Análise descritiva - Uma visão univariada

Giovanna Segantini

giovanna.ufrn@gmail.com

Introdução

A análise univariada consiste basicamente em, para cada uma das variáveis individualmente:

Classificar a variável quanto a seu tipo:

- **Qualitativas (categóricas)**

-nominais

-ordinais

- **Quantitativas**

-discretas

-contínuas

Obter tabelas, gráficos e/ou medidas que resumam a variável A partir destes resultados pode-se montar um resumo geral dos dados.

Importando dados

O livro "Estatística Básica" de W. O. Bussab e P. A. Morettin traz no segundo capítulo um conjunto de dados hipotético de atributos de 36 funcionários da companhia "Milsa". Consulte o livro para mais detalhes sobre este dados.

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


View(milsa)

class(milsa)
```

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

E para conferir a estrutura dos dados podemos usar algumas funções como:

```
str(milsa)
```

```
## 'data.frame':    36 obs. of  8 variables:
##  $ Funcionario: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Est.civil  : chr  "solteiro" "casado" "casado" "solte
##  $ Inst       : chr  "1o Grau" "1o Grau" "1o Grau" "2o (
##  $ Filhos     : int  NA 1 2 NA NA 0 NA NA 1 NA ...
##  $ Salario    : num  4 4.56 5.25 5.73 6.26 6.66 6.86 7.3
##  $ Anos       : int  26 32 36 20 40 28 41 43 34 23 ...
##  $ Meses      : int  3 10 5 10 7 0 0 4 10 6 ...
##  $ Regiao     : chr  "interior" "capital" "capital" "out
```

Podemos classificar todas as variáveis desse conjunto de dados como:

Variável	Classificação
`Funcionario`	Quantitativa discreta
`Est.civil`	Qualitativa nominal
`Inst`	Qualitativa ordinal
`Filhos`	Quantitativa discreta
`Salario`	Quantitativa contínua
`Anos`	Quantitativa contínua
`Meses`	Quantitativa contínua
`Regiao`	Qualitativa nominal

Como a variável Inst é qualitativa ordinal, podemos indicar que ela deve ser tratada como ordinal:

```
class(milsa$Inst)

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

levels(milsa$Inst)
```

já notamos que a ordenação está correta (da esquerda para a direita), pois sabemos que a classificação interna dos níveis é por ordem alfabética, e nesse caso, por coincidência, a ordem já está na sequência correta. Mesmo assim, podemos indicar que este fator é ordinal, usando o argumento *ordered* da função *factor()*

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

Criando variável

Podemos ainda definir uma nova variável, chamada **Idade**, a partir das variáveis **Anos** e **Meses**:

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

Análise univariada

A seguir vamos mostrar como obter tabelas, gráficos e medidas com o R. Para isto vamos selecionar uma variável de cada tipo para que o leitor possa, por analogia, obter resultados para as demais.

Variável qualitativa nominal

A variável *Est.civil* é uma qualitativa nominal. Desta forma podemos obter: (i) uma tabela de frequências (absolutas e/ou relativas), (ii) um gráfico de setores, (iii) a "moda", i.e. o valor que ocorre com maior frequência.

```
class(milsa$Est.civil)
```

```
## [1] "character"
```

Variável qualitativa nominal

- Frequência

```
## Frequência absoluta
civil.tb <- table(milsa$Est.civil)
civil.tb
```

```
##
##   casado solteiro
##       20       16
```

```
## Frequência relativa, calculando manualmente
civil.tb/length(milsa$Est.civil)
```

```
##
##    casado  solteiro
## 0.5555556 0.4444444
```

```
## Frequência relativa, com a função prop.table()
prop.table(civil.tb)
```

Variável qualitativa nominal

- Gráficos

Os gráficos de barras e de setores são adequados para representar esta variável.

```
barplot(civil.tb)
pie(civil.tb)
```

Variável qualitativa nominal

-Moda

```
# Moda
getmode <- function(x) {
  na.x<- na.omit(x)
  ux <- unique(na.x)
  tab <- tabulate(match(na.x, ux))
  ux[tab == max(tab)]
}

getmode(milsa$Est.civil)
```

```
## [1] "casado"
```

Variável Qualitativa Ordinal

Para exemplificar como obter análises para uma variável qualitativa ordinal vamos selecionar a variável Inst.

- Frequências

```
## Frequência absoluta
inst.tb <- table(milsa$Inst)
inst.tb
```

```
##
##  1o Grau  2o Grau Superior
##       12       18        6
```

```
## Frequência relativa
prop.table(inst.tb)
```

```
##
##   1o Grau   2o Grau  Superior
## 0.3333333 0.5000000 0.1666667
```

Variável Qualitativa Ordinal

O gráfico de setores não é adequado para este tipo de variável por não expressar a ordem dos possíveis valores. Usamos então apenas um gráfico de barras conforme mostrado abaixo:

```
barplot(inst.tb)

## Menor para maior
barplot(sort(inst.tb))

## Maior para menor
barplot(sort(inst.tb, decreasing = TRUE))
```

Variável Qualitativa Ordinal

- Moda

```
getmode(milsa$Inst)
```

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

```
# Moda
names(inst.tb)[which.max(inst.tb)]
```

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

Variável quantitativa discreta

Vamos agora usar a variável Filhos (número de filhos) para ilustrar algumas análises que podem ser feitas com uma quantitativa discreta.

- Frequências

Frequências absolutas e relativas são obtidas como anteriormente.

```
## Frequência absoluta
filhos.tb <- table(milsa$Filhos)
filhos.tb
```

```
##
## 0 1 2 3 5
## 4 5 7 3 1
```

```
## Frequência relativa
filhos.tbr <- prop.table(filhos.tb)
filhos.tbr
```

```
##
```

Também vamos calcular a frequência acumulada, onde a frequência em uma classe é a soma das frequências das classes anteriores. Para isso usamos a função *cumsum()*, que já faz a soma acumulada.

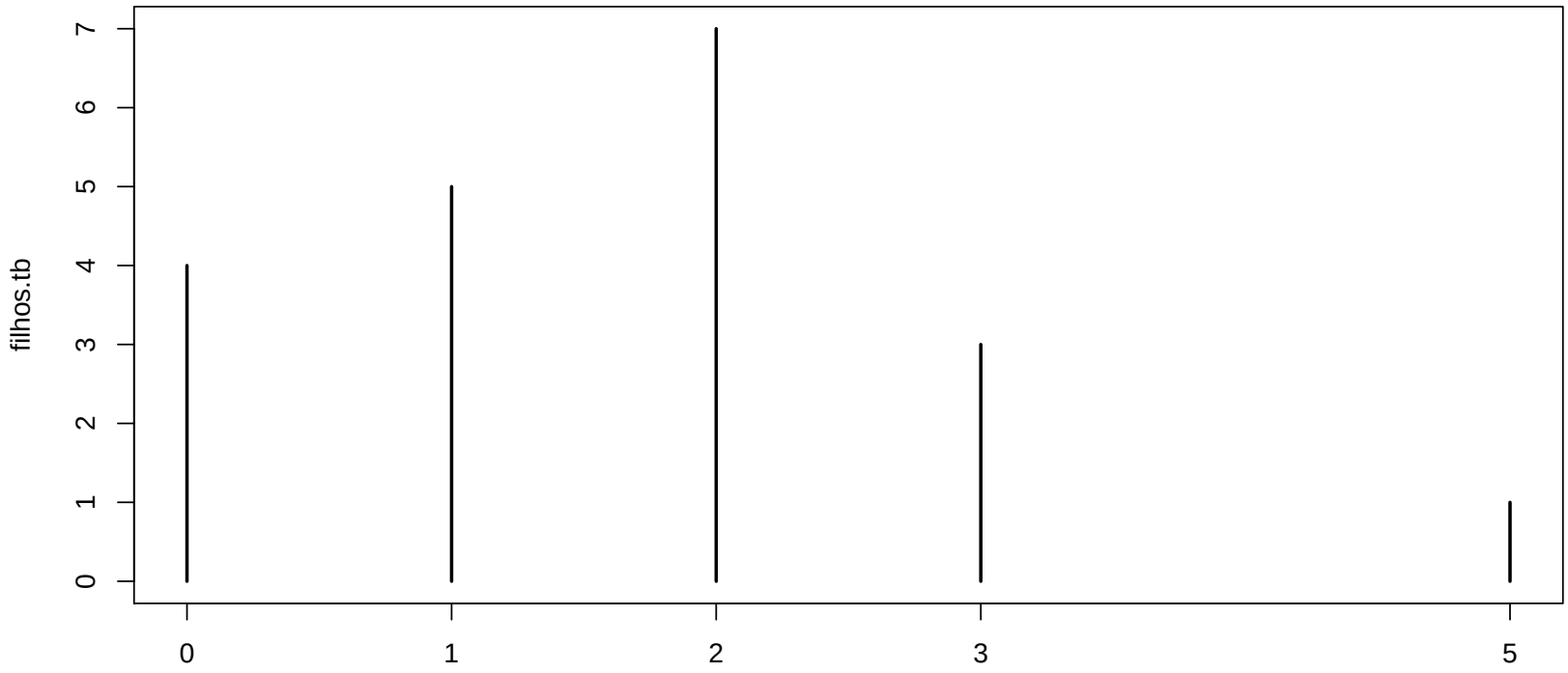
```
## Frequência acumulada
filhos.tba <- cumsum(filhos.tbr)
filhos.tba
```

```
##    0    1    2    3    5
## 0.20 0.45 0.80 0.95 1.00
```

Variável quantitativa discreta

Para a representação gráfica de frequências absolutas de uma variável discreta usarems um gráfico semelhante ao de barras, mas nesse caso, as frequências são indicadas por linhas.

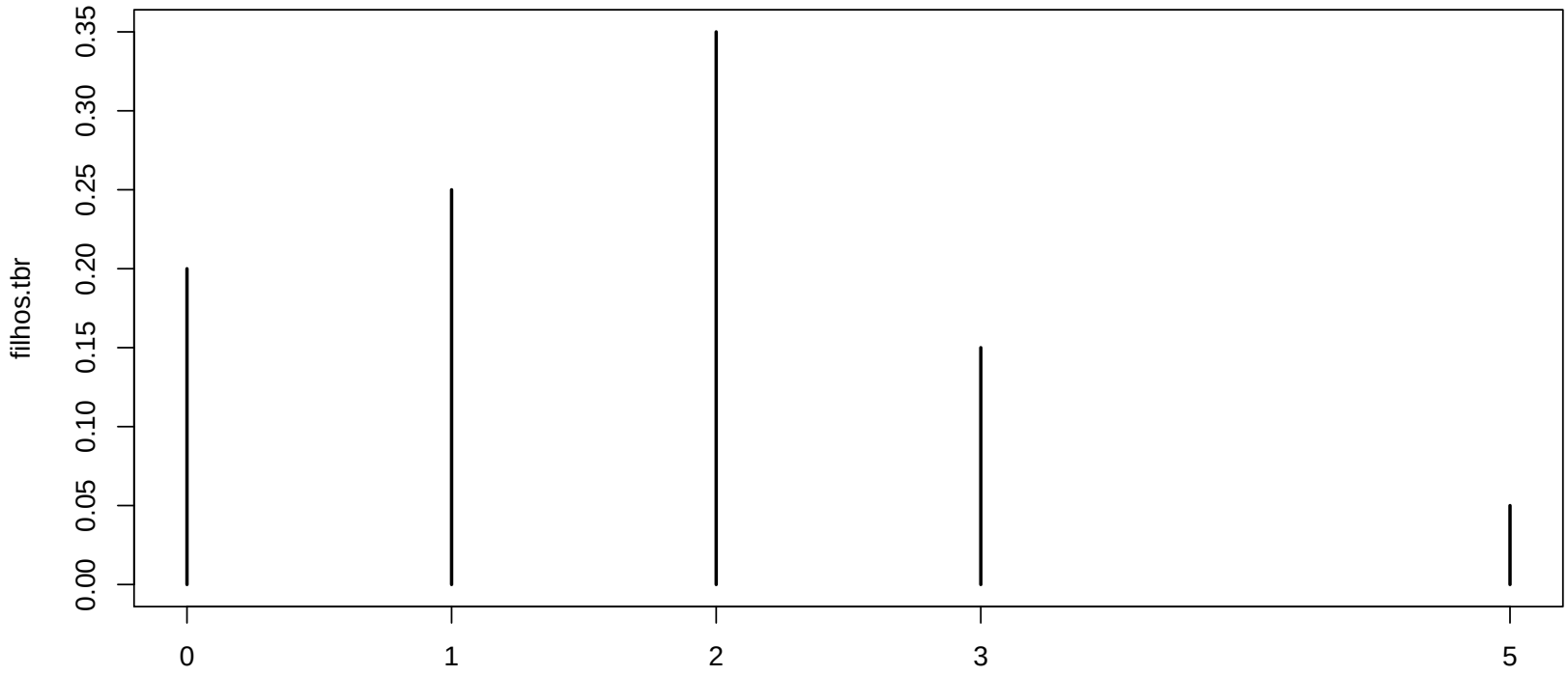
```
plot(filhos.tb)
```



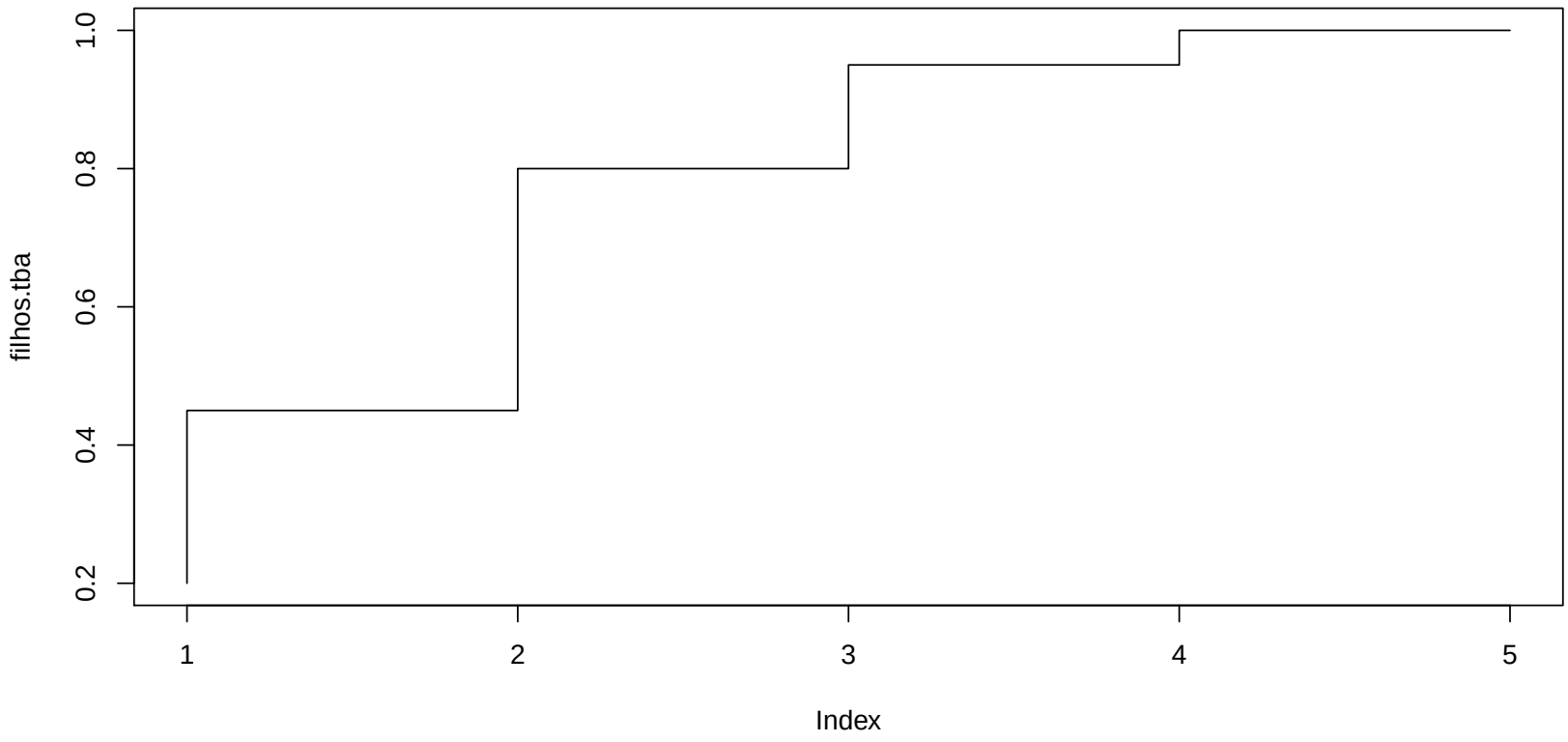
Variável quantitativa discreta

Outra possibilidade seria fazer gráficos de frequências relativas e de frequências acumuladas conforme mostrado na

```
## Frequência relativa
plot(filhos.tbr)
```



```r
## Frequência relativa acumulada
plot(filhos.tba, type = "S") # tipo step (escada)
```



Medidas resumo

A seguir mostramos como obter algumas medidas de posição: moda, mediana, média. Note que o argumento na.rm = TRUE é necessário porque não há informação sobre número de filhos para alguns indivíduos (NA)

```
## Moda
names(filhos.tb)[which.max(filhos.tb)]
```

```
## [1] "2"
```

```
## Mediana
median(milsa$Filhos, na.rm = TRUE)
```

```
## [1] 2
```

```
## Média
mean(milsa$Filhos, na.rm = TRUE)
```

```
## [1] 1.65
```

Pode-se calcular a média aparada, na qual usamos o argumento trim $= 0.1$ que indica que a média deve ser calculada excluindo-se 10% dos menores e 10% dos maiores valores do vetor de dados.

```
## Média aparada
mean(milsa$Filhos, trim = 0.1, na.rm = TRUE)
```

```
## [1] 1.5625
```

- Quartis

```
## Quartis
quantile(milsa$Filhos, na.rm = TRUE)
```

```
##   0%  25%  50%  75% 100%
##    0    1    2    2    5
```

Passando agora para medidas de dispersão, vejamos como obter máximo e mínimo, e com isso a amplitude.

```
## Máximo e mínimo
max(milsa$Filhos, na.rm = TRUE)
```

```
## [1] 5
```

```
min(milsa$Filhos, na.rm = TRUE)
```

```
## [1] 0
```

```
## As duas informações juntas
range(milsa$Filhos, na.rm = TRUE)
```

```
## [1] 0 5
```

```
## Amplitude é a diferença entre máximo e mínimo
diff(range(milsa$Filhos, na.rm = TRUE))
```

```
## [1] 5
```

A variância, desvio padrão, e coeficiente de variação.

```
## Variância
var(milsa$Filhos, na.rm = TRUE)
```

```
## [1] 1.607895
```

```
## Desvio-padrão
sd(milsa$Filhos, na.rm = TRUE)
```

```
## [1] 1.268028
```

```
## Coeficiente de variação
sd(milsa$Filhos, na.rm = TRUE)/
  + mean(milsa$Filhos, na.rm = TRUE)
```

```
## [1] 0.7685018
```

Também obtemos os quartis para calcular a amplitude interquartílica.

```
## Quartis
(filhos.qt <- quantile(milsa$Filhos, na.rm = TRUE))
```

```
##   0%  25%  50%  75% 100%
##    0    1    2    2    5
```

```
## Amplitude interquartílica
filhos.qt[4] - filhos.qt[2]
```

```
## 75%
##   1
```

Finalmente, podemos usar a função genérica *summary()* para resumir od dados de uma só vez

```
summary(milsa$Filhos)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    1.00    2.00    1.65    2.00    5.00      16
```

Variável quantitativa contínua

Vamos considerar a variável quantitativa contínua Salario.

- Frequência

Para se fazer uma tabela de frequências de uma VA contínua, é preciso primeiro agrupar os dados em classes. Nos comandos mostrados a seguir verificamos inicialmente os valores máximo e mínimo dos dados, depois usamos o critério de Sturges para definir o número de classes. Usamos a função *cut()* para agrupar os dados em classes e finalmente obtemos as frequências absolutas e relativas.

```r
## Máximo e mínimo
range(milsa$Salario)
```

```
## [1]  4.0 23.3
```

```r
## Número de classes estimado, com base
## no critério de Sturges.
## outras opções em ?nclass
nclass.Sturges(milsa$Salario)
```

```
## [1] 7
```

```r
## Criando as classes com a função cut(),
## usando os valores mínimos e
## máximos dados em range()
salario.cut <- cut(milsa$Salario, breaks =
                     + seq(4, 23.3, length.out = 8))
```

```r
## Tabela com as frequencias absolutas por classe
salario.tb <- table(salario.cut)
salario.tb
```

```
## salario.cut
##    (4,6.76] (6.76,9.51] (9.51,12.3]   (12.3,15]   (15,17
##           5          10           7           6
## (20.5,23.3]
##           1
```

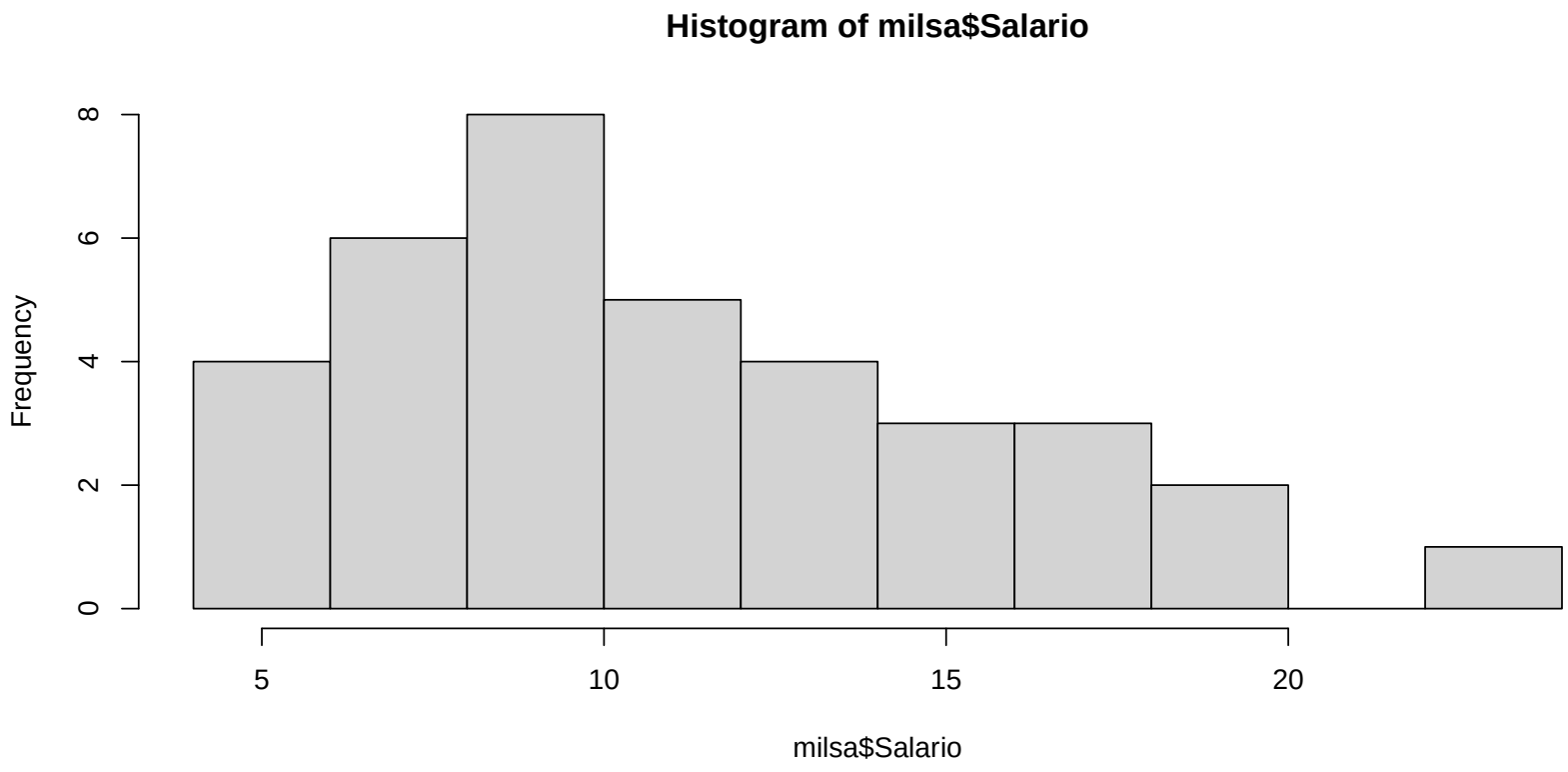
```r
## Tabela com as frequências relativas
prop.table(salario.tb)
```

```
## salario.cut
##    (4,6.76] (6.76,9.51] (9.51,12.3]   (12.3,15]   (15,17
##  0.14285714  0.28571429  0.20000000  0.17142857  0.11428
## (20.5,23.3]
##  0.02857143
```

Variável quantitativa contínua

Na sequência vamos mostrar dois possíveis gráficos para variáveis contínuas: o histograma e o box-plot.

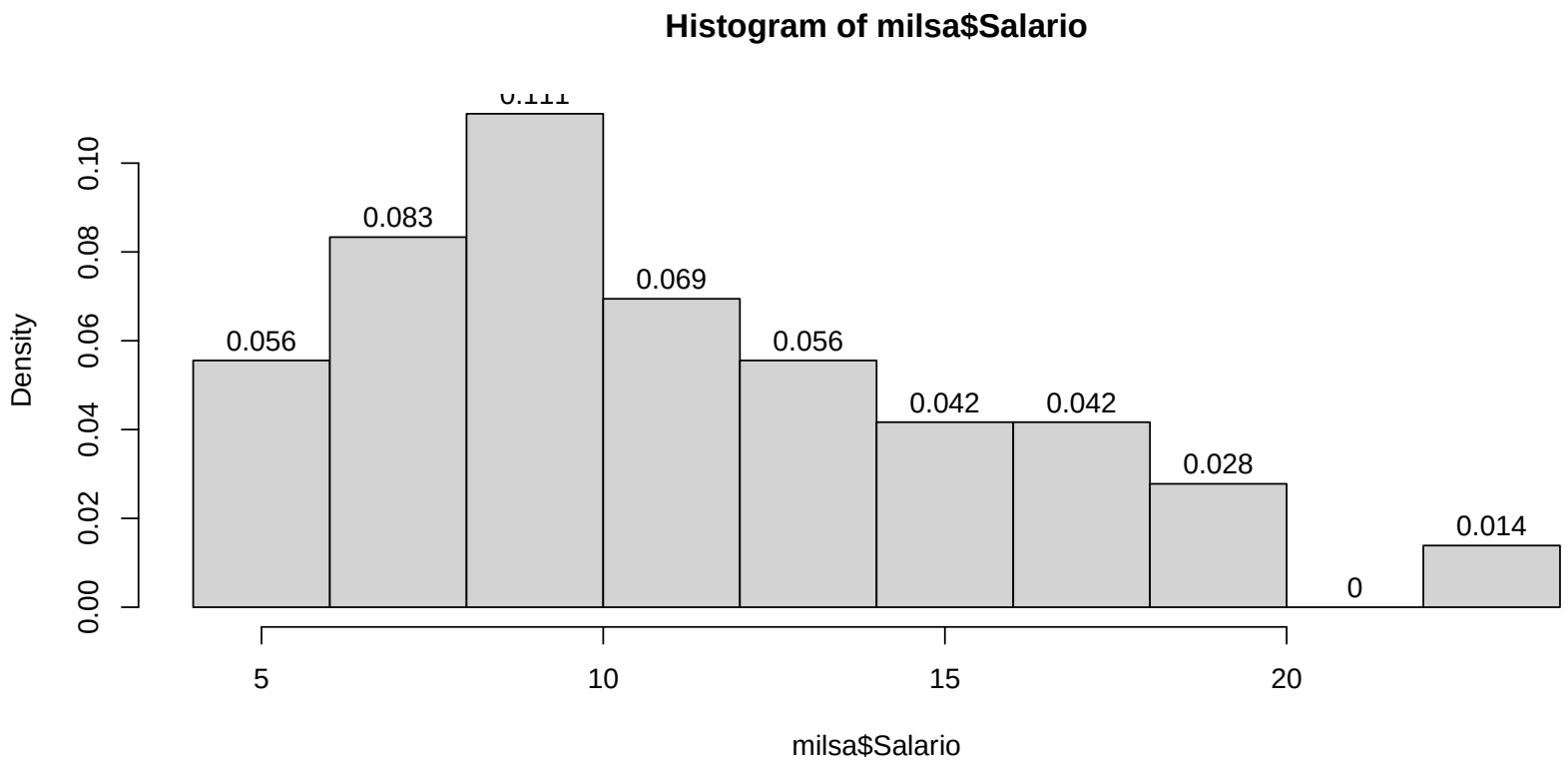
```
hist(milsa$Salario)
```



Variável quantitativa contínua

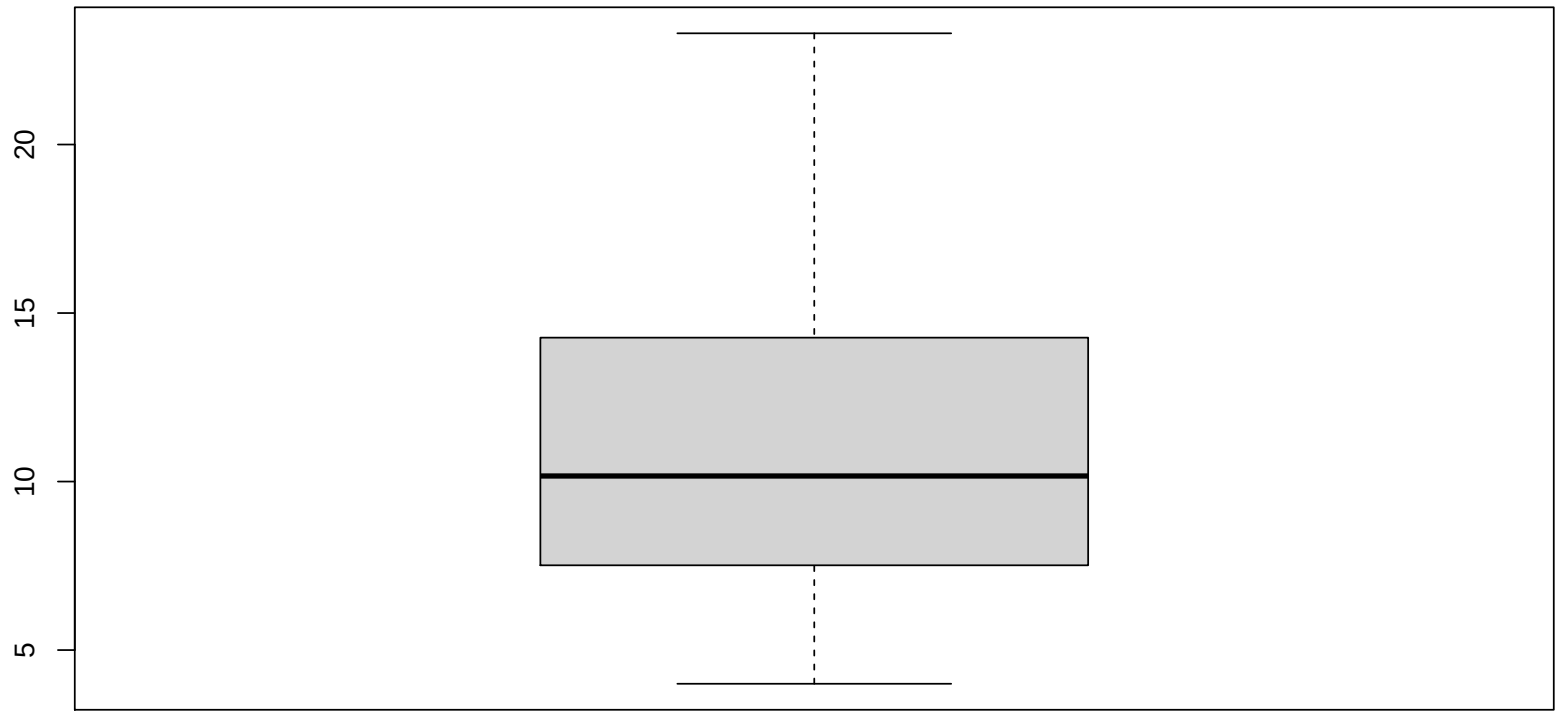
A função hist() possui vários argumentos para alterar o comportamento da saída do gráfico. Por exemplo, com labels = TRUE as frequências são mostradas acima de cada barra. Com freq = FALSE, o gráfico é feito com as frequências relativas.

```
hist(milsa$Salario, freq = FALSE, labels = TRUE)
```



Os boxplots são úteis para revelar o centro, a dispersão e a distribuição dos dados, além de outliers. São construídos da seguinte forma:

```
boxplot(milsa$Salario)
```



Finalmente, podemos obter as medidas de posição e dispersão da mesma forma que para variáveis discretas. Veja alguns exemplos a seguir.

```
summary(milsa$Salario)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.000   7.553  10.165  11.122  14.060  23.300
```

```
var(milsa$Salario)
```

```
## [1] 21.04477
```

```
sd(milsa$Salario)
```

```
## [1] 4.587458
```

```
sd(milsa$Salario)/mean(milsa$Salario)
```

```
## [1] 0.4124587
```