

8. Medidas de associação

Parte 2

2012



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} | $f_{1\bullet}$ |
| \vdots | \vdots | ... | \vdots | ... | \vdots | \vdots |
| x_i | f_{i1} | ... | f_{ij} | ... | f_{im} | $f_{i\bullet}$ |
| \vdots | \vdots | ... | \vdots | ... | \vdots | \vdots |
| x_k | f_{k1} | ... | f_{kj} | ... | f_{km} | $f_{k\bullet}$ |
| Totais | $f_{\bullet 1}$ | ... | $f_{\bullet j}$ | ... | $f_{\bullet m}$ | n |

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

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

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

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



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•} |
| ... | ... | ... | ... | ... | ... | ... |
| x _i | f _{i1} | ... | f _{ij} | ... | f _{im} | f _{i•} |
| ... | ... | ... | ... | ... | ... | ... |
| x _k | f _{k1} | ... | f _{kj} | ... | f _{km} | f _{k•} |
| Totais | f _{•1} | ... | f _{•j} | ... | f _{•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•} |
| ... | ... | ... | ... | ... | ... | ... |
| x _i | f _{i1} | ... | f _{ij} | ... | f _{im} | f _{i•} |
| ... | ... | ... | ... | ... | ... | ... |
| x _k | f _{k1} | ... | f _{kj} | ... | f _{km} | f _{k•} |
| Totais | f _{•1} | ... | f _{•j} | ... | f _{•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 (nº de observações = n),
- (b) em relação ao total de cada linha ($f_{i\bullet}$) e
- (c) em relação ao total de cada coluna ($f_{\bullet j}$).

(a)

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

Distribuição marginal de x

Distribuição marginal de y

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



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** ($f_{i\bullet}$).

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

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,

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

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

| | | 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 ($f_{ij}^{ind} = f_{i\cdot} \cdot f_{\cdot j} / n$):

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

$Q^2=0 \Rightarrow$ ausência de associação entre x e y

$Q^2>0 \Rightarrow$ comparar com o quantil de uma v.a. com distribuição $\chi^2_{(k-1)(m-1)}$

Coeficiente de Contingência $C = \sqrt{\frac{Q^2}{Q^2 + n}}$

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

Coeficiente de Tschuprow $T = \sqrt{\frac{Q^2}{n \sqrt{(k-1)(m-1)}}$

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, $Q^2 = n(k - 1)$. 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 $\text{province} \rightarrow \text{urbanity}$, 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(tablrel, 2)
```

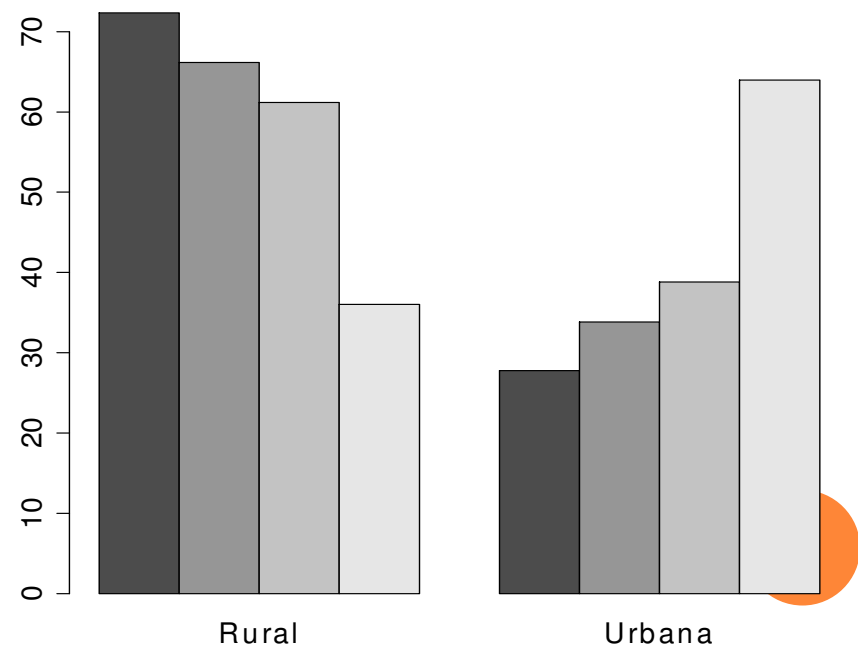
| province | urbanity | | Sum |
|--------------|-----------|-----------|-----------|
| | Rural | Urbana | |
| Ilocos Norte | 0.7230769 | 0.2769231 | 1.0000000 |
| Ilocos Sur | 0.6617647 | 0.3382353 | 1.0000000 |
| La Union | 0.6120690 | 0.3879310 | 1.0000000 |
| Pangasinan | 0.3603133 | 0.6396867 | 1.0000000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

```
> print(addmargins(tablrel, 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 |

```
> tablrelp = tablrel * 100
```

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

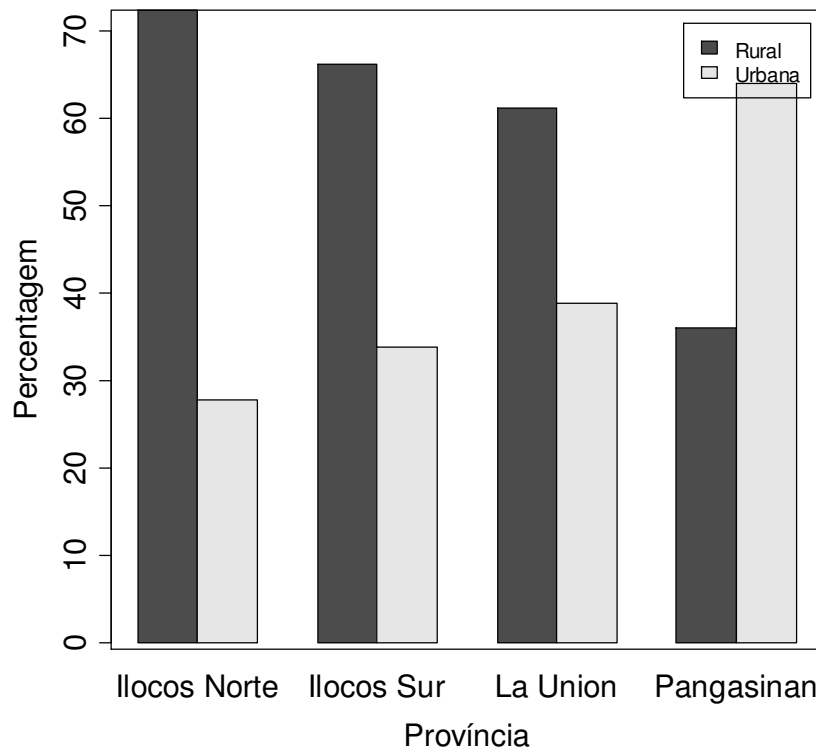


Era o gráfico que esperávamos?

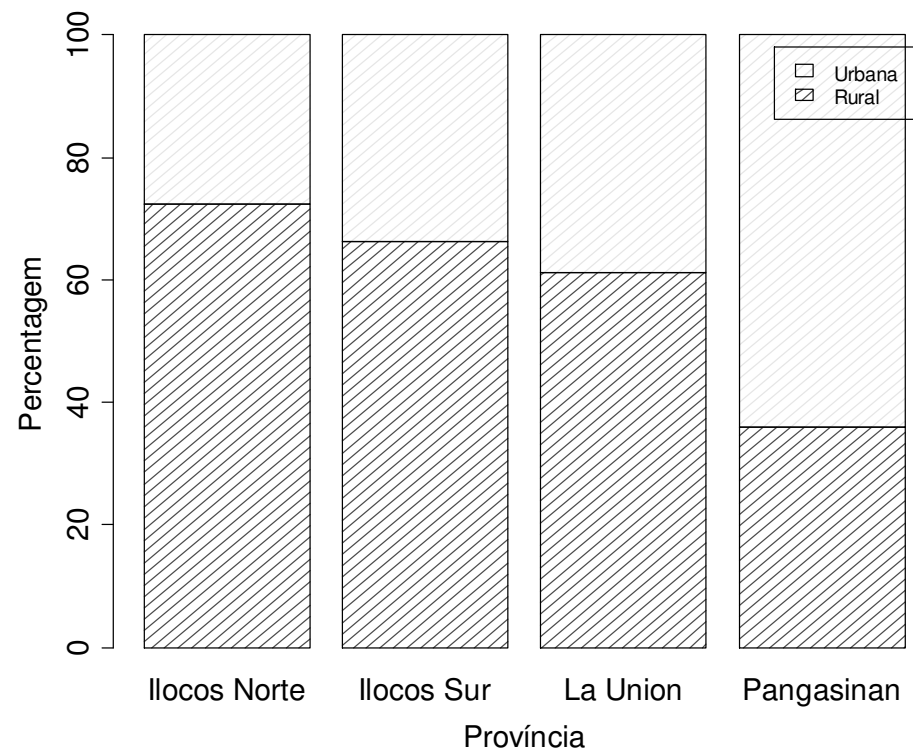
Funções em R

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

```
> box()
```



```
> barplot(t(tablrelp), 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)
```

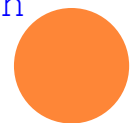
| province | sex | | Sum |
|--------------|----------|-----------|-------|
| | Feminino | Masculino | |
| Ilocos Norte | 12.3 | 87.7 | 100.0 |
| Ilocos Sur | 26.5 | 73.5 | 100.0 |
| La Union | 15.5 | 84.5 | 100.0 |
| Pangasinan | 18.3 | 81.7 | 100.0 |

| 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)
```

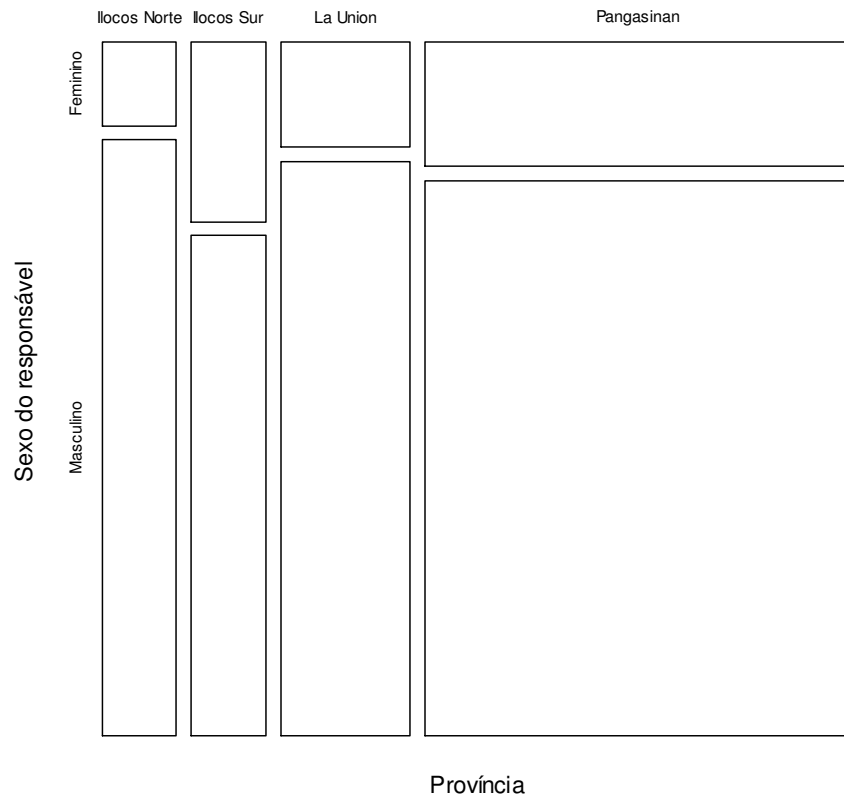
| tab2ind | Feminino Masculino | |
|------------|--------------------|-------|
| | Ilocos Norte | 11.7 |
| Ilocos Sur | 12.3 | 55.7 |
| La Union | 20.9 | 95.1 |
| Pangasinan | 69.1 | 313.9 |



Funções em R

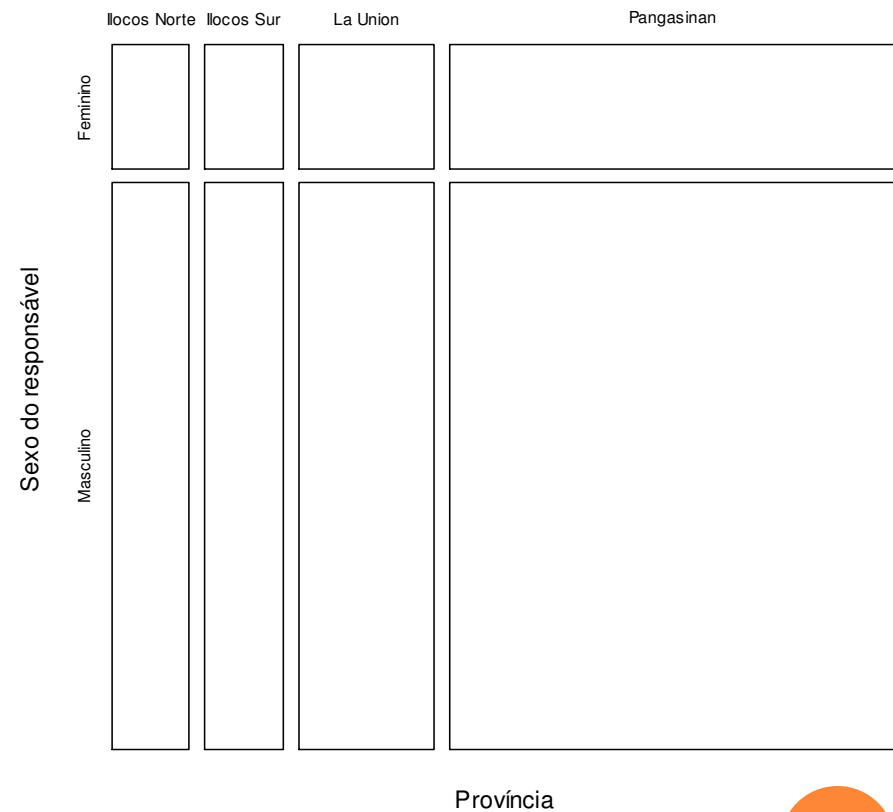
```
> mosaicplot(tab2, ylab = "Sexo do responsável", xlab = "Província", col = "white", main = "Dados observados")
```

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 Q^2 e a tabela supondo independência (`tab2ind`) podem ser obtidos usando a função `chisq.test`.

Um gráfico **não** muito recomendado:

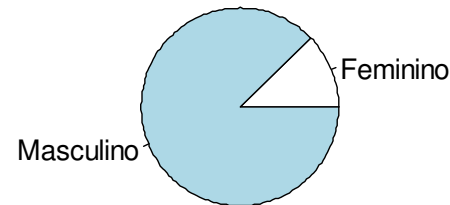
```
> nlinhas = ceiling(k / 2)
```

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

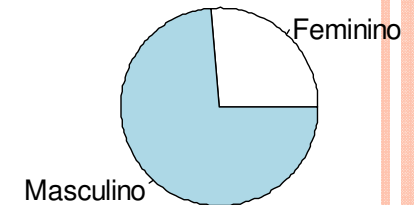
```
> for (i in 1:k) pie(tab2[i,],  
main = rownames(tab2rel)[i])
```

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

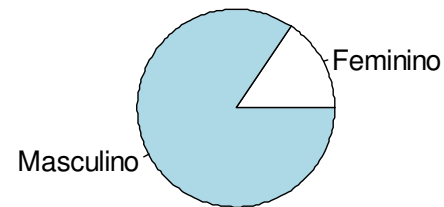
Ilocos Norte



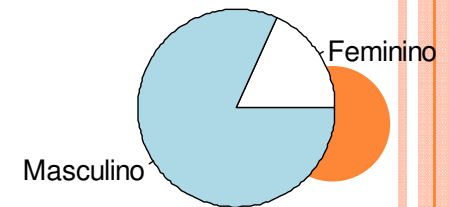
Ilocos Sur



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)
```

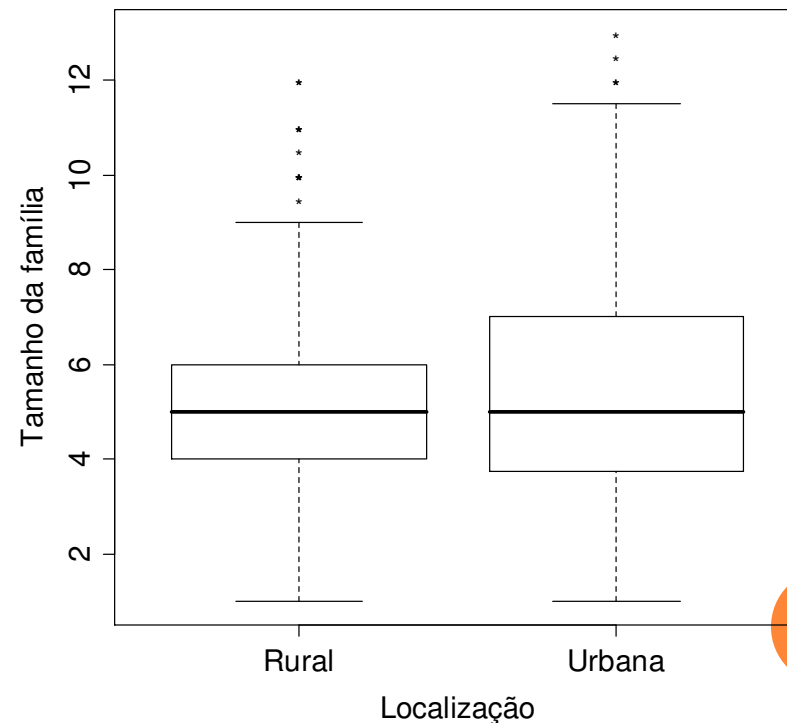
```
  y      x      y      x      x      "AP.income"  
"income" "sex" "family.size" "urbanity" "province" "AP.income"  
"AP.family.size" "AP.weight"
```

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

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

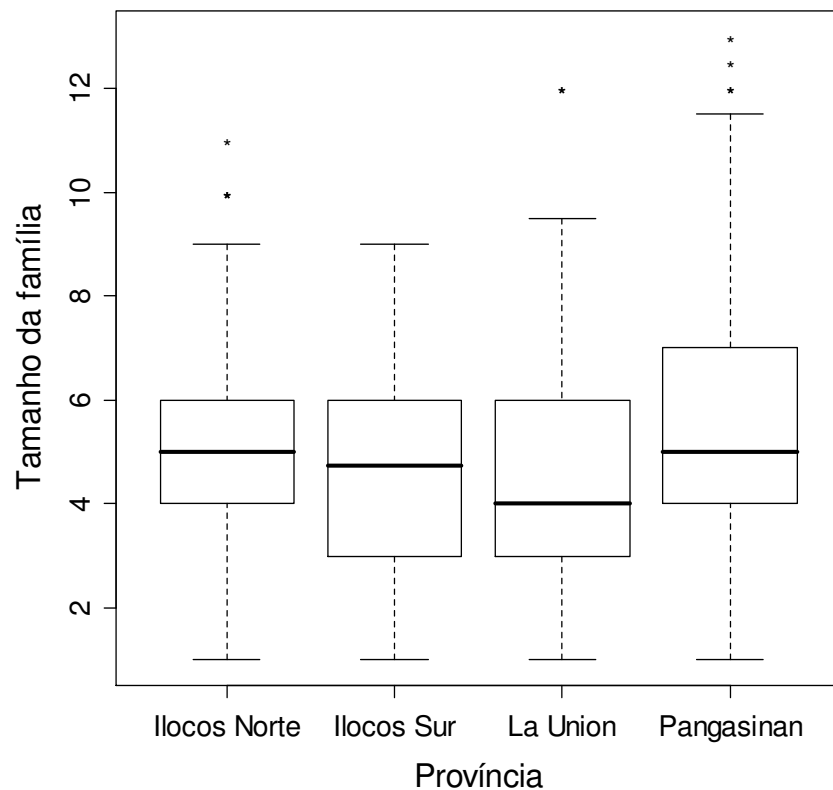
| income | family.size |
|-----------------------|----------------|
| Min. : 6067 | Min. : 1.000 |
| 1st Qu.: 48001 | 1st Qu.: 3.875 |
| <u>Median : 75926</u> | Median : 5.000 |
| <u>Mean : 112292</u> | Mean : 5.193 |
| 3rd Qu.: 137068 | 3rd Qu.: 6.500 |
| Max. : 835742 | Max. : 13.000 |

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

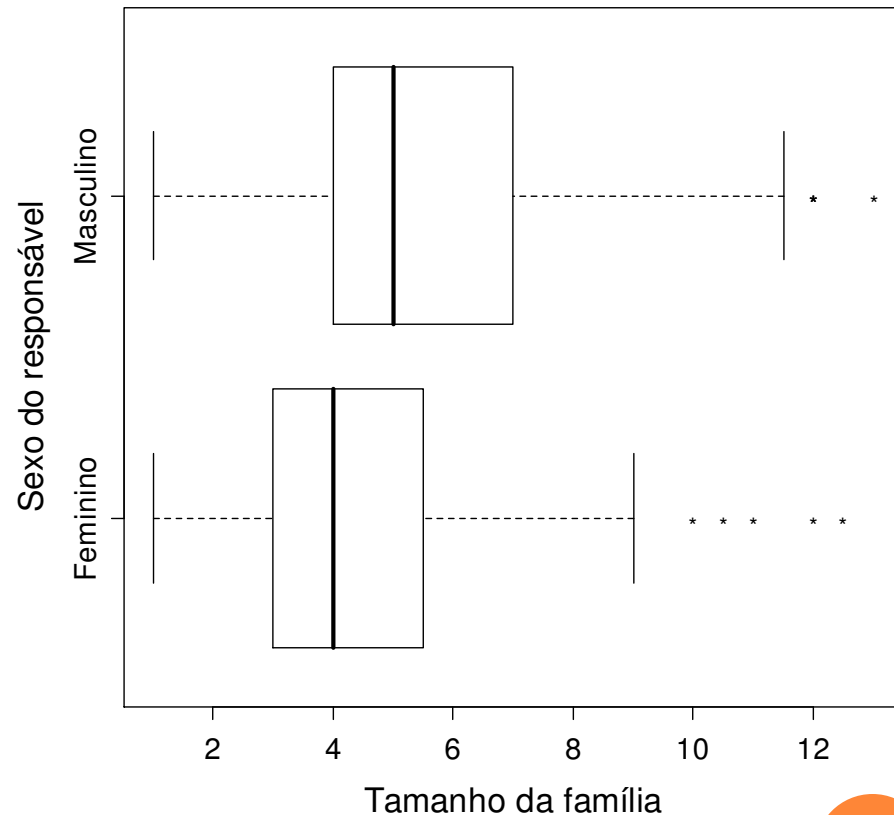


Funções em R

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



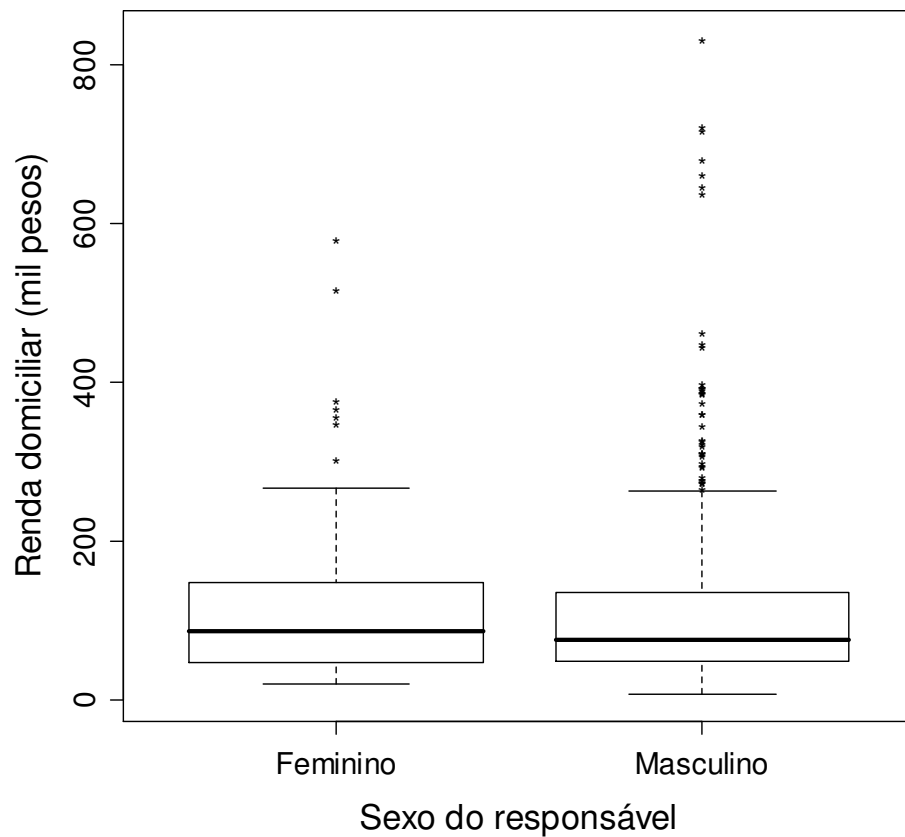
```
> plot(family.size ~ sex,  
xlab = "Sexo do responsável", ylab = "Tamanho da família",  
pch = "*", horizontal = TRUE)
```



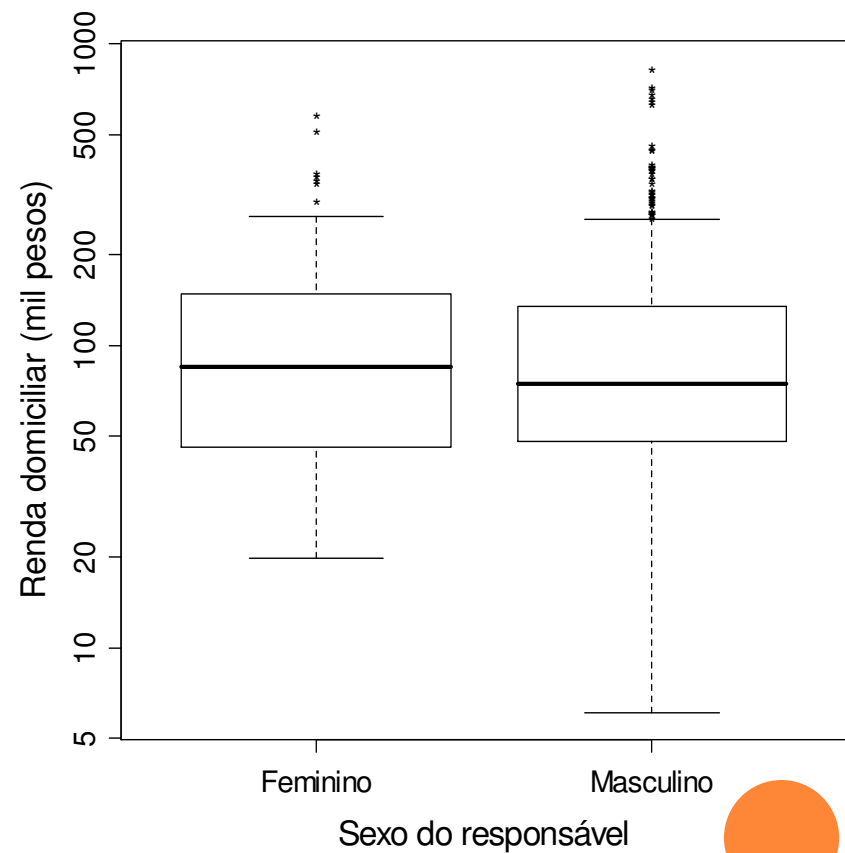
Exercício. Apresente o gráfico à esquerda com níveis em ordem decrescente da mediana.

Funções em R

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



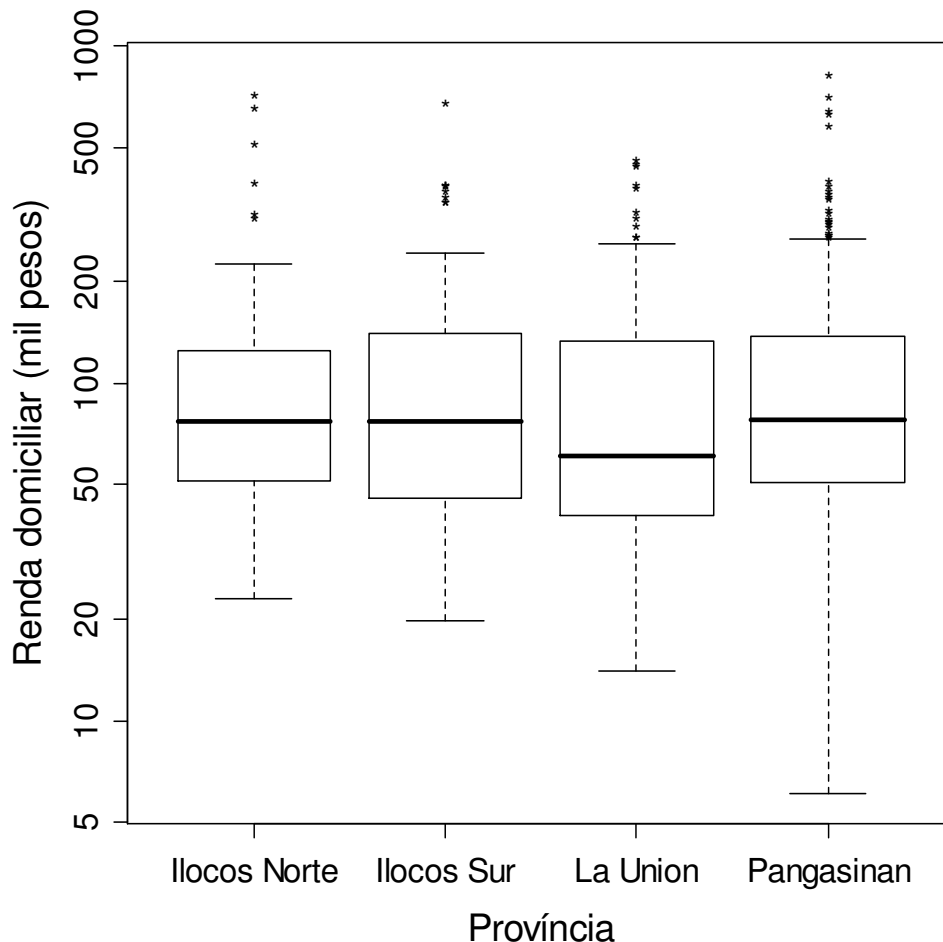
```
> plot(income / 1000 ~ sex,  
xlab = "Sexo do responsável",  
log = "y", ylab = "Renda  
domiciliar (mil pesos)", pch =  
"*)
```



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(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)
```

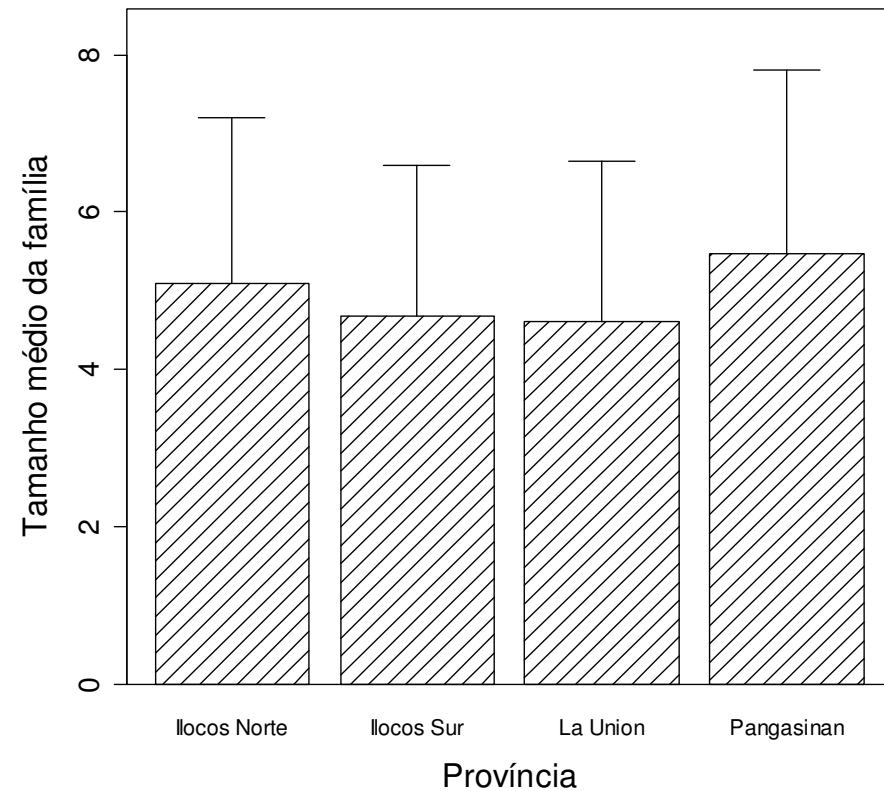
```
> box()
```

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.