Análise descritiva bivariada

Giovanna Segantini

giovanna.ufrn@gmail.com

Referência

Capítulo 3 - Estatística descritiva bivariada: FÁVERO, Luiz Paulo; BELFIORE, Patrícia. Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®. Elsevier Brasil, 2017.

Objetivos

- Compreender os principais conceitos de esatística descritiva bivariada
- Escolher o(s) método(s) adequado(s) para descrever o comportamento das variáveis
- Gerar tabelas, gráficos e medidas-resumo por meio da linguagem R.

Introdução

Na análise bivariada, procuramos identificar relações entre duas variáveis.

O tipo de resumo estatístico informativo vai depender dos tipos das variáveis envolvidas.

A seguir, mostramos algumas possibilidades desse tipo de análise. Lembrando que as apresentadas não esgotam as possibilidades de análise envolvendo duas variáveis.

Obs. Salientamos que as relações entre duas variáveis devem ser examinadas com cautela, pois podem ser mascaradas por variáveis adicionais, não consideradas na análise

Banco de dados

```
milsa <- read.delim("C:/Users/gato/Desktop/Github/analise-d
#milsa <- read.delim("C:/Users/gato/
#Desktop/Github/analise-de-dados/milsa.txt")

milsa$Inst <- as.factor(milsa$Inst)

levels(milsa$Inst)
```

```
## [1] "1o Grau"  "2o Grau"  "Superior"
```

```
milsa$Inst <- factor(milsa$Inst, ordered = TRUE)

milsa$Idade <- milsa$Ano + milsa$Meses/12
```

Associação entre duas variáveis Qualitativas

O Objetivo é avaliar se existe relação entre as variáveis qualitativas ou categóricas estudadas, além do grau de associação entre elas.

- Tabelas de frequência cruzadas
- Medidas-resumo: -chi-quadrado (variáveis nominais e ordinais); -coeficiente Phi(variáveis nominais) -coeficiente de contingência (variáveis nominais) -coeficiente de V de Cramer (variáveis nominais) -coeficiente de Spearman (variáveis ordinais)

Vamos considerar as variáveis civil (estado civil) e instrução (grau de instrução).

QUALITATIVA VS QUALITATIVA

Inicialmente obteremos a tabela de Frequências absolutas e relativas

```
civ.gi.tb <- table(milsa$Est.civil, milsa$Inst)
civ.gi.tb
```

```
##
##             1o Grau 2o Grau Superior
##   casado          5      12        3
##   solteiro        7       6        3
```

```
addmargins(civ.gi.tb)
```

```
##
##             1o Grau 2o Grau Superior Sum
##   casado          5      12        3  20
##   solteiro        7       6        3  16
##   Sum            12      18        6  36
```

Tabelas de frequências relativas são obtidas com *prop.table()*

```
prop.table(civ.gi.tb)
```

```
##
##                1o Grau    2o Grau   Superior
##   casado   0.13888889 0.33333333 0.08333333
##   solteiro 0.19444444 0.16666667 0.08333333
```

Existe também a possibilidade de fazer tabelas de frequência:
- Em relação aos totais por linha (margin = 1)

```
prop.table(civ.gi.tb, margin = 1)
```

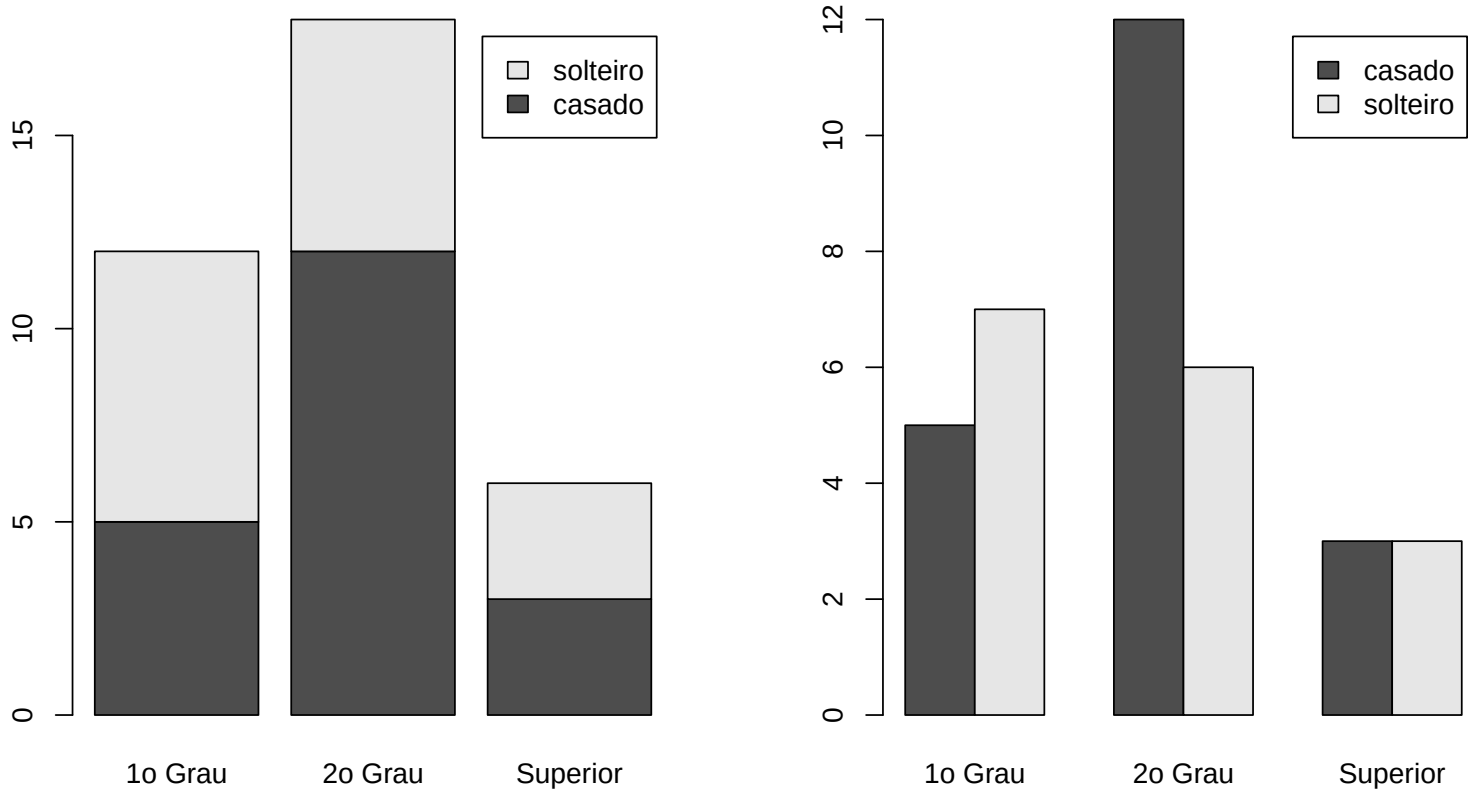
```
##
##              1o Grau 2o Grau Superior
##    casado     0.2500  0.6000   0.1500
##    solteiro   0.4375  0.3750   0.1875
```

- Em relação aos totais por coluna (margin = 2)

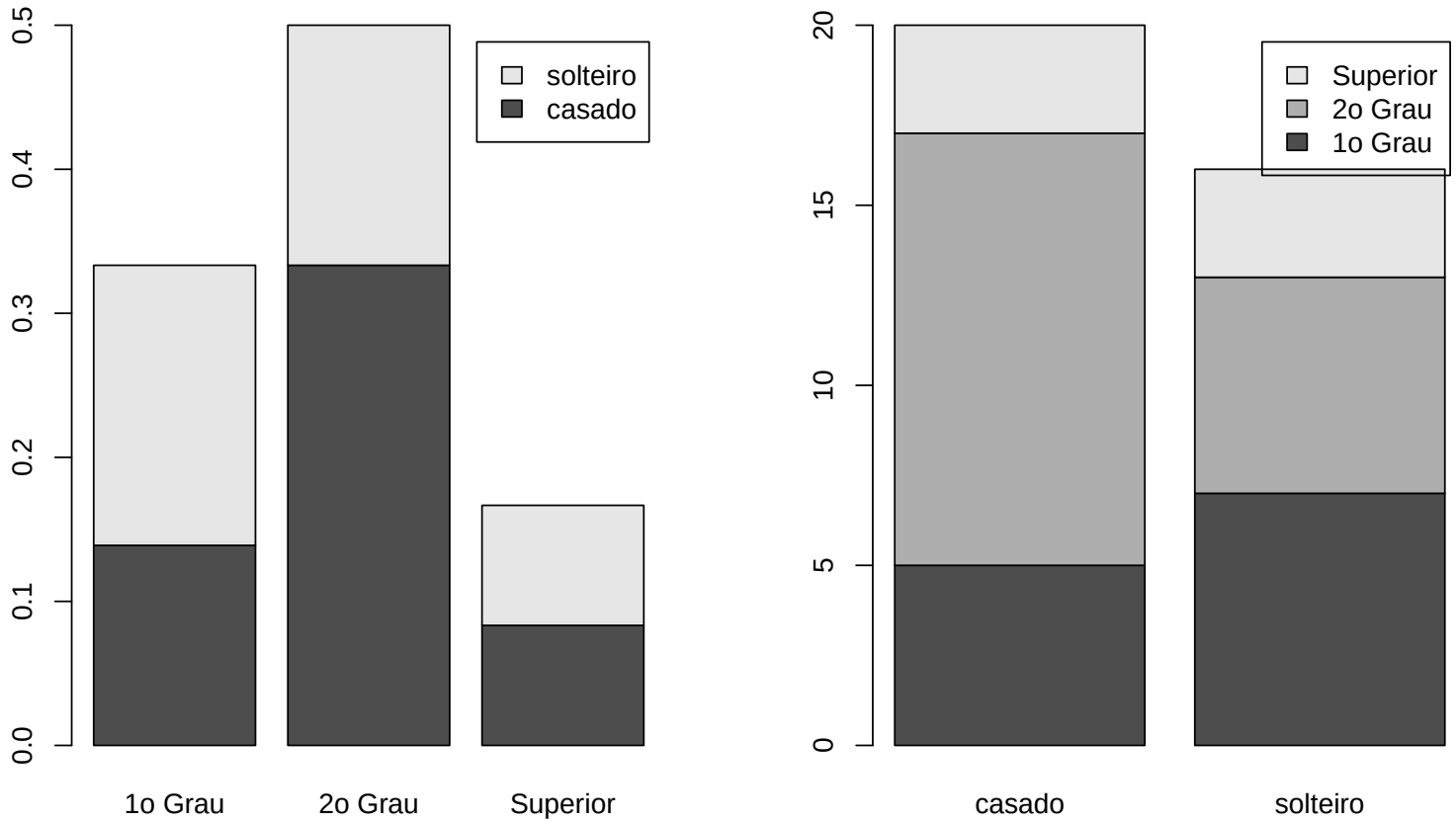
```
prop.table(civ.gi.tb, margin = 2)
```

```
##
##                1o Grau   2o Grau  Superior
##    casado    0.4166667 0.6666667 0.5000000
##    solteiro  0.5833333 0.3333333 0.5000000
```

```r
par(mfrow = c(1,2), oma = c(4,1,1,1))

barplot(civ.gi.tb, legend = T)

barplot(civ.gi.tb, beside = T, legend = T)
```



```r
par(mfrow = c(1,2), oma = c(3,1,1,1))

barplot(prop.table(civ.gi.tb), legend = T)

barplot(t(civ.gi.tb), legend = T)
```



QUALITATIVA VS QUALITATIVA

- Medidas de associação

As principais medidas que representam a associação entre duas variáveis qualitativas são:

- a estatística qui-quadrado χ^2, utilizada para variáveis qualitativas nominais e ordinais;
- ϕ (phi) (é o R de pearson quando aplicado a tabelas 2x2), V de Crámer, Coeficiente de contingência, baseado no χ^2, para variáveis nominais
- o coeficiente de Spearman para variáveis ordinais

QUALITATIVA VS QUALITATIVA

- Estatística qui-quadrado

A estatística qui-quadrado mede a discrepância entre uma frequência observada e uma frequência esperada, partindo da hipótese de que não há associação entre as variáveis estudadas. Assim, um valor baixo do qui-quadrado indica independência entre as variáveis.

Hipóteses a serem testadas – Teste de independência:

H: A e B são variáveis independentes

A: As variáveis A e B não são independentes

```r
summary(civ.gi.tb)
```

```
## Number of cases in table: 36
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 1.9125, df = 2, p-value = 0.3843
##  Chi-squared approximation may be incorrect
```

```r
chisq.test(milsa$Est.civil, milsa$Inst)
```

```
## Warning in chisq.test(milsa$Est.civil, milsa$Inst): Chi-
## may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  milsa$Est.civil and milsa$Inst
## X-squared = 1.9125, df = 2, p-value = 0.3843
```

Diante dos resultados a hipótese nula não pode ser rejeitada, ou seja, o estado civil dos funcionários e o grau de instrução são independentes.

O teste qui-quadrado quando aplicado a amostras pequenas, como por exemplo com tamanho inferior a 20, veja:

```
fisher.test(civ.gi.tb)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  civ.gi.tb
## p-value = 0.4044
## alternative hypothesis: two.sided
```

Para saber a internsidade desta relação, utilizam-se medidas de associação.

Considere as seguintes medidas:

-ϕ (phi) (é o R de pearson quando aplicado a tabelas 2x2) -V de Crámer -Coeficiente de contingência

Ambos variam de 0 (ausência de associação) a 1 (associação muito forte).

```r
#install.packages("vcd")
library(vcd)

summary(assocstats(civ.gi.tb))
```

Correlação entre as variáveis Qualitativa e quantitativas

Para exemplificar este caso vamos considerar as variáveis Inst e Salario.

Para se obter uma tabela de frequências é necessário agrupar a variável quantitativa em classes. No exemplo a seguir vamos agrupar a variável salário em 4 classes definidas pelos quartis usando a função cut(). Lembre-se que as classes são definidas por intervalos abertos à esquerda, então usamos o argumento include.lowest = TRUE para garantir que todos os dados, incluve o menor (mínimo) seja incluído na primeira classe. Após agrupar esta variável, obtemos a(s) tabela(s) de cruzamento como mostrado no caso anterior.

```r
## Quartis de salario
quantile(milsa$Salario)
```

```
##      0%     25%     50%     75%    100%
##  4.0000  7.5525 10.1650 14.0600 23.3000
```

```r
## Classificação de acordo com os quartis
salario.cut <- cut(milsa$Salario,
                   breaks = quantile(milsa$Salario),
                   include.lowest = TRUE)



salario.cut
```

```
##  [1] [4,7.55]     [4,7.55]     [4,7.55]     [4,7.55]     [4,
##  [7] [4,7.55]     [4,7.55]     (7.55,10.2] [4,7.55]     (7.
## [13] (7.55,10.2] (7.55,10.2] (7.55,10.2] (7.55,10.2] (7.
## [19] (10.2,14.1] (10.2,14.1] (10.2,14.1] (10.2,14.1] (10
## [25] (10.2,14.1] (10.2,14.1] (10.2,14.1] (14.1,23.3] (14
```

```r
## Tabela de frequências absolutas
inst.sal.tb <- table(milsa$Inst, salario.cut)
inst.sal.tb
```

```
##             salario.cut
##              [4,7.55] (7.55,10.2] (10.2,14.1] (14.1,23.3]
##    1o Grau          7           3           2           0
##    2o Grau          2           6           5           5
##    Superior         0           0           2           4
```

```r
prop.table(inst.sal.tb)
```

```
##             salario.cut
##                [4,7.55] (7.55,10.2] (10.2,14.1] (14.1,23.3
##    1o Grau  0.19444444  0.08333333  0.05555556  0.0000000
##    2o Grau  0.05555556  0.16666667  0.13888889  0.1388888
##    Superior 0.00000000  0.00000000  0.05555556  0.1111111
```

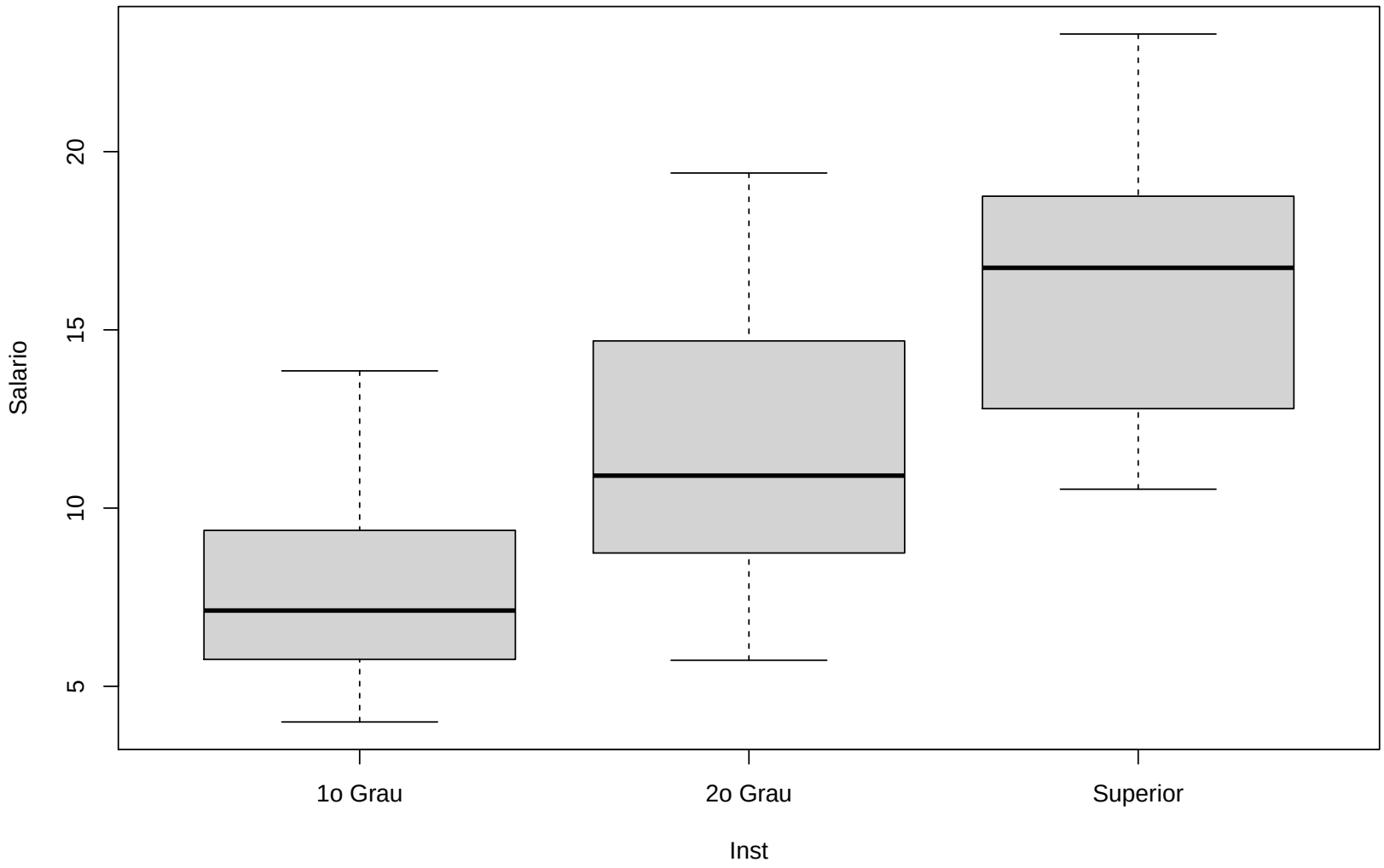
Qualitativa vs Quantitativa

- Gráficos

No gráfico vamos considerar que neste exemplo a instrução deve ser a variável explicativa e portanto colocada no eixo X, e o salário é a variável resposta, e portanto deve ser colocada no eixo Y. Isto é, consideramos que a instrução deve explicar, ainda que parcialmente, o salário.

Vamos então obter um boxplot dos salários para cada nível de instrução. Note que na função abaixo, usamos a notação de fórmula do R, indicando que a variável Salario é explicada, ou descrita, pela variável Inst.

```r
boxplot(Salario ~ Inst, data = milsa)
```



Qualitativa vs Quantitativa

Para as medidas descritivas, o usual é obter um resumo da variável quantitativa como mostrado na análise univariada, porém agora informando este resumo para cada nível do fator qualitativo de interesse.

```
with(milsa, tapply(Salario, Inst, summary))

with(milsa, tapply(Salario, Inst, sd))

with(milsa, tapply(Salario, Inst, var))

with(milsa, tapply(Salario, Inst, qunatile))
```

Qualitativa vs Quantitativa

- Medidas de associação

Independente de ser normal ou nao

- Spearman (amostras maiores)
- Kendall (amostras pequenas)

```
educacao<- as.numeric(milsa$Inst)
```

```r
#Exemplo de uso de spearman:

cor(educacao, milsa$Salario, method = "spearman")
```

```
## [1] 0.6291094
```

```r
cor.test(educacao, milsa$Salario, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  educacao and milsa$Salario
## S = 2881.8, p-value = 3.961e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.6291094
```

```r
#Exemplo de uso de kendall com uma amostra menor:

cor(educacao, milsa$Salario, method = "kendall")
```

```
## [1] 0.5165375
```

```r
cor.test(educacao, milsa$Salario, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  educacao and milsa$Salario
## z = 3.8667, p-value = 0.0001103
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.5165375
```

Correlação entre as variáveis Quatitativa e quantitativa

Para ilustrar este caso vamos considerar as variáveis Salario e Idade. Para se obter uma tabela é necessário agrupar as variáveis em classes conforme fizemos no caso anterior.

Nos comandos abaixo, agrupamos as duas variáveis em classes definidas pelos respectivos quartis, gerando portanto uma tabela de cruzamento 4×4.

```
## Classes de Idade
idade.cut <- with(milsa, cut(Idade,
                  breaks = quantile(Idade),


table(idade.cut)
```

```
## idade.cut
## [20.8,30.7] (30.7,34.9] (34.9,40.5] (40.5,48.9]
##           9           9           9           9
```

```r
## Classes de salario
salario.cut <- with(milsa, cut(Salario,
                    breaks = quantile(Salario),
                    include.lowest = TRUE))

table(salario.cut)
```

```
## salario.cut
##    [4,7.55] (7.55,10.2] (10.2,14.1] (14.1,23.3]
##           9           9           9           9
```

```r
## Tabela cruzada
table(idade.cut, salario.cut)
```

```
##                salario.cut
## idade.cut       [4,7.55] (7.55,10.2] (10.2,14.1] (14.1,23.
##   [20.8,30.7]          4            2            2
##   (30.7,34.9]          1            3            3
##   (34.9,40.5]          1            3            2
##   (40.5,48.9]          3            1            2
```

```r
prop.table(table(idade.cut, salario.cut))
```

```
##                salario.cut
## idade.cut         [4,7.55] (7.55,10.2] (10.2,14.1] (14.1,2
##   [20.8,30.7] 0.11111111  0.05555556  0.05555556  0.027
##   (30.7,34.9] 0.02777778  0.08333333  0.08333333  0.055
##   (34.9,40.5] 0.02777778  0.08333333  0.05555556  0.083
##   (40.5,48.9] 0.08333333  0.02777778  0.05555556  0.083
```

Caso queiramos definir um número menor de classes podemos fazer como no exemplo a seguir onde cada variável é dividida em 3 classes e gerando um tabela de cruzamento 3×3.

```
idade.cut2 <- with(milsa, cut(Idade,
                      breaks = quantile(Idade,
                      seq(0, 1, length = 4)),
                      include.lowest = TRUE))
salario.cut2 <- with(milsa, cut(Salario,
                      breaks = quantile(Salario,
                      seq(0, 1, length = 4)),
                      include.lowest = TRUE))

table(idade.cut2, salario.cut2)
```

```
##                salario.cut2
## idade.cut2     [4,8.65] (8.65,12.9] (12.9,23.3]
##    [20.8,32.1]        5           5            2
##    (32.1,37.8]        4           3            5
##    (37.8,48.9]        3           4            5
```

Quantitativa vs Quantitativa

- Diagrama de dispersão

```
plot(Salario ~ Idade, data = milsa)
```