação da reta e estime o coeficiente de natalidade para um país cujo consumo diário de proteína seja igual a 65,0.

Exercício 2: *Considerando a prevalência de anemia em catadores de material reciclável no município de Santos em julho de 2005, e o seu consumo de ovos, verifique se existe associação entre essas variáveis. O que se pode dizer com os valores obtidos.*

Consumo de Ovos	Anemia		
	Sem anemia	**Com anemia**	**Total**
Nunca	42	32	**74**
1x / semana	25	19	**44**
2 a 3x /semana	64	30	**94**
Diariamente	18	11	**29**
Total	**149**	**92**	**241**

Fonte: PEREZ JUNIOR, Valmir, *et all*. Anemia em catadores de material reciclável que utilizam carrinho de propulsão humana no município de Santos. Santos: Núcleo de Estudos Epidemiológicos, Universidade Católica de Santos, **SciELO - Scientific Electronic Library Online**. 2010. Disponível em: <https://www.scielosp.org/article/rbepid/2010.v13n2/326-336/>. Acesso em: 10 set. 2020.