

8. Medidas de associação

2010

Dados: (x_i, y_i) , $i = 1, \dots, n$. n pares de observações das variáveis x e y , que podem ser qualitativas ou quantitativas.

Os pares representam a ocorrência simultânea de x e y . Cada par (x_i, y_i) é indissociável.

Em várias situações há interesse em estudar a relação entre x e y , se existir.

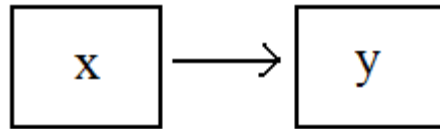
Uma possível relação entre x e y pode ser quantificada por uma medida resumo: medida de associação.

A associação entre variáveis pode ser negativa ou positiva (sentido). E fraca ou forte (intensidade ou força).

Gráficos são úteis no estudo de associação entre variáveis.

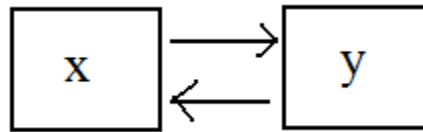
Tipos de relação

(a) **Causal unilateral**. y depende de x (ou x depende de y).



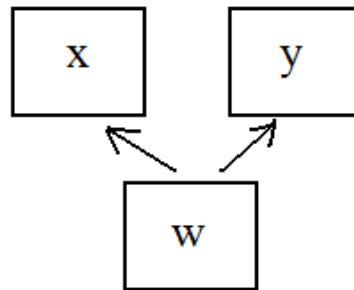
Exemplo. O preço de venda de um produto (y) depende da distância entre o local de produção e o local de venda (x). x depende de y ?

(b) **Causal bilateral** ou **interdependência**. y depende de x e x depende de y .



Exemplo. Relação entre peso e altura de uma pessoa.

(c) **Dependência indireta** (ou **associação espúria**). x e y são influenciadas por outra(s) variável(is).

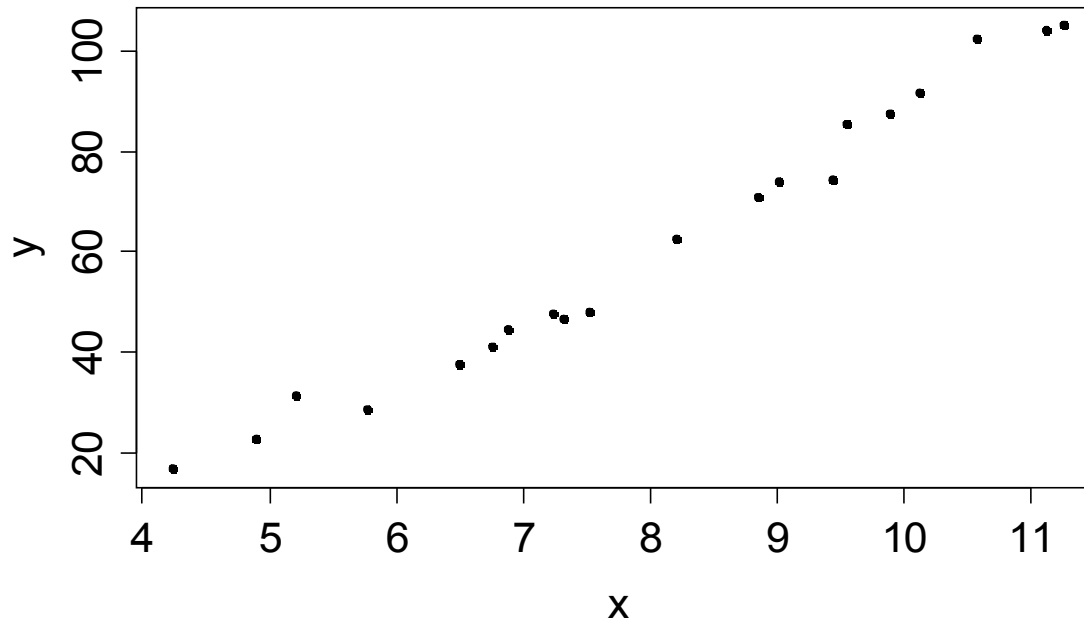


Exemplo. Relação entre o número anual de casos de insolação e a produção anual de trigo. Causa comum: temperatura.

8.1. Variáveis quantitativas

$(x_1, y_1), \dots, (x_n, y_n)$: conjunto de dados **bivariado**.

Representação gráfica: **gráfico de dispersão** (*scatter plot*). Gráfico cartesiano dos pares (x_i, y_i) , $i = 1, \dots, n$.



Covariância entre x e y : medida da variação **conjunta** (ou concomitante ou simultânea) de x e y em relação às suas médias.

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad -\infty < \text{cov}(x, y) < \infty.$$

8.1. Variáveis quantitativas

Obs. (a) $\text{cov}(x, y) = \text{cov}(y, x)$ e (b) $\text{cov}(x, x) = s_x^2$.

Coeficiente de correlação linear de Pearson (r):

$$\text{cor}(x, y) = r = \frac{\text{cov}(x, y)}{s_x s_y} = \frac{1}{n-1} \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y},$$

sendo que s_x e s_y denotam os desvios padrão de x e y . Se $s_x = 0$ e/ou $s_y = 0$, r não está definido.

Propriedades: **P1.** $\text{cor}(x, x) = 1$. **P2.** $-1 \leq r \leq 1$.

P3. $r = 1$ se, e somente se, a relação entre x e y for **linear** ($y = a + bx$) e $b > 0$.

P4. $r = -1$ se, e somente se, a relação entre x e y for **linear** ($y = a + bx$) e $b < 0$.

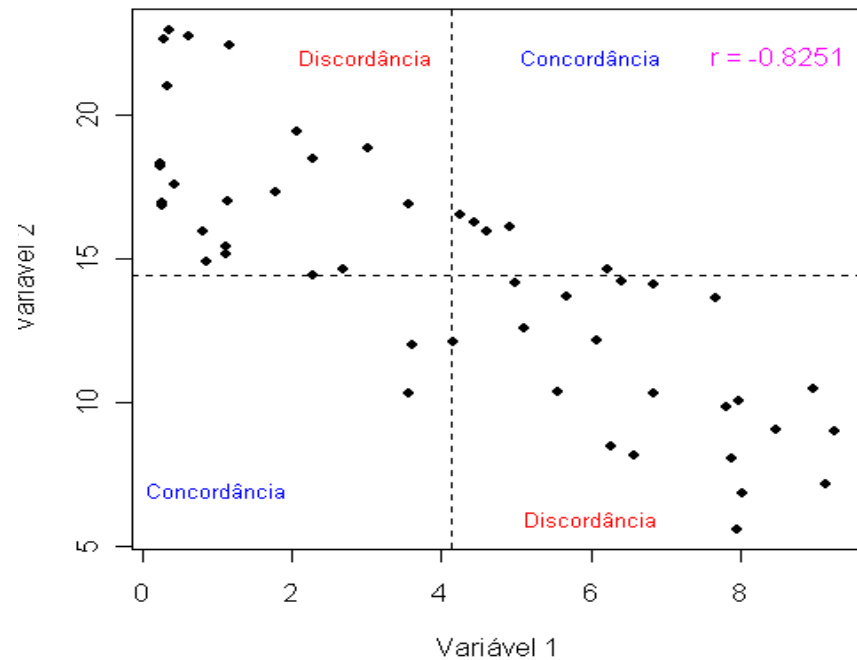
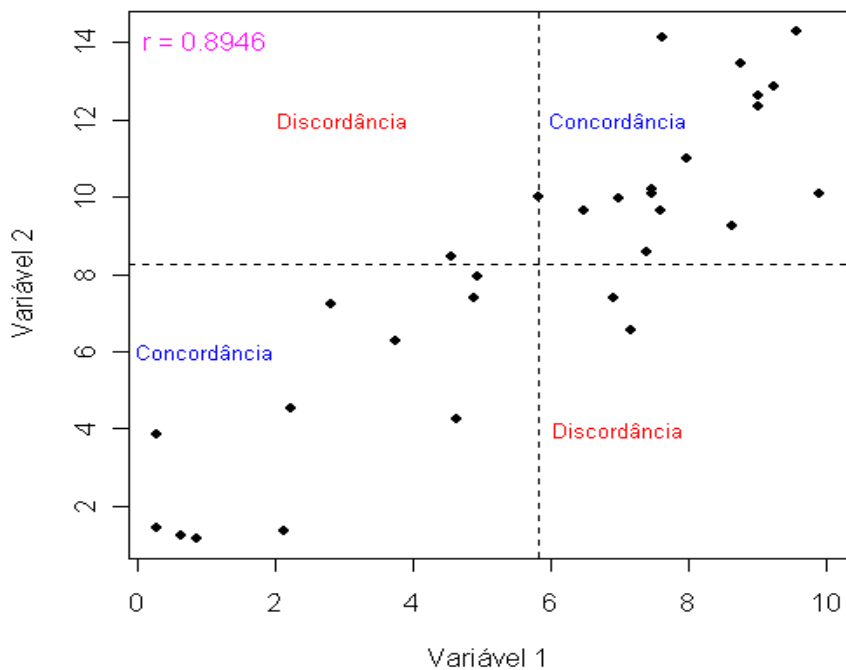
P5. Invariância. Se $b_1 > 0$ e $b_2 > 0$, então $\text{cor}(x, y) = \text{cor}(a_1 + b_1 x, a_2 + b_2 y)$, em que a_1 e a_2 são reais quaisquer.

Exercício. Se $b_1 < 0$ e $b_2 > 0$ ou $b_1 > 0$ e $b_2 < 0$ ou $b_1 < 0$ e $b_2 < 0$, o que se pode afirmar sobre $\text{cor}(a_1 + b_1 x, a_2 + b_2 y)$?

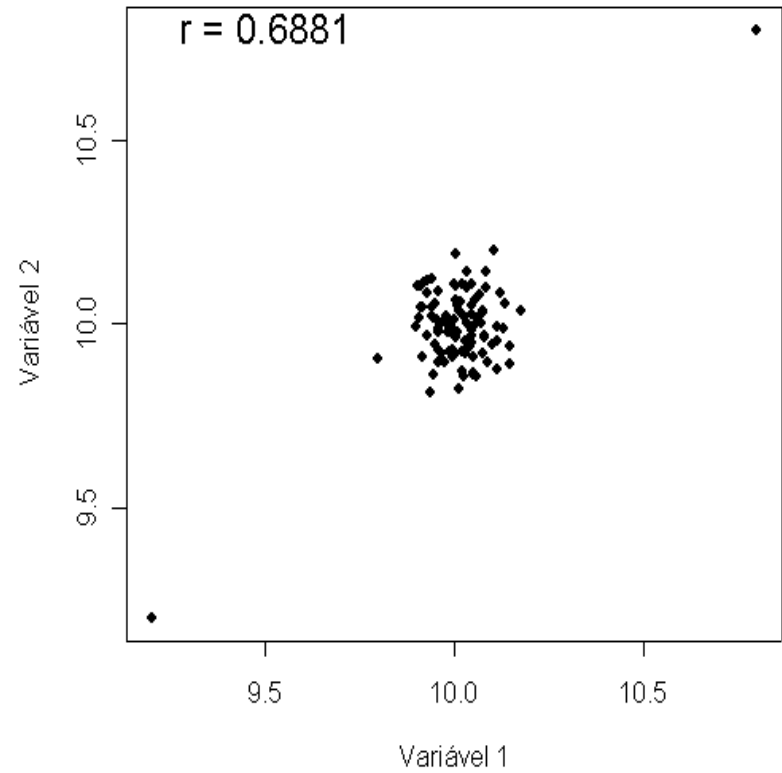
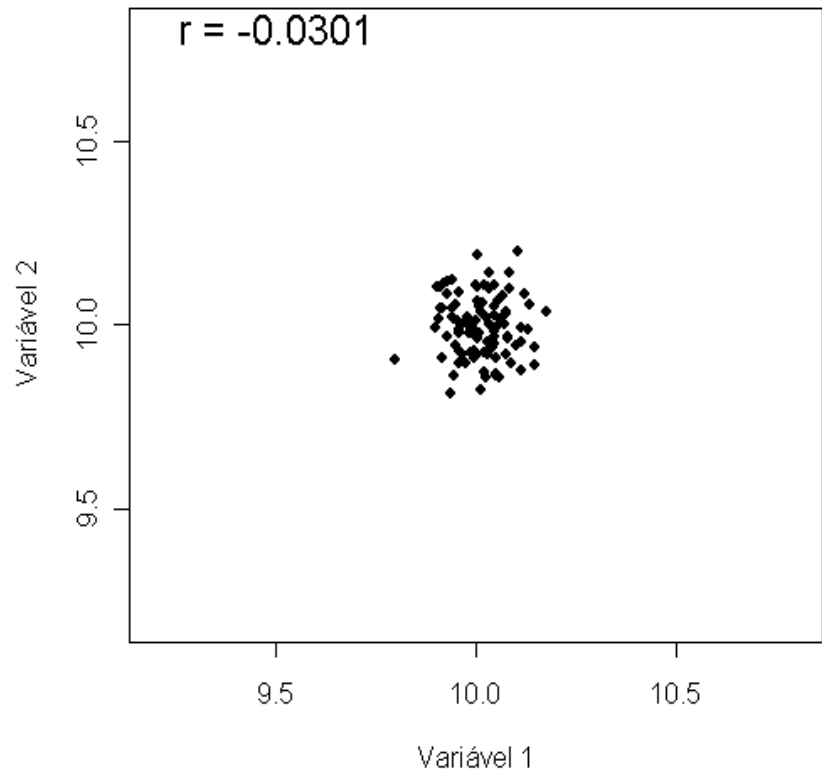
8.1. Variáveis quantitativas

Sentido e força de r

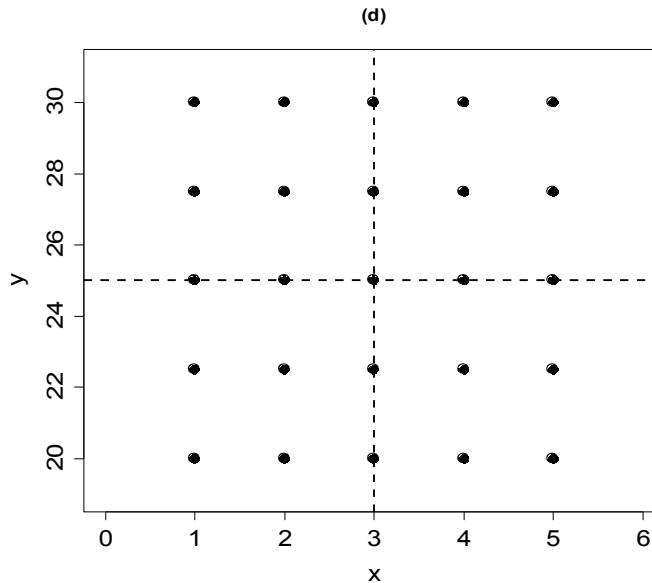
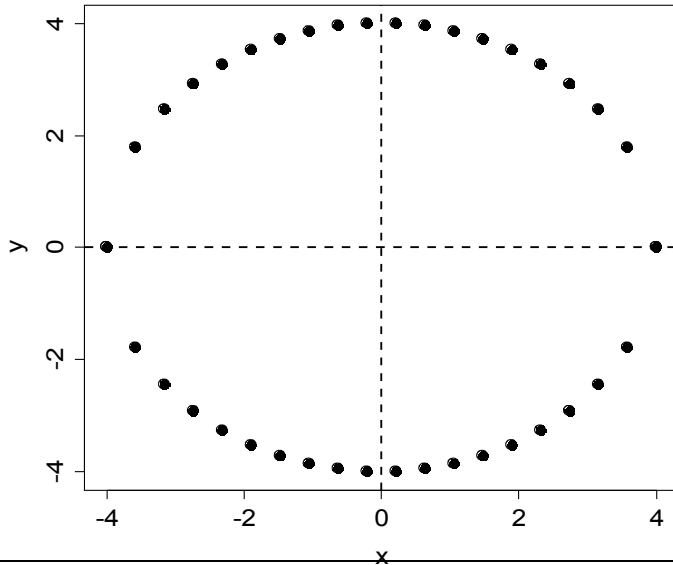
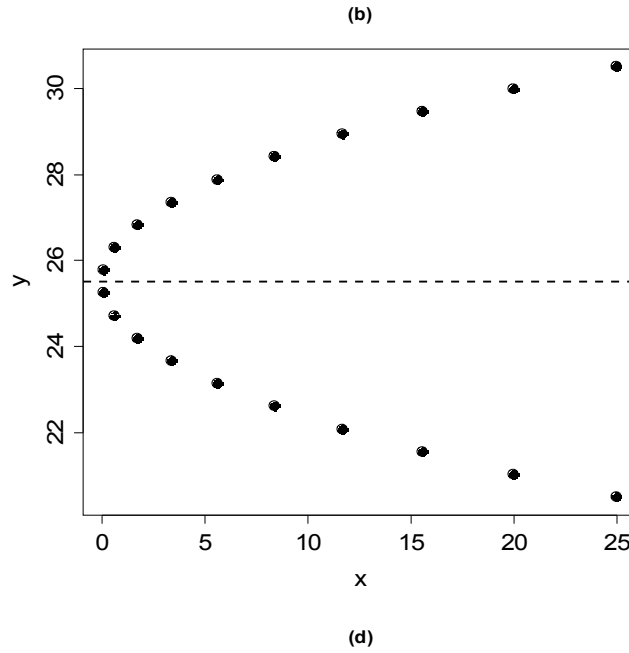
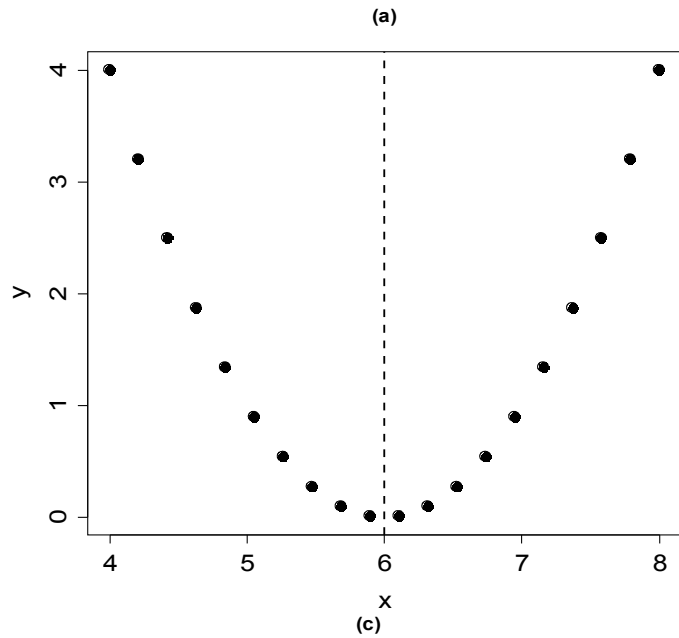
negativa forte negativa fraca positiva fraca positiva forte
-1 negativa moderada 0 positiva moderada 1
ausência



8.1. Variáveis quantitativas



8.1. Variáveis quantitativas

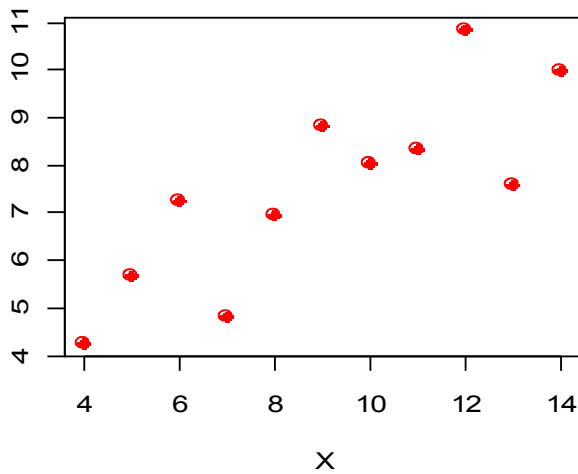


Exercício. Prove que se houver simetria em x e/ou y , então $r = 0$.

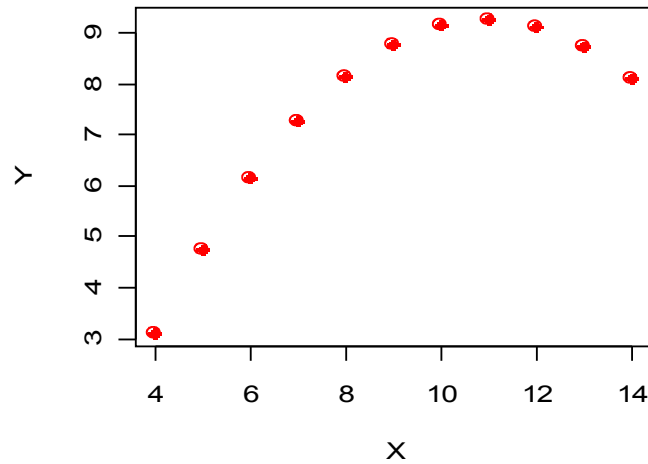
Obs. $r = 0$ não significa ausência de associação.

8.1. Variáveis quantitativas

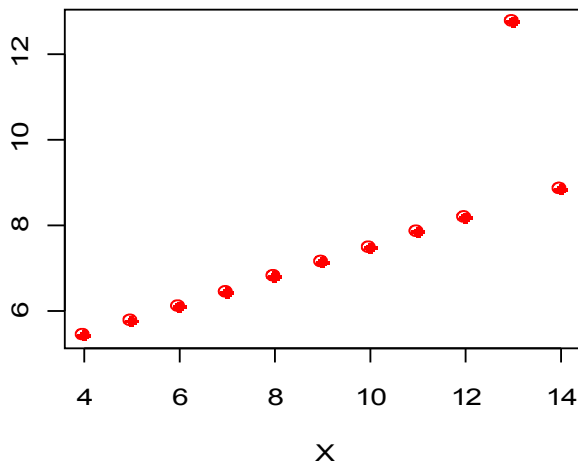
Exemplo 1



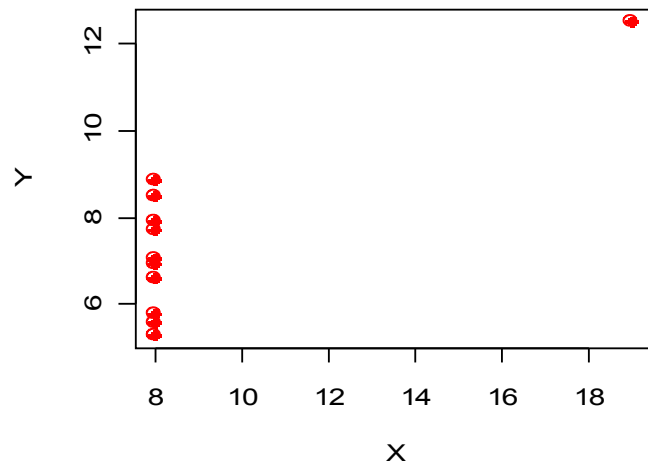
Exemplo 2



Exemplo 3



Exemplo 4



Dados anscombe em R

```
> ?anscombe
```

Valores de r:

Exemplo 1: 0,8164

Exemplo 2: 0,8162

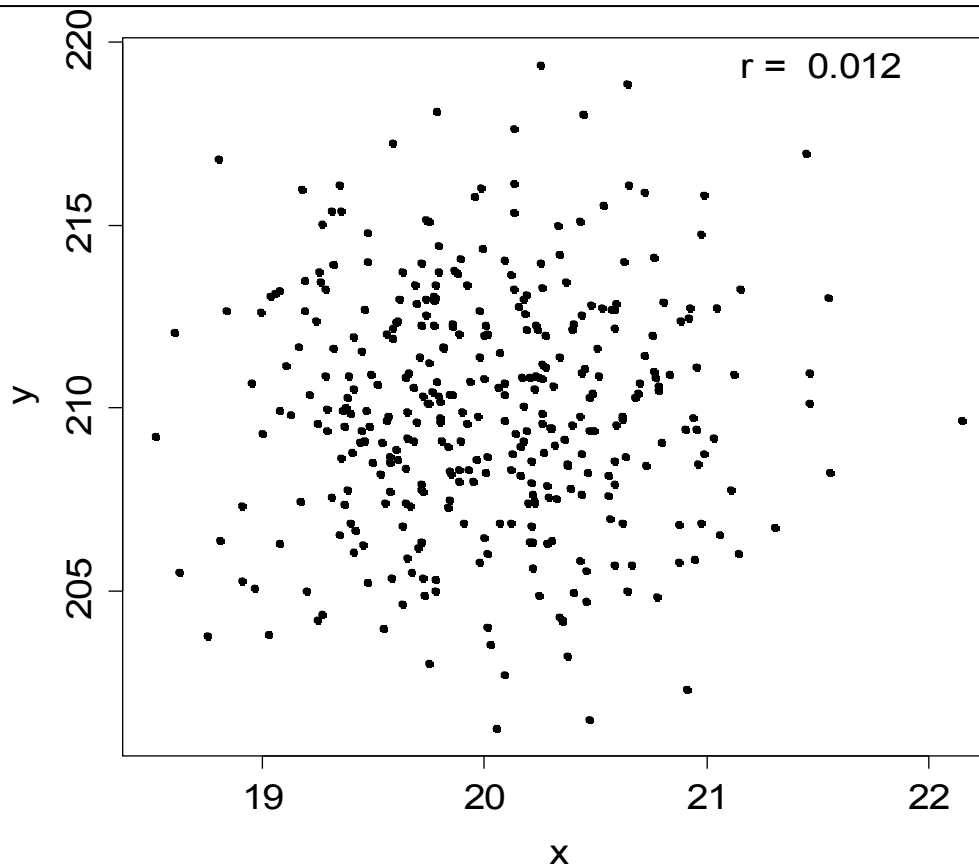
Exemplo 3: 0,8163

Exemplo 4: 0,8165

Veja também <http://www.jerrydallal.com/LHSP/corr.htm>

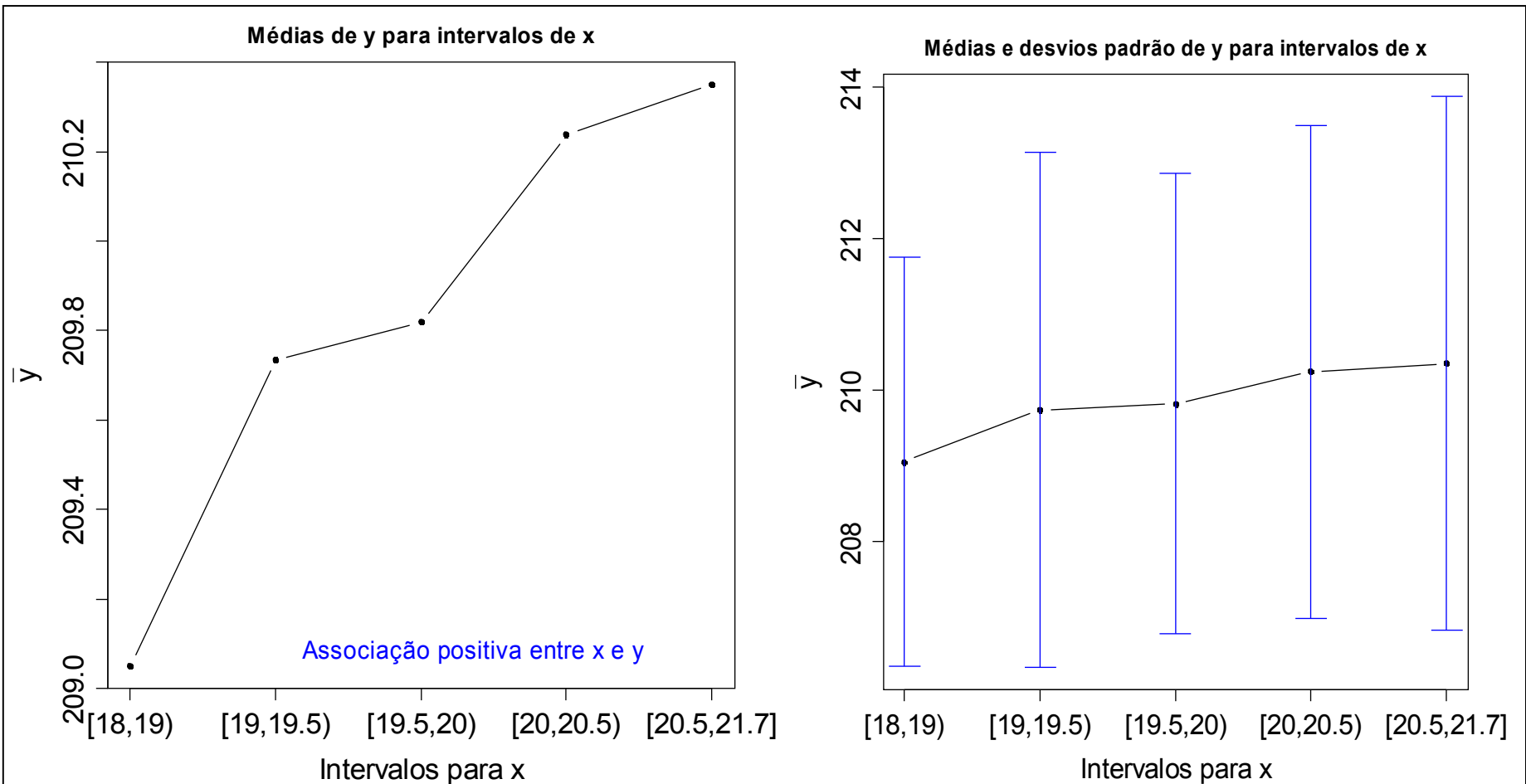
Efeito do agrupamento

Exemplo.
n = 360.



Intervalo para x	[18;	[19;	[19,5;	[20;	[20,5;
Frequênc	23	54	95	14	71
Média	19,9	209,7	209,8	210,2	210,3
Desvio de y padrão	2,7	3,4	3,0	3,2	3,5

Efeito do agrupamento



Mesmo **não existindo** associação entre x e y, se o número de observações (**n**) for suficientemente **grande**, é possível encontrar intervalos para x que resultam em **qualquer tipo** de associação.

Correlação em R

Funções `cor`, `cov` e `cov2cor`.

```
> x = c(5.5,6.7,9.5,4.2,9.0,11.6,4.5,9.6,6.2,11.6,8.8,8.6,7.8,4.8,  
10.1)
```

```
> y = c(11.6,11.3,17.5,9.1,15.7,16.9,8.1,21.2,11.7,18.7,13.9,15.0,  
11.6,7.0, 15.6)
```

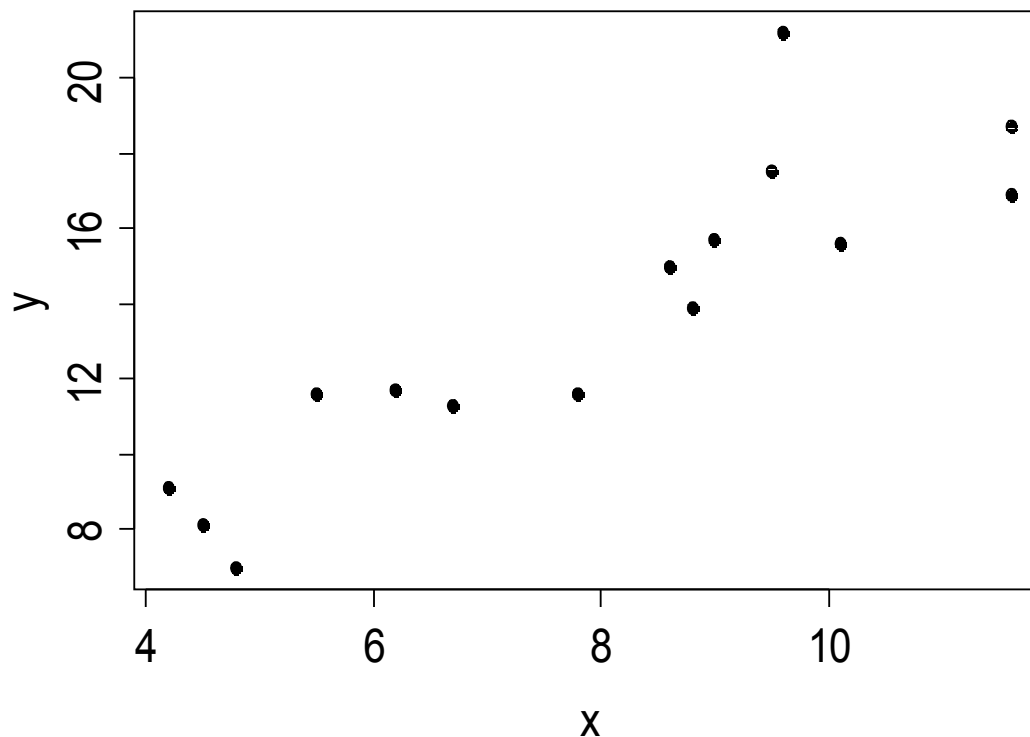
```
> length(x)
```

```
[1] 15
```

```
> cor(x, y)
```

```
[1] 0.8908723
```

```
> plot(x, y, pch = 20)
```



Correlação em R

```
> ? USArrests
```

Description

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

Número de prisões por assalto, homicídio e estupro por 100000 hab. e proporção da população urbana.

```
> names(USArrests)
```

```
[1] "Murder" "Assault" "UrbanPop" "Rape"
```

```
> rownames(USArrests)
```

```
[1] "Alabama" "Alaska" "Arizona" "Arkansas" "California" etc  
[50] "Wyoming"
```

```
> summary(USArrests)
```

Murder	Assault	UrbanPop	Rape
Min. : 0.800	Min. : 45.0	Min. :32.00	Min. : 7.30
1st Qu.: 4.075	1st Qu.:109.0	1st Qu.:54.50	1st Qu.:15.07
Median : 7.250	Median :159.0	Median :66.00	Median :20.10
Mean : 7.788	Mean :170.8	Mean :65.54	Mean :21.23
3rd Qu.:11.250	3rd Qu.:249.0	3rd Qu.:77.75	3rd Qu.:26.18
Max. :17.400	Max. :337.0	Max. :91.00	Max. :46.00

```
> class(USArrests)
```

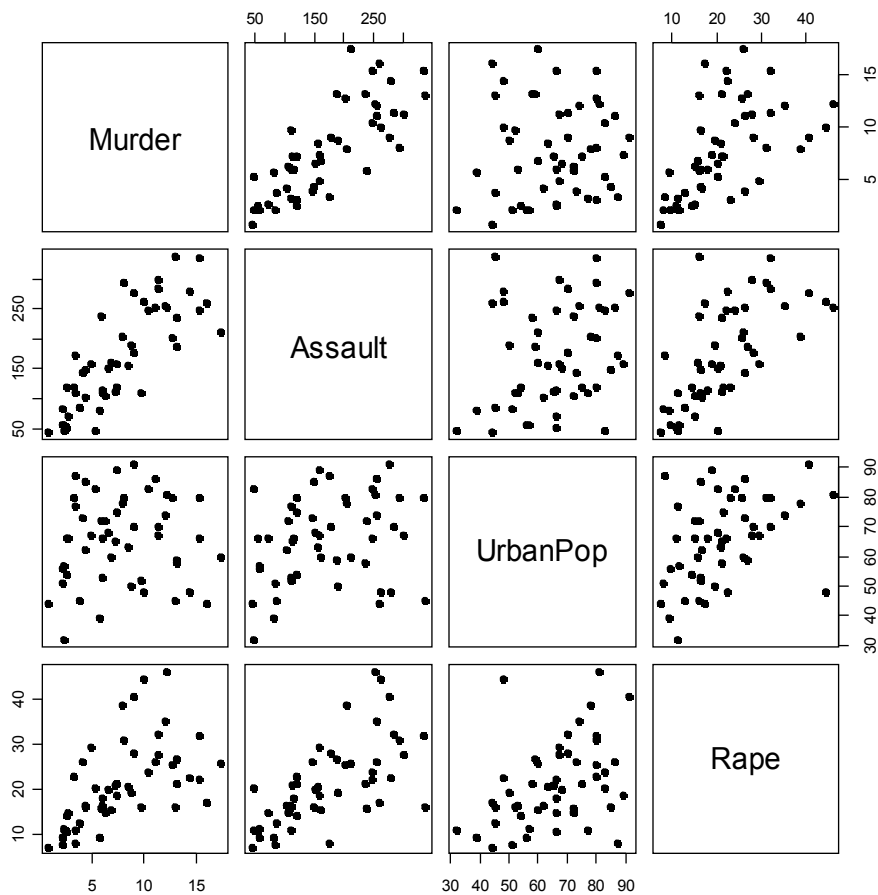
```
[1] "data.frame"
```

Classe “folha de dados”.

Correlação em R

Gráficos de dispersão: função `pairs`.

```
> pairs(USArrests, pch = 20)
```

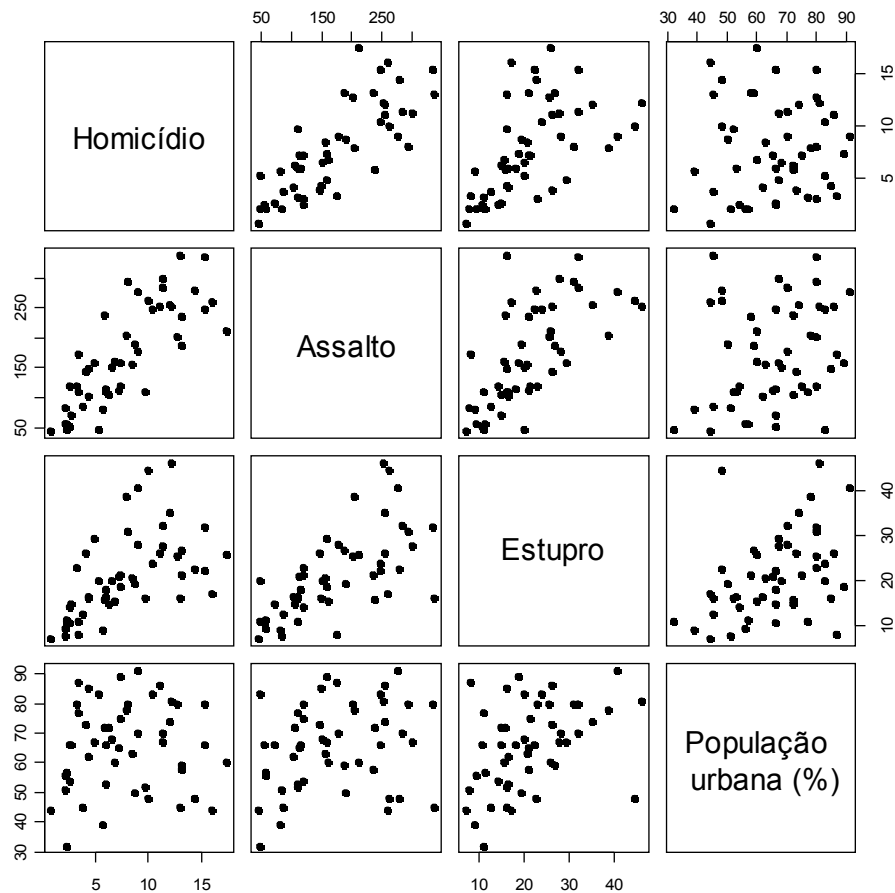


Matriz de gráficos de dispersão
(*scatter plot matrix*).

```
> ordem = c("Murder", "Assault",  
"Rape", "UrbanPop")
```

```
> nomes = c("Homicídio", "Assalto",  
"Estupro", "População \n urbana (%)")
```

```
> pairs(USArrests[, ordem], pch = 20,  
labels = nomes)
```



Correlação em R

Matriz de covariâncias:

```
> cov(USArrests[, ordem])
```

	Murder	Assault	Rape	UrbanPop
Murder	18.970465	291.0624	22.99141	4.386204
Assault	291.062367	6945.1657	519.26906	312.275102
Rape	22.991412	519.2691	87.72916	55.768082
UrbanPop	4.386204	312.2751	55.76808	209.518776

Obs. É uma matriz simétrica com as variâncias na diagonal principal.

Matriz de correlações:

```
> cor(USArrests[, ordem])
```

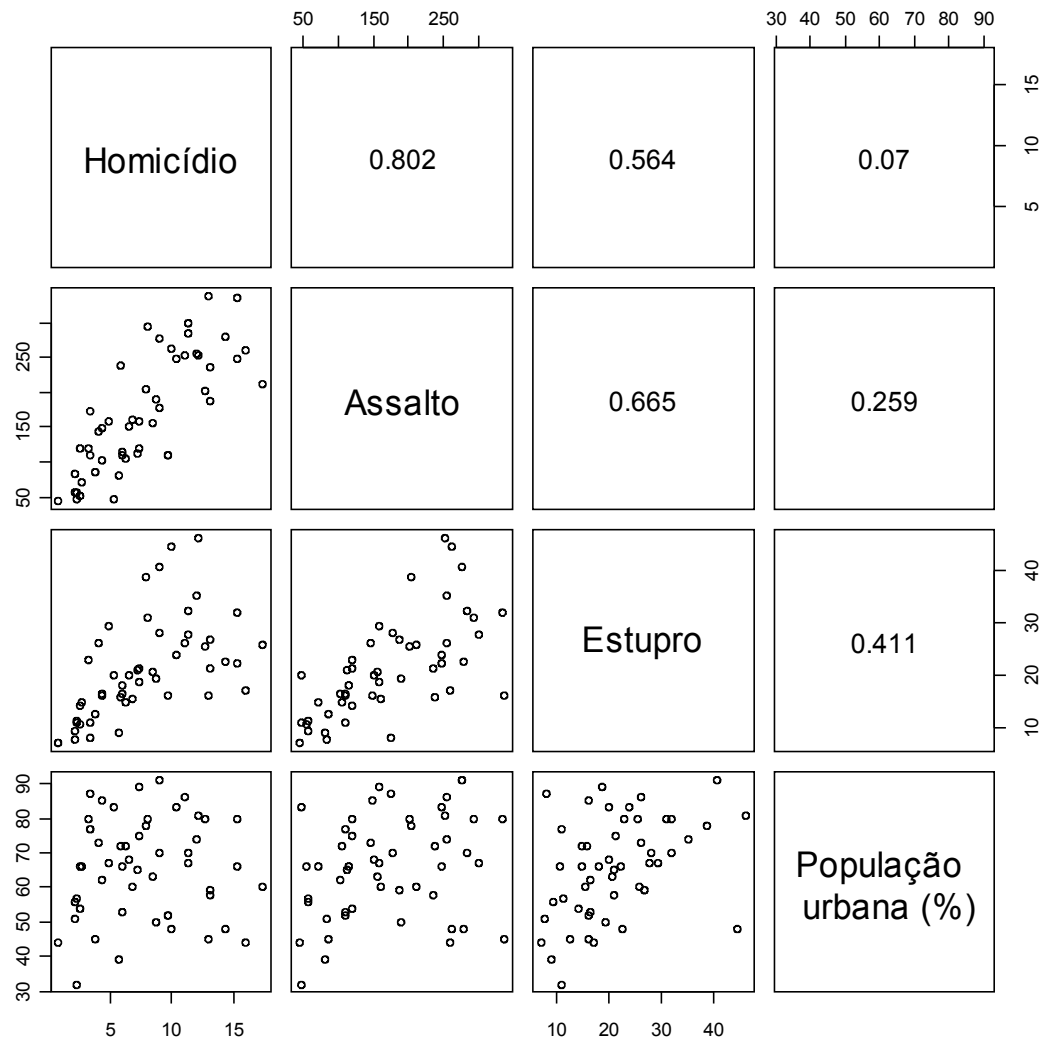
	Murder	Assault	Rape	UrbanPop
Murder	1.00000000	0.8018733	0.5635788	0.06957262
Assault	0.80187331	1.00000000	0.6652412	0.25887170
Rape	0.56357883	0.6652412	1.00000000	0.41134124
UrbanPop	0.06957262	0.2588717	0.4113412	1.00000000

Obs. A função `cov2cor` transforma uma matriz de covariâncias em uma matriz de correlações.

Correlação em R

```
> panel.cor = function(x, y,
digits = 3)
{
  usr = par("usr")
  on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r = cor(x, y)
  text(0.5, 0.5, round(r,
digits), cex = 1.5)
}

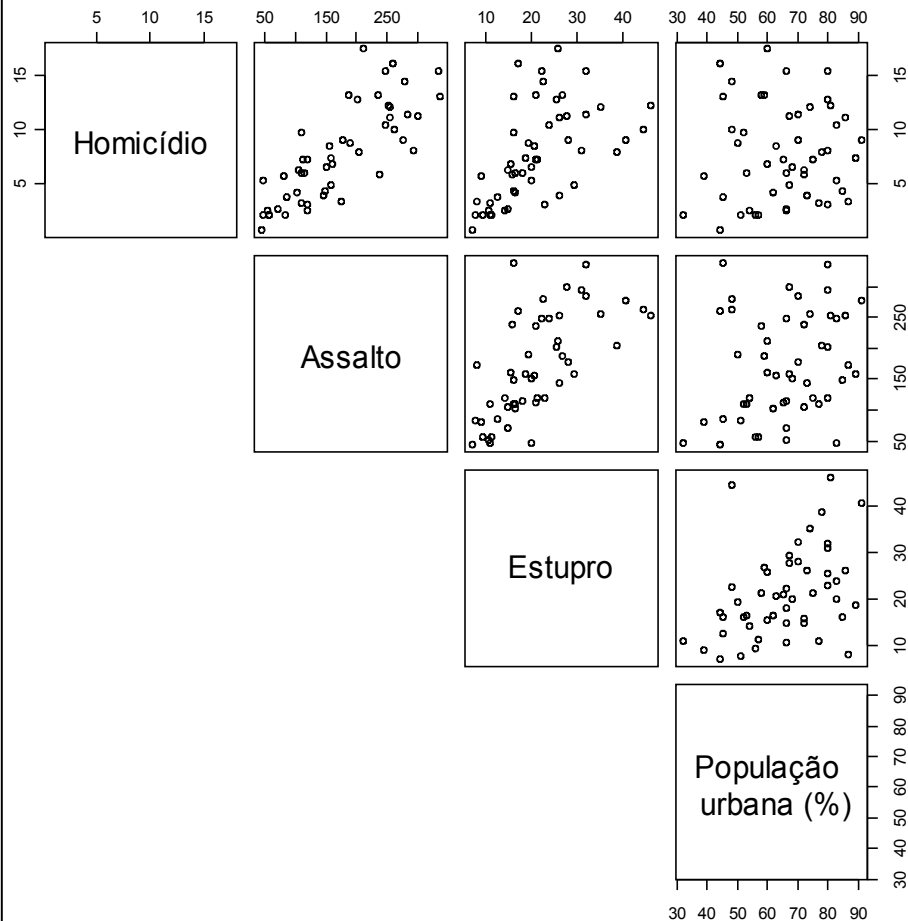
> pairs(USArrests[, ordem],
labels = nomes, upper.panel
= panel.cor)
```



Correlação em R

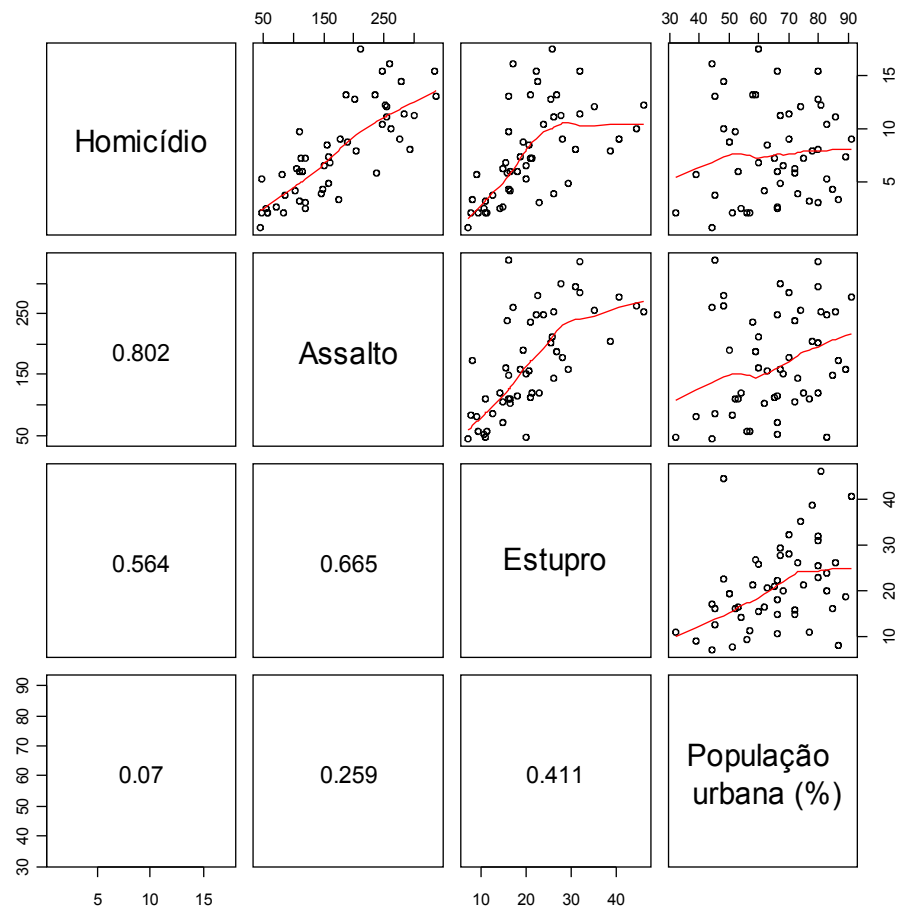
Omitindo a parte inferior da matriz:

```
> pairs(USArrests[, ordem],  
labels = nomes, lower.panel =  
NULL)
```



Correlações e linhas de tendência:

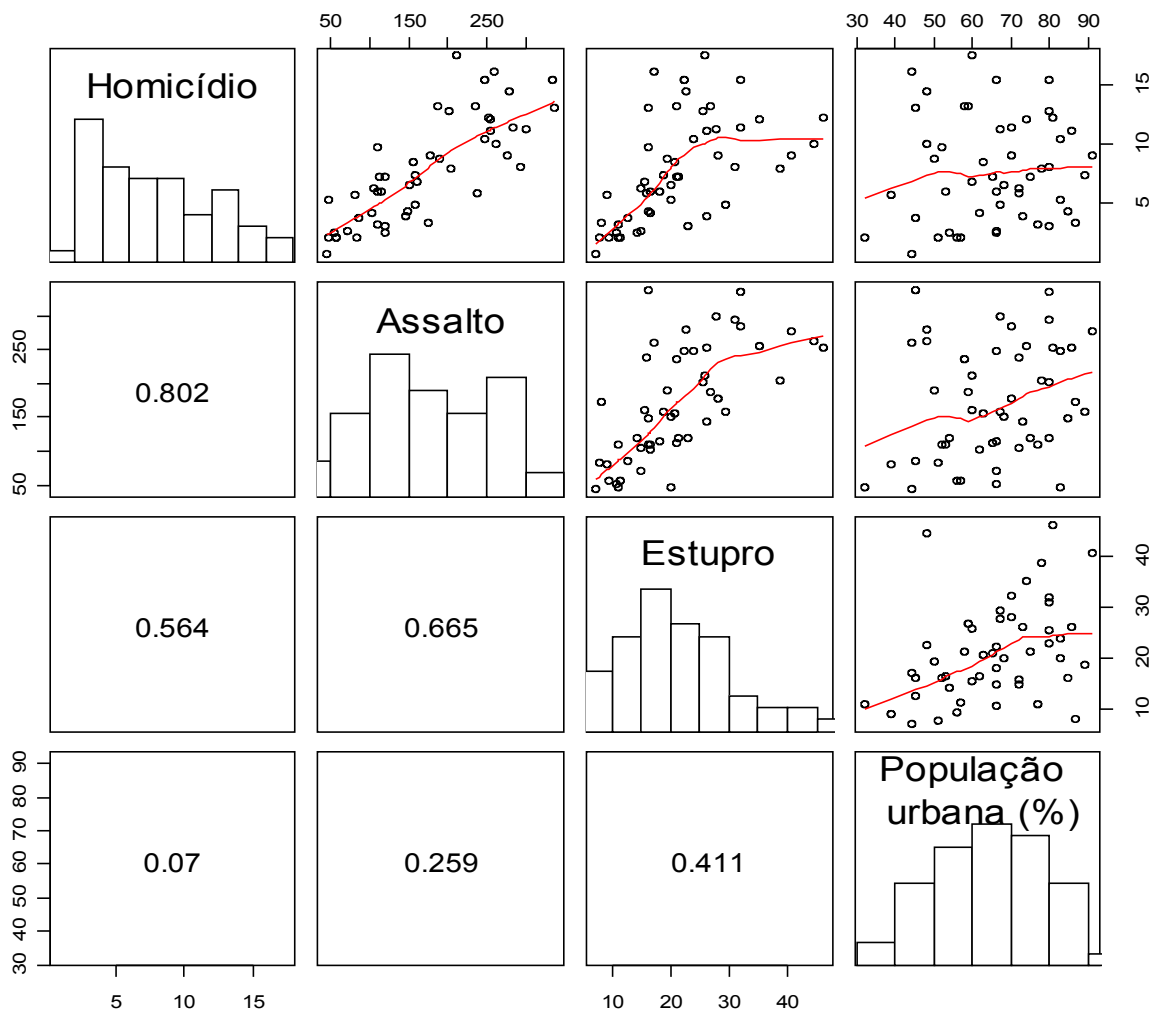
```
> pairs(USArrests[, ordem],  
labels = nomes, upper.panel =  
panel.smooth, lower.panel =  
panel.cor)
```



Correlação em R

Correlações, linhas de tendência e histogramas (utilize `?pairs`):

```
> pairs(USArrests[, ordem], labels = nomes, upper.panel =  
panel.smooth, lower.panel = panel.cor, diag.panel = panel.hist)
```



Quais pares apresentam as correlações mais fracas e mais fortes?

O efeito de urbanização está mais associado a qual tipo de crime?

Uma grande quantidade de assaltos resultou em homicídios?

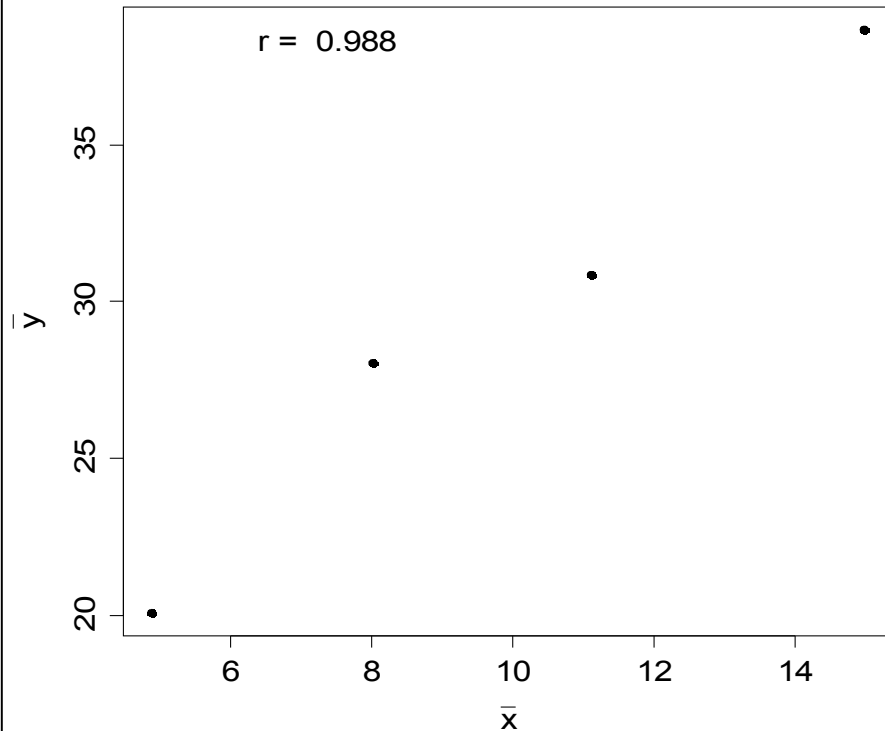
Que outras variáveis poderiam estar relacionadas à ocorrência dos crimes?

Falácia ecológica (*ecological fallacy*)

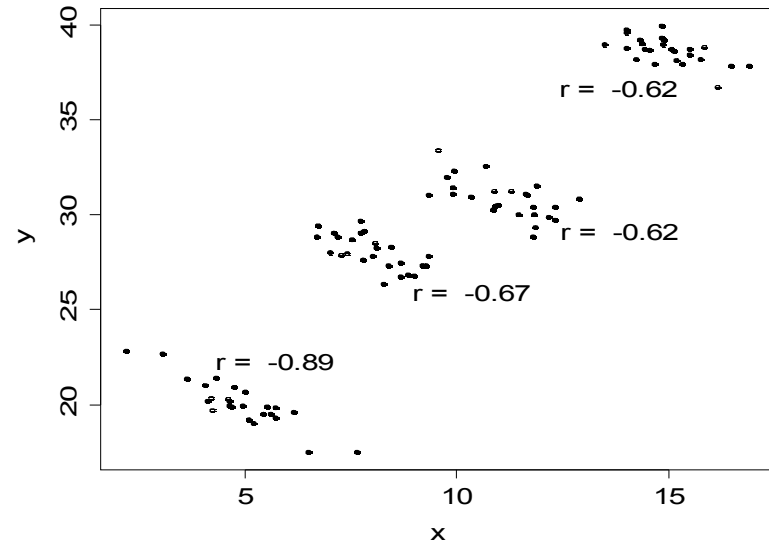
Ocorre quando conclusões obtidas para dados **agregados** não podem ser estendidas para dados **individuais** (microdados).

Exemplo. Forte correlação positiva entre médias de grupos (dados agregados) implica forte correlação positiva entre as observações (microdados)? **Não.**

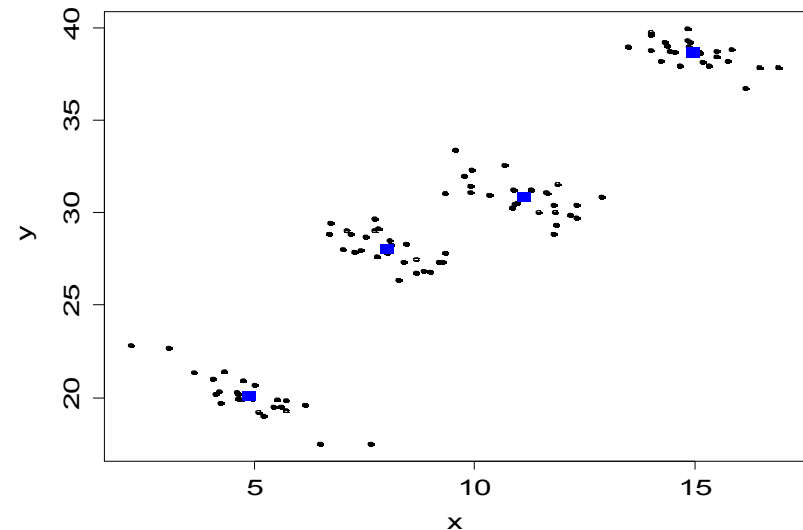
Médias de quatro grupos



Dados individuais



Dados individuais e médias dos grupos



Número médio de **peças por cômodo** em **60** países ou regiões.

Dados: <http://unstats.un.org/unsd/demographic/products/socind/housing.htm>

```
> dados = read.csv("Housing_Dec2009.csv", header = TRUE, sep = ";")
```

```
> names(dados)
```

```
[1] "countryarea" "year" "total" "urban" "rural"
```

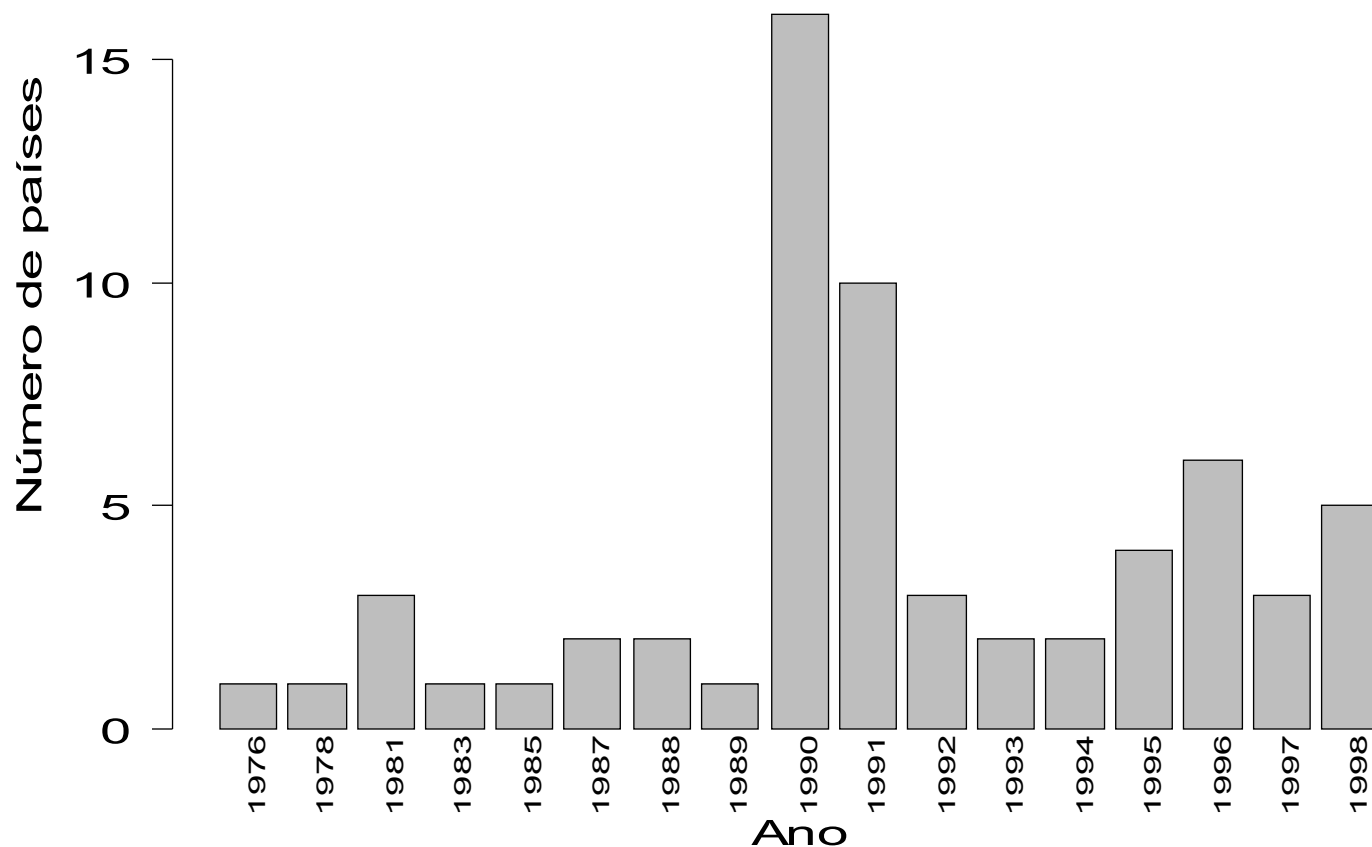
```
> summary(dados[, -1])
```

year	total	urban	rural
Min. :1976	Min. :0.500	Min. :0.500	Min. : 0.400
1st Qu.:1990	1st Qu.:0.700	1st Qu.:0.700	1st Qu.: 0.700
Median :1991	Median :1.000	Median :1.000	Median : 1.000
Mean :1991	Mean :1.141	Mean :1.153	Mean : 1.230
3rd Qu.:1995	3rd Qu.:1.300	3rd Qu.:1.300	3rd Qu.: 1.400
Max. :1998	Max. :3.000	Max. :3.100	Max. : 3.300
	<u>NA's :2.000</u>	<u>NA's :8.000</u>	<u>NA's :10.000</u>

É possível comparar dados coletados de **1976** com os de **1998**?

Exemplo em Wainer (2009)

```
> attach(dados)
> table(year)
> barplot(table(year), xlab = "Ano", ylab = "Número de países",
           las = 2)
```



Exemplo em Wainer (2009)

```
> countryarea[year == 1976]
```

```
[1] Cameroon
```

```
> countryarea[year == 1998]
```

```
[1] Azerbaijan Brazil
```

```
Finland Netherlands
```

```
Pakistan
```

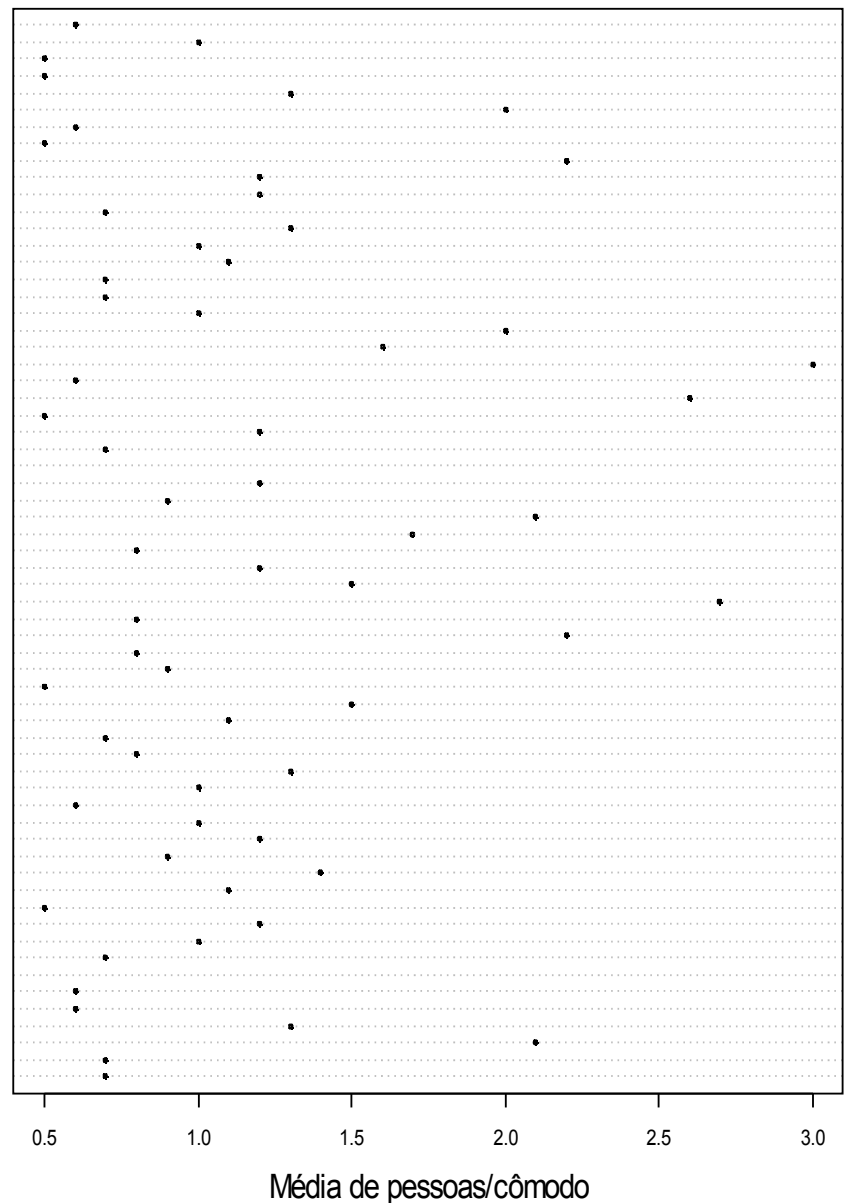
```
> dotchart(total, labels =  
countryarea, xlab = "Média  
de pessoas/cômodo", pch =  
20, cex = 0.7, cex.lab =  
1.5)
```

Por que utilizar a ordem
alfabética?

```
> ordem = order(total,  
decreasing = TRUE)
```

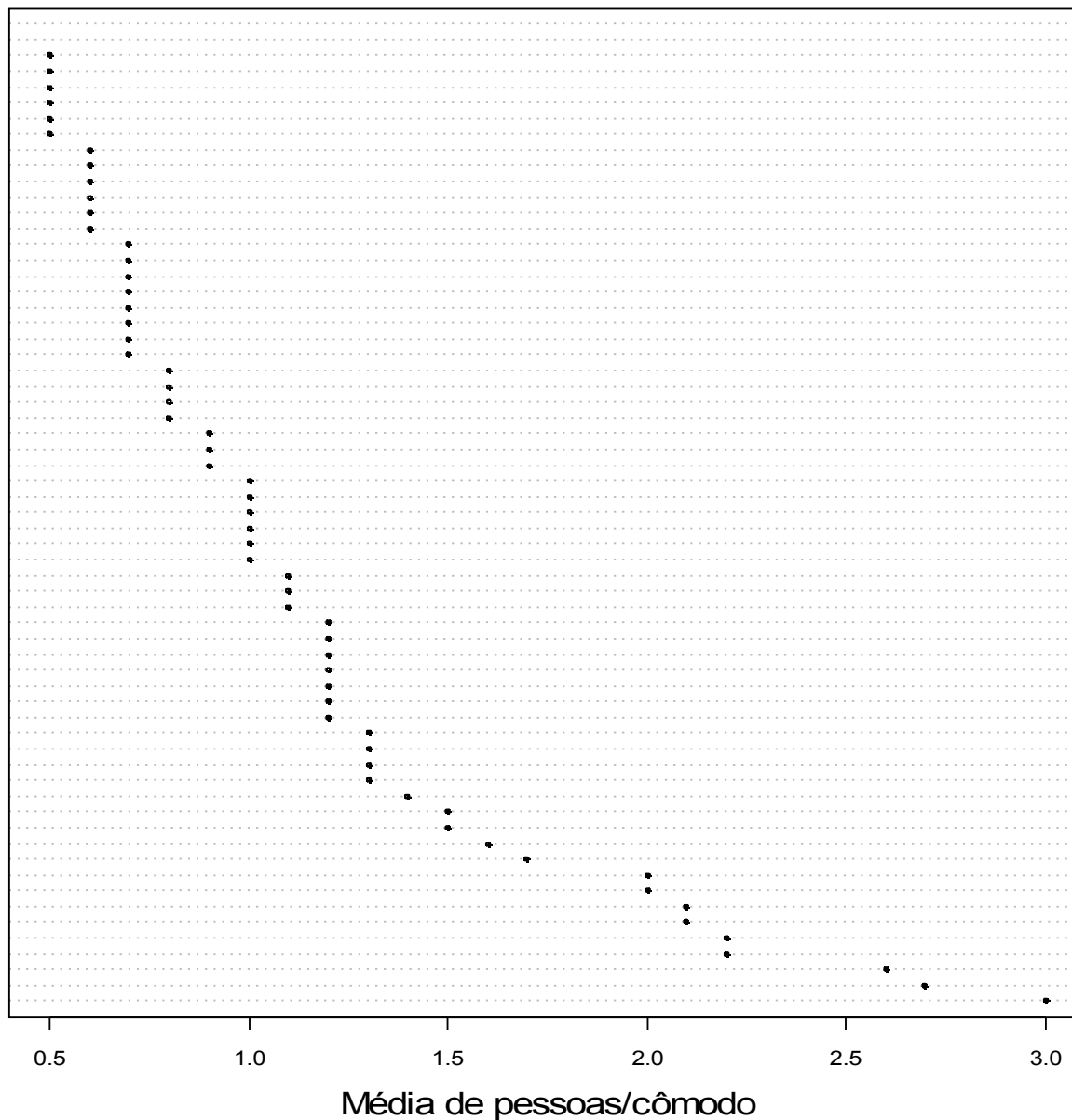
```
> dotchart(total[ordem],  
labels =  
countryarea[ordem], xlab =  
"Média de pessoas/cômodo",  
pch = 20, cex = 0.7,  
cex.lab = 1.5)
```

US Virgin Islands
Uruguay
United States
United Kingdom
Turkey
Syrian Arab Republic
Switzerland
Sweden
Sri Lanka
Slovakia
Serbia
San Marino
Romania
Reunion
Republic of Korea
Puerto Rico
Portugal
Poland
Peru
Panama
Pakistan
Norway
Nicaragua
New Zealand
New Caledonia
Netherlands
Mexico
Mauritius
Martinique
Lesotho
Kuwait
Japan
Israel
Iraq
India
Hungary
Honduras
Guam
Guadeloupe
Germany
Gambia
French Guiana
France
Finland
Egypt
Czech Republic
Cyprus
Cuba
Croatia
Costa Rica
Colombia
China, Macao SAR
Canada
Cameroon
Bulgaria
Brazil
Bolivia
Bermuda
Belgium
Bahamas
Azerbaijan
Austria
Aruba



Exemplo em Wainer (2009)

Mexico
Bolivia
United States
United Kingdom
Sweden
New Zealand
Germany
Canada
US Virgin Islands
Switzerland
Norway
Cyprus
Bermuda
Belgium
San Marino
Puerto Rico
Portugal
Netherlands
France
Brazil
Austria
Aruba
Japan
Hungary
Guam
Finland
Martinique
Guadeloupe
Costa Rica
Uruguay
Reunion
Poland
Czech Republic
Cuba
Bulgaria
Republic of Korea
French Guiana
China, Macao SAR
Slovakia
Serbia
New Caledonia
Mauritius
Israel
Croatia
Cameroon
Turkey
Romania
Egypt
Bahamas
Colombia
Iraq
Gambia
Panama
Kuwait
Syrian Arab Republic
Peru
Lesotho
Azerbaijan
Sri Lanka
Honduras
Nicaragua
India
Pakistan



Exemplo em Wainer (2009)

```
> plot(year, total, xlab = "Ano", ylab = "Média de pessoas/cômodo",  
pch = 20)
```

```
> abline(lm(total ~  
year), lty = 2)
```

```
> cor(year, total)
```

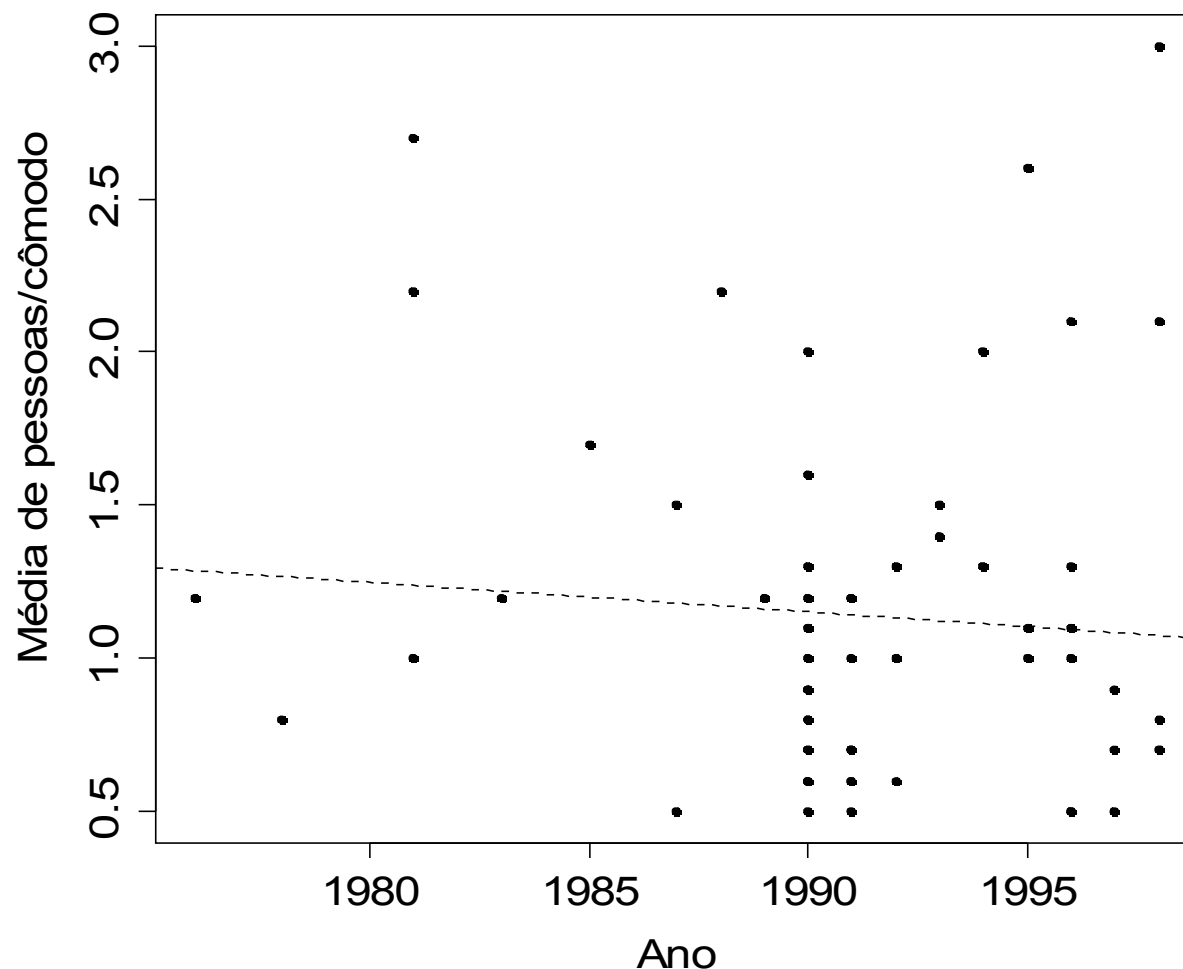
```
[1] NA
```

```
> cor(year, total,  
use = "complete")
```

```
[1] -0.07985232
```

Não há indício de relação entre a densidade de ocupação e o ano em que o dado foi coletado.

Há diferença entre a ocupação nos meios rural e urbano?



Se a resposta for não, podemos trabalhar com a média geral (`total`).

Exemplo em Wainer (2009)

```
> plot(rural, urban, xlab = "Média de pessoas/cômodo - rural",  
      ylab = "Média de pessoas/cômodo - urbano", pch = 20)
```

```
> abline(0, 1, lty = 2)
```

```
> cor(rural, urban,  
     use = "complete")
```

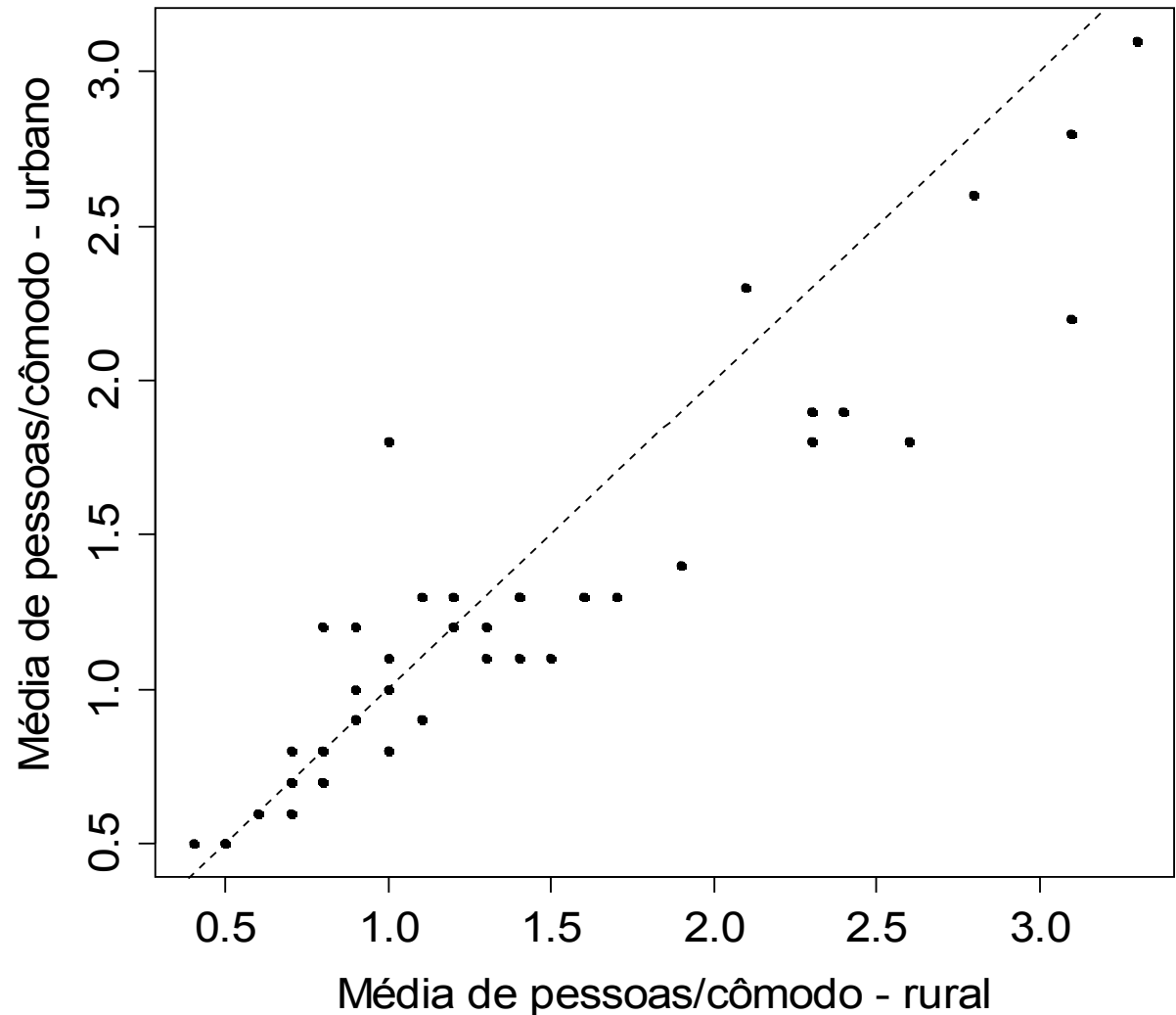
```
[1] 0.9385013
```

Correlação positiva forte.

Tendência de **maiores** médias no meio **rural**.

Situação econômica pode estar associada à densidade de ocupação?

Variável: PIB *per capita*.



Exemplo em Wainer (2009)

```
> pib = read.csv("Income_Dec2009.csv", header = TRUE, sep = ";")
```

```
> names(pib)
```

```
[1] "countryarea" "year" "GDPcapita"
```

```
> summary(pib)
```

countryarea	year	GDPcapita
Afghanistan: 1	Min. 2008	Min. : 138
Albania : 1	1st Qu. 2008	1st Qu.: 1218
Algeria : 1	Median 2008	Median : 4874
Andorra : 1	Mean 2008	Mean : 15772
Angola : 1	3rd Qu. 2008	3rd Qu.: 19291
Anguilla : 1	Max. 2008	Max. : 211501
(Other) :209	<u>NA's</u> : 6	<u>NA's</u> : 6

```
> pib$country[which.min(pib$GDPcapita)]
```

```
[1] Burundi
```

```
> pib$country[which.max(pib$GDPcapita)]
```

```
[1] Monaco
```

```
> dim(pib)
```

```
[1] 215 3
```

Dados de 2008
serão utilizados
apenas como
ilustração.

GDP: *per capita*
gross domestic
product (em US\$).

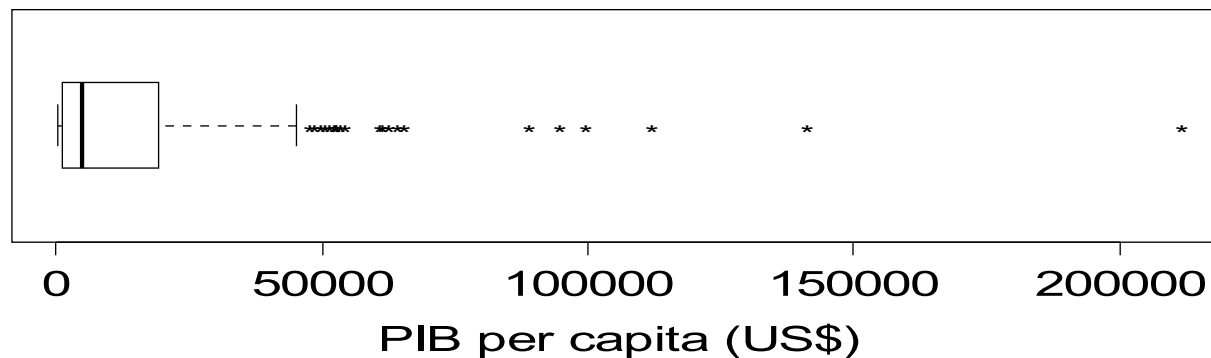
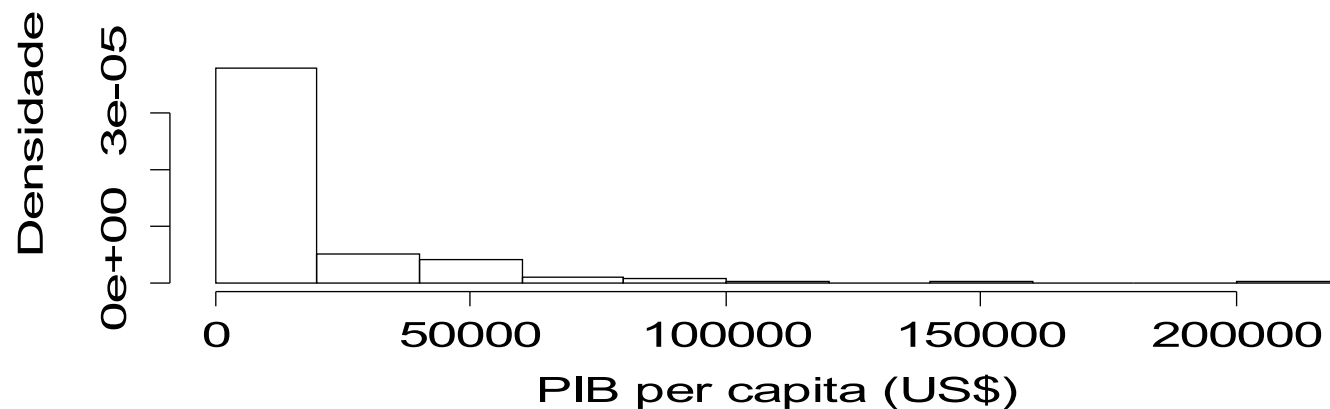
[http://unstats.un.org/
unsd/snaama/dnllist.
asp](http://unstats.un.org/unsd/snaama/dnllist.asp)

```
> pib$GDP[pib$country ==  
"Brazil"]
```

```
[1] 8311
```

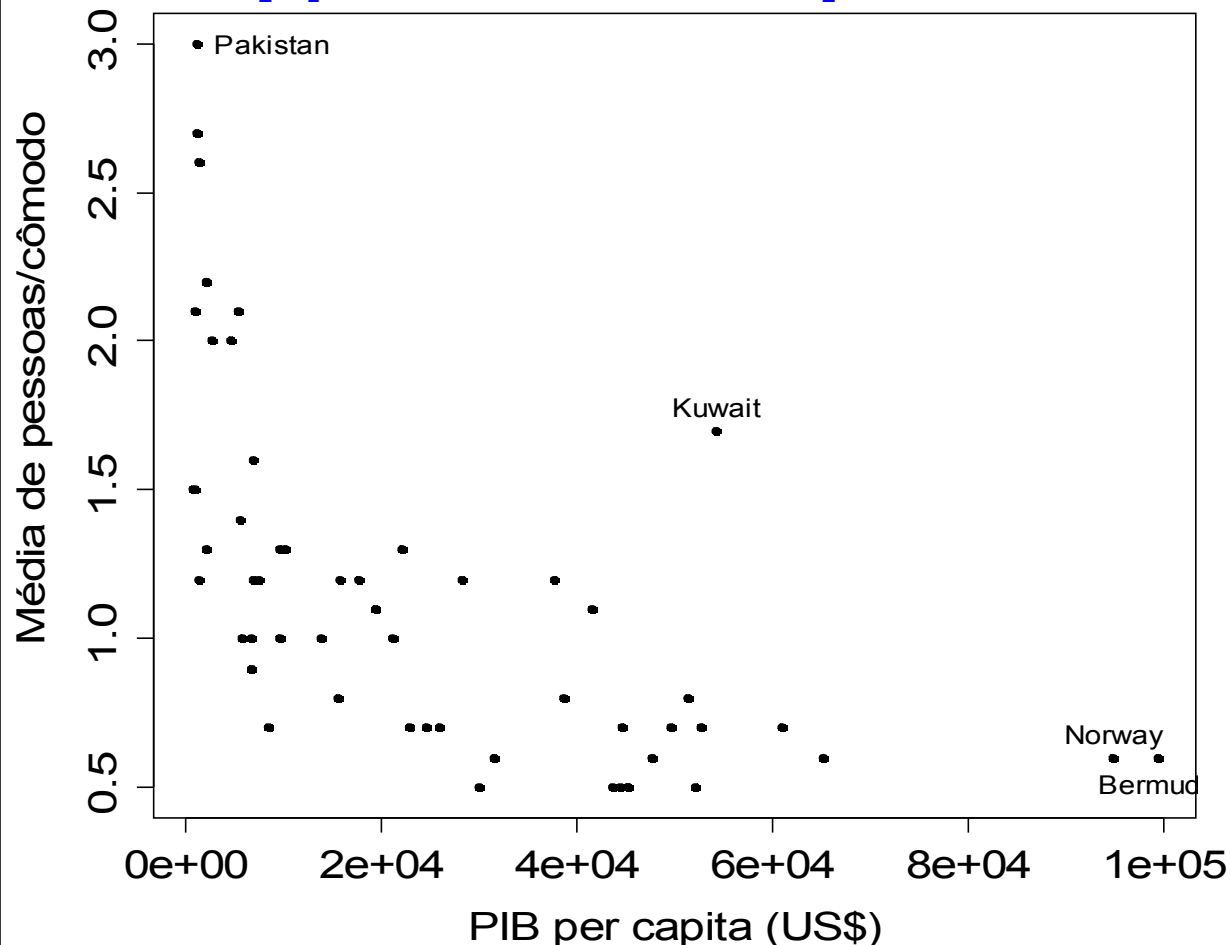
Exemplo em Wainer (2009)

```
> par(mfrow = c(2, 1))  
> hist(pib$GDP, freq = FALSE, xlab = "PIB per capita (US$)", ylab =  
"Densidade", main = "")  
> boxplot(pib$GDP, xlab = "PIB per capita (US$)", pch = "*",  
horizontal = TRUE)
```



Exemplo em Wainer (2009)

```
> pib60 = pib$GDP[match(countryarea, pib$countryarea)]  
> plot(pib60, total, pch = 20, ylab = "Média de pessoas/cômodo",  
      xlab = "PIB per capita (US$)")  
> identify(pib60, total, countryarea)
```



Associação **negativa**.

Assimetria em PIB
per capita.

Transformações de variáveis

Alguns objetivos: (a) **simetrizar** os dados e (b) **linearizar** a relação entre as variáveis.

$$\text{Família de transformações: } t = t(x) = \begin{cases} x^\lambda, & \text{se } \lambda \neq 0, \\ \log(x), & \text{se } \lambda = 0, \end{cases}, \text{ se } x > 0.$$

λ deve ser **escolhido** de modo a atingir o(s) objetivo(s), pelo menos aproximadamente.

$t(x)$ é monótona em x :

$$(1) \lambda \geq 0. \quad x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)} \Leftrightarrow t(x_{(1)}) \leq t(x_{(2)}) \leq \dots \leq t(x_{(n)}).$$

$$(2) \lambda < 0. \quad x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)} \Leftrightarrow t(x_{(n)}) \leq t(x_{(n-1)}) \leq \dots \leq t(x_{(1)}).$$

Posições são **preservadas** em (1) e são **invertidas** em (2).

Obs. Se M é a mediana de x , então $t(M)$ é a mediana de t .

Transformações **comuns**: $\log(x)$, $x^{1/2}$, $1/x$ e $1/x^2$.

Exemplo em Wainer (2009)

Transformação **logarítmica** da variável PIB *per capita*.

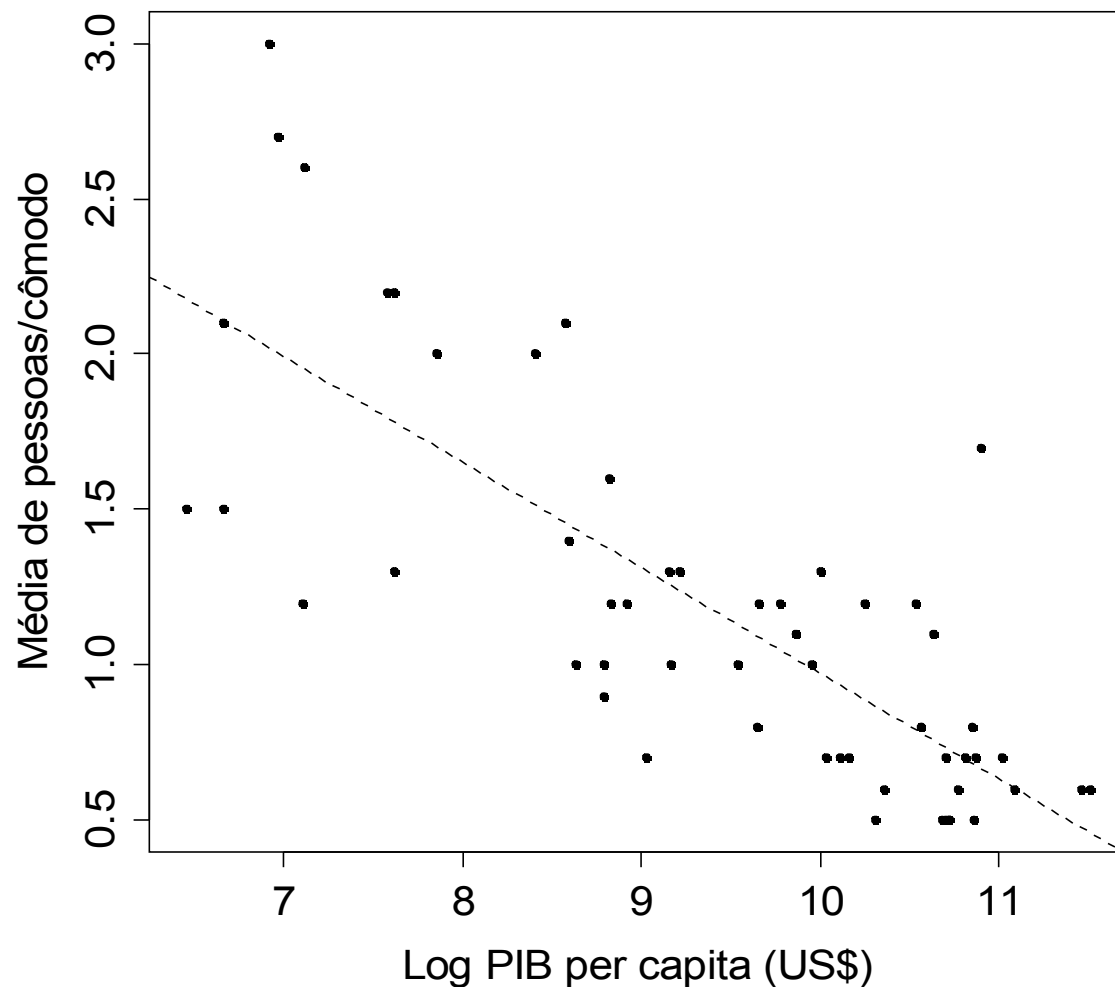
```
> plot(log(pib60),  
total, pch = 20, ylab =  
"Média de  
pessoas/cômodo", xlab =  
"Log PIB per capita  
(US$) ")
```

```
> abline(lm(total ~  
log(pib60)), lty = 2)
```

```
> cor(log(pib60),  
total, use =  
"complete")
```

```
[1] -0.7787283
```

Outras variáveis:
fertilidade e
desemprego
feminino.



8.2. Variáveis qualitativas

$x \in \{x_1, \dots, x_k\}$ e $y \in \{y_1, \dots, y_m\}$, $1 < k \leq n$ e $1 < m \leq n$.

f_{ij} : **frequência absoluta** do par (x_i, y_j) , $i = 1, \dots, k$ e $j = 1, \dots, m$.

Tabela de contingências (*contingency table*) ou tabela de **dupla entrada**: tabela com os diferentes pares (x_i, y_j) e suas frequências f_{ij} .

x	y					Totais
	y_1	...	y_j	...	y_m	
x_1	f_{11}	...	f_{1j}	...	f_{1m}	$f_{1\bullet}$
...
x_i	f_{i1}	...	f_{ij}	...	f_{im}	$f_{i\bullet}$
...
x_k	f_{k1}	...	f_{kj}	...	f_{km}	$f_{k\bullet}$
Totais	$f_{\bullet 1}$...	$f_{\bullet j}$...	$f_{\bullet m}$	n

$$\sum_{i=1}^k \sum_{j=1}^m f_{ij} = n.$$

$$f_{i\bullet} = \sum_{j=1}^m f_{ij}, \quad i = 1, \dots, k \text{ e}$$

$$\sum_{i=1}^k f_{i\bullet} = n.$$

$$f_{\bullet j} = \sum_{i=1}^k f_{ij}, \quad j = 1, \dots, m \text{ e } \sum_{j=1}^m f_{\bullet j} = n.$$

8.2. Variáveis qualitativas

Tabela de contingências:
distribuição de frequências conjunta de x e y .

Distribuição marginal de x
(frequências absolutas)

x	y					Totais
	y_1	...	y_j	...	y_m	
x_1	f_{11}	...	f_{1j}	...	f_{1m}	$f_{1\bullet}$
...
x_i	f_{i1}	...	f_{ij}	...	f_{im}	$f_{i\bullet}$
...
x_k	f_{k1}	...	f_{kj}	...	f_{km}	$f_{k\bullet}$
Totais	$f_{\bullet 1}$...	$f_{\bullet j}$...	$f_{\bullet m}$	n

Distribuição marginal de y
(frequências absolutas)

x	y					Totais
	y_1	...	y_j	...	y_m	
x_1	f_{11}	...	f_{1j}	...	f_{1m}	$f_{1\bullet}$
...
x_i	f_{i1}	...	f_{ij}	...	f_{im}	$f_{i\bullet}$
...
x_k	f_{k1}	...	f_{kj}	...	f_{km}	$f_{k\bullet}$
Totais	$f_{\bullet 1}$...	$f_{\bullet j}$...	$f_{\bullet m}$	n

8.2. Variáveis qualitativas

Frequências **relativas** (f^*) são bastante utilizadas em tabelas de contingências.

Três possibilidades de cálculo: (a) em relação ao total geral (n^o de observações = n), (b) em relação ao total de cada linha ($f_{i\cdot}$) e (c) em relação ao total de cada coluna ($f_{\cdot j}$).

(a)

x	y					Totais
	y_1	...	y_j	...	y_m	
x_1	f_{11} / n	...	f_{1j} / n	...	f_{1m} / n	$f_{1\cdot} / n$
...
x_i	f_{i1} / n	...	f_{ij} / n	...	f_{im} / n	$f_{i\cdot} / n$
...
x_k	f_{k1} / n	...	f_{kj} / n	...	f_{km} / n	$f_{k\cdot} / n$
Totais	$f_{\cdot 1} / n$...	$f_{\cdot j} / n$...	$f_{\cdot m} / n$	1

Distribuição marginal de x

Distribuição marginal de y

$$\sum_{i=1}^k \sum_{j=1}^m \frac{f_{ij}}{n} = 1.$$

8.2. Variáveis qualitativas

(b)

x	y					Totais
	y ₁	...	y _j	...	y _m	
x ₁	f ₁₁ / f _{1•}	...	f _{1j} / f _{1•}	...	f _{1m} / f _{1•}	1
...
x _i	f _{i1} / f _{i•}	...	f _{ij} / f _{i•}	...	f _{im} / f _{i•}	1
...
x _k	f _{k1} / f _{k•}	...	f _{kj} / f _{k•}	...	f _{km} / f _{k•}	1

Distribuição
condicional de y
dado $x = x_i$.

k distribuições
condicionais de y .

(c)

x	y				
	y ₁	...	y _j	...	y _m
x ₁	f ₁₁ / f _{•1}	...	f _{1j} / f _{•j}	...	f _{1m} / f _{•m}
...
x _i	f _{i1} / f _{•1}	...	f _{ij} / f _{•j}	...	f _{im} / f _{•m}
...
x _k	f _{k1} / f _{•1}	...	f _{kj} / f _{•j}	...	f _{km} / f _{•m}
Totais	1	...	1	...	1

Distribuição
condicional de x
dado $y = y_j$.

m distribuições
condicionais de x .

8.2. Variáveis qualitativas

Que frequência relativa utilizar?

(a) Relação causal **bilateral** ($x \leftrightarrow y$): em relação ao **total geral** (n).

(b) Relação causal **unilateral** ($x \rightarrow y$): em relação ao **total** de cada **linha** ($f_{i\cdot}$).

(c) Relação causal **unilateral** ($y \rightarrow x$): em relação ao **total** de cada **coluna** ($f_{\cdot j}$).

Obs. 1. Em (b) temos k distribuições **condicionais** de y . Quanto **mais semelhantes** forem estas distribuições, **mais fraca** é a **associação** entre x e y .

Obs. 2. Em (c) é usual mudar **intercambiar** os nomes, de modo que x ocupe as **linhas** e y ocupe as **colunas** da tabela de contingências.

Exemplo

Intenção de voto (%) para presidente de acordo com o domicílio eleitoral, 20 e 21/5/2010.

Região	Candidato(a)					Total
	Serra	Dilma	Marina	Em branco, nulo ou nenhum	Não sabe	
SE	40	33	12	7	8	100
S	38	35	12	4	10	99
NE	33	44	8	1	11	97
N e CO	34	40	14	5	7	100

Fonte. DataFolha (http://datafolha.folha.uol.com.br/po/ver_po.php?session=971).

Sugestão. Quanto um total é diferente de 100%, a compensação é efetuada nas frequências de maiores valores.

A região do domicílio eleitoral (x) influencia a intenção de voto (y) ?

Como quantificar?

Independência

x e y são independentes se, e somente se,

$$f_{ij} = \frac{f_{i\cdot} \cdot f_{\cdot j}}{n}, \quad j = 1, \dots, m \text{ e } i = 1, \dots, k.$$

De forma equivalente, $\frac{f_{ij}}{n} = \frac{f_{i\cdot}}{n} \frac{f_{\cdot j}}{n}, \quad j = 1, \dots, m \text{ e } i = 1, \dots, k.$

		y					
x	y ₁	...	y _j	...	y _m	Totais	
x ₁	f ₁₁ / n	...	f _{1j} / n	...	f _{1m} / n	f _{1·} / n	
...	Conjunta	
x _i	f _{i1} / n	...	f _{ij} / n	...	f _{im} / n	f _{i·} / n	
...	
x _k	f _{k1} / n	...	f _{kj} / n	...	f _{km} / n	f _{k·} / n	
Totais	f _{·1} / n	...	f _{·j} / n	...	f _{·m} / n	1	

Marginal de y

Marginal de x

Justificativa. Adaptação do conceito de independência entre as v.a. discretas X e Y: $P(X = a, Y = b) = P(X = a) P(Y = b).$

Coeficientes de associação

Uma das **várias** medidas de associação entre variáveis qualitativas.

Baseado nas **diferenças** entre as frequências absolutas **observadas** (f_{ij}) e as frequências **calculadas** supondo **independência** entre x e y ($f_{ij}^{\text{ind}} = f_{i\cdot} \cdot f_{\cdot j} / n$):

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^m \frac{(f_{ij} - f_{ij}^{\text{ind}})^2}{f_{ij}^{\text{ind}}} : \text{qui - quadrado de Pearson.}$$

$\chi^2 = 0 \Leftrightarrow$ **ausência** de associação entre x e y.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} : \text{coeficiente de contingência.}$$

O valor **máximo** de C depende de k e m.

$$T = \sqrt{\frac{\chi^2}{n \sqrt{(k-1)(m-1)}}} : \text{coeficiente de Tschuprow.}$$

Obs. $0 \leq T \leq 1$.

Coeficientes de associação

Exemplo. Tabela $k \times k$ ($m = k$).

x	y					Totais
	y ₁	...	y _i	...	y _k	
x ₁	f ₁₁	...	—	...	—	f ₁₁
...
x _i	—	...	f _{ii}	...	—	f _{ii}
...
x _k	—	...	—	...	f _{kk}	f _{kk}
Totais	f ₁₁	...	f _{ii}	...	f _{kk}	n

Exercício. Provar que, neste caso, $\chi^2 = n(k - 1)$. Logo, $T = 1$.

Apresente outros exemplos nos quais $T = 1$.

Funções em R

```
> library(ineq)
```

```
> ?Ilocos
```

Dados coletados em domicílios nas Filipinas.

```
Ilocos {ineq}
```

R Documentation

```
Income Metadata from Ilocos, Philippines
```

```
Description
```

```
Income metadata from surveys conducted by the Philippines' National Statistics Office.
```

```
Usage
```

```
data(Ilocos)
```

```
> data(Ilocos)
```

```
> dados = Ilocos
```

```
> dim(dados) [1] 632 8 n = 632 observações de 8 variáveis.
```

```
> names(dados)
```

```
"income" "sex" "family.size" "urbanity" "province" "AP.income"  
"AP.family.size" "AP.weight"
```

```
> summary(dados[, c("sex", "urbanity", "province")])
```

```
sex                urbanity                province  
female:114        rural:301          Ilocos Norte: 65  
male :518         urban:331          Ilocos Sur   : 68  
La Union         :116  
Pangasinan      :383
```

```
> class(dados$province)
```

```
[1] "factor"
```

Variável qualitativa: fator (*factor*).

Funções em R

```
> attach(dados)
> levels(urbanity) = c("Rural", "Urbana") Um fator tem níveis (levels).
> (tab1 = table(province, urbanity))
```

```
      urbanity
province Rural Urbana
Ilocos Norte    47    18
Ilocos Sur      45    23
La Union        71    45
Pangasinan     138   245
```

x: province

y: urbanity

Tabela 4×2 com f_{ij} , $i = 1, \dots, 4$ ($k = 4$) e $j = 1, 2$ ($m = 2$).

```
> addmargins(tab1, 1) > addmargins(tab1, 2) > addmargins(tab1, 1:2)
```

```
      urbanity
province Rural Urbana
Ilocos Norte    47    18
Ilocos Sur      45    23
La Union        71    45
Pangasinan     138   245
Sum            301   331
```

```
      urbanity
province Rural Urbana Sum
Ilocos Norte    47    18  65
Ilocos Sur      45    23  68
La Union        71    45 116
Pangasinan     138   245 383
```

```
      urbanity
province Rural Urbana Sum
Ilocos Norte    47    18  65
Ilocos Sur      45    23  68
La Union        71    45 116
Pangasinan     138   245 383
Sum            301   331 632
```

Para estudar a relação `province` \rightarrow `urbanity`, qual das três tabelas é mais útil?

Funções em R

```
> margin.table(tab1, margin = 1)
```

```
province
```

```
Ilocos Norte      Ilocos Sur      La Union      Pangasinan
                65                68                116                383
```

```
urbanity
```

```
> margin.table(tab1, margin = 2)
```

```
Rural Urbana
```

```
301  331
```

```
> prop.table(tab1)
```

```
> (tab1rel = prop.table(tab1,
margin = 1))
```

```
                urbanity
province        Rural    Urbana
Ilocos Norte 0.07436709 0.02848101
Ilocos Sur   0.07120253 0.03639241
La Union     0.11234177 0.07120253
Pangasinan   0.21835443 0.38765823
```

Frequências relativas em relação
ao total geral (soma = 1).

```
                urbanity
province        Rural    Urbana
Ilocos Norte 0.7230769 0.2769231
Ilocos Sur   0.6617647 0.3382353
La Union     0.6120690 0.3879310
Pangasinan   0.3603133 0.6396867
```

Distribuição condicional
de urbanity | province.

Funções em R

```
> addmargins(tab1rel, 2)
```

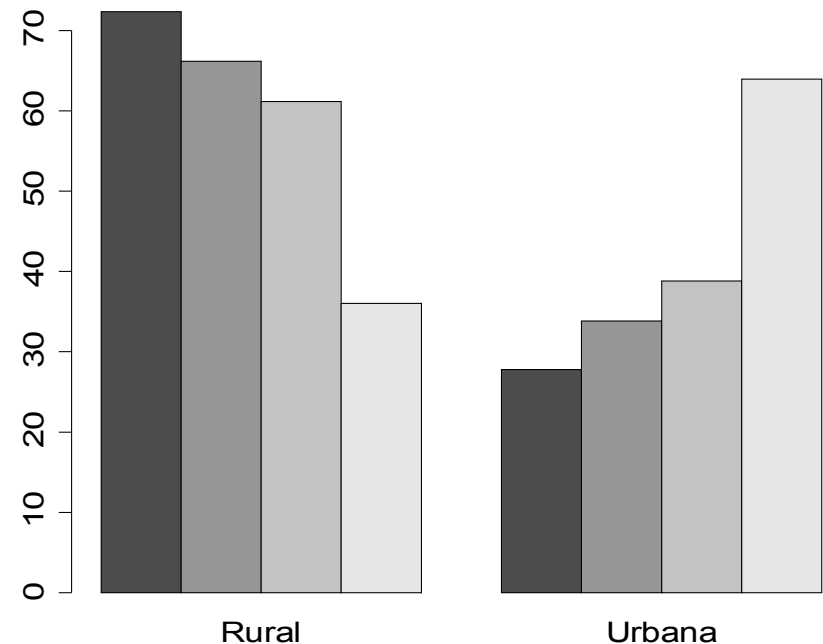
province	urbanity		Sum
	Rural	Urbana	
Ilocos Norte	0.7230769	0.2769231	1.0000000
Ilocos Sur	0.6617647	0.3382353	1.0000000
La Union	0.6120690	0.3879310	1.0000000
Pangasinan	0.3603133	0.6396867	1.0000000

```
> print(addmargins(tab1rel, 2)  
* 100, digits = 3)
```

province	urbanity		Sum
	Rural	Urbana	
Ilocos Norte	72.3	27.7	100.0
Ilocos Sur	66.2	33.8	100.0
La Union	61.2	38.8	100.0
Pangasinan	36.0	64.0	100.0

```
> tab1relp = tab1rel * 100
```

```
> barplot(tab1relp, beside = TRUE)
```

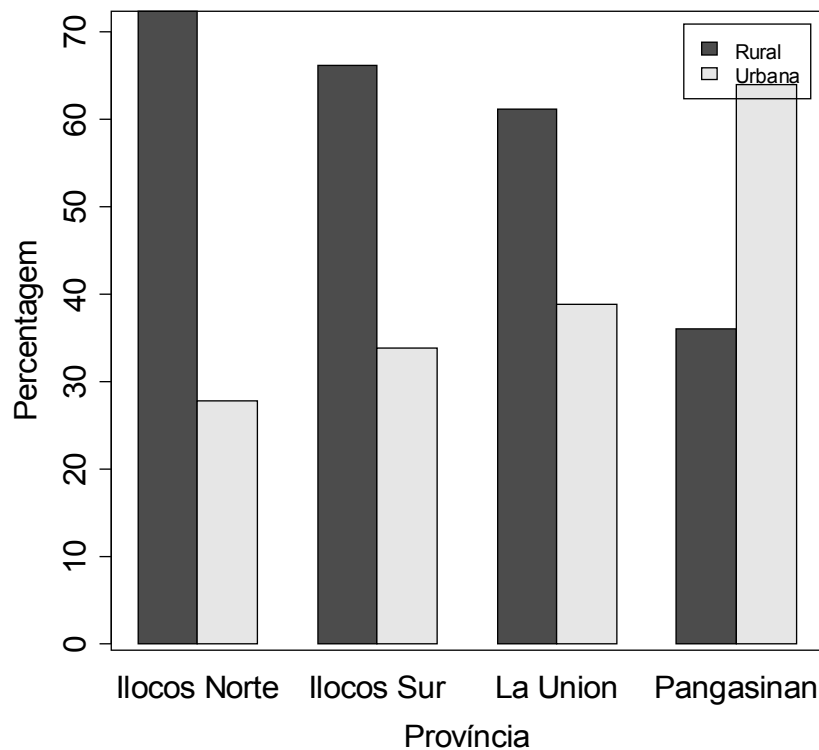


Era o gráfico que esperávamos?

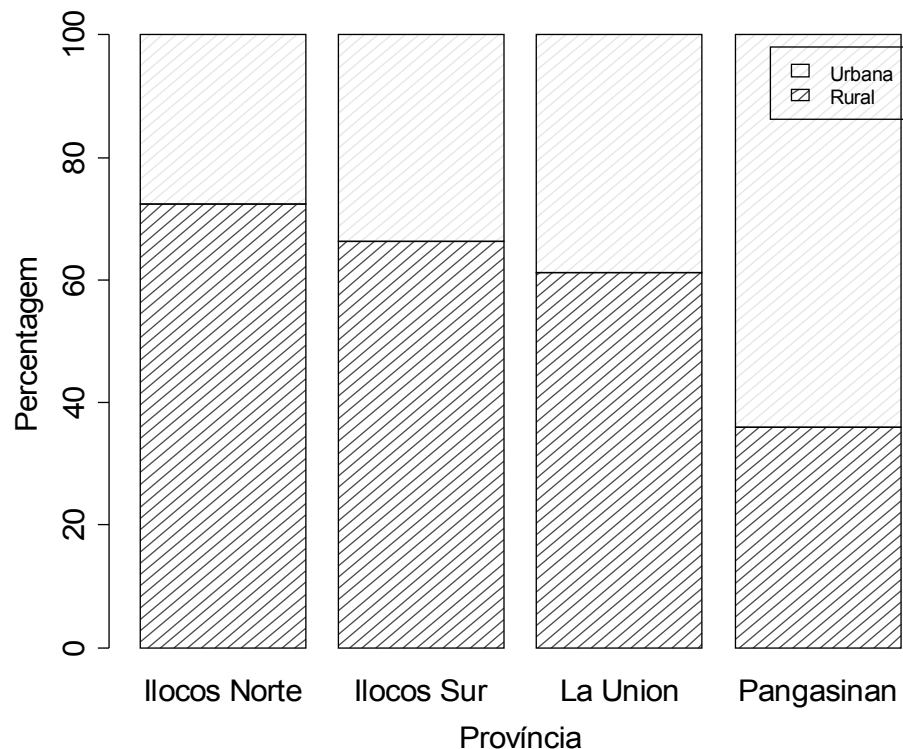
Funções em R

```
> barplot(t(tab1relp), beside =  
TRUE, xlab = "Província", ylab =  
"Porcentagem", legend.text =  
TRUE)
```

```
> box()
```



```
> barplot(t(tab1relp), xlab =  
"Província", ylab =  
"Porcentagem", density = 15,  
legend.text = TRUE)
```



Exercício. Verificar a utilização de cores e a posição da legenda.

Funções em R

Gráfico de **mosaico** (*mosaic plot*). Representação de uma tabela de contingências usando retângulos com **áreas** proporcionais às **frequências**.

```
> levels(sex) = c("Feminino", "Masculino")
> tab2 = table(province, sex)
> tab2rel = prop.table(tab2, margin = 1)
> print(addmargins(tab2rel, 2) * 100, digits = 3)
```

province	sex		Sum
	Feminino	Masculino	
Ilocos Norte	12.3	87.7	100.0
Ilocos Sur	26.5	73.5	100.0
La Union	15.5	84.5	100.0
Pangasinan	18.3	81.7	100.0

province	sex	
	Feminino	Masculino
Ilocos Norte	8	57
Ilocos Sur	18	50
La Union	18	98
Pangasinan	70	313

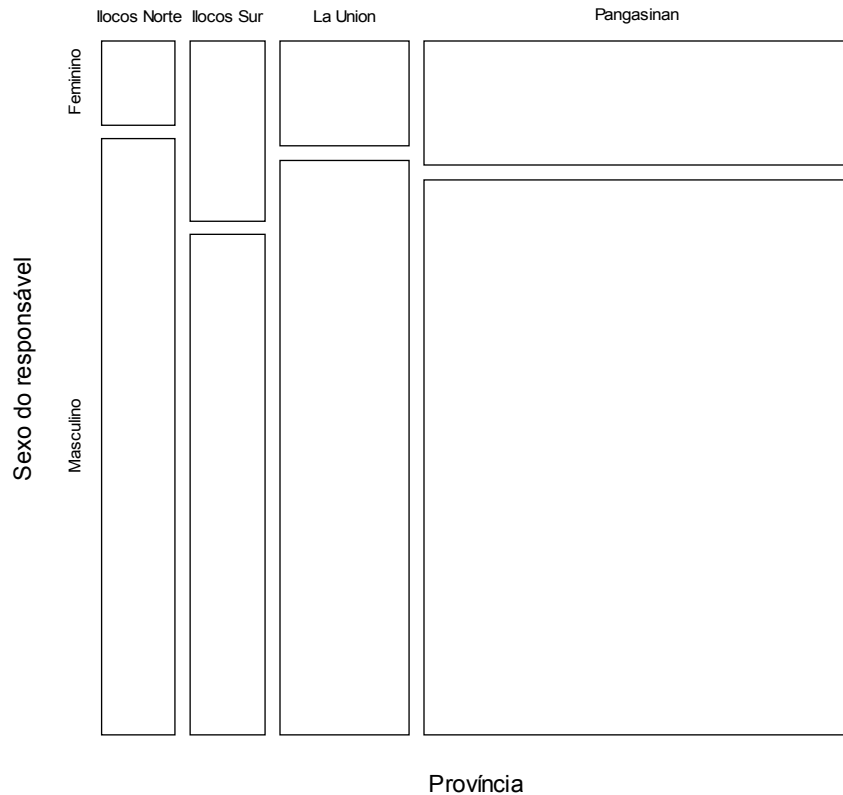
Supondo **independência** entre province e sex:

```
> tab2marg = addmargins(tab2, 1:2)
> k = nrow(tab2marg) - 1
> m = ncol(tab2marg) - 1
> n = sum(tab2)
> tab2ind = tab2marg[1:k, m + 1] %*% t(tab2marg[k + 1, 1:m]) / n
> rownames(tab2ind) = rownames(tab2)
> colnames(tab2ind) = colnames(tab2)
```

Funções em R

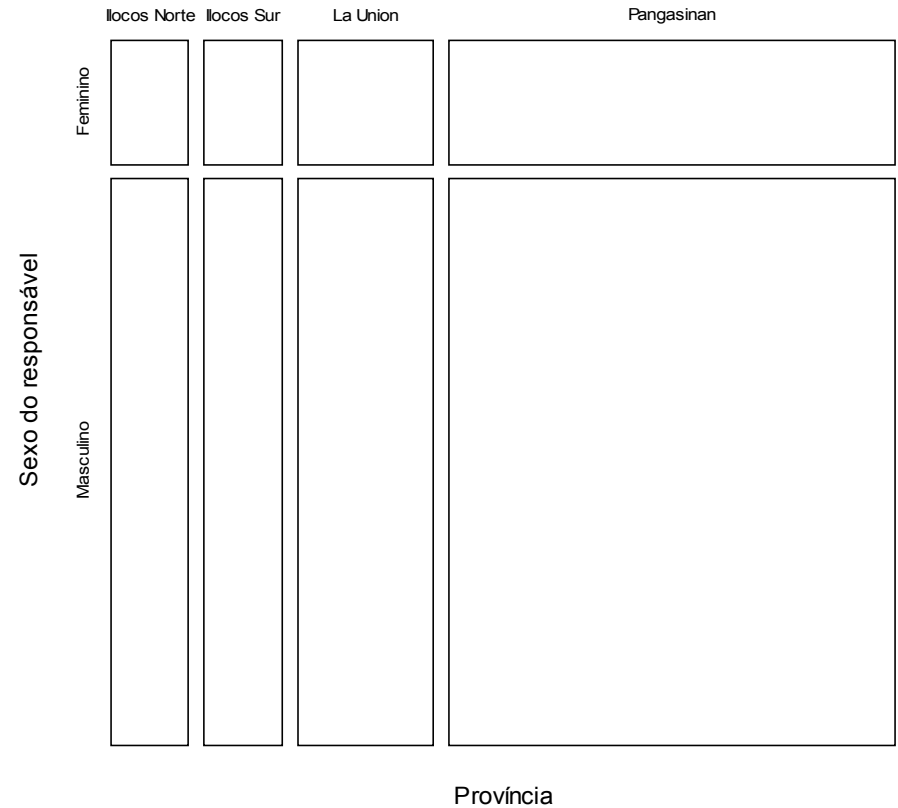
```
> mosaicplot(tab2, ylab = "Sexo do responsável", xlab = "Província", col = "white", main = "Dados observados")
```

Dados observados



```
> mosaicplot(tab2ind, ylab = "Sexo do responsável", xlab = "Província", col = "white", main = "Independência")
```

Independência



Retângulos com **bases** proporcionais às frequências da variável `province` e **alturas** proporcionais às frequências da variável `sex`.

Funções em R

Obs. Substitua `mosaicplot` por `plot` na lâmina anterior. O resultado é diferente? Como explicar?

```
> X2 = sum((tab2 - tab2ind)^2 / tab2ind)
```

```
> (TproW = sqrt(X2 / (n * sqrt((k - 1) * (m - 1))))))
```

```
[1] 0.06910562 Coeficiente de Tschuprow
```

Obs. O valor de χ^2 e a tabela supondo independência (`tab2ind`) podem ser obtidos usando a função `chisq.test`.

Um gráfico **não** muito recomendado:

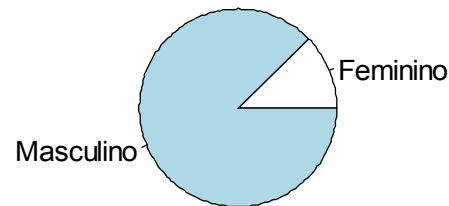
```
> nlinhas = ceiling(k / 2)
```

```
> par(mfrow = c(nlinhas, 2))
```

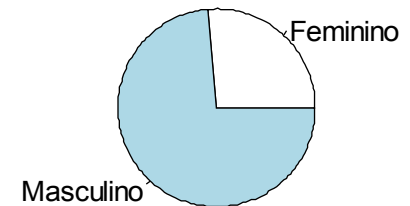
```
> for (i in 1:k) pie(tab2[i,],  
main = rownames(tab2rel)[i])
```

Parece **mais difícil** comparar áreas de setores do que alturas de retângulos (em um gráfico de barras).

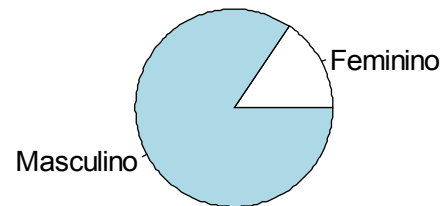
Ilocos Norte



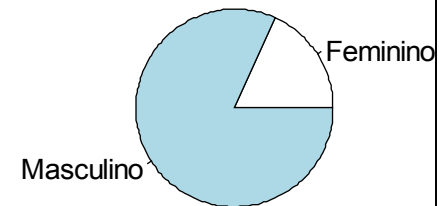
Ilocos Sur



La Union



Pangasinan



8.3. Variáveis qualitativas e quantitativas

$x \in \{x_1, \dots, x_k\}$, $1 < k \leq n$, é uma variável **qualitativa** e y é uma variável **quantitativa**.

Dados observados: n pares de valores (x_j, y_j) , sendo que $x_j \in \{x_1, \dots, x_k\}$, $j = 1, \dots, n$.

É muito **comum** o interesse na relação causal unilateral $x \rightarrow y$.

Apresentação dos dados: medidas resumo e gráficos de y para cada **nível** de x .

Cada nível x_i ocorre f_i vezes (frequência). Para cada nível x_i calculamos a variância s_i^2 dos valores y_j para os quais $x_j = x_i$, $j = 1, \dots, n$ e $i = 1, \dots, k$.

Média **ponderada** das variâncias:

$$\overline{s^2} = \frac{\sum_{i=1}^k f_i s_i^2}{\sum_{i=1}^k f_i} = \frac{\sum_{i=1}^k f_i s_i^2}{n}.$$

Variância de y :

$$s^2 = \frac{1}{n-1} \sum_{j=1}^n (y_j - \bar{y})^2.$$

Obs. Podemos ter $s_i^2 = 0$, mas $s^2 > 0$.

Ganho na variância: $s^2 - \overline{s^2}$. **Ganho** relativo na variância: $R^2 = \frac{s^2 - \overline{s^2}}{s^2}$, $0 \leq R^2 \leq 1$.

Quanto **maior** R^2 , **mais forte** a **associação** entre x e y .

Quanto **maior** R^2 , **maior** o **poder de explicação** de x para y (em termos de variabilidade).

Funções em R

Dados Ilocos na lâmina 40.

```
> names(dados)
```

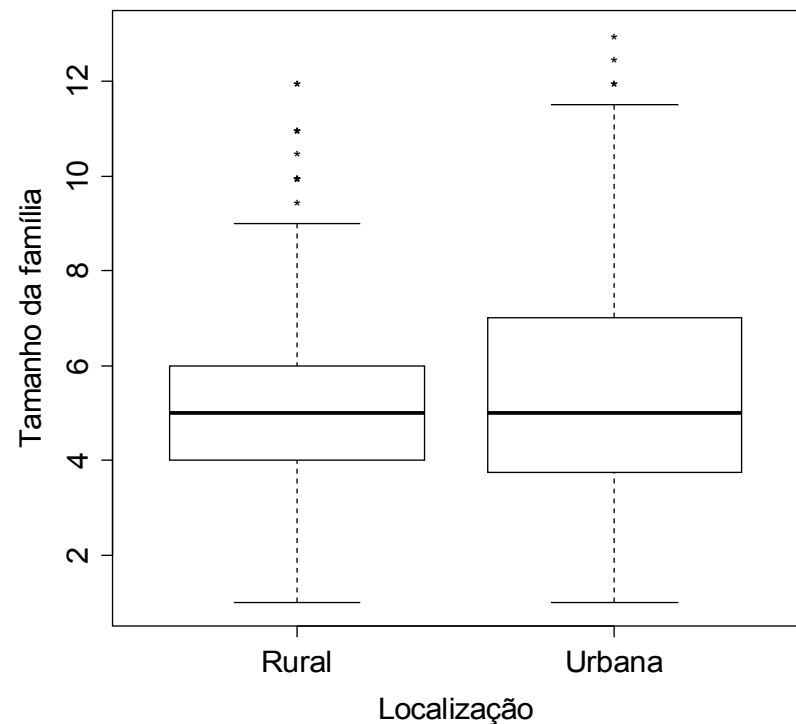
```
      y      x      y      x      x      "AP.income"  
"income" "sex" "family.size" "urbanity" "province" "AP.income"  
"AP.family.size" "AP.weight"
```

Fórmula: $y \sim x$ (y como função de x ou y depende de x).

```
> summary(dados[, c("income", "family.size")])
```

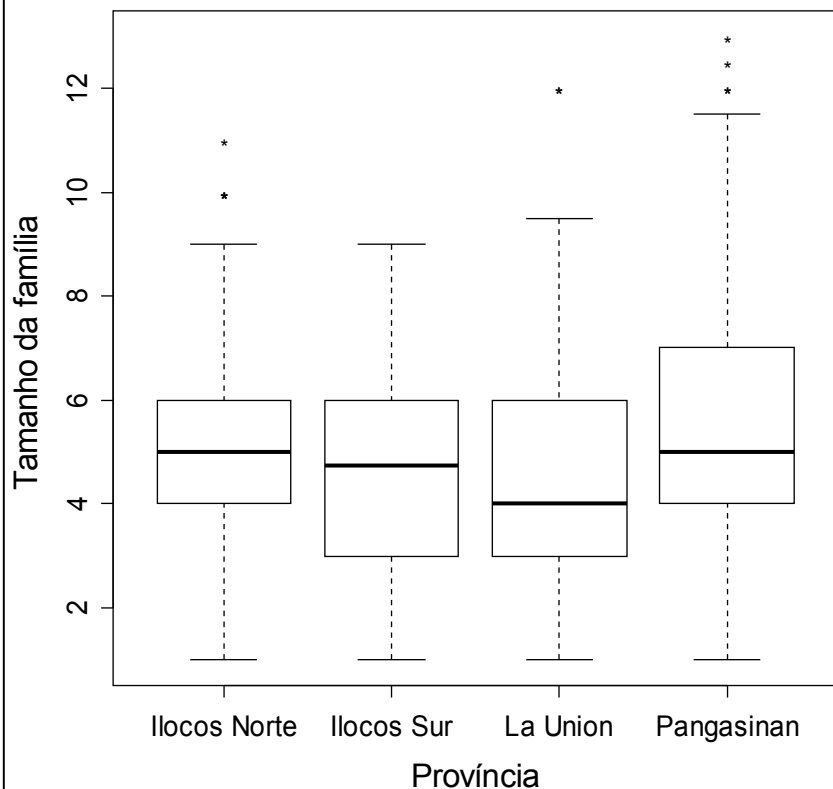
income	family.size
Min. : 6067	Min. : 1.000
1st Qu.: 48001	1st Qu.: 3.875
<u>Median : 75926</u>	Median : 5.000
<u>Mean : 112292</u>	Mean : 5.193
3rd Qu.: 137068	3rd Qu.: 6.500
Max. : 835742	Max. : 13.000

```
> plot(family.size ~ urbanity,  
xlab = "Localização", ylab =  
"Tamanho da família", pch =  
"*")
```

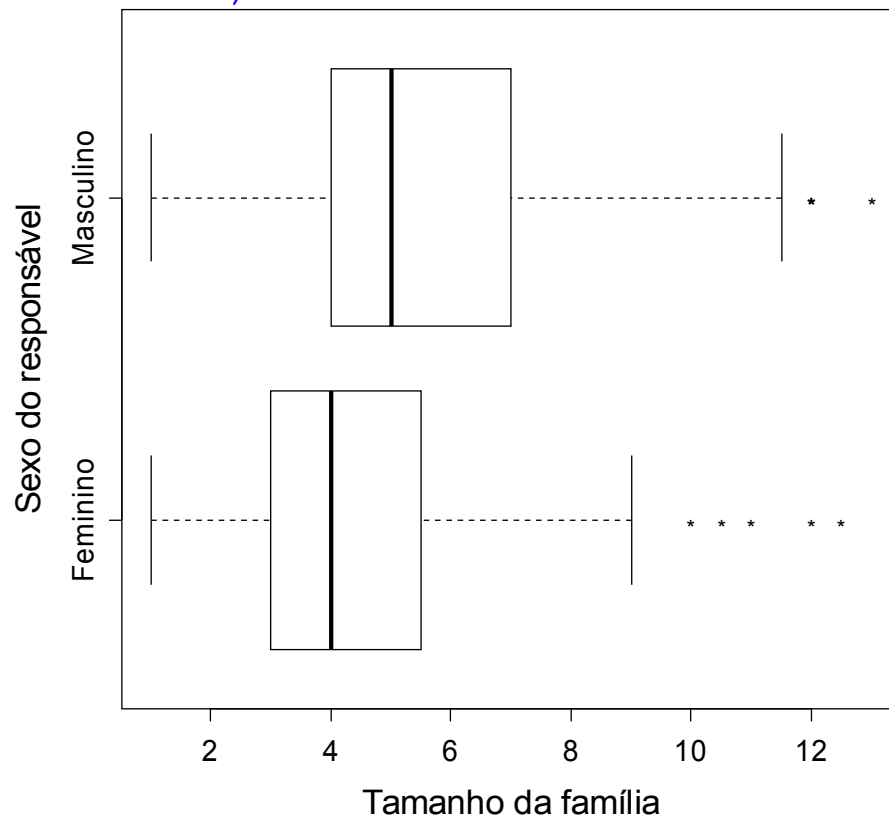


Funções em R

```
> plot(family.size ~ province,  
xlab = "Província", ylab =  
"Tamanho da família", pch =  
"*")
```



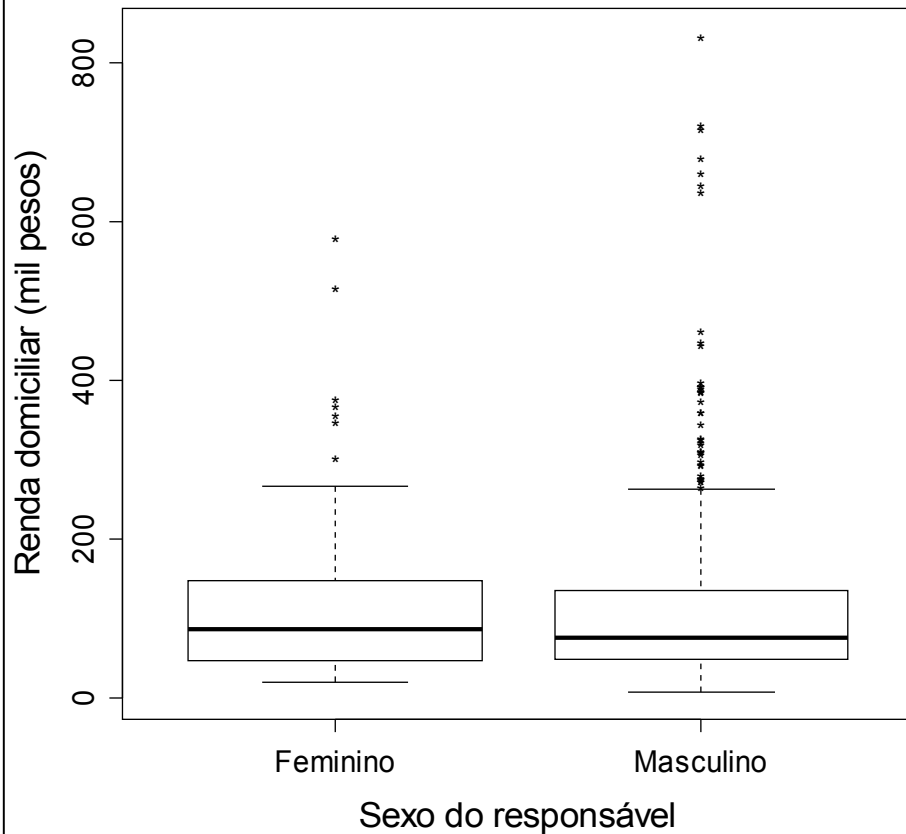
```
> plot(family.size ~ sex,  
xlab = "Sexo do responsável", ylab = "Tamanho da família",  
pch = "*", horizontal =  
TRUE)
```



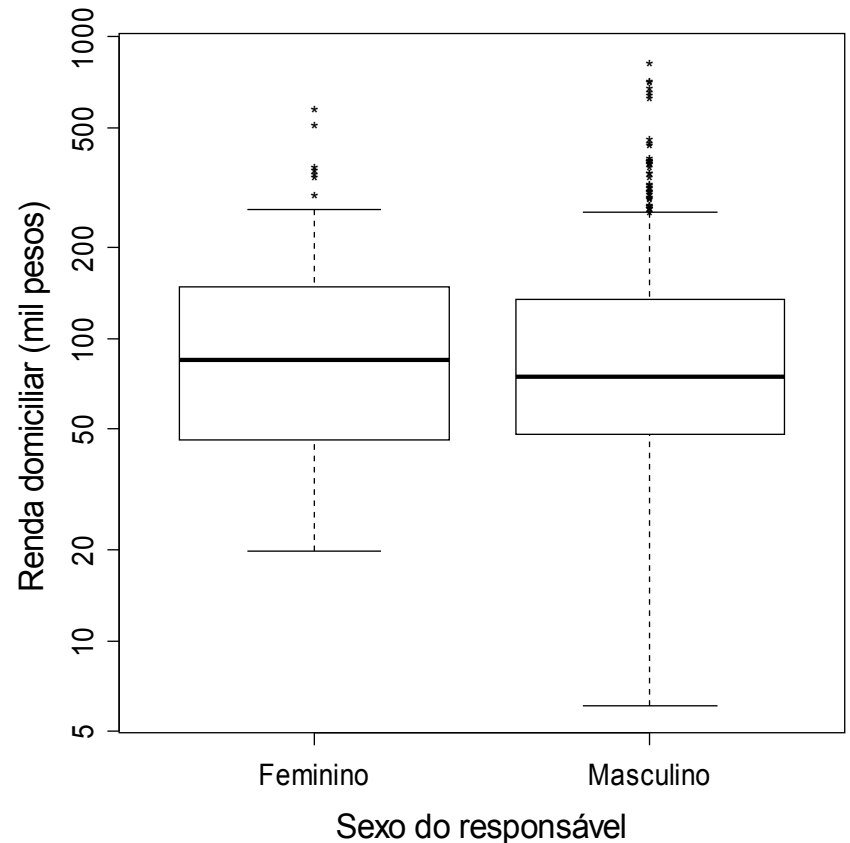
Exercício. Apresente o gráfico à esquerda com níveis em ordem decrescente da mediana.

Funções em R

```
> plot(income / 1000 ~ sex,  
xlab = "Sexo do responsável",  
ylab = "Renda domiciliar (mil pesos)",  
pch = "*")
```



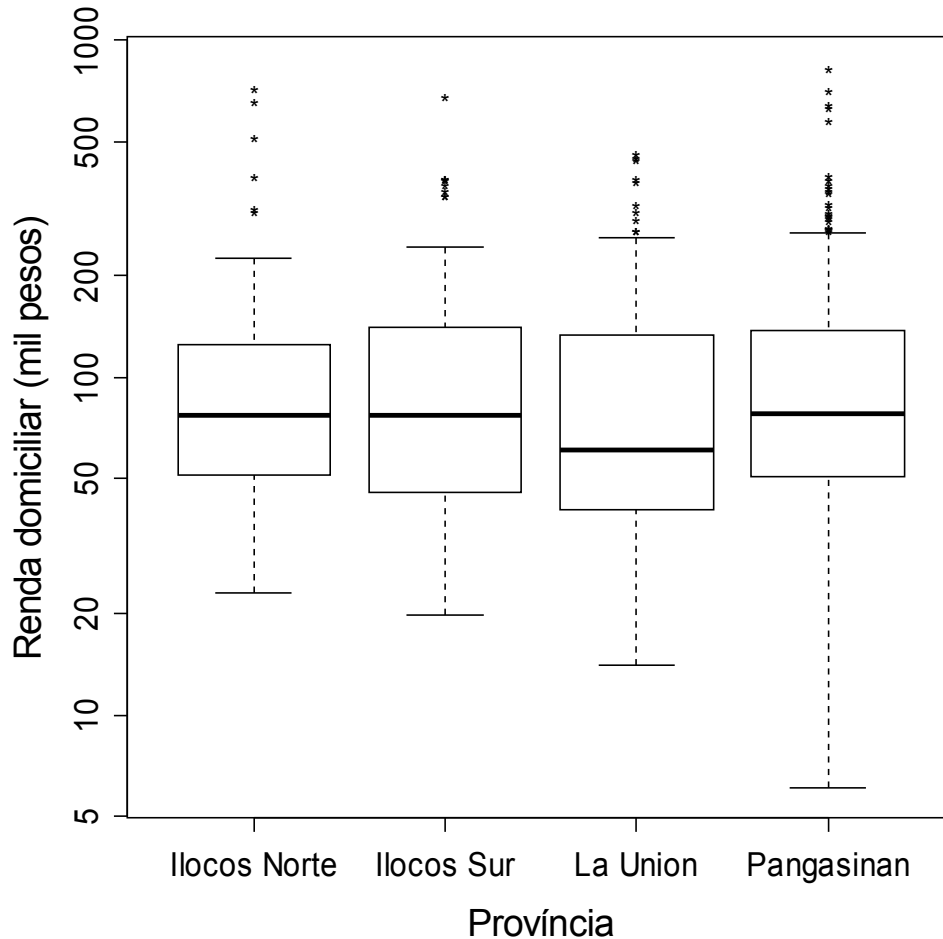
```
> plot(income / 1000 ~ sex,  
xlab = "Sexo do responsável",  
log = "y", ylab = "Renda  
domiciliar (mil pesos)", pch =  
"*)
```



Distribuição da renda é assimétrica. [Exercício](#). Apresente medidas de assimetria.

Funções em R

```
> plot(income / 1000 ~ province,  
xlab = "Província", log = "y",  
ylab = "Renda domiciliar (mil  
pesos)", pch = "*")
```



Médias e variâncias do tamanho da família por província:

```
> (tabmed = tapply(family.size,  
province, "mean"))
```

```
Ilocos Norte    Ilocos Sur  
5.084615       4.683824
```

```
La Union        Pangasinan  
4.607759       5.479112
```

```
> (tabvar = tapply(family.size,  
province, "var"))
```

```
Ilocos Norte    Ilocos Sur  
4.504447       3.618690
```

```
La Union        Pangasinan  
4.186113       5.376526
```

```
> (s2 = var(family.size))
```

```
[1] 5.000712
```

Funções em R

Gráfico de médias e desvios padrão do tamanho da família por província:

```
> limy = c(0, 1.1 * max(tabmed +  
sqrt(tabvar)))
```

```
> gbarras = barplot(tabmed, xlab =  
"Província", ylab = "Tamanho  
médio da família", ylim = limy,  
col = "black", density = 10)
```

```
> arrows(gbarras, tabmed,  
gbarras, tabmed + sqrt(tabvar),  
angle = 90)
```

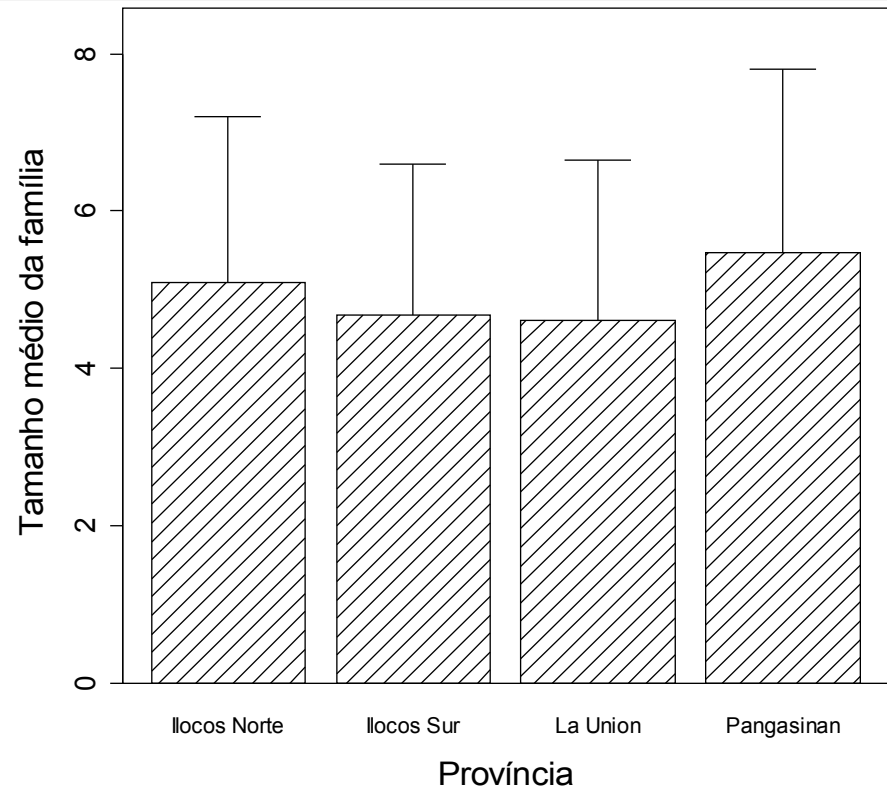
```
> box()
```

Exercício. Apresente o gráfico com níveis em ordem decrescente da média.

```
> fprov = table(province)
```

```
> (s2barra = weighted.mean(tabvar,  
fprov))
```

```
[1] 4.879207
```



```
> (R2 = 1 - s2barra / s2)
```

```
[1] 0.02429767
```

A variável `province` explica cerca de **2,4%** da variabilidade do tamanho da família.