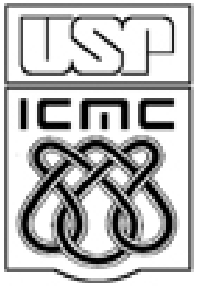


Medidas de associação

8.2. Associação entre Variáveis qualitativas

Exemplo: Queremos verificar se existe ou não associação entre sexo e a carreira escolhida por 200 alunos de Economia e Administração. Esses dados são dados a continuação

Distribuição conjunta de alunos segundo o sexo (x) e curso escolhido

Curso (Y)	Sexo (X)		Total
	Masculino	Feminino	
Economia	85	35	120
Administração	55	25	80
Total	140	60	200

Difícil de tirar alguma conclusão devido à diferença entre os totais marginais

Deve-se construir as proporções segundo as linhas ou colunas para podermos fazer comparações

Fixamos os totais das colunas; a distribuição conjunta em proporções são dadas na seguinte tabela.

Curso (Y)	Sexo (X)		Total
	Masculino	Feminino	
Economia	61%	58%	60%
Administração	39%	42%	40%
Total	100%	100%	100%

- Independente do sexo, 60% das pessoas preferem economia e 40% administração.
- Não havendo dependência entre as variáveis, espera-se essas mesmas proporção para cada sexo.
- As proporções do sexo masculino (61% e 39%) e do sexo feminino (60% e 40%) são próximas das marginais.
- Esses resultados parecem indicar que não há relação de dependência entre as duas variáveis.

Exemplo: Queremos verificar se existe ou não associação entre sexo e a carreira escolhida por 200 alunos de Física e Ciências sociais. Esses dados são dados a continuação

Distribuição conjunta de alunos segundo o sexo (x) e curso escolhido

Curso (Y)	Sexo (X)		Total
	Masculino	Feminino	
Física	100 (71%)	20 (33%)	60%
Ciências Sociais	40 (29%)	40 (67%)	40%
Total	140 (100%)	60 (100%)	100%

Comparando a distribuição das proporções pelos cursos, independente do sexo (coluna dos totais), com as distribuições diferenciadas por sexo (colunas de masculino e feminino), observamos uma disparidade bem acentuada nas proporções.

8.2. Associação entre Variáveis qualitativas

$x \in \{ x_1,...,x_k \}$ e $y \in \{ y_1,...,y_m \}$, $1 < k \leq n$ e $1 < m \leq n$.
f_{ij} : frequencia absoluta do par (x_i, y_j), $i = 1,...,k$ e $j = 1,...,m$.

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

x	y					
	y_1	...	y_j	...	y_m	Totais
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

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

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

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

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

8.2. Associação entre Variáveis qualitativas

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

Distribuição marginal de x (frequencias absolutas)

x	y					
	y_1	...	y_j	...	y_m	Totais
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 (frequencias absolutas)

x	y					
	y_1	...	y_j	...	y_m	Totais
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. Associação entre Variáveis qualitativas

Frequencias 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				
	y_1	...	y_j	...	y_m	Totais
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. Associação entre Variáveis qualitativas

(b)

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

Distribuição condicional de x dado $y = y_j$.

m distribuições condicionais de x.

8.2. Associação entre Variáveis qualitativas

Que frequencia 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 domicílo 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 frequencias 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,

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

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

x	y					
	y_1	...	y_j	...	y_m	Totais
x_1	f_{11}/n	...	f_{1j}/n	...	f_{1m}/n	$f_{1\bullet}/n$
...	...	...	Conjunta	...	...	...
x_i	f_{i1}/n	...	f_{ij}/n	...	f_{im}/n	$f_{i\bullet}/n$ Marginal de x
...	...	...	...	...	...	...
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

Marginal de y

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 frequencias absolutas observadas (f_{ij}) e as frequencias calculadas supondo independência entre x e y ($f_{ij}^{ind} = f_{i\bullet} f_{\bullet j} / n$):

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

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

$Q^2>0$ => 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 × k (m = k).

x	y					
	y_1	...	y_i	...	y_k	Totais
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.

Exemplo: Suponha a seguinte tabela de contingência

Ao examinar 400 estudantes de certa Instituição distribuídos pelos cursos de Estatística e Engenharia, obteve-se:

sexo	Curso 1 Estatística	Curso 2 Engenharia	total
Homens	40	200	240
Mulheres	60	100	160
total	100	300	400

Valores esperados sob independência

- Como são 100 alunos em Estatística e 300 alunos em Engenharia, (240 do sexo masculino e 160 do sexo feminino) esperaria-se, em caso de independência, ter a seguinte tabela de contingência:

sexo	Curso 1 Estatística	Curso 2 Engenharia	total
Homens	60	180	**240**
Mulheres	40	120	**160**
total	**100**	**300**	**400**

Tabela com as freqüências observadas:

sexo	Curso 1 Estatística	Curso 2 Engenharia	total
Homens	**40**	**200**	240
Mulheres	**60**	**100**	160
total	100	300	400

Tabela com as freqüências esperadas no caso de não associação:

sexo	Curso 1 Estatística	Curso 2 Engenharia	total
Homens	**60**	**180**	**240**
Mulheres	**40**	**120**	**160**
total	**100**	**300**	**400**

$$Q^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(f_{ij} - f_{ij}^{\text{ind}})^2}{f_{ij}^{\text{ind}}}$$

Tabela com as freqüências observadas:

sexo	Curso 1 Estatística	Curso 2 Engenharia	total
Homens	40	200	240
Mulheres	60	100	160
total	100	300	400

$$\frac{(200-180)^2}{180}$$

$$\frac{(100-120)^2}{120}$$

Tabela com as freqüências esperadas no caso de não associação:

$$\frac{(40-60)^2}{60}$$

$$\frac{(60-40)^2}{40}$$

sexo	Curso 1 Estatística	Curso 2 Engenharia	total
Homens	60	180	**240**
Mulheres	40	120	**160**
total	**100**	**300**	**400**

CÁLCULO DO COEFICIENTE DE CONTINGÊNCIA

- O qui-quadrado é, então,

$$Q^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(f_{ij} - f_{ij}^{\text{ind}})^2}{f_{ij}^{\text{ind}}}$$

$$Q^2 = \frac{(40-60)^2}{60} + \frac{(200-180)^2}{180} + \frac{(60-40)^2}{40} + \frac{(100-120)^2}{120} \cong 22{,}22$$

$$T = \sqrt{\frac{Q^2}{n\sqrt{(k-1)(m-1)}}} = \sqrt{\frac{22{,}22}{400(2-1)(2-1)}} = 0{,}23$$

Funções em R

```
> library(ineq)
> ?Ilocos
```

Ilocos {ineq} R Documentation

Income Metadata from Ilocos, Philippines

Description

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

Usage

```
data(Ilocos)
```

Dados coletados em domicílios nas Filipinas.

```
> data(Ilocos)
> dados = Ilocos
> dim(dados)  [1] 632   8
> names(dados)
"income" "sex" "family.size" "urbanity" "province" "AP.income"
"AP.family.size" "AP.weight"
```

n = 632 observações de 8 variáveis.

```
> 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")
> (tab1 = table(province, urbanity))
```

Um fator tem níveis (*levels*).

province	urbanity: 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,...,4 (k = 4) e j = 1,2 (m = 2).

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

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

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

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

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

province	urbanity: 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 → 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
```

```
> margin.table(tab1, margin = 2)
urbanity
Rural Urbana
  301    331
```

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

province	urbanity 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

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

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

province	urbanity 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

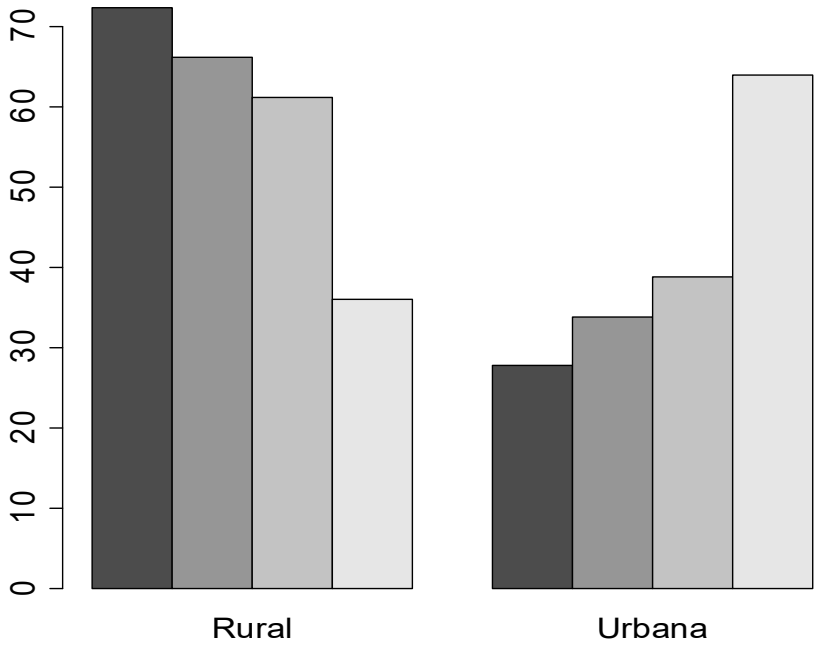
```
> addmargins(tab1rel, 2)
```

	urbanity		
province	Rural	Urbana	Sum
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)
```

	urbanity		
province	Rural	Urbana	Sum
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)
```

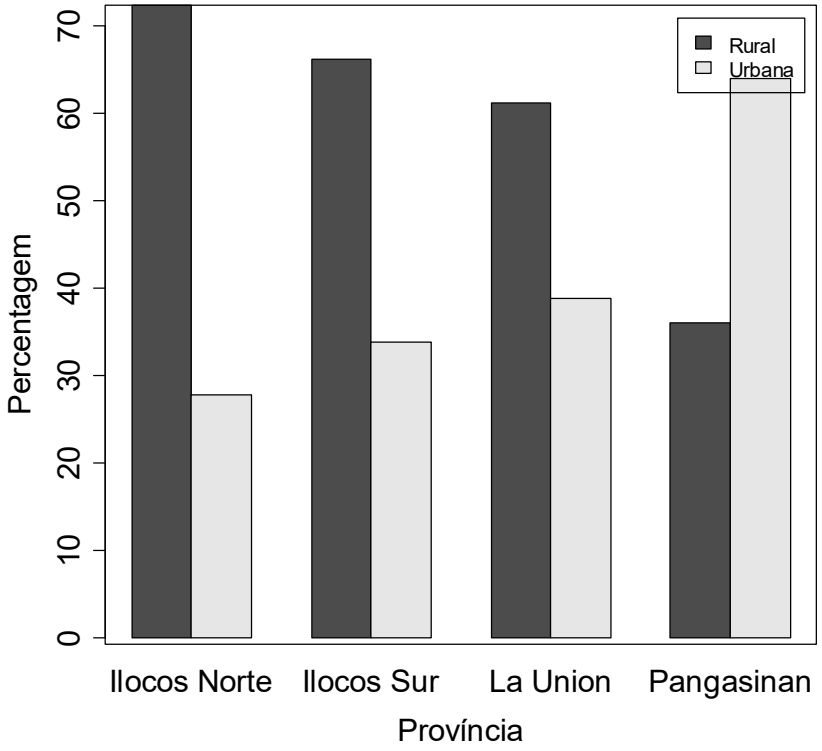


Era o gráfico que esperávamos?

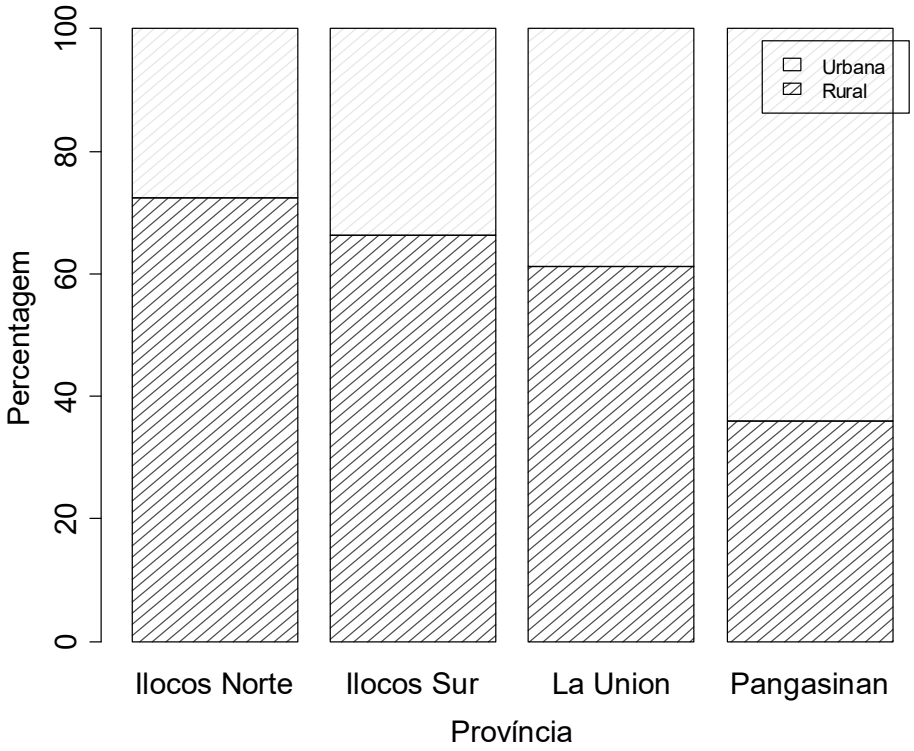
Funções em R

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

```
> box()
```




```
> barplot(t(tab1relp), xlab = "Província", ylab = "Percentagem", 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 frequencias.

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

province	sex: Feminino	Masculino	Sum
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

tab2

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`:

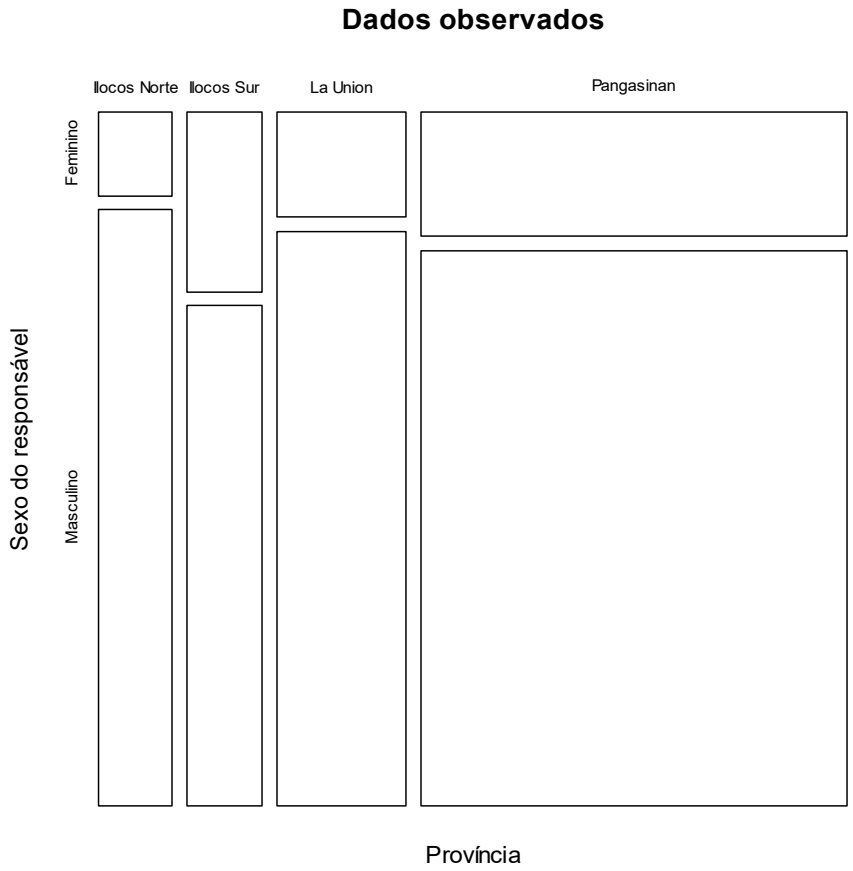
```
> tab2marg = addmargins(tab2, 1:2)
> k = nrow(tab2marg) - 1
> m = ncol(tab2marg) - 1
> n = sum(tab2)
> tab2ind = tab2marg[1:k, m + 1] %*% t(tab2marg[k + 1, 1:m]) / n
> rownames(tab2ind) = rownames(tab2)
> colnames(tab2ind) = colnames(tab2)
```

tab2ind

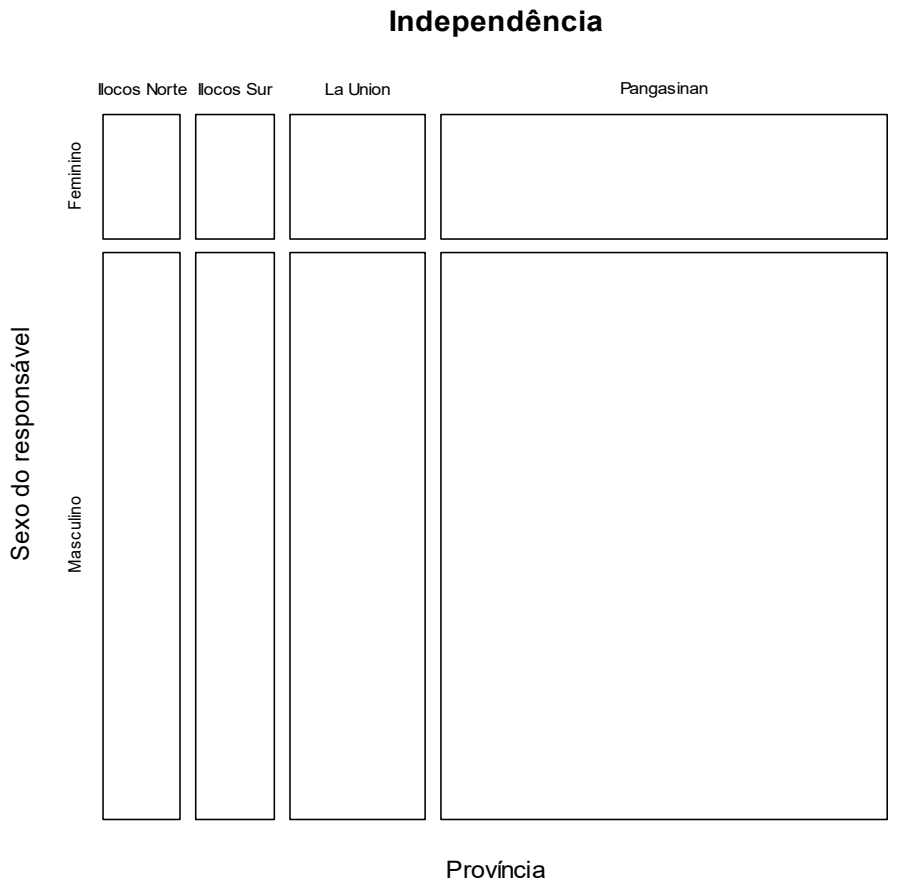
	Feminino	Masculino
Ilocos Norte	11.7	53.3
Ilocos Sur	12.3	55.7
La Union	20.9	95.1
Pangasinan	69.1	313.9

Funções em R

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




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



Retângulos com bases proporcionais às frequencias da variável `province` e alturas proporcionais às frequencias 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

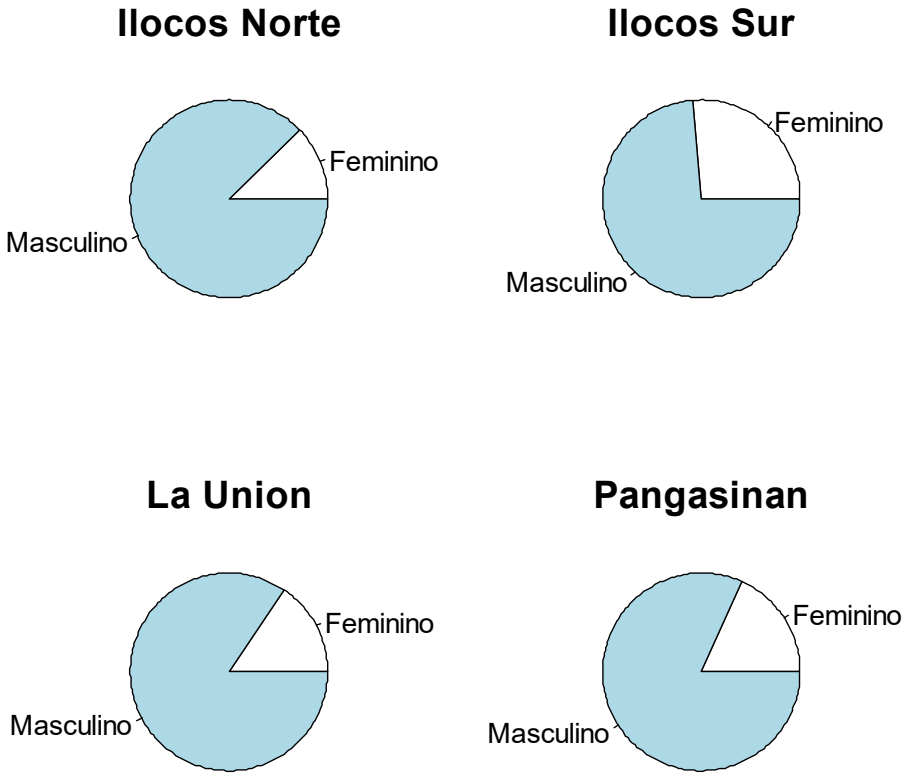
Obs. O valor de Q^2 e a tabela supondo independência (`tab2ind`) podem ser obtidos usando a função `chisq.test`.

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).

8.3. Variáveis qualitativas e quantitativas

8.3. Variáveis qualitativas e quantitativas

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

Dados observados: n pares de valores (x_j, y_j), sendo que $x_j \in \{ x_1,...,x_k \}$, j = 1,...,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 (frequencia). Para cada nível x_i calculamos a variância s_i^2 dos valores y_j para os quais $x_j = x_i$, j = 1,...,n e i = 1,...,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

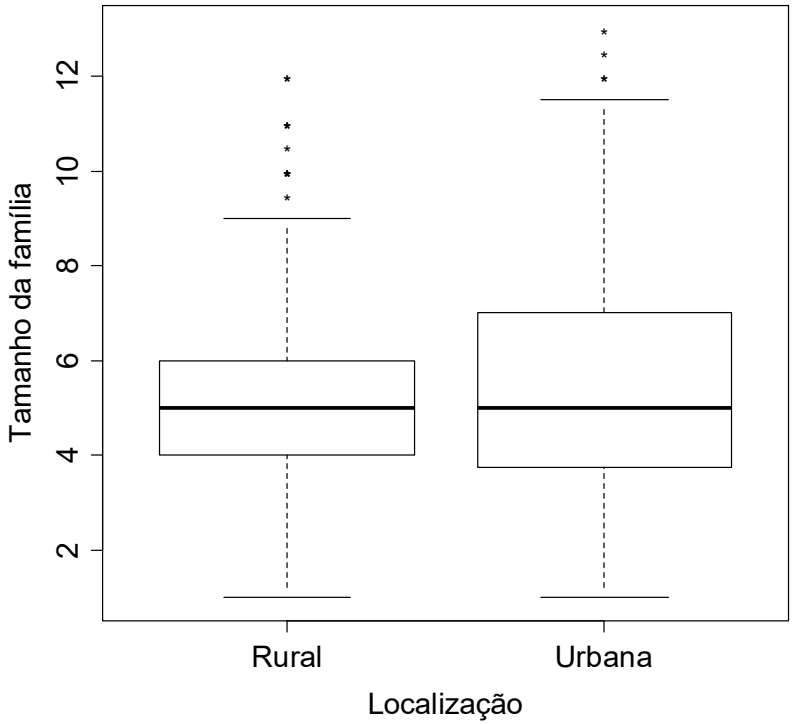
```
"income" "sex" "family.size" "urbanity" "province" "AP.income"
"AP.family.size" "AP.weight"
```

Fórmula: `y ~ x` (y como função de x ou y depende de x).

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

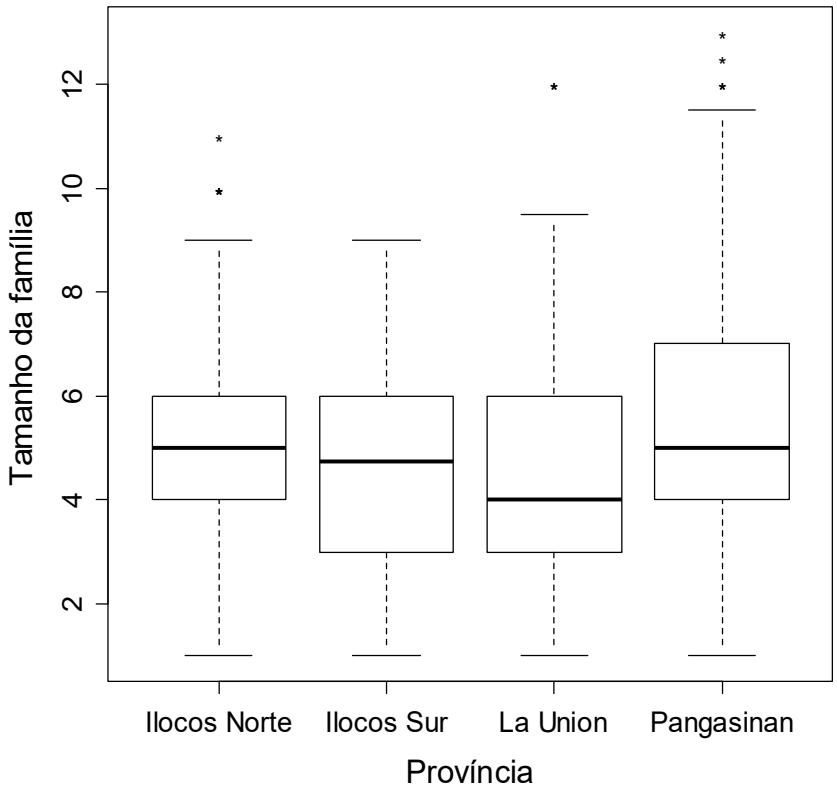
```
     income        family.size
 Min.   :  6067   Min.   : 1.000
 1st Qu.: 48001   1st Qu.: 3.875
 Median : 75926   Median : 5.000
 Mean   :112292   Mean   : 5.193
 3rd Qu.:137068   3rd Qu.: 6.500
 Max.   :835742   Max.   :13.000
```

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

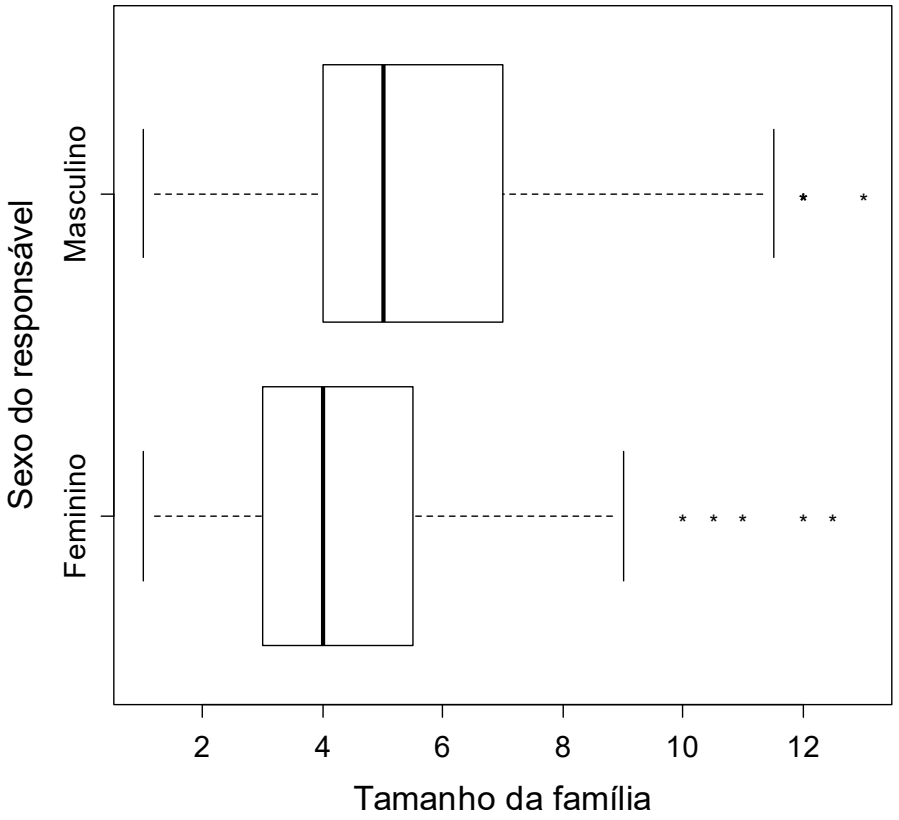


Funções em R

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



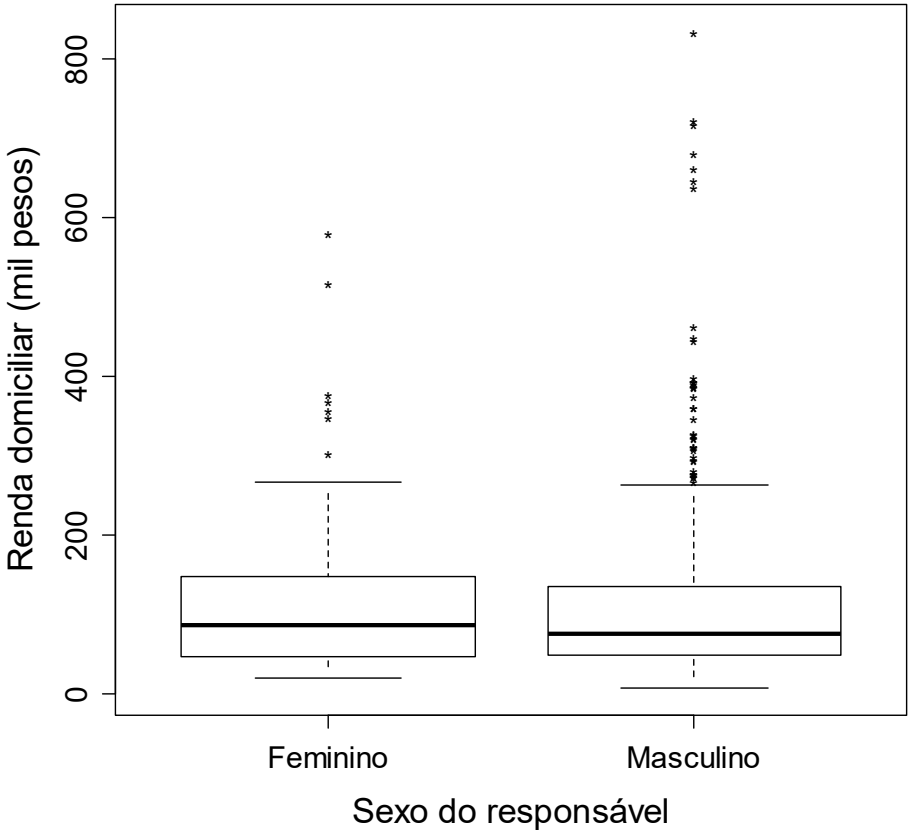

```
> plot(family.size ~ sex, xlab = "Sexo do responsável", ylab = "Tamanho da família", pch = "*", horizontal = TRUE)
```



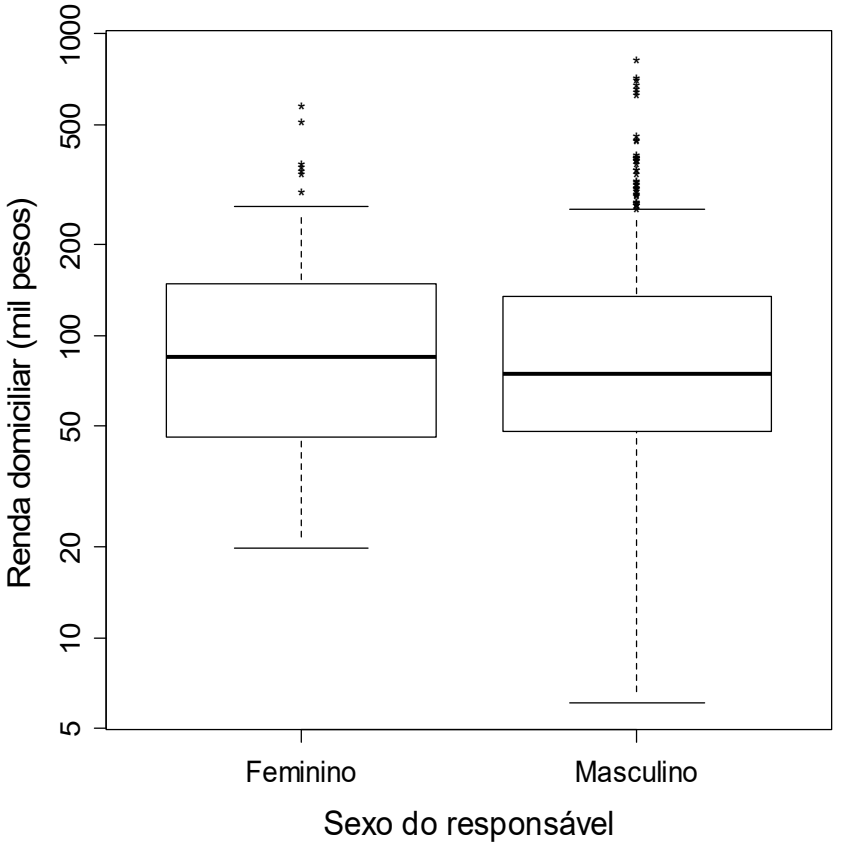
Exercício. Apresente o gráfico à esquerda com níveis em ordem decrescente da mediana.

Funções em R

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



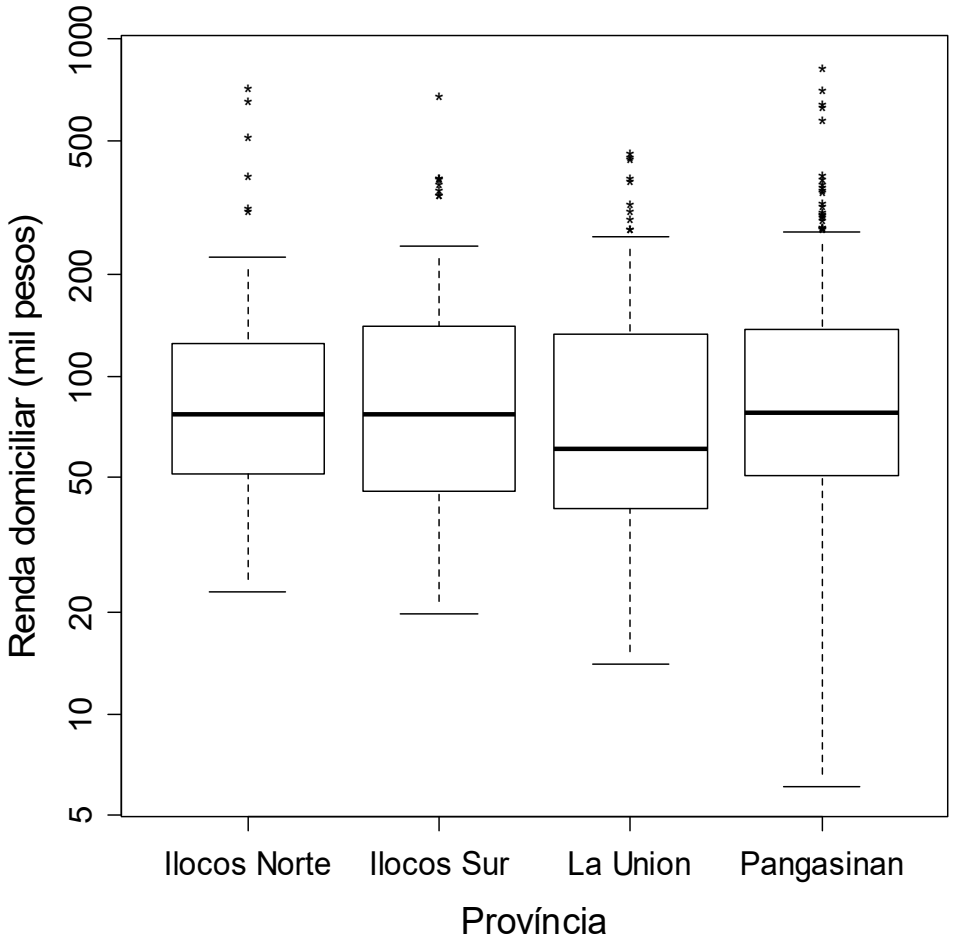

```
> plot(income / 1000 ~ sex,
xlab = "Sexo do responsável",
log = "y", ylab = "Renda
domiciliar (mil pesos)", pch =
"*")
```



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

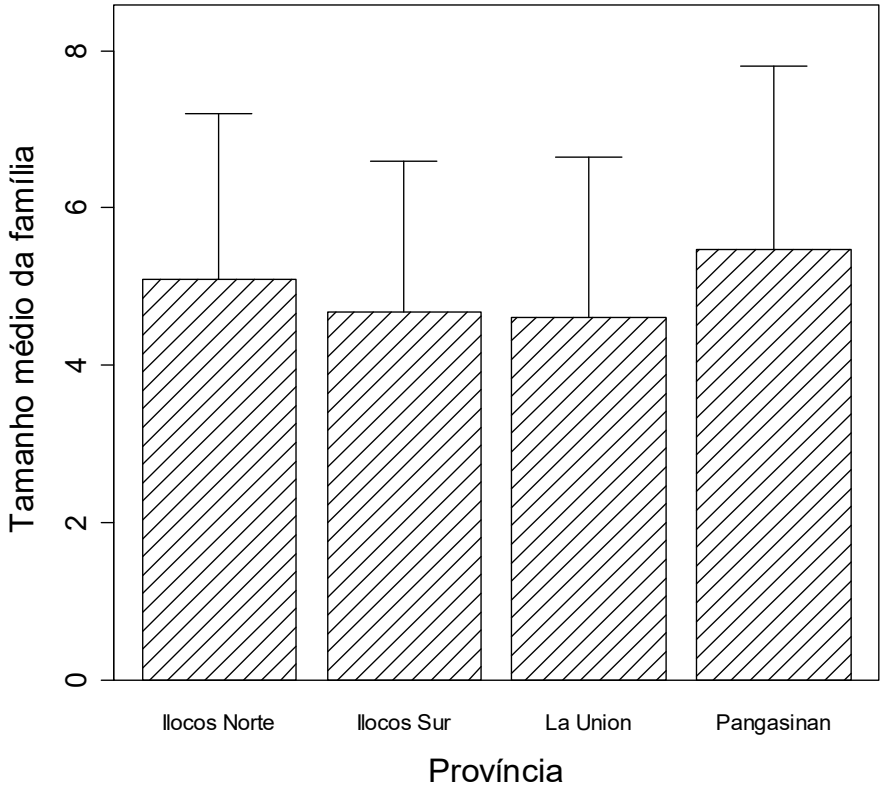
Gráfico de médias e desvios padrão do tamanho da família por província:

```
> 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.