Vizualização de dados multidimensionais e Análise Descritiva

Vicente G. Cancho
garibay@icmc.usp.br

Departamento de Matemática Aplicada e Estatística
Universidade de São Paulo

Exemplo 1-Eleições americanas 1996

- Dados de 944 eleitores na eleição nacional americana de 1996 (Rosenstone et al., 1997).
  - X_1 : voto (Democratas, Republicanos, Independentes(D+R))
  - X_2 : Idade (anos)
  - X_3 : Educação
  - X_4 : renda anual em milhares de dólares
- Interesse: dado as característica dos eleitores predizer o voto do eleitor.
- O resumo dos dados

```
> summary(rnes96)
       Partido        Renda          Educação       Idade
Democrata   :380   Min.   :  1.50   MS     : 13   Min.   :19.00
Independente:239   1st Qu.: 23.50   HSdrop: 52   1st Qu.:34.00
Republicano :325   Median : 37.50   HS     :248   Median :44.00
                     Mean   : 46.58   Coll   :187   Mean   :47.04
                   3rd Qu.: 67.50   CCdeg : 90   3rd Qu.:58.00
                   Max.   :115.00   BAdeg :227   Max.   :91.00
                                    MAdeg :127
```

Exemplo 2- Ganho de Empresas

Apresenta-se dados relativos de 12 empresas no que se refere a 3 variáveis (medidas em unidas monetárias): ganho bruto (X_1), ganho liquido (X_2) e patrimônio acumulado (X_3)

Empresa	G_Bruto	G_líquido	Patrimonio
E1	9893	564	17689
E2	8776	389	17359
E3	13572	1103	18597
E4	6455	743	8745
E5	5129	203	14397
E6	5432	215	3467
E7	3807	385	4679
E8	3423	187	6754
E9	3708	127	2275
E10	3294	297	6754
E11	5433	432	5589
E12	6287	451	8972

Suponha que são observadas $p \geq 1$ variáveis em n indivíduos, itens ou unidades experimentais (observações), cujos componentes podem ser: p variáveis quantitativas, p variáveis qualitativas ou de ambos os tipos.

Notação

Seja x_{jk}: medição da k-ésima variável na j-ésima unidade experimental, com $j = 1, \ldots, n$ e $k = 1, \ldots, p$.

Podemos representar os dados construindo a matriz de dados

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}_{n \times p}$$

A **média amostral** da k-ésima variável, $k = 1, \ldots, p$ é dada por

$$\bar{x}_k = \frac{1}{n} \sum_{j=1}^{n} x_{jk}$$

A **variância amostral** da k-ésima variável, $k = 1, \ldots, p$ é dada por

$$s_k^2 = \frac{1}{n} \sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2$$

A **covariância amostral** entre a i-ésima e k-ésima variáveis, $i, k = 1, \ldots, p; i \neq k$, é dada por

$$s_{ik} = \frac{1}{n} \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

A **correlação amostral** entre a i-ésima e k-ésima variáveis, $i, k = 1, \ldots, p; i \neq k$, é dada por

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_i^2}\sqrt{s_k^2}} = \frac{\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)} \sqrt{\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)}}$$

O **vetor de médias amostrais** é dado por

$$\bar{\boldsymbol{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}$$

A **matriz de variâncias e covariâncias amostrais** ou simplesmente matriz de covariâncias amostrais é dada por

$$S = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{12} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \cdots & s_p^2 \end{pmatrix}$$

A **matriz de correlações amostrais** é dada por

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{12} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \dots & 1 \end{pmatrix}$$

Obs: $-1 \leq r_{ik} \leq 1, \ \forall i, k = 1, \dots, p, i \neq k.$

Exemplo 3: vendas de livros

São coletadas informações a respeito de 4 registros de vendas de livros:
Variável 1 (valor da nota): 42, 52, 48, 58
Variável 2 (número de livros): 2, 3, 2, 3
Neste exemplo, temos $p = 2$ variáveis e $n = 4$ observações.
Como fica a matriz de dados neste caso?

$$X = \begin{pmatrix} 42 & 2 \\ 52 & 3 \\ 48 & 2 \\ 58 & 3 \end{pmatrix}_{4\times 2}$$

Exemplo 1: vendas de livros

Temos interesse em resumir a informação dos dados em medidas-resumo.
Para isso, utilizamos o vetor de médias amostrais e a matriz de variâncias e covariâncias amostrais de $\boldsymbol{X}$.

O vetor de médias e a matriz de covariâncias são respectivamente

$$\bar{\boldsymbol{x}} = \begin{pmatrix} 50.0 \\ 2.5 \end{pmatrix} \text{ e } S = \begin{pmatrix} 45.3 & 3.3 \\ 3.3 & 0.33 \end{pmatrix}$$

Exemplo 1: vendas de livros

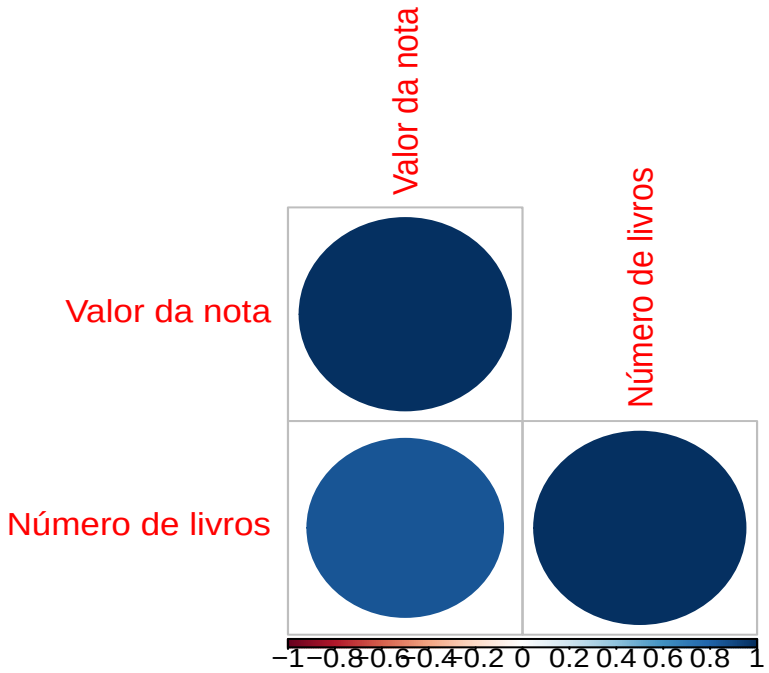
Além disso, é de se esperar que, quanto maior o número de livros no pedido, maior será o valor da compra. Essa ideia pode ser representada por meio da matriz de correlações amostrais de X.

$$R = \begin{pmatrix} 1.000 & 0.858 \\ 0.858 & 1.000 \end{pmatrix}$$

Exemplo 1: vendas de livros

Além disso, é de se esperar que, quanto maior o número de livros no pedido, maior será o valor da compra. Essa ideia pode ser representada por meio da matriz de correlações amostrais de X.

$$R = \begin{pmatrix} 1.000 & 0.858 \\ 0.858 & 1.000 \end{pmatrix}$$



Códigos em R

```
X<-matrix(c(42,52,48,58,2,3,2,3), nrow=4,
ncol=2, byrow=FALSE)
Xbarra<-apply(X, 2, mean)
Xbarra
S<-cov(X)
S
R<-cor(X)
Rinstall.packages("corrplot")
library(corrplot)
corrplot(R, type="lower")
```

Exemplo 2- Ganho de Empresas

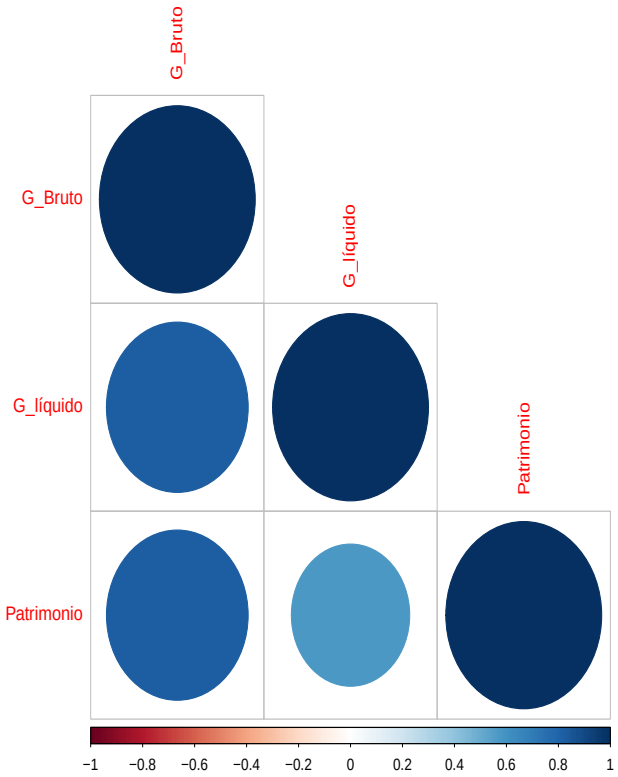
Para os dados do Exemplo 2, a matriz de covariâncias amostral das variáveis aleatórias X_1, X_2 e X_3 é dada por

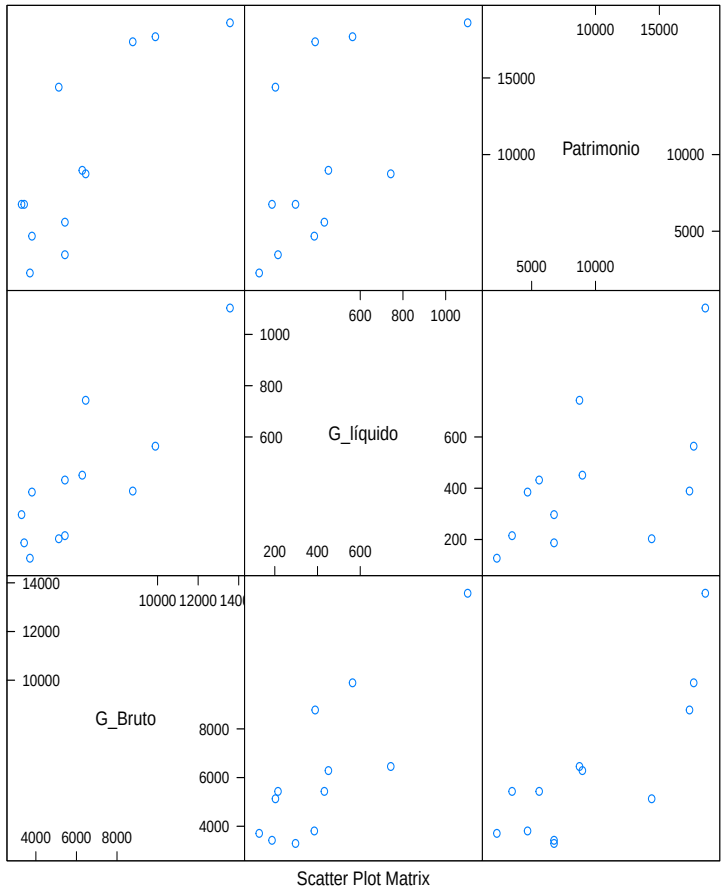
$$\bar{\boldsymbol{X}} = \begin{pmatrix} 6267.4167 \\ 424.6667 \\ 9606.4167 \end{pmatrix}, \quad S = \begin{pmatrix} 9550609 & 706121 & 14978233 \\ 706121 & 76270 & 933915 \\ 14978233 & 9339150 & 34408113 \end{pmatrix}.$$

A matriz de correlação amostral é dada por

$$R = \begin{pmatrix} 1.000 & 0.827 & 0.826 \\ 0.827 & 1.000 & 0.577 \\ 0.826 & 0.577 & 1.000 \end{pmatrix}$$

Exemplo 2- Ganho de Empresas



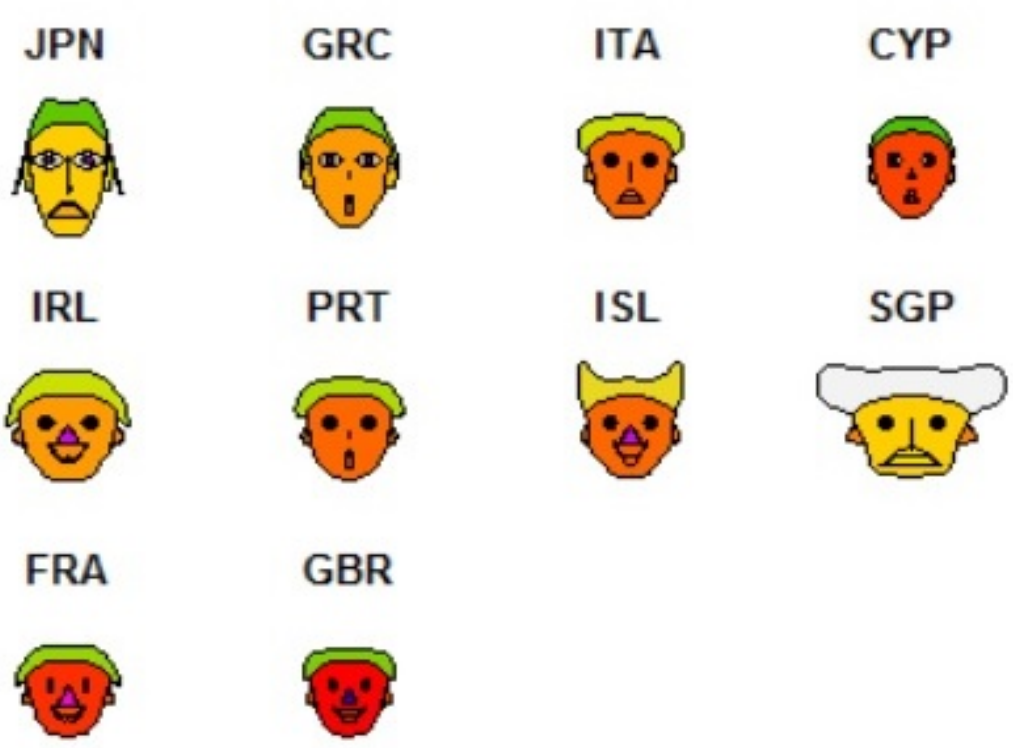


Faces de Chernoff

- Um dos métodos para visualizar dados multivariados é o uso de faces de Chernoff (veja, Chernoff, 1973).
- Cada variável no conjunto de dados é usada para representar uma característica da face.
- Chernoff usou 18 variáveis para representar diferentes características faciais, como cabeça, nariz, olhos, sobrancelhas, boca e orelhas

Faces de Chernoff

Variáveis macroeconômicas específicas para um conjunto de países.



Faces de Chernoff

Para Exemplo 2- Ganho de Empresas um comparativo das empresas.

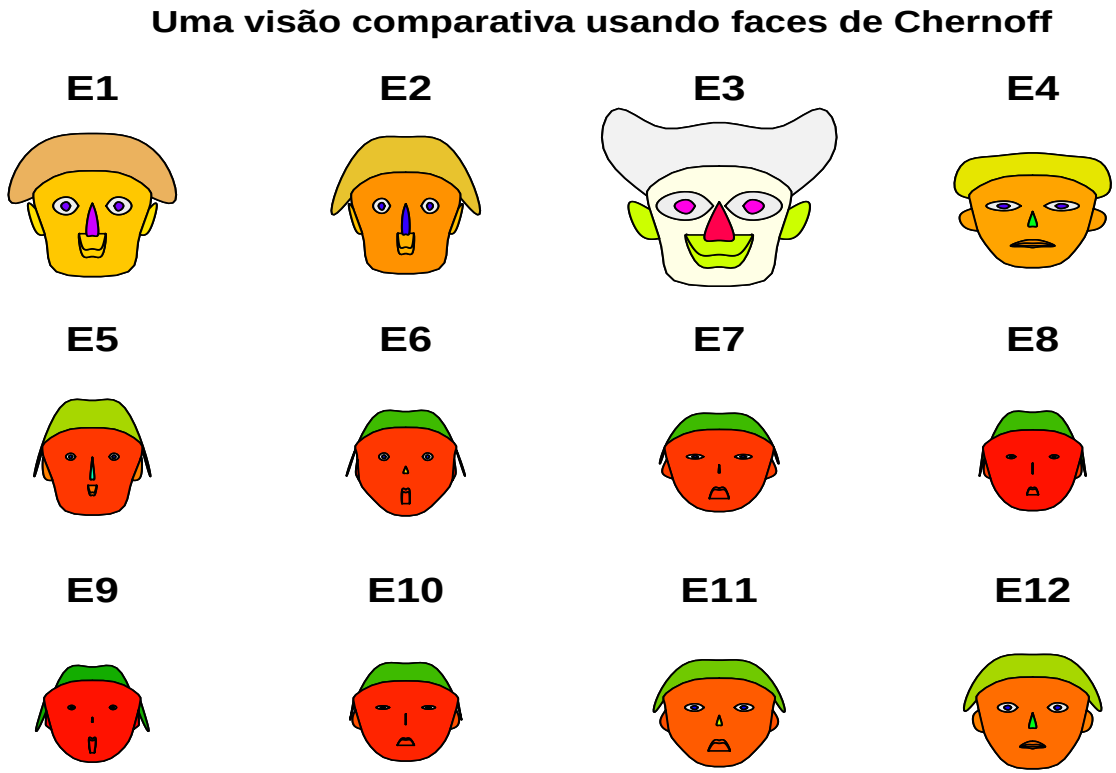
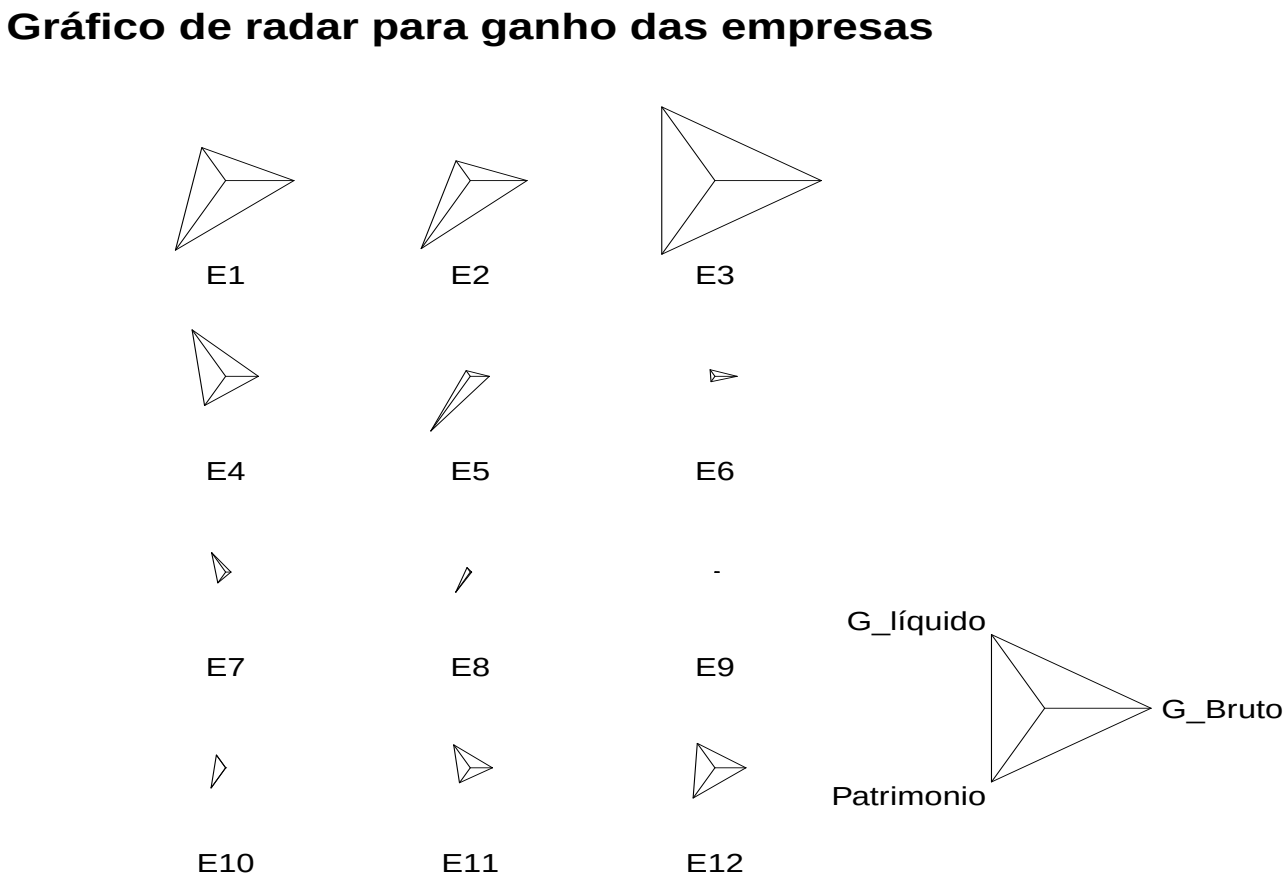


Gráfico de radar

- É um método gráfico de apresentar dados multivaridos na forma de um gráfico bidimensional de três ou mais variáveis quantitativas representadas em eixos que partem do mesmo ponto.
- A posição relativa e o ângulo dos eixos normalmente é pouco informativo.
- O gráfico de radar é também conhecido como gráfico de teia, gráfico de aranha, gráfico de estrela, polígono irregular ou gráfico polar.
- Ele é equivalente a um gráfico de coordenadas paralelas em coordenadas polares.

Gráfico de radar

Para Exemplo 2- Ganho de Empresas um comparativo das empresas.



Code R

```
dados=read.table("dados1.txt",header=T)
library(aplpack)
faces(dados[,-1], labels = dados$Empresa,
main = "Uma visão comparativa usando faces de Chernoff")
stars(dados[,-1],labels = dados$Empresa,key.loc=c(10,3),
main='Gráfico de radar para ganho das empresas')
```

Variáveis Qualitativas

- Tabelas de contingência multidimensionais;
- Gráficos de mosaico.

Considere dados coletados em domicílios nas Filipinas.

```
> library(ineq)
> data(Ilocos)
> dados = Ilocos
> dim(dados)
[1] 632   8
> 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
```

Variáveis Qualitativas

```
(tab = ftable(urbanity, province, sex))
```

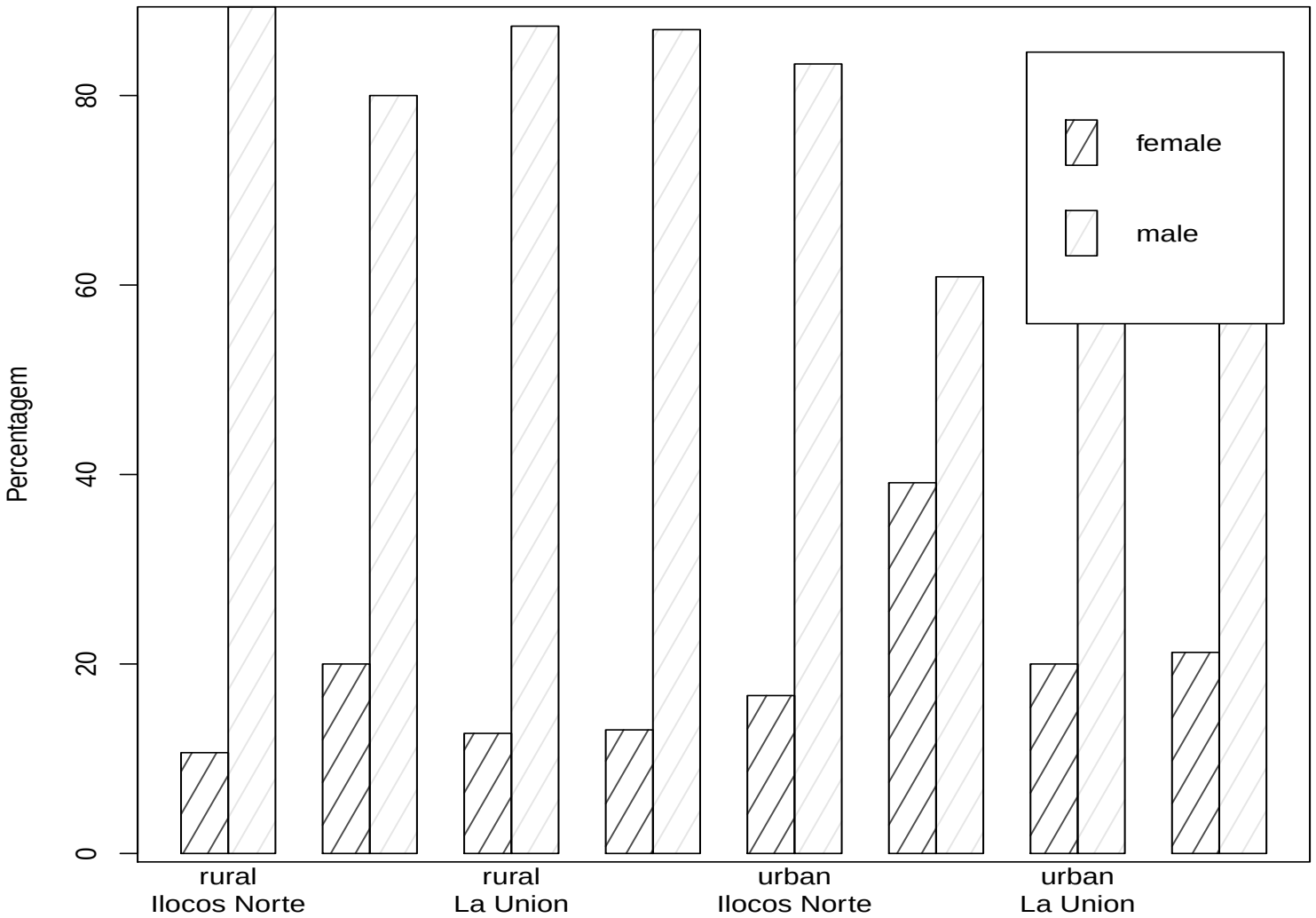
		sex	female	male
urbanity	province			
rural	Ilocos Norte		5	42
	Ilocos Sur		9	36
	La Union		9	62
	Pangasinan		18	120
urban	Ilocos Norte		3	15
	Ilocos Sur		9	14
	La Union		9	36
	Pangasinan		52	193

Variáveis Qualitativas

```
tabrel = prop.table(tab, margin = 1)
(tabrelp = tabrel * 100)
```

```
                    sex    female     male
urbanity province
rural    Ilocos Norte      10.63830 89.36170
         Ilocos Sur        20.00000 80.00000
         La Union          12.67606 87.32394
         Pangasinan        13.04348 86.95652
urban    Ilocos Norte      16.66667 83.33333
         Ilocos Sur        39.13043 60.86957
         La Union          20.00000 80.00000
         Pangasinan        21.22449 78.77551
```

Variáveis Qualitativas



Variáveis Qualitativas

Função xtabs: tabelas multidimensionais utilizando uma formula.

```
(tab3var = xtabs(~ urbanity + province + sex))
, , sex = female

province
urbanity Ilocos Norte Ilocos Sur La Union Pangasinan
rural              5           9         9         18
urban              3           9         9         52

, , sex = male

province
urbanity Ilocos Norte Ilocos Sur La Union Pangasinan
rural             42          36        62        120
urban             15          14        36        193
```

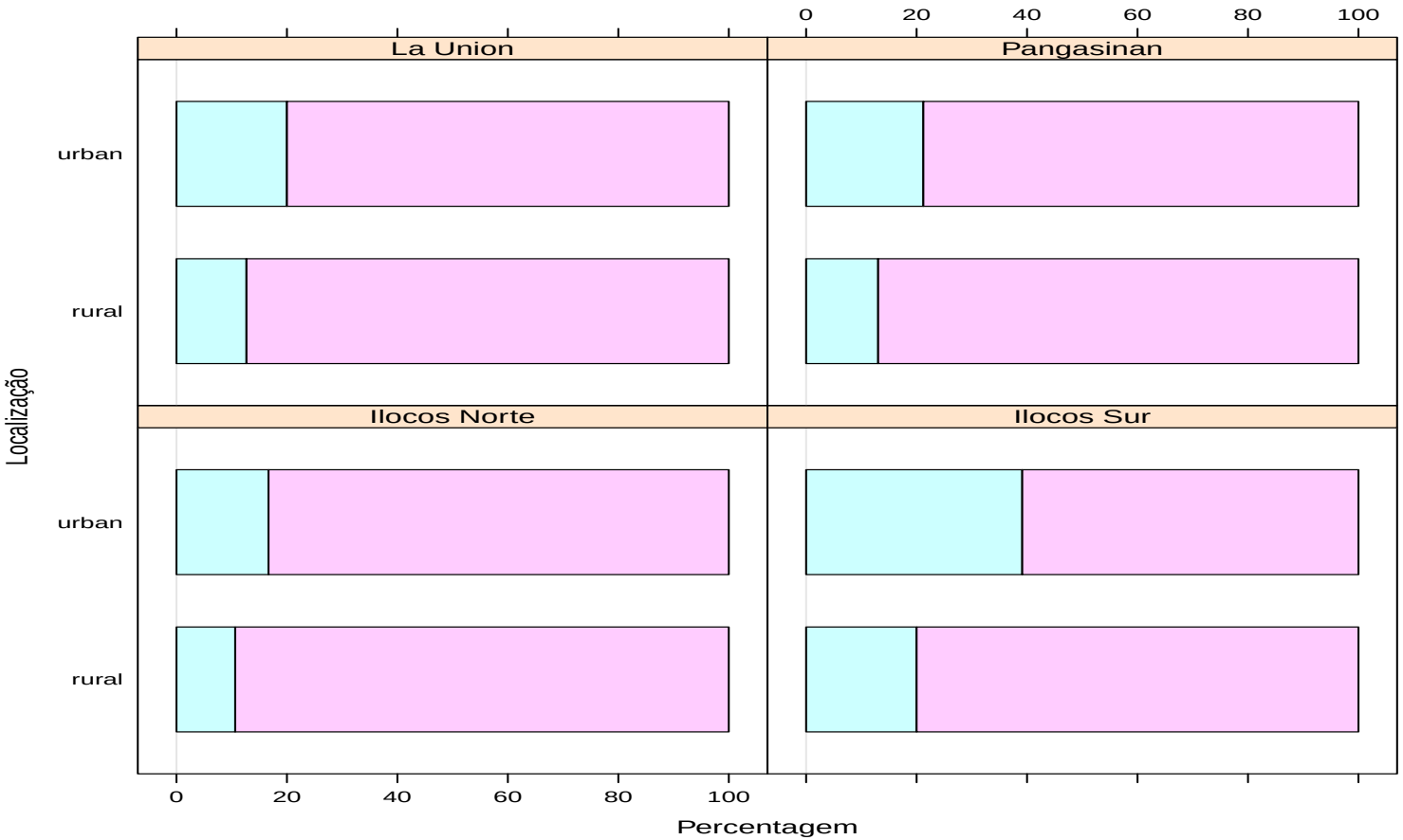
Tabela na forma de uma folha de dados (data frame)

```
 as.data.frame(tab3var)
    urbanity     province    sex Freq
1     rural Ilocos Norte female    5
2     urban Ilocos Norte female    3
3     rural   Ilocos Sur female    9
4     urban   Ilocos Sur female    9
5     rural     La Union female    9
6     urban     La Union female    9
7     rural   Pangasinan female   18
8     urban   Pangasinan female   52
9     rural Ilocos Norte   male   42
10    urban Ilocos Norte   male   15
11    rural   Ilocos Sur   male   36
12    urban   Ilocos Sur   male   14
13    rural     La Union   male   62
14    urban     La Union   male   36
15    rural   Pangasinan   male  120
16    urban   Pangasinan   male  193
```

Variáveis Qualitativas

Gráfico de barras de sex com frequencias relativas ao par (urbanity, province). Função barchart (lattice)

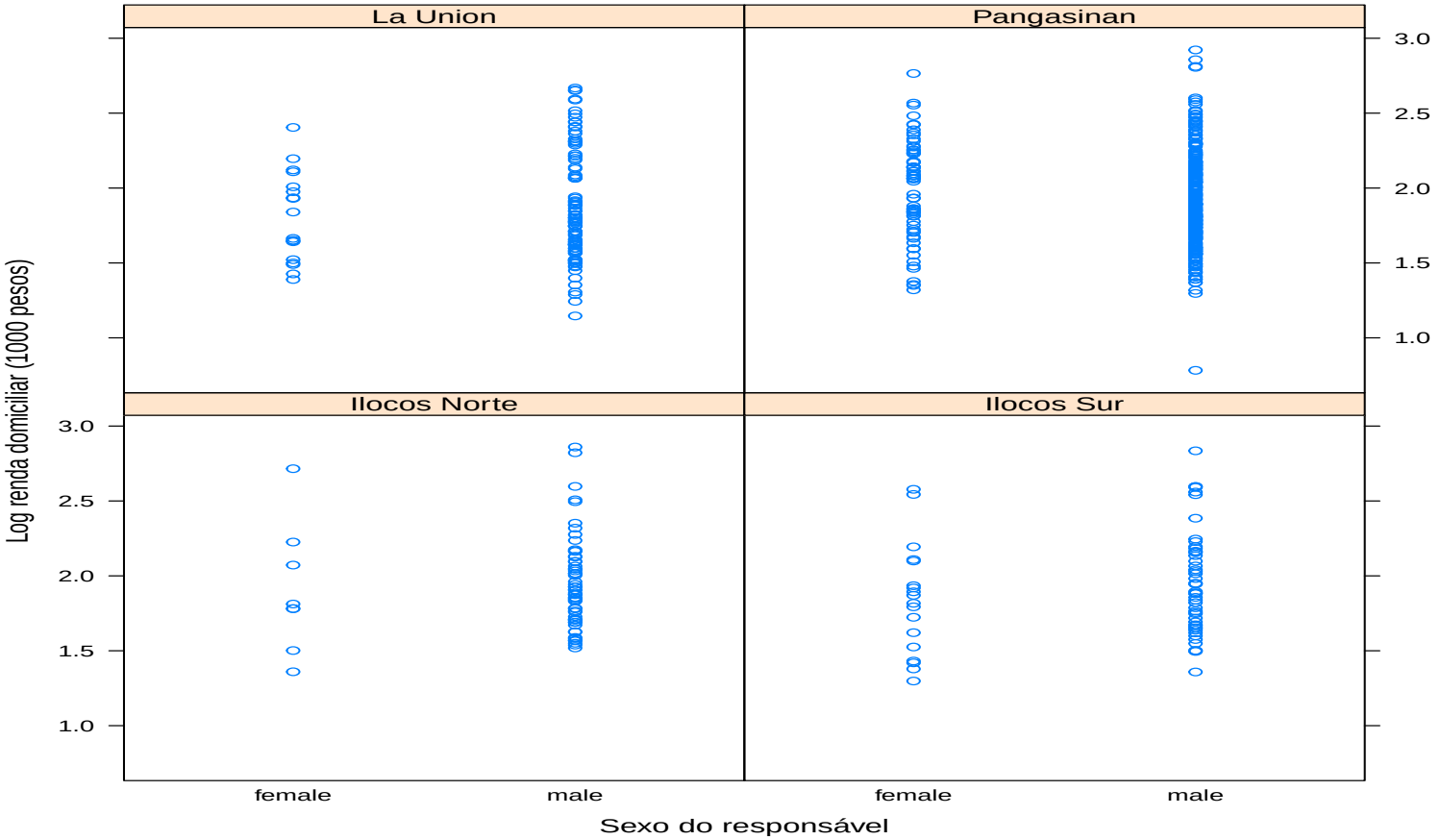
```
barchart(prop.table( tab3var, margin = c(1, 2)) * 100, xlab ="Pe
 ylab = "Localização")
```



Variáveis quantitativas e qualitativas

Gráfico de pontos: Função stripplot (lattice)

```
stripplot(log(income /1000) ~ sex | province,xlab ="Sexo do responsável",
ylab = " renda domiciliar (1000 pesos)")
```



Variáveis quantitativas e qualitativas

Duas variáveis condicionantes e acréscimo de ruído

```
stripplot(log(income/1000,10)~sex|province + urbanity,
xlab = "Sexo do responsável",ylab = "Log renda domiciliar (1000 pesos)",
jitter.data = TRUE)
```

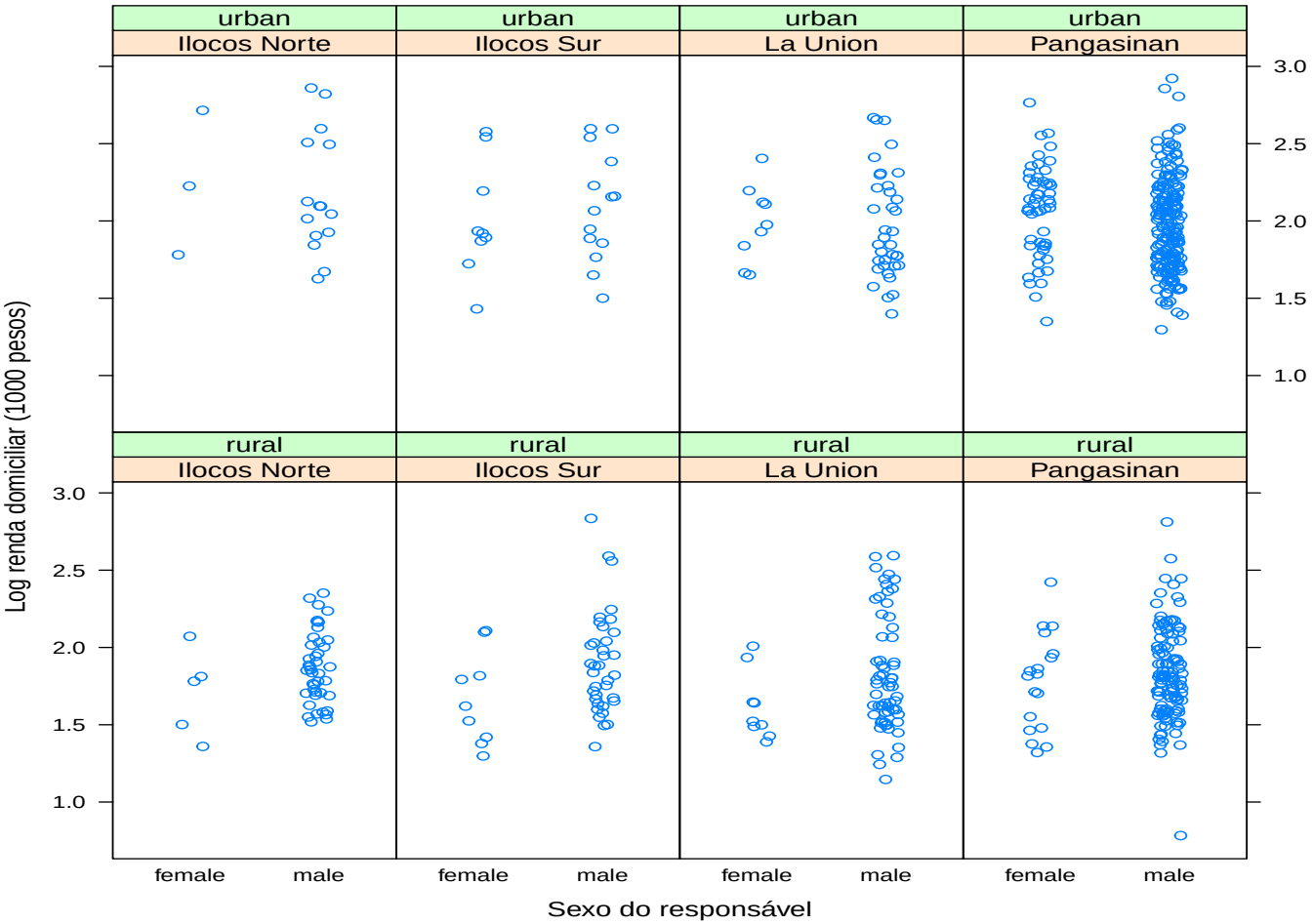
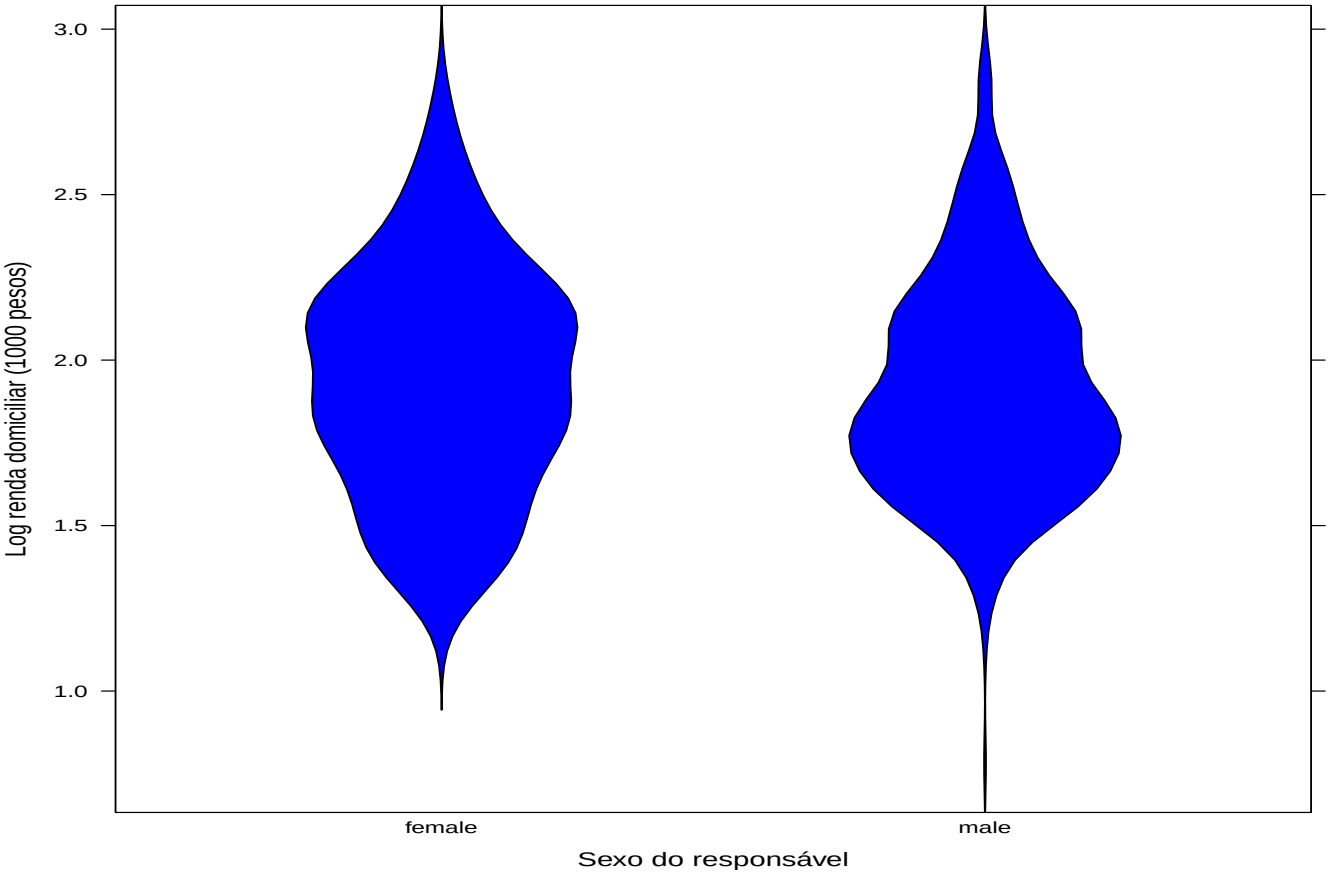


Gráfico de violino Função bwplot (lattice)

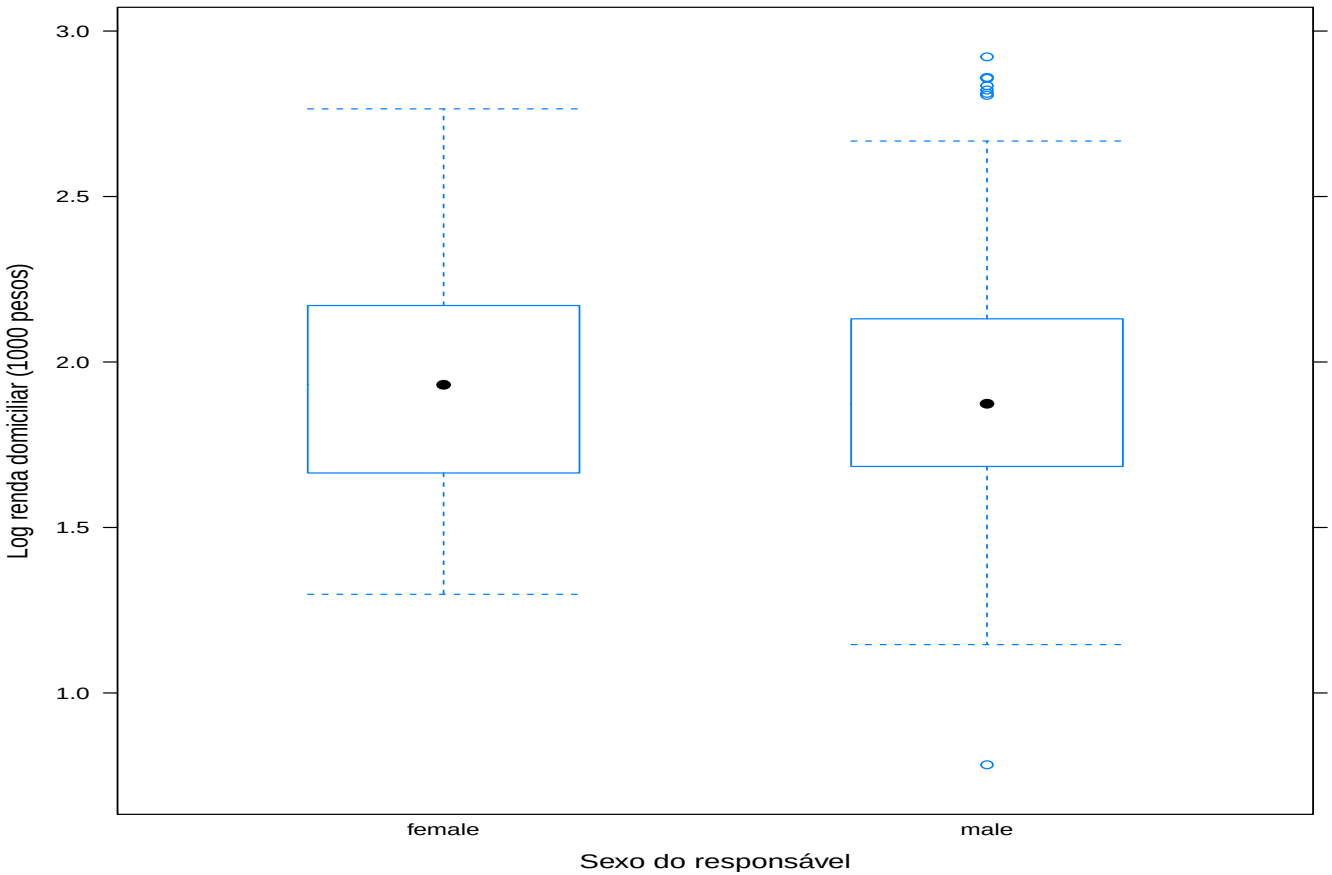
```
bwplot(log(income / 1000, 10) ~ sex, panel = panel.violin,
xlab = "Sexo do responsável", ylab = "Log renda domiciliar (1000 pesos)",
col = "blue")
```



Variáveis quantitativas e qualitativas

Gráfico de caixas Função bwplot (lattice)

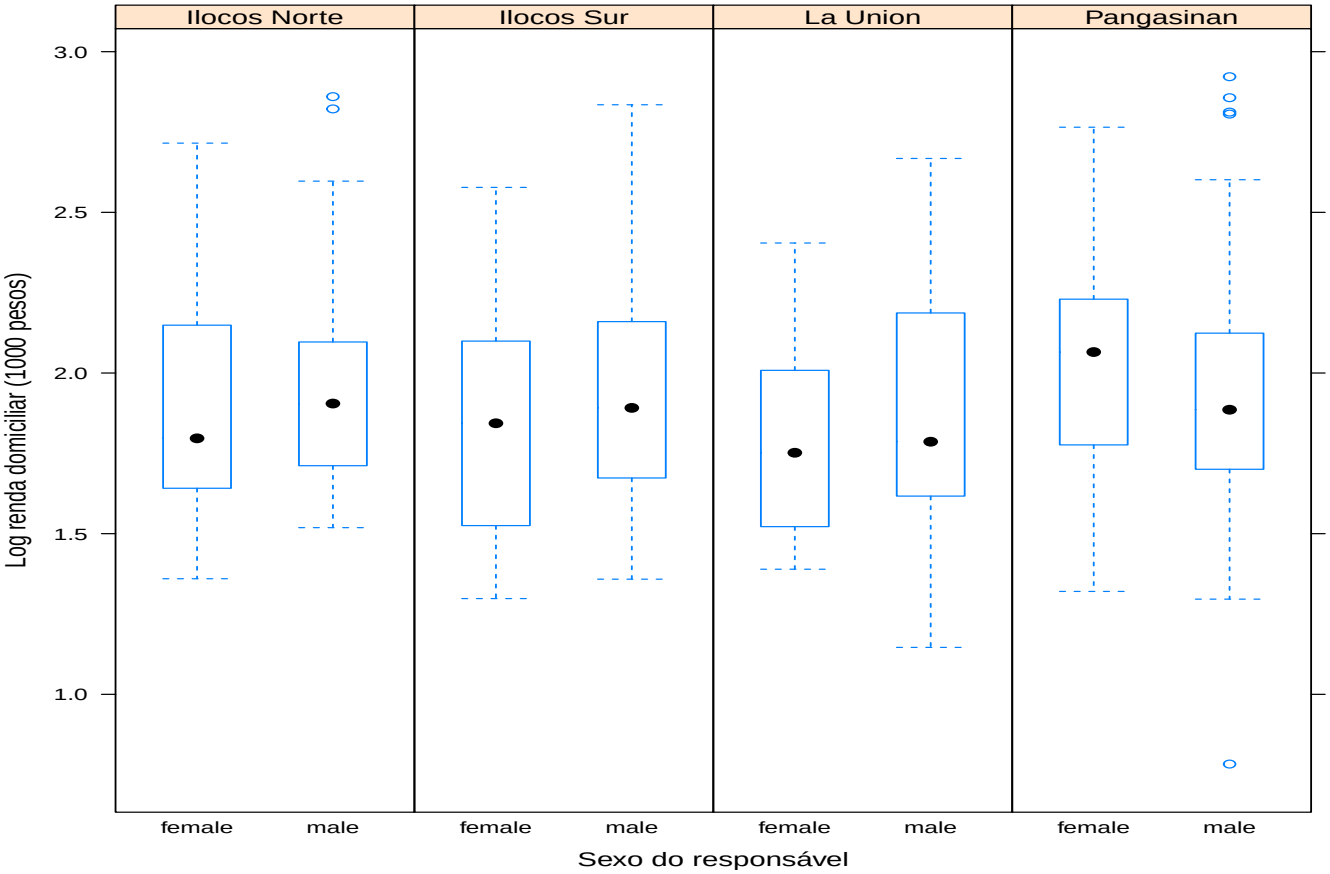
```
bwplot(log(income / 1000, 10) ~ sex, xlab = "Sexo do responsável",  ylab = "Lo
```



Variáveis quantitativas e qualitativas

Uma variável condicionante

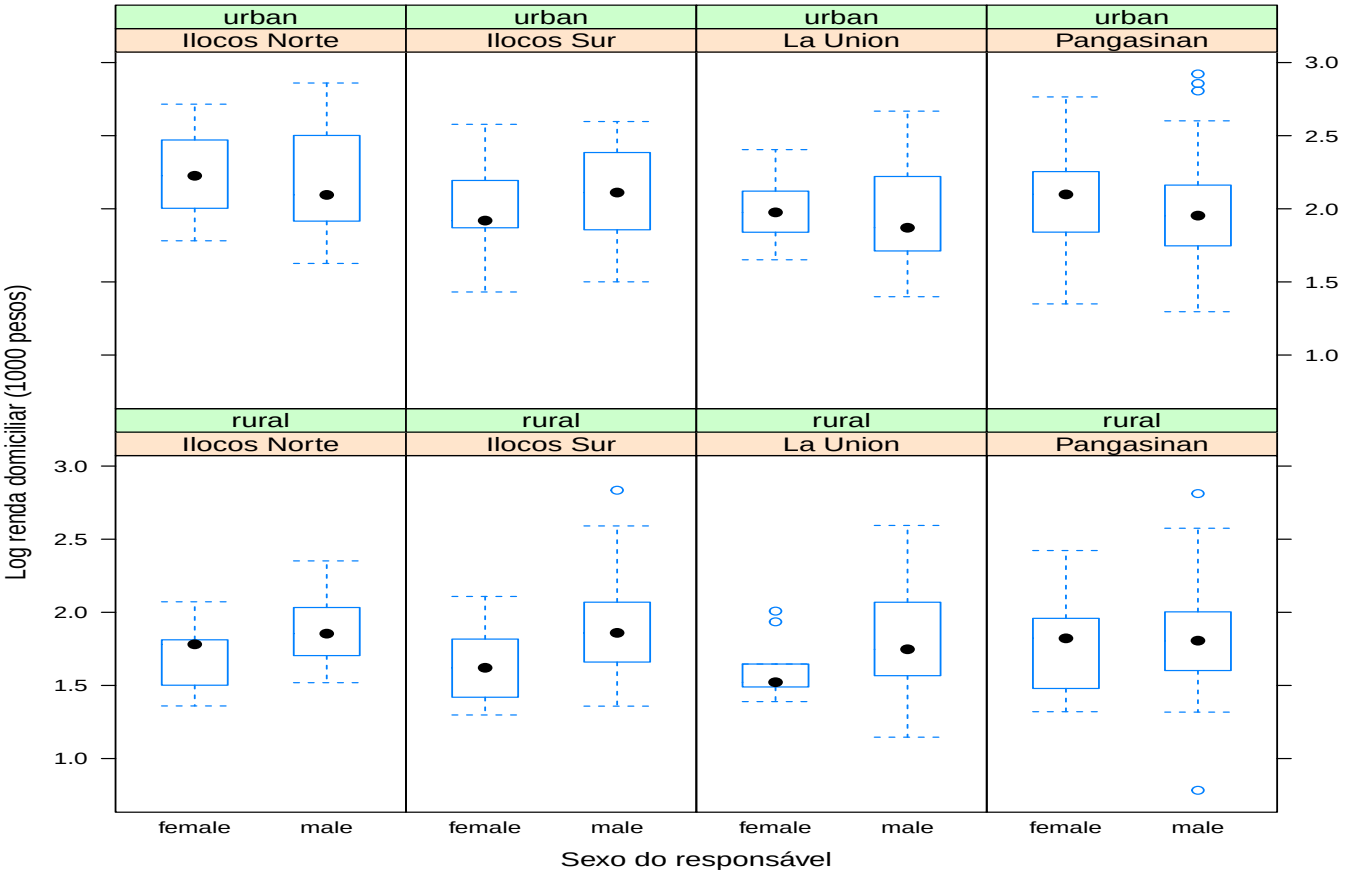
```
bwplot(log(income / 1000, 10) ~ sex | province, xlab = "Sexo do responsável",
ylab = "Log renda domiciliar (1000 pesos)", layout = c(4, 1),data=dados)
```



Variáveis quantitativas e qualitativas

Duas variáveis condicionantes

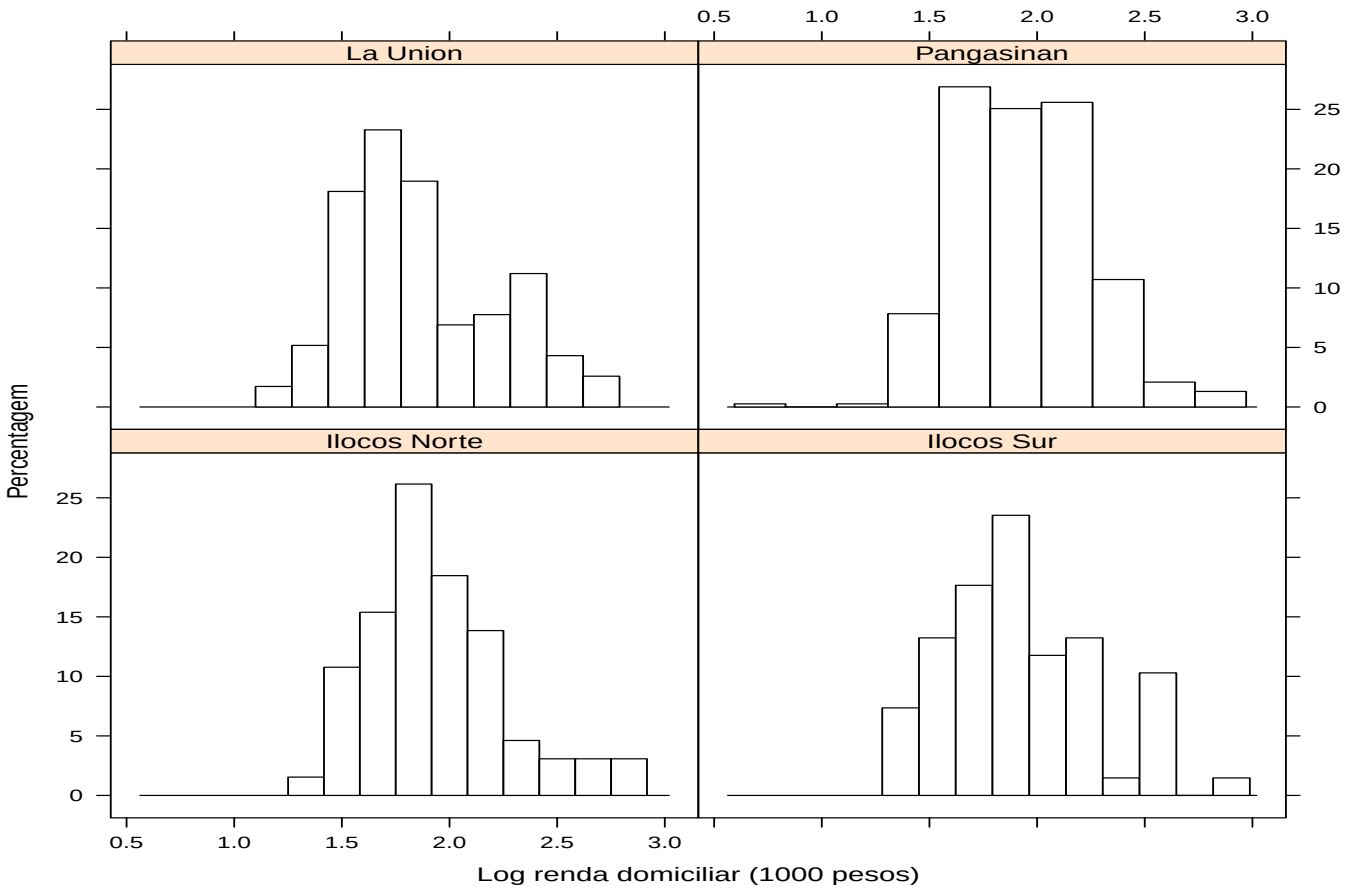
```
bwplot(log(income / 1000, 10) ~ sex | province, xlab = "Sexo do responsável",
ylab = "Log renda domiciliar (1000 pesos)", layout = c(4, 1),data=dados)
```



Variáveis quantitativas e qualitativas

Histograma: Função histogram (lattice)

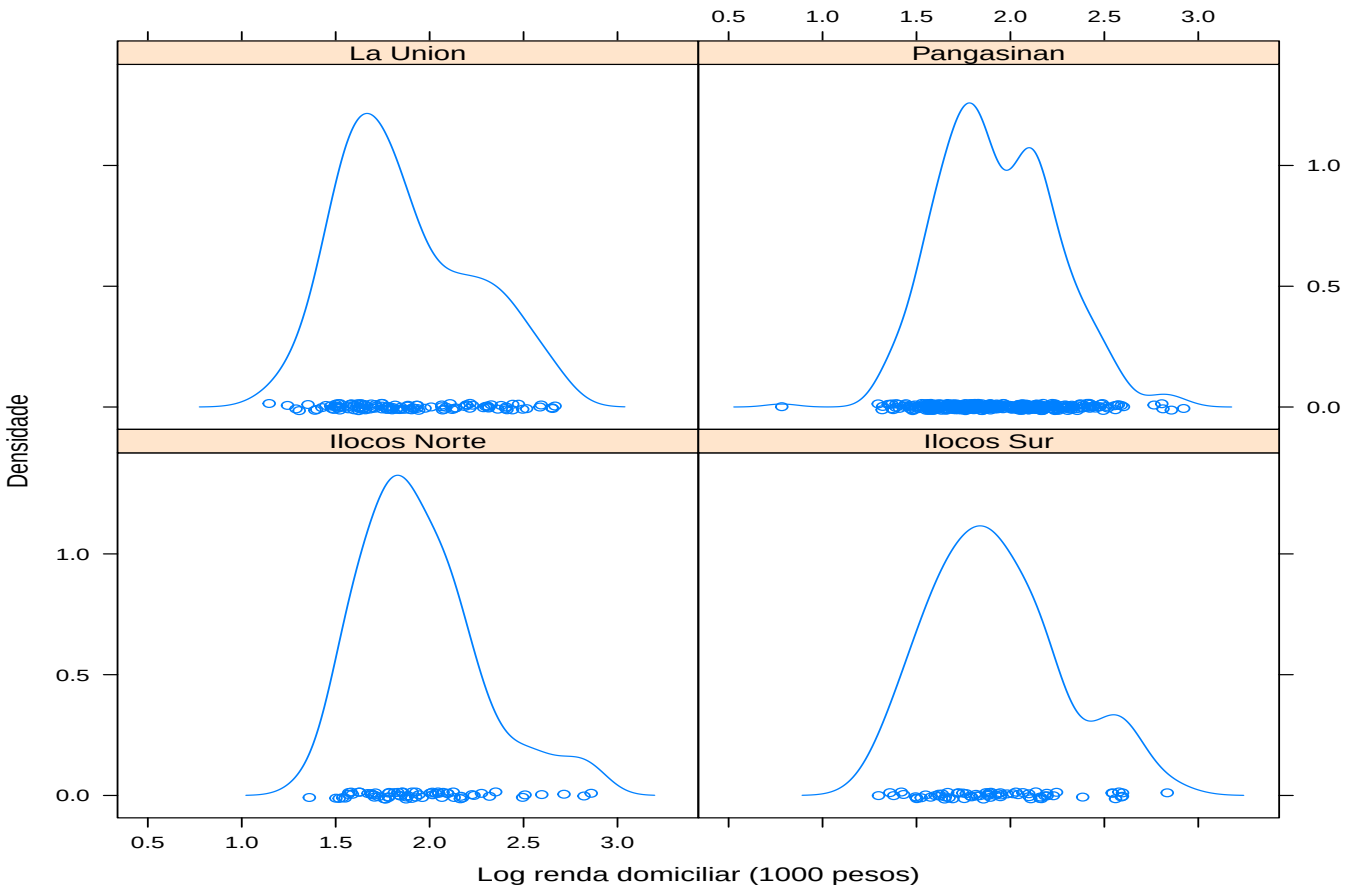
```
histogram(~ log(income / 1000, 10) | province, type = "percent",
 ylab = "Percentagem",xlab = "Log renda domiciliar (1000 pesos)",
  col = "white",data=dados)
```



Variáveis quantitativas e qualitativas

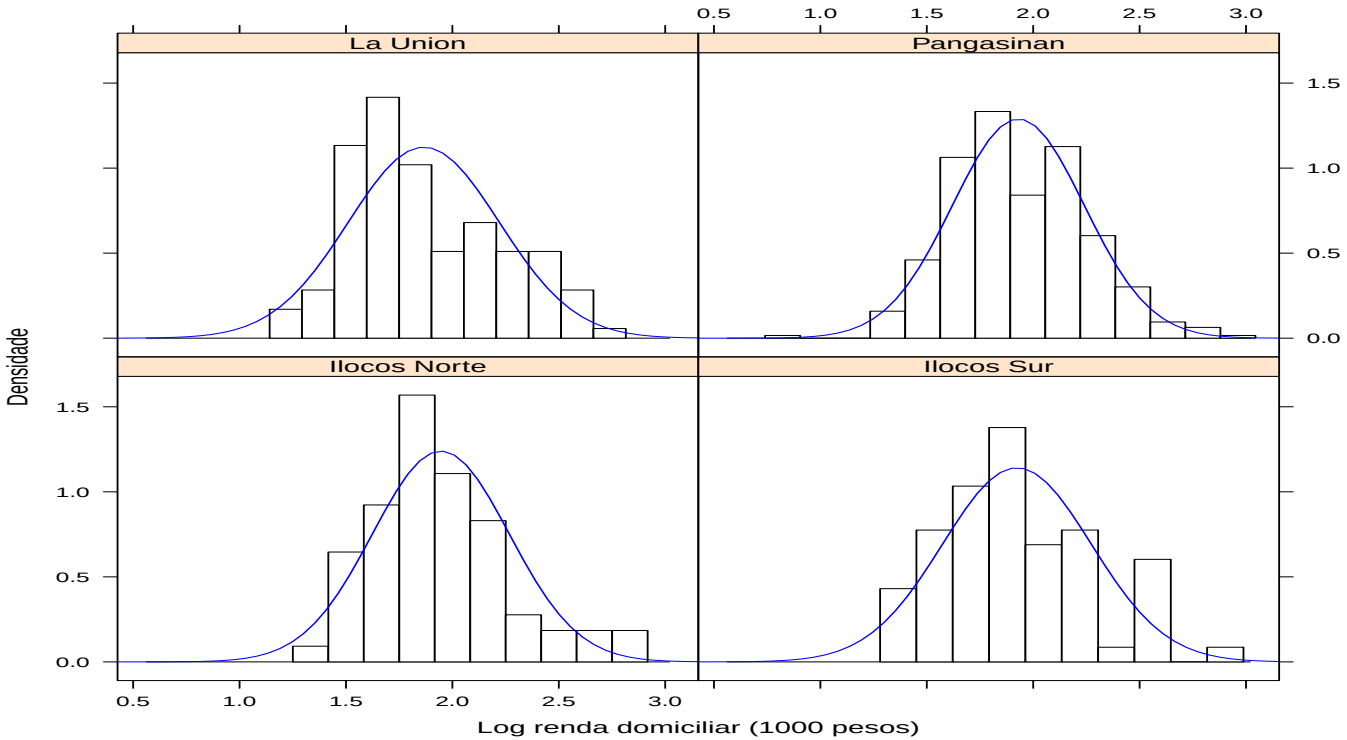
Gráfico de densidade Função densityplot (lattice)

```
densityplot(~ log(income / 1000, 10) | province, type = "percent",
ylab = "Percentagem",xlab = "Log renda domiciliar (1000 pesos)",
col = "white",data=dados)
```



Histograma e função densidade normal

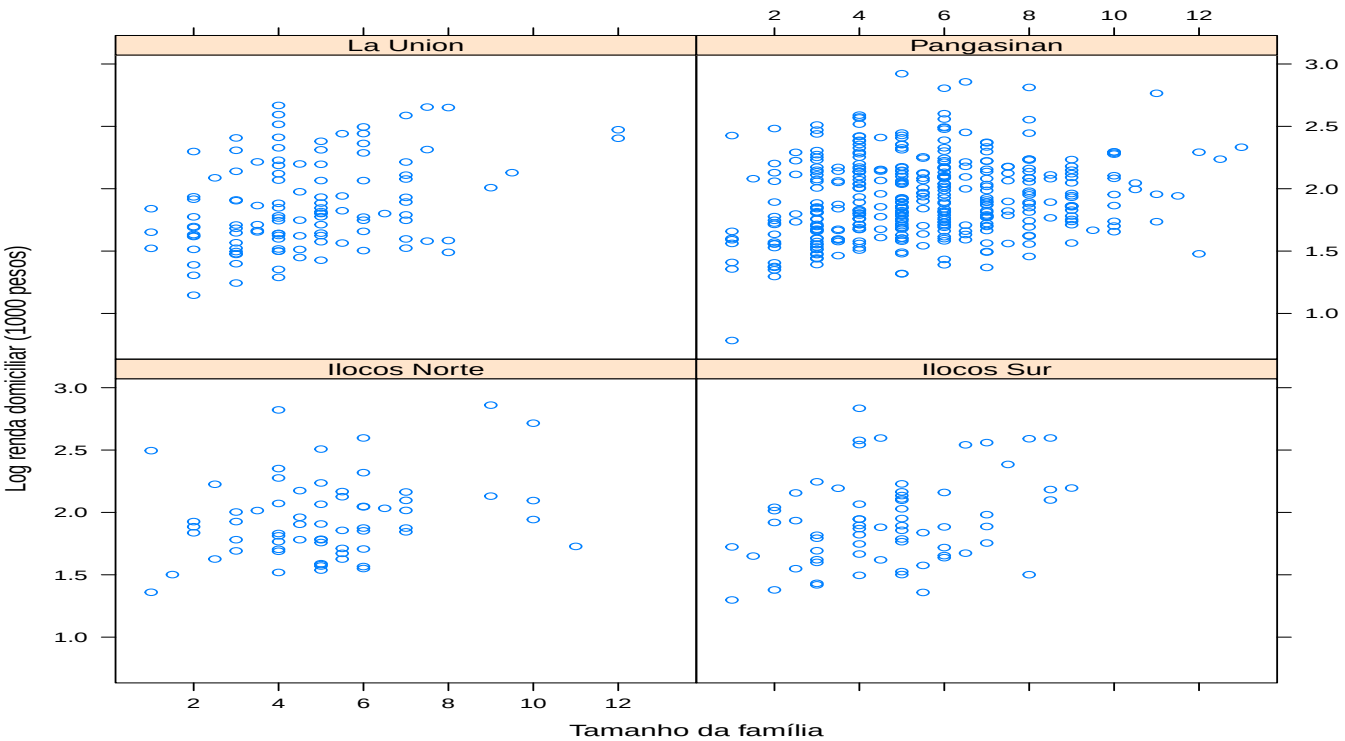
```
histogram(~ log(income / 1000, 10) | province, type = "density",
ylab = "Densidade",  xlab = "Log renda domiciliar (1000 pesos)",
col = "white",
panel = function(x, ...){ panel.histogram(x, ...)
panel.mathdensity(dmath = dnorm, col = "blue",
args = list(mean  = mean(x),sd = sd(x))) },data=dados)
```



Variáveis quantitativas e qualitativas

Gráfico de dispersão Função xyplot (lattice)

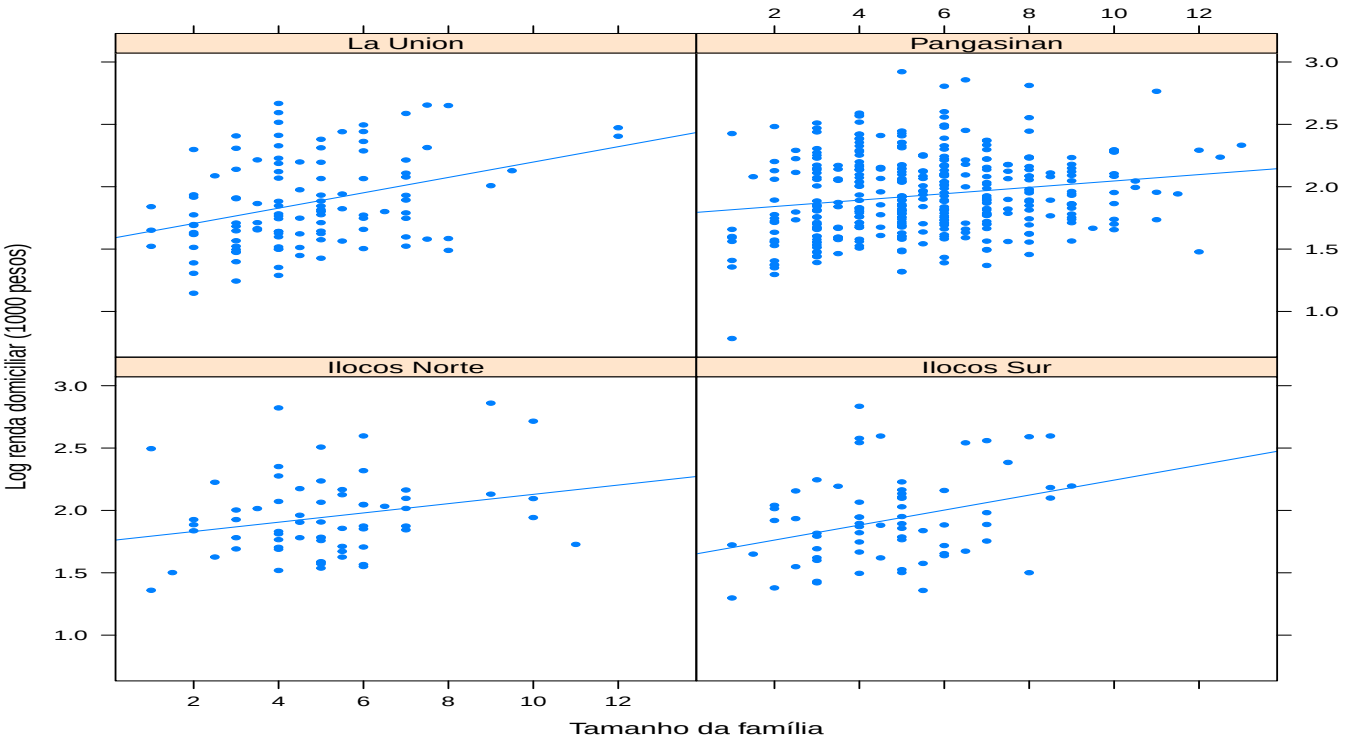
```
xyplot(log(income / 1000, 10) ~ family.size | province,
xlab = "Tamanho da família",
ylab = "Log renda domiciliar (1000 pesos)",data=dados)
```



Variáveis quantitativas e qualitativas

Gráfico com pontos (p) e reta ajustada (r)

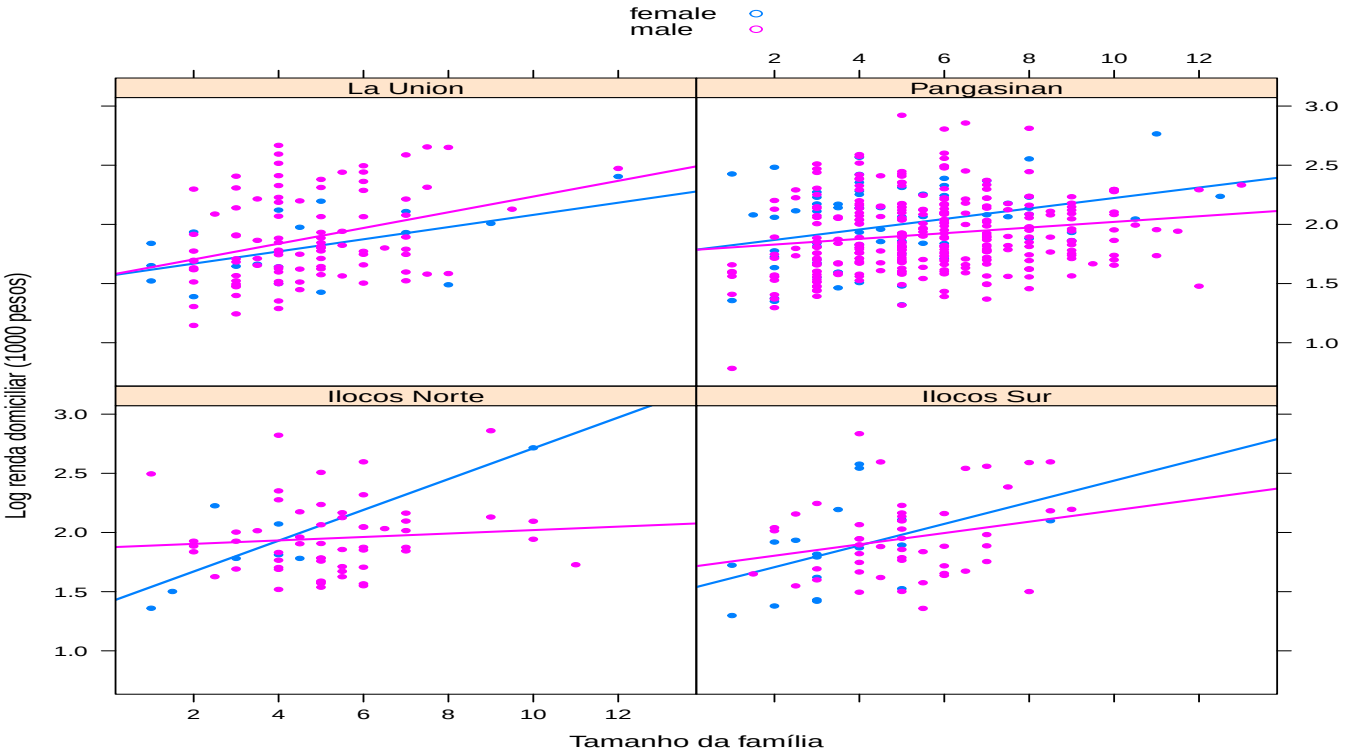
```
xyplot(log(income / 1000, 10) ~ family.size | province,
xlab = "Tamanho da família",ylab = "Log renda domiciliar (1000 pesos)",
pch = 20, type = c("p", "r"),data=dados)
```



Variáveis quantitativas e qualitativas

Grupos de acordo com a variável sex

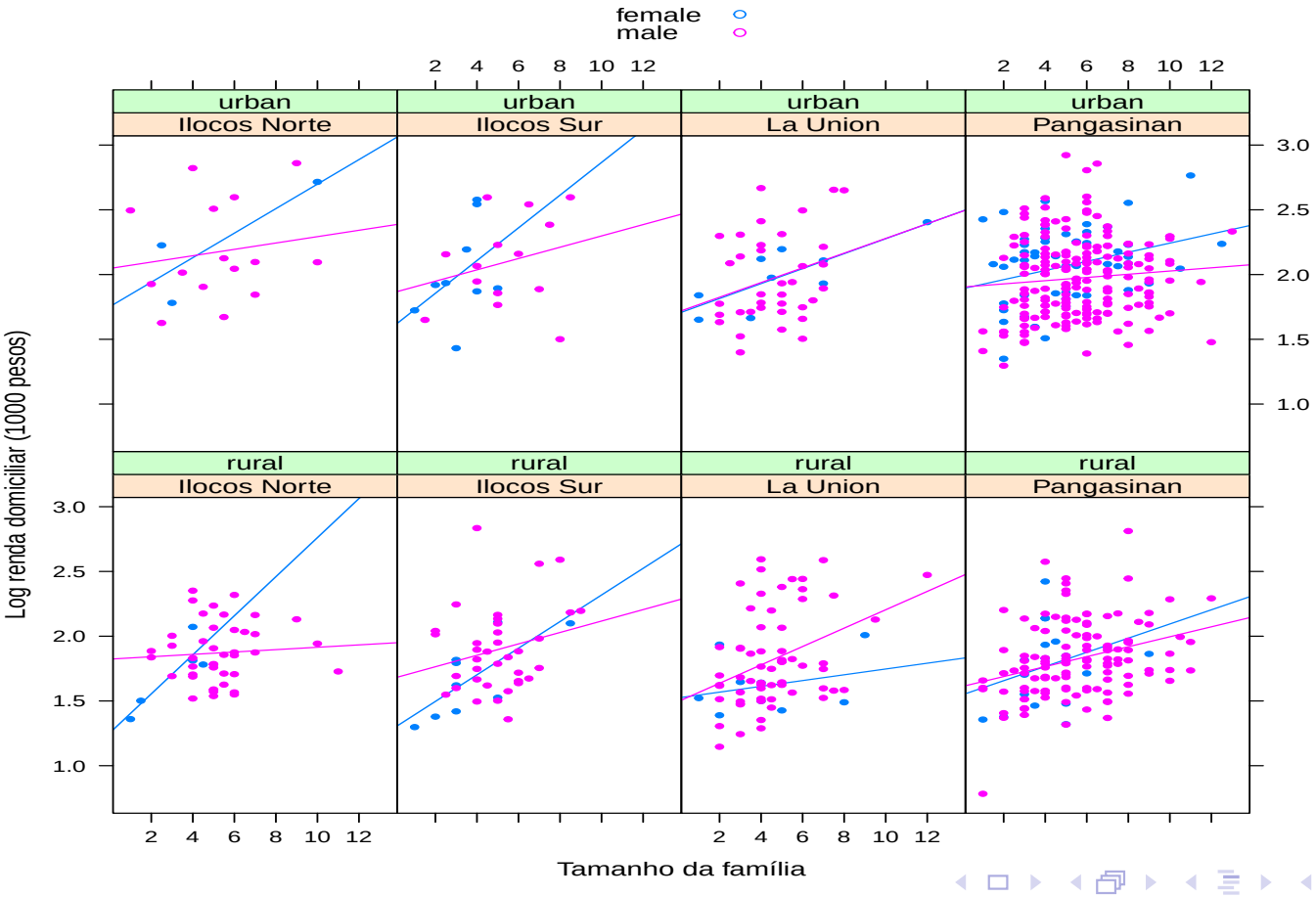
```
xyplot(log(income / 1000, 10) ~ family.size | province, group = sex,
auto.key = TRUE, xlab = "Tamanho da família",
ylab = "Log renda domiciliar (1000 pesos)",
 pch = 20, type = c("p", "r"),data=dados)
```



Variáveis quantitativas e qualitativas

Duas variáveis condicionantes

```
xyplot(log(income / 1000, 10) ~ family.size | province + urbanity, group = sex,
ylab = "Log renda domiciliar (1000 pesos)",
pch = 20, type = c("p", "r"),data=dados)
```



Variáveis quantitativas e qualitativas

Para os dados do Exemplos2- Eleições Americanas, 1996

```
data(nes96, package="faraway")
Partido <-  nes96$PID
levels(Partido) <- c("Democrata","Democrata","Independente","Independente", "In
"Republicano","Republicano")
inca <- c(1.5,4,6,8,9.5,10.5,11.5,12.5,13.5,14.5,16,18.5,21,23.5,
27.5,32.5,37.5,42.5,47.5,55,67.5,82.5,97.5,115)
income <- inca[unclass(nes96$income)]
rnes96 <- data.frame(Partido, Renda=income,
 Educação=nes96$educ, Idade=nes96$age)
head(rnes96)
        Partido Renda Educação Idade
1 Republicano   1.5       HS    36
2   Democrata   1.5     Coll    20
3   Democrata   1.5    BAdeg    24
4   Democrata   1.5    BAdeg    28
5   Democrata   1.5    BAdeg    68
6   Democrata   1.5     Coll    21
```

Variáveis quantitativas e qualitativas