Estatística Descritiva (III)

Associação entre Variáveis

Associação entre variáveis qualitativas

Tabelas de Contingência

Podemos construir tabelas de frequências conjuntas (***tabelas de contingência***), relacionando duas variáveis qualitativas.

Exemplo 1: Dados ***CEA06P24***, do projeto ***Caracterização Postural de Crianças de 7 e 8 anos das Escolas Municipais da Cidade de Amparo/SP***

- Estudo realizado pelo Departamento de Fisioterapia, Fonoaudiologia e Terapia Ocupacional da Faculdade de Medicina da *USP*;
- Ano de realização: 2006;
- Finalidade: mestrado;
- Análise estatística: Centro de Estatística Aplicada (*CEA*), *IME-USP*.

Objetivo: caracterizar a postura de crianças da cidade de Amparo/SP, entre sete e oito anos, de ambos os sexos

Amostra: 230 crianças com 7 e 8 anos.

Algumas variáveis coletadas:

- **Sexo** (feminino, masculino);
- **Peso** (em kg);
- **Altura** (em metros);
- **Índice de Massa Corpórea** – IMC (em kg/m^2);
- **Atividade Física** (em horas/semana);
- **Tipo de Mochila Utilizada** (com fixação escapular, com fixação lateral, de carrinho, outros);
- **Dominância** (destro, canhoto);
- **Região da escola**.

Algumas variáveis relativas à postura:

- **Postura do ombro no plano frontal** (cm): avaliado pelo desnível entre os ombros, conforme figura; anota-se a diferença Direito-Esquerdo;

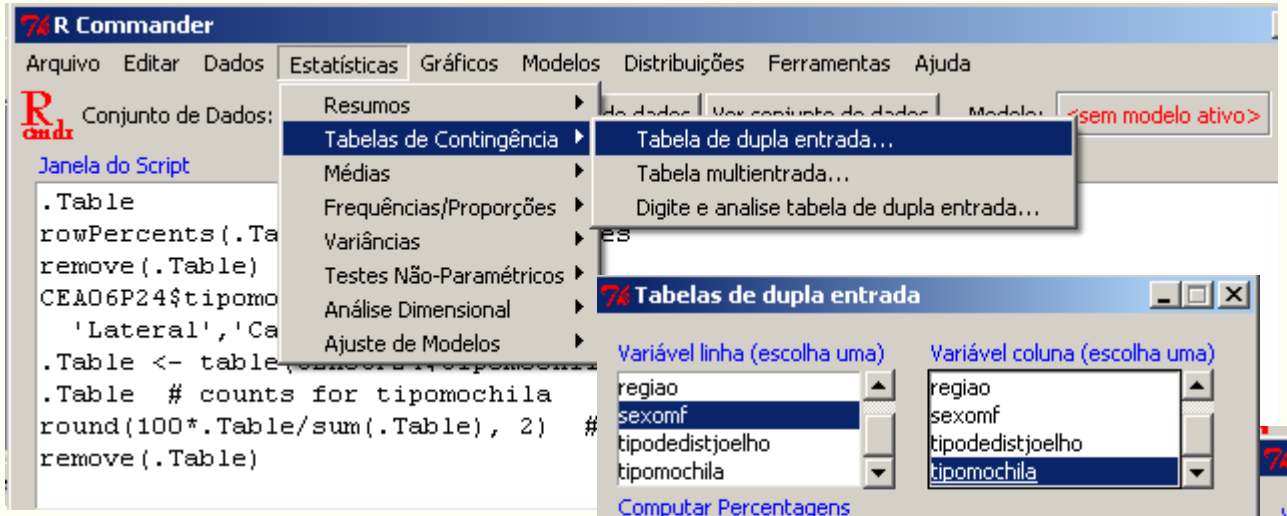
- **Lordose Lombar** (graus): avaliada pelo aumento e diminuição (retificação) da lordose lombar, medindo-se o ângulo formado entre os pontos de maior convexidade da coluna torácica e da região glútea e o ponto de maior concavidade da coluna lombar, em ambos lados (Direito e Esquerdo).

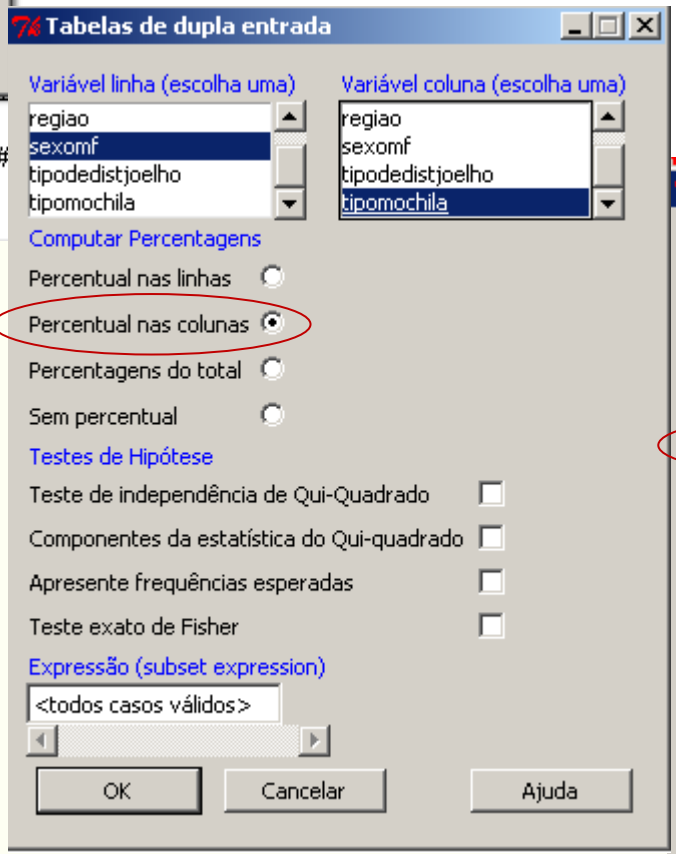
- **Lado da escoliose**

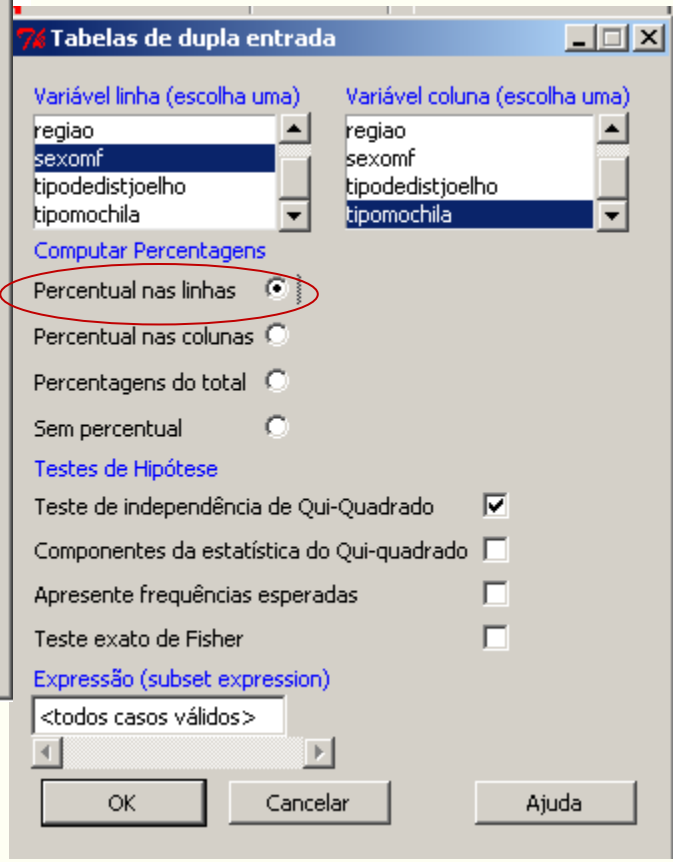
A) Há indícios de associação entre Lado da escoliose e Tipo de mochila?

Tipo de Mochila	Lado da Escoliose			Total
	Ausente	Direito	Esquerdo	
Carrinho	8	37	35	80
Escapular	16	35	72	123
Lateral	2	10	11	23
Total	26	82	118	226

Qual é o significado dos valores desta tabela?







Criando tabelas de contingência

No ***R***:

- **Dados → Importar arquivos de dados →**

→ de conjunto de dados do Excel, Access ou dBase...

(Defina o nome do conjunto de dados: *dados*)

- **Estatísticas → Tabelas de Contingência → Tabelas de dupla entrada**

(Variável linha: *tipomochila* ; Variável coluna: *escollado*)

Saída editada do *software R*

	Lado da escoliose			
Tipo de mochila	**Ausente**	**Direito**	**Esquerdo**	**Total**
Carrinho	**8**	**37**	**35**	**80**
Escapular	**16**	**35**	**72**	**123**
Lateral	**2**	**10**	**11**	**23**
Total	**26**	**82**	**118**	**226**

Verificar associação através da:

- porcentagem segundo as colunas, ou
- porcentagem segundo as linhas.

	Lado da Escoliose			
Tipo de Mochila	Ausente	Direito	Esquerdo	Total
Carrinho	10,0%	46,2%	43,8%	100,0%
Escapular	13,0%	28,5%	58,5%	100,0%
Lateral	8,7%	43,5%	47,8%	100,0%
Total	11,5%	36,3%	52,2%	100,0%

→ Como concluir? Será que o Tipo de Mochila utilizada influencia o Lado da Escoliose (caso tenha) de uma criança

Comparando as porcentagens de cada uma das linhas, observamos uma diferença com relação à porcentagem total. Aparentemente, há influência do tipo de mochila utilizada no lado de ocorrência da escoliose.

• **Estatísticas → Tabelas de Contingência → Tabelas de dupla entrada**

(Variável linha: *tipomochila*; Variável coluna: *escollado*)

Marcar opção *Percentual nas linhas*

Saída editada do *software R*

	Lado escoliose			
Tipo de mochila	**Ausente**	**Direito**	**Esquerdo**	**Total**
Carrinho	**10,0**	**46,2**	**43,8**	**100**
Escapular	**13,0**	**28,5**	**58,5**	**100**
Lateral	**8,7**	**43,5**	**47,8**	**100**
Total	**11,5**	**36,3**	**52,2**	**100**

B) Será que existe relação entre o Sexo das crianças e o Tipo de Mochila utilizada por elas?

Sexo	Tipo de Mochila			Total
	Carrinho	Escapular	Lateral	
Feminino	53 (41,4%)	59 (46,1%)	16 (12,5%)	128 (100%)
Masculino	27 (27,6%)	64 (65,3%)	7 (7,1%)	98 (100%)
Total	80 (35,4%)	123 (54,4%)	23 (10,2%)	226 (100%)

Parece existir relação entre Sexo e Tipo de Mochila.
A maioria dos meninos (65,3%) prefere mochila escapular.
Por outro lado, a preferência da maioria das meninas é dividida entre mochila escapular (46,1%) e carrinho (41,4%).

Associação entre variáveis quantitativas

Correlação e Regressão

Objetivo

Estudar a relação entre duas variáveis quantitativas.

Exemplos:

Idade e altura das crianças

Tempo de prática de esportes e ritmo cardíaco

Tempo de estudo e nota na prova

Taxa de desemprego e taxa de criminalidade

Expectativa de vida e taxa de analfabetismo

Investigaremos a presença ou ausência de relação linear sob dois pontos de vista:

a) Quantificando a força dessa relação: correlação.

b) Explicitando a forma dessa relação: regressão.

Representação gráfica de duas variáveis quantitativas:

Diagrama de dispersão

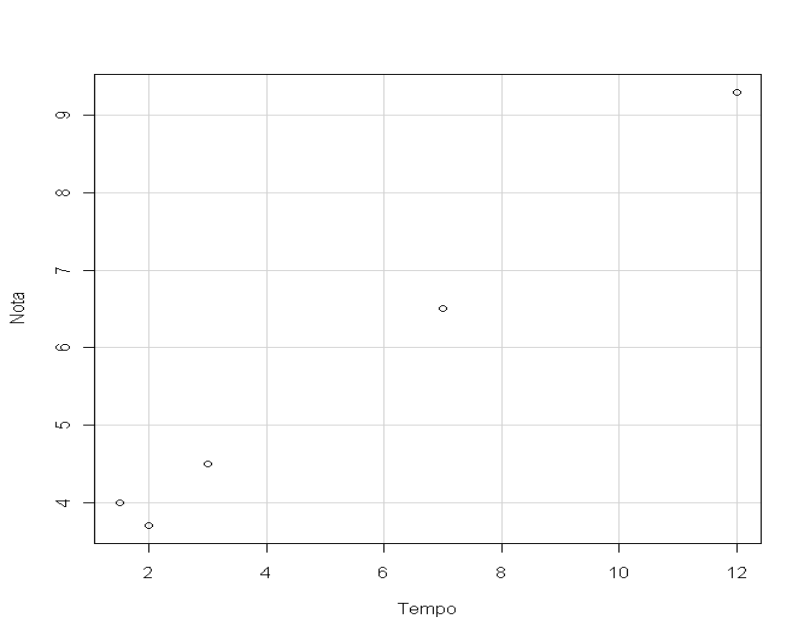
Exemplo 2: nota da prova e tempo de estudo

X : tempo de estudo (em horas)

Y : nota da prova

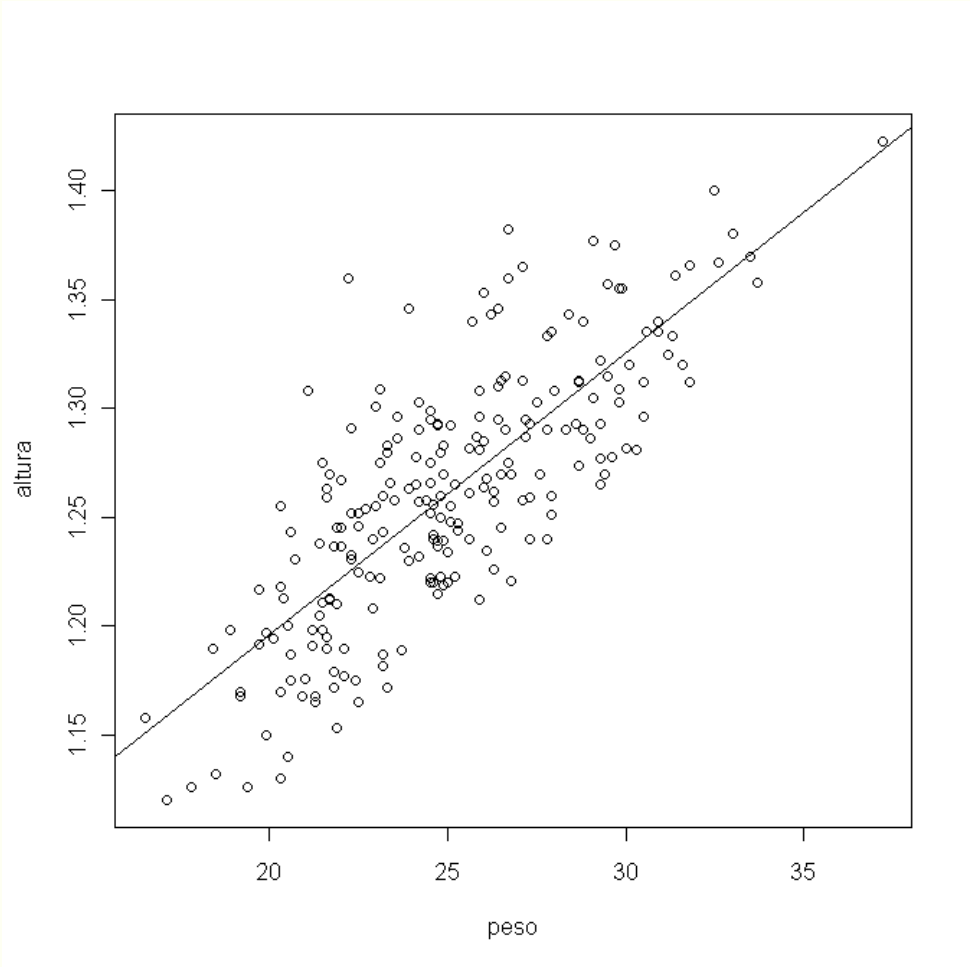
Pares de observações (X_i , Y_i) para cada estudante

Tempo(X)	Nota(Y)
3,0	4,5
7,0	6,5
2,0	3,7
1,5	4,0
12,0	9,3



Coeficiente de correlação linear

É uma medida que avalia o quanto a “nuvem de pontos” no diagrama de dispersão aproxima-se de uma reta.



O **coeficiente de correlação linear de *Pearson*** é calculado por:

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{(n-1) S_X S_Y}$$

sendo que

$\overline{X}$ e $\overline{Y}$ são as médias amostrais de X e Y, respectivamente,

S_X e S_Y são os desvios padrão de X e Y, respectivamente.

Fórmula alternativa para o coeficiente de correlação:

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{(n-1) S_X S_Y},$$

sendo

$$S_X^2 = \frac{\sum_{i=1}^{n} X_i^{\,2} - n\overline{X}^{\,2}}{n-1}.$$

Voltando ao Exemplo 2:

Tempo (*X*)	Nota (*Y*)	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$
3,0	4,5	-2,1	-1,1	2,31
7,0	6,5	1,9	0,9	1,71
2,0	3,7	-3,1	-1,9	5,89
1,5	4,0	-3,6	-1,6	5,76
12,0	9,3	6,9	3,7	25,53
25,5	28,0	0	0	41,2

$\bar{X} = 5,1$ $\bar{Y} = 5,6$

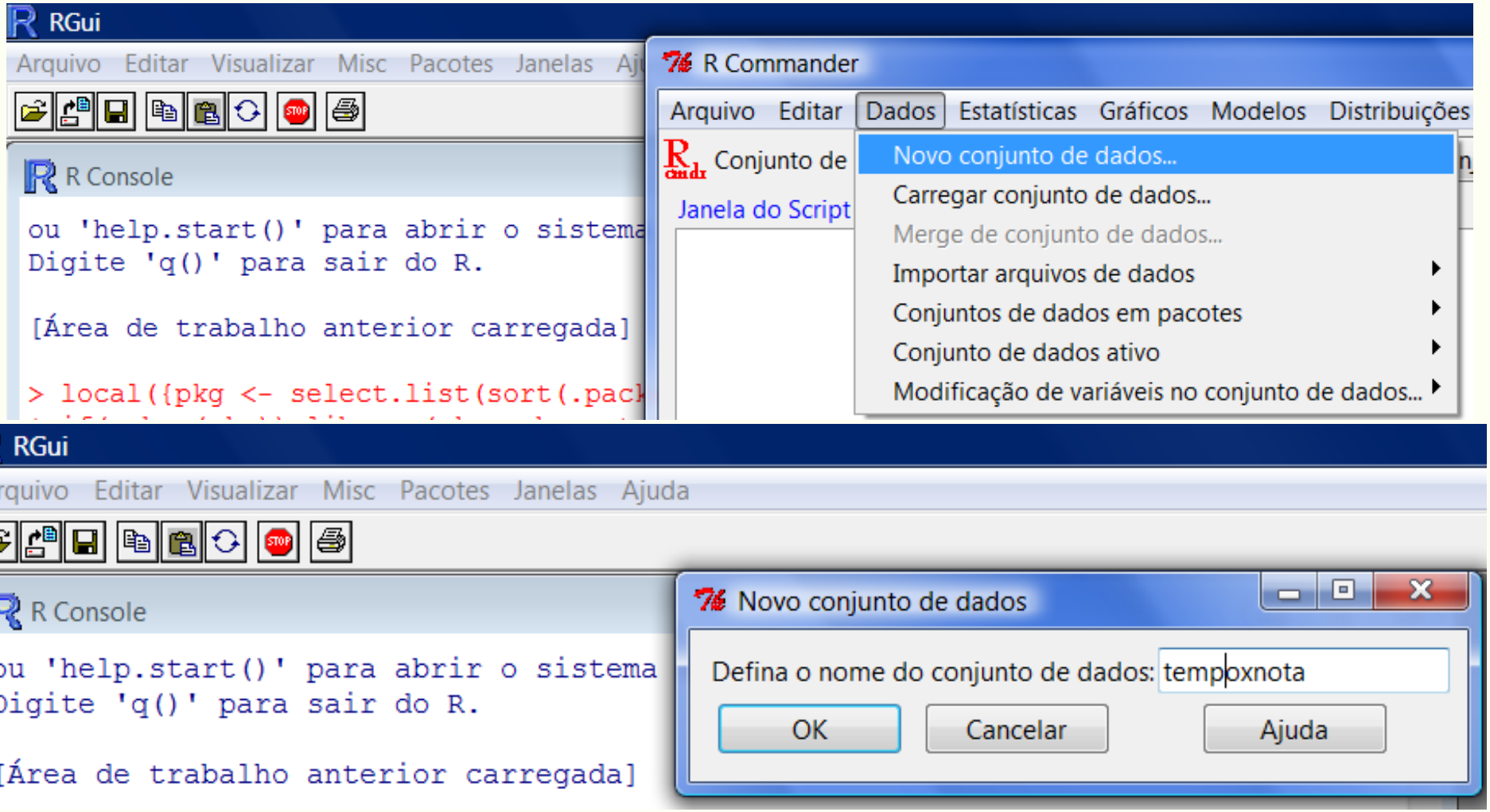
$$S_x^2 = \frac{(-2,1)^2 + \ldots + (6,9)^2}{4} = \frac{78,2}{4} = 19,55 \Rightarrow S_x = 4,42$$

$$S_y^2 = \frac{(-1,1)^2 + \ldots + (3,7)^2}{4} = \frac{21,9}{4} = 5,47 \Rightarrow S_y = 2,34$$

Então,

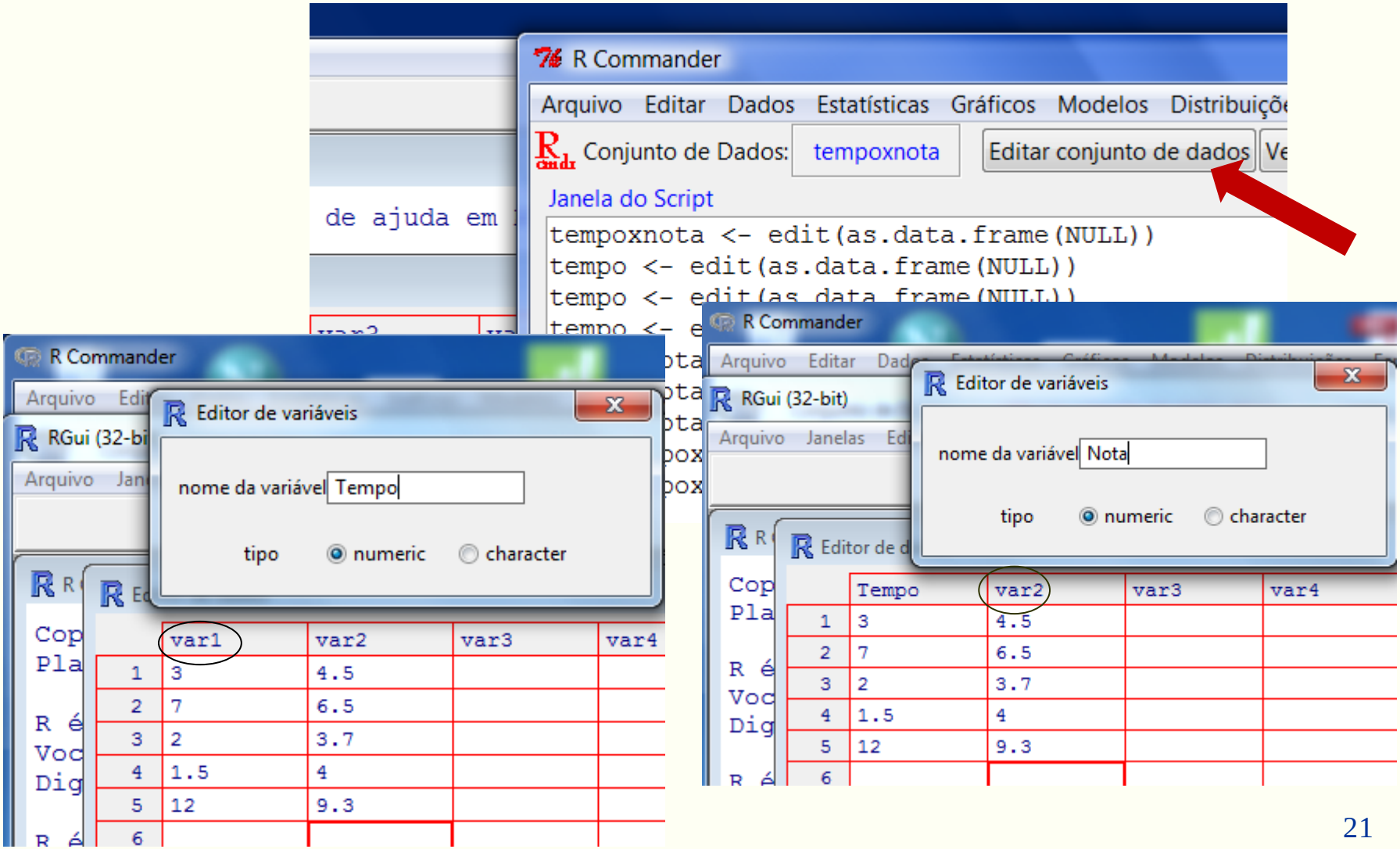
$$r = \frac{41,2}{4 \,.\, 4,42 \,.\, 2,34} = 0,9959$$

Criando arquivo de dados no *R*



Criando arquivo de dados no *R*

Digitar os dados na janela do editor e dar nomes (“Tempo” e “Nota”) às variáveis e fechar.



No *R* temos:

```
> cor(tempoxnota$Tempo, tempoxnota$Nota)
```

```
[1] 0.9960249
```

ou ainda

- **Estatísticas → Resumos → Matriz de Correlação**

(Selecione *Tempo* e *Nota* no conjunto de dados *tempoxnota*)

```
           Nota     Tempo
Nota   1.0000000 0.9960249
Tempo  0.9960249 1.0000000
```

Exemplo no *R*

O arquivo ***CEA05P11.xls*** contém dados sobre o projeto "Avaliação de um trabalho de Ginástica Laboral implantado em algumas unidades da USP".

Amostra: 143 funcionários que participaram de atividades de Ginástica Laboral.

Algumas variáveis registradas no estudo são:

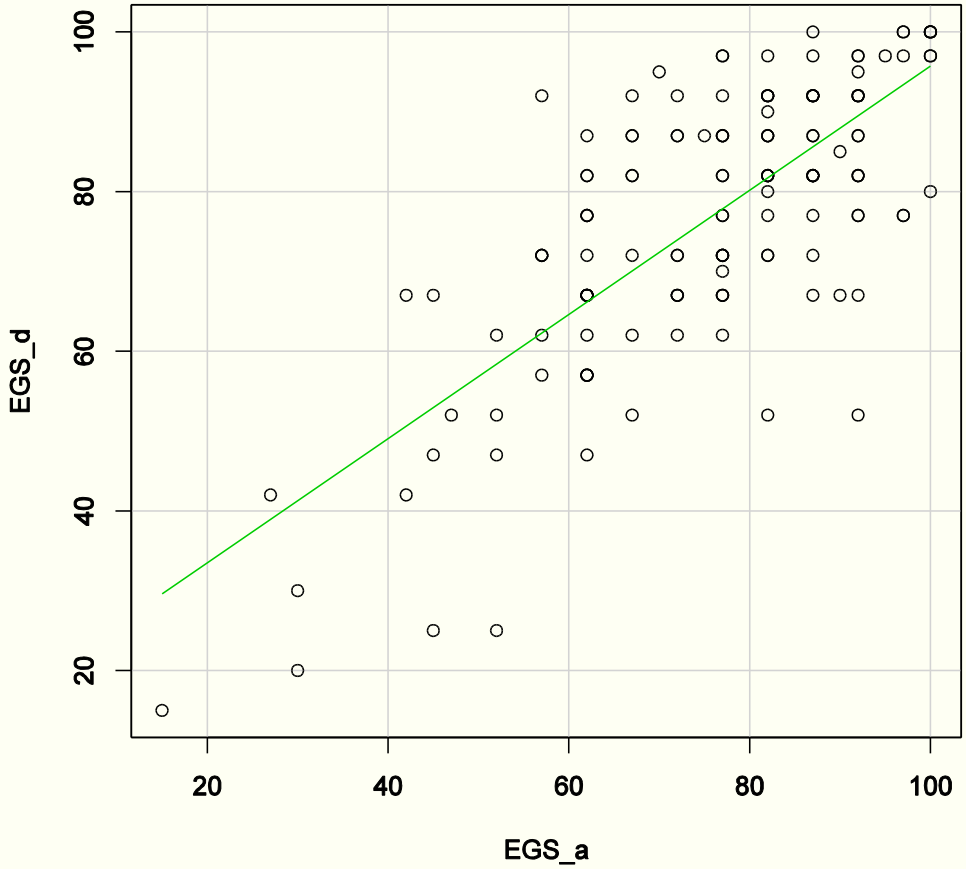
Sexo: Feminino e Masculino;

Idade: idade do funcionário, em anos;

Unidade da USP: EP, FAU, IAG, IF, IO e Reitoria

Estado Geral de Saúde antes (EGS_a) e **Estado Geral de Saúde depois** (EGS_d): auto-avaliação do funcionário a respeito do seu estado de saúde antes e depois do início das atividades respectivamente. Quanto maior o índice, melhor a avaliação.

Gráficos → Diagrama de Dispersão (variável-x: EGS_a ; variável-y: EGS_d; marcar opção *Linha de quadrados mínimos*)



Estatísticas → Ajuste de Modelos → Regressão Linear

(variável resposta: EGS_d ; variável explicativa: EGS_a)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.94397    4.54712   3.946 0.000125 ***
EGS_a        0.77791    0.05894  13.198  < 2e-16 ***
```

$a = 17{,}94397, \; b = 0{,}77791$

Reta ajustada :

$$\hat{Y} = 17{,}94397 + 0{,}77791\, EGS\_a$$