ME111 - Laboratório de Estatística

Aula 2 - Análise de Dados

Profa. Larissa Avila Matos

Análise de Dados

- Na aula de hoje iremos ver alguns comandos do R para fazer análise descritiva de um conjunto de dados.
- Uma boa forma de iniciar uma análise descritiva adequada é verificar os tipo de variáveis disponíveis no conjunto de dados. Variáveis podem ser classificadas da seguinte forma: qualitativas e quantitativas.
- Qualitativas: São aquelas que se baseiam em qualidades e não podem ser mensuradas numericamente. Uma variável é qualitativa quando seus possíveis valores são categorias.
  - Nominais: Não existe ordenaçãao nas possíveis respostas (ex: sexo, estado civil);
  - Ordinais: Existe uma certa ordem nas possíveis respostas (ex: escolaridade).

- Quantitativas: São aquelas que são numericamente mensuráveis, ou seja, que seus possíveis valores podem ser numéricos ou de contagem.
  - Discretas: Os possíveis valores formam um conjunto finito ou enumerável de números, são variáveis de contagem (ex: número de filhos),
  - Contínuas: Os possíveis valores estão dentro de um intervalo, aberto ou fechado, dos números reais (ex: peso de um indivíduo).
- Para cada tipo de variável existem técnicas apropriadas para resumir as informações. Vamos ver que técnicas usadas num caso podem ser adaptadas para outros.
- Essas variáveis podem ser resumidas por tabelas, gráficos e/ou medidas.

O conjunto de dados `SleepStudy`

- O pacote `Lock5Data` do `R` contém o conjunto de dados `SleepStudy`. Esse conjunto de dados se refere a um estudo de padrões de sono para estudantes universitários.
- Os dados foram obtidos de uma amostra de alunos que fizeram testes de habilidades para medir a função cognitiva. Todos os alumos na pesquisa registraram o tempo e a qualidade do sono em um diário do sono durante um período de duas semanas.
- Nesse conjunto de dados encontamos todos os tipos de variáveis.
- Objetivo:
  - entrar com os dados; e
  - fazer uma análise descritiva.

```r
library(Lock5Data)
data(SleepStudy)
attach(SleepStudy)
names(SleepStudy)
```

```
 [1] "Gender"           "ClassYear"        "LarkOwl"
 [4] "NumEarlyClass"    "EarlyClass"       "GPA"
 [7] "ClassesMissed"    "CognitionZscore"  "PoorSleepQuality"
[10] "DepressionScore"  "AnxietyScore"     "StressScore"
[13] "DepressionStatus" "AnxietyStatus"    "Stress"
[16] "DASScore"         "Happiness"        "AlcoholUse"
[19] "Drinks"           "WeekdayBed"       "WeekdayRise"
[22] "WeekdaySleep"     "WeekendBed"       "WeekendRise"
[25] "WeekendSleep"     "AverageSleep"     "AllNighter"
```

```r
dim(SleepStudy)
```

```
[1] 253  27
```

```r
str(SleepStudy)
```

```
'data.frame':   253 obs. of  27 variables:
 $ Gender          : int  0 0 0 0 0 1 1 0 0 0 ...
 $ ClassYear       : int  4 4 4 1 4 4 2 2 1 4 ...
 $ LarkOwl         : Factor w/ 3 levels "Lark","Neither",..: 2 2 3 1 3 2 1 1 2 2 ...
 $ NumEarlyClass   : int  0 2 0 5 0 0 2 0 2 2 ...
 $ EarlyClass      : int  0 1 0 1 0 0 1 0 1 1 ...
 $ GPA             : num  3.6 3.24 2.97 3.76 3.2 3.5 3.35 3 4 2.9 ...
 $ ClassesMissed   : int  0 0 12 0 4 0 2 0 0 0 ...
 $ CognitionZscore : num  -0.26 1.39 0.38 1.39 1.22 -0.04 0.41 -0.59 1.03 0.72 ...
 $ PoorSleepQuality: int  4 6 18 9 9 6 2 10 5 2 ...
 $ DepressionScore : int  4 1 18 1 7 14 1 2 12 6 ...
 $ AnxietyScore    : int  3 0 18 4 25 8 0 2 16 11 ...
 $ StressScore     : int  8 3 9 6 14 28 1 3 20 31 ...
 $ DepressionStatus: Factor w/ 3 levels "moderate","normal",..: 2 2 1 2 2 1 2 2 1 2 ...
 $ AnxietyStatus   : Factor w/ 3 levels "moderate","normal",..: 2 2 3 2 3 1 2 2 3 1 ...
 $ Stress          : Factor w/ 2 levels "high","normal": 2 2 2 2 2 1 2 2 1 1 ...
 $ DASScore        : int  15 4 45 11 46 50 2 7 48 48 ...
 $ Happiness       : int  28 25 17 32 15 22 25 29 29 30 ...
 $ AlcoholUse      : Factor w/ 4 levels "Abstain","Heavy",..: 4 4 3 3 4 1 4 3 3 4 ...
 $ Drinks          : int  10 6 3 2 4 0 6 3 3 6 ...
 $ WeekdayBed      : num  25.8 25.7 27.4 23.5 25.9 ...
 $ WeekdayRise     : num  8.7 8.2 6.55 7.17 8.67 8.95 8.48 9.07 8.75 8 ...
 $ WeekdaySleep    : num  7.7 6.8 3 6.77 6.09 9.05 7.73 9.02 8.25 6.6 ...
 $ WeekendBed      : num  25.8 26 28 27 23.8 ...
 $ WeekendRise     : num  9.5 10 12.6 8 9.5 ...
 $ WeekendSleep    : num  5.88 7.25 10.09 7.25 7 ...
 $ AverageSleep    : num  7.18 6.93 5.02 6.9 6.35 9.04 7.52 9.01 8.54 6.68 ...
 $ AllNighter      : int  0 0 0 0 0 0 1 0 0 0 ...
```

Análise Univariada

- A análise univariada consiste basicamente em, para cada uma das variáveis individualmente:
  - classificar a variável quanto a seu tipo: qualitativa (nominal ou ordinal) ou quantitativa (discreta ou contínua)
  - obter tabela, gráfico e/ou medidas que resumam a variável
- A partir destes resultados pode-se montar um resumo geral dos dados.

- A seguir vamos mostrar como obter tabelas, gráficos e medidas com o R. Para isto, selecionaremos do conjunto de dados `SleepStudy` uma variável de cada tipo.

```r
# Transformando a variável Gender em fator
dados<-SleepStudy
dados$Gender[dados$Gender==1]<-"Male"
dados$Gender[dados$Gender==0]<-"Female"
dados$Gender<-as.factor(dados$Gender)
# Transformando a variável AnxietyStatus em fator ordenado,
# ou seja, em variável qualitativa ordinal
dados$AnxietyStatus<-ordered(dados$AnxietyStatus)
```

- As varáveis de interesse serão:
  - `Gender`: Sexo (Homem ou Mulher) - Qualitativa Nominal;
  - `AnxietyStatus`: Pontuação de ansiedade codificada em mormal, moderada ou grave - Qualitativa Ordinal;
  - `Drinks`: Número de bebidas alcoólicas por semana - Quantitativa discreta;
  - `AverageSleep`: Média de horas de sono para todos os dias - Quantitativa Contínua.

Variável Qualitativa Nominal

- A variável **Gender** é uma variável qualitativa nominal. Para fazer uma análise descritiva para esse tipo de variável podemos resumir os dados através de uma tabela de frequências e/ou um gráfico de setores/barras.
- Vamos primeiro listar os dados e checar se estão na forma de um fator, que é adequada para variáveis deste tipo.

```
is.factor(dados$Gender)
```

```
[1] TRUE
```

```
dados$Gender[1:8]
```

```
[1] Female Female Female Female Female Male   Male   Female
Levels: Female Male
```

- Para resumir numericamente os dados de uma variável precisamos olhar para todos os valores possíveis e contar quantas vezes cada um aparece, podemos fazer isso no `R` com o comando `table()`.

```
Gender.table <- table(dados$Gender) # Frequência absoluta
Gender.table
```

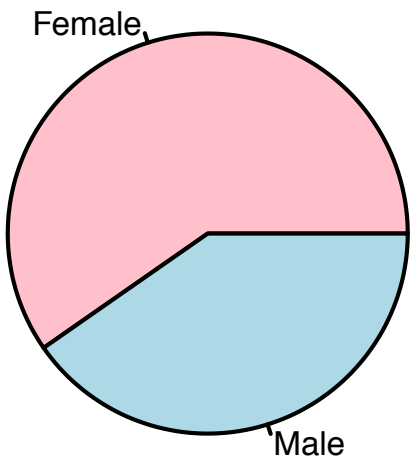
```
Female   Male
   151    102
```

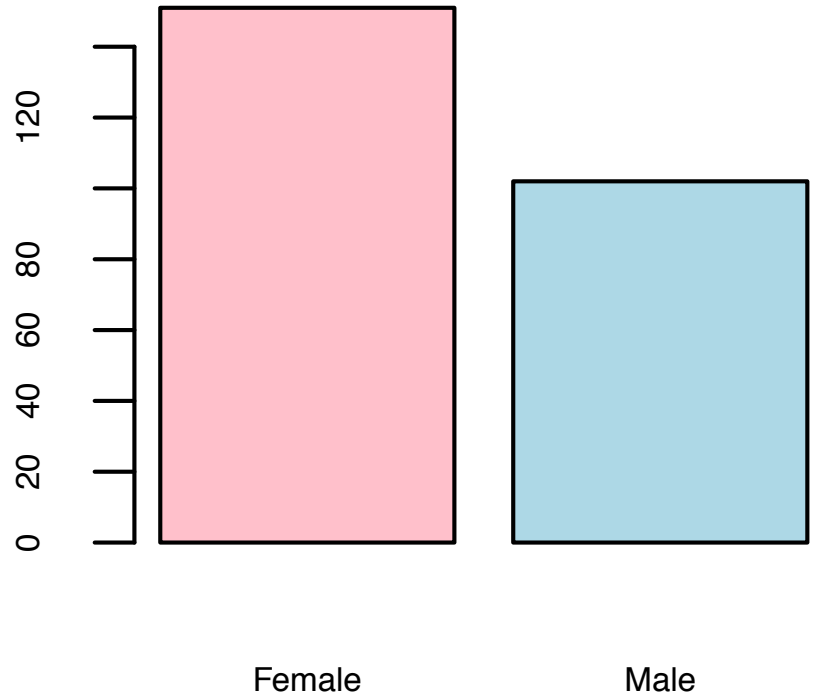
```
prop.table(Gender.table) # Frequência relativa
```

```
   Female      Male
0.5968379 0.4031621
```

- Podemos representar graficamente esta variável através de um gráfico de setores ou de barras.

```
par(mfrow=c(1,2))
pie(table(dados$Gender),cex=0.5,col=c("pink","lightblue"))
barplot(table(dados$Gender),cex.axis = 0.7,cex.names = 0.7,
        col=c("pink","lightblue"))
```





Variável Qualitativa Ordinal

- A variável `AnxietyStatus` é uma variável qualitativa ordinal.

```
is.factor(dados$AnxietyStatus)
```

```
[1] TRUE
```

```
dados$AnxietyStatus[1:5]
```

```
[1] normal normal severe normal severe
Levels: moderate < normal < severe
```

- Assim como na variável qualitativa nominal podemos utilizar as frequências absolutas e relativas para resumir os dados.

```
Anxiety.table <- table(dados$AnxietyStatus)
Anxiety.table
```

```
moderate   normal   severe
      56      181       16
```

```
prop.table(Anxiety.table)
```

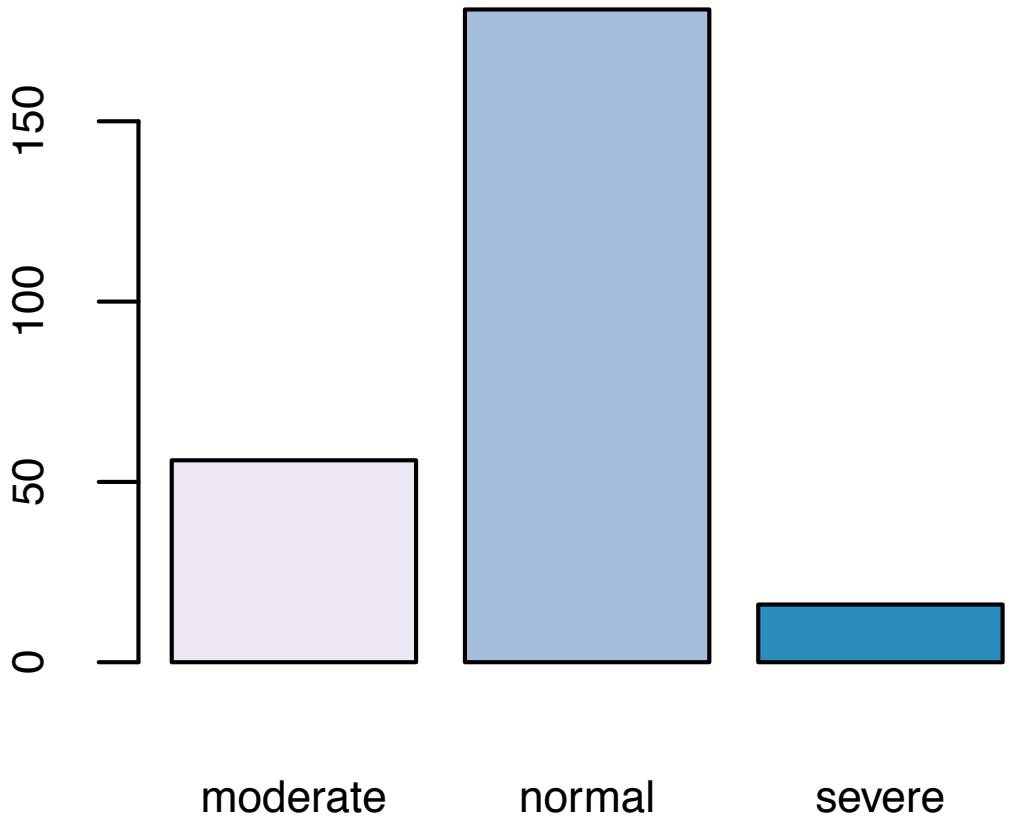
```
  moderate     normal     severe
0.22134387 0.71541502 0.06324111
```

```
100*table(dados$AnxietyStatus)/length(dados$AnxietyStatus)
```

```
 moderate    normal    severe
22.134387 71.541502  6.324111
```

- Para esse tipo de variável podemos usar o gráfico de barras.

```
barplot(Anxiety.table,cex.axis = 0.7,cex.names = 0.7,
        col = c("#ece7f2","#a6bddb","#2b8cbe"))
```



Variável Quantitativa Discreta

- Como vimos, a variável **Drinks** (número de bebidas alcoólicas por semana) é uma variável quantitativa discreta. Note que esta deve ser uma variável numérica, e não um fator.

```
str(dados$Drinks)
```

```
 int [1:253] 10 6 3 2 4 0 6 3 3 6 ...
```

```
is.numeric(dados$Drinks)
```

```
[1] TRUE
```

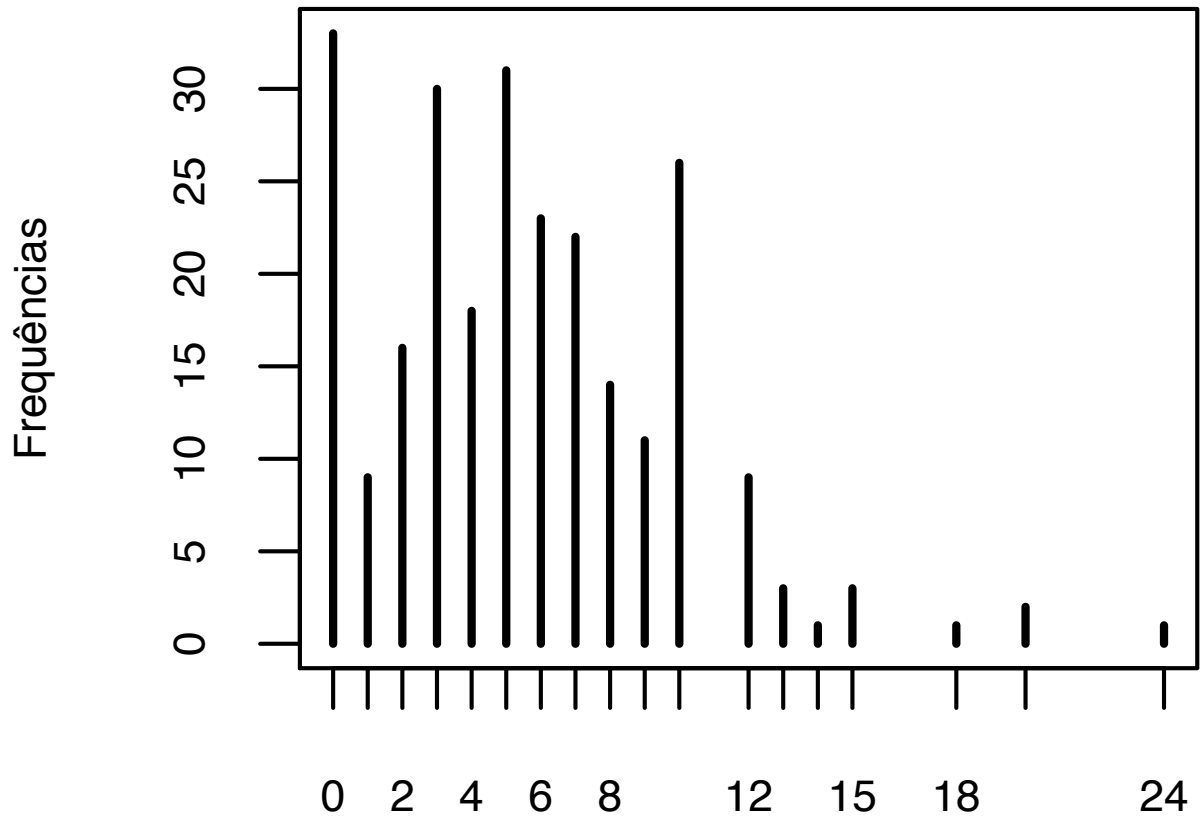
- Como nas variáveis qualitativas podemos obter as frequências absolutas e relativas.

```
Drinks.table <- table(dados$Drinks)
Drinks.table
```

```
 0  1  2  3  4  5  6  7  8  9 10 12 13 14 15 18 20 24
33  9 16 30 18 31 23 22 14 11 26  9  3  1  3  1  2  1
```

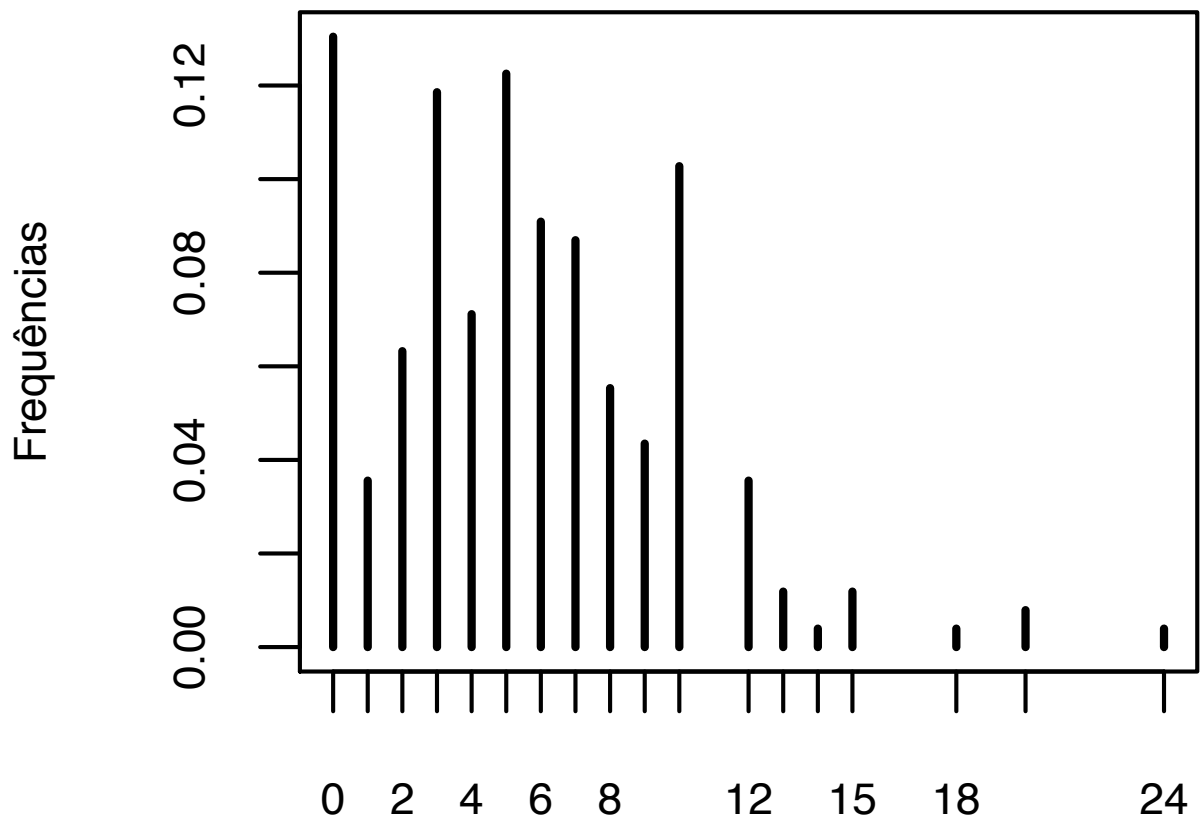
- O gráfico adequado para frequências absolutas de uma variável discreta é dado com o comando `plot()`.

```
plot(Drinks.table,cex.axis=0.7,cex.lab=0.7,ylab="Frequências")
```

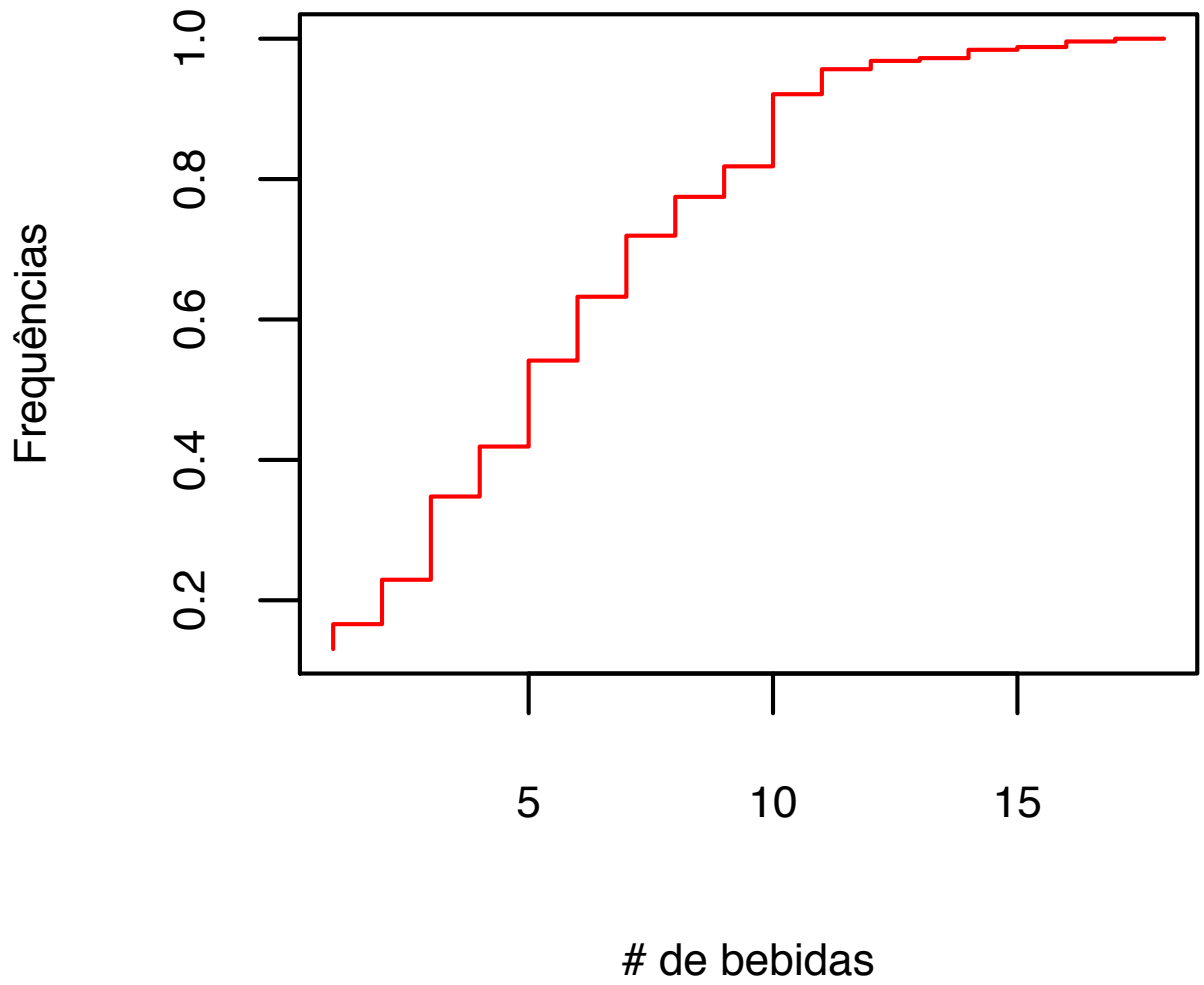


- Outra possibilidade seria fazer gráficos de frequências relativas e de frequências acumuladas.

```
Drinks.FreqR <- prop.table(Drinks.table)
plot(Drinks.FreqR,cex.axis=0.7,cex.lab=0.7,ylab="Frequências")
```



```r
Drinks.fac <- cumsum(Drinks.FreqR) # frequências acumuladas
plot(Drinks.fac,type="S",cex.axis=0.7,cex.lab=0.7, col="red",
     ylab="Frequências", xlab="# de bebidas")
```



- Sendo a variável `Drinks` numérica há uma maior diversidade de medidas estatísticas que podem ser calculadas.
- Podemos calcular algumas das chamadas medidas de posição, bem como, algumas medidas de dispersão, consideradas mais importantes no campo da aplicabilidade prática do nosso dia a dia.
- Tais medidas servem para
  - Localizar uma distribuição;
  - Caracterizar sua variabilidade.

- Medidas de Posição: Servem para localizar a distribuição dos dados brutos (ou das freqüências) sobre o eixo de variação da variável em questão. As principais de medidas de posição são: média aritmética, mediana e moda.

```
Drinks.media <- mean(dados$Drinks)
Drinks.media # média aritmética
```

```
[1] 5.56917
```

```
Drinks.mediana <- median(dados$Drinks)
Drinks.mediana # mediana
```

```
[1] 5
```

```
Drinks.moda <- names(Drinks.table)[which.max(Drinks.table)]
Drinks.moda # moda
```

```
[1] "0"
```

- Medidas de Dispersão: A informação fornecida pelas Medidas de Posição em geral necessitam de ser complementas pelas Medidas de Dispersão. As Medidas de Dispersão servem para indicar o "quanto os dados se apresentam dispersos em torno da região central". Portanto caracterizam o grau de variação existente em um conjunto de dados. As principais de medidas de dispersão são: amplitude, variância, desvio padrão e coeficiente de variação.

```
range(dados$Drinks)
```

```
[1]  0 24
```

```
Drinks.amplitude <- diff(range(dados$Drinks))
Drinks.amplitude # amplitude
```

```
[1] 24
```

```
max(dados$Drinks)-min(dados$Drinks) # amplitude
```

```
[1] 24
```

```r
var(dados$Drinks) # variância
```

```
[1] 16.77
```

```r
Drinks.dp <- sd(dados$Drinks)
Drinks.dp # desvio padrão
```

```
[1] 4.095119
```

```r
Drinks.cv <- 100 * Drinks.dp/Drinks.media
Drinks.cv # coeficiente de variação
```

```
[1] 73.53194
```

```r
Drinks.qt <- quantile(dados$Drinks)
Drinks.qt # quartis
```

```
  0%  25%  50%  75% 100%
   0    3    5    8   24
```

```r
Drinks.ai <- Drinks.qt[4] - Drinks.qt[2]
Drinks.ai # intervalo interquartil
```

```
75%
  5
```

Variável Quantitativa Contínua

- Neste caso, a variável **AverageSleep** é uma variável quantitativa contínua. Para esse tipo de variável podemos usar todas as medidas calculadas para a variável discreta. Além disso, podemos criar diferentes gráficos.

```
dados$AverageSleep[1:10]
```

```
 [1] 7.18 6.93 5.02 6.90 6.35 9.04 7.52 9.01 8.54 6.68
```

```
is.factor(dados$AverageS)
```

```
[1] FALSE
```

```
is.numeric(dados$AverageS)
```

```
[1] TRUE
```

- Para fazer a tabela de frequências de uma variável 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 cut() para agrupar os dados em classes e finalmente obtemos as frequências absolotas e relativas.

```
range(dados$AverageSleep) # variação (mínimo e máximo)
```

```
[1]  4.95 10.62
```

```
# Calcula o número de classes para um histograma.
nclass.Sturges(dados$AverageSleep)
```

```
[1] 9
```

```
k=1+3.32*log10(dim(dados)[1]) # fórmula de Sturges
round(k,0)
```

```
[1] 9
```

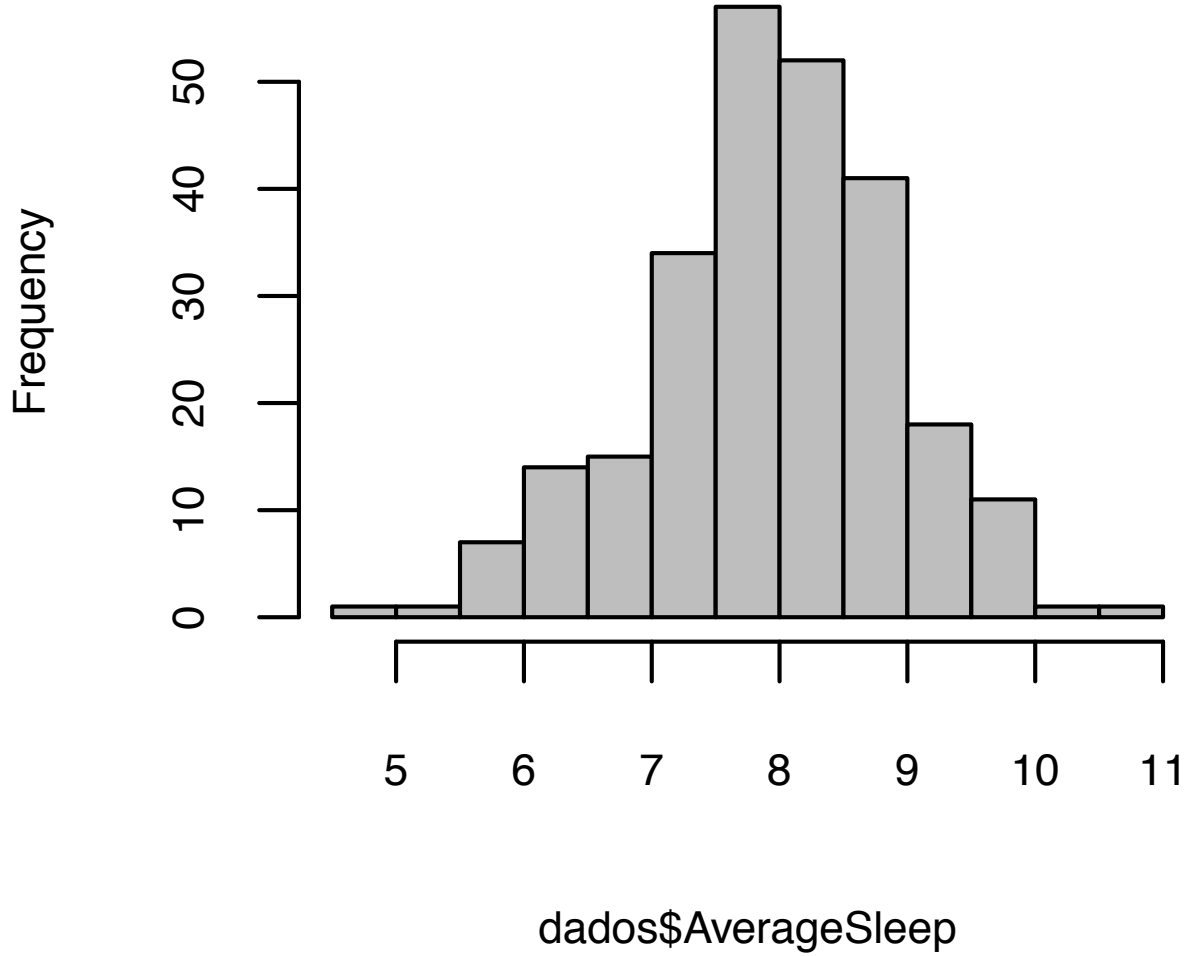
```
AverageSleep.table <- table(cut(dados$AverageSleep,
                                seq(4.9,10.7,l=10),dig.lab=4))
AverageSleep.table
```

```
  (4.9,5.544] (5.544,6.189] (6.189,6.833] (6.833,7.478] (7.478,8.122]
            2             11             17             40             66
(8.122,8.767] (8.767,9.411] (9.411,10.06]  (10.06,10.7]
           68             36             11              2
```

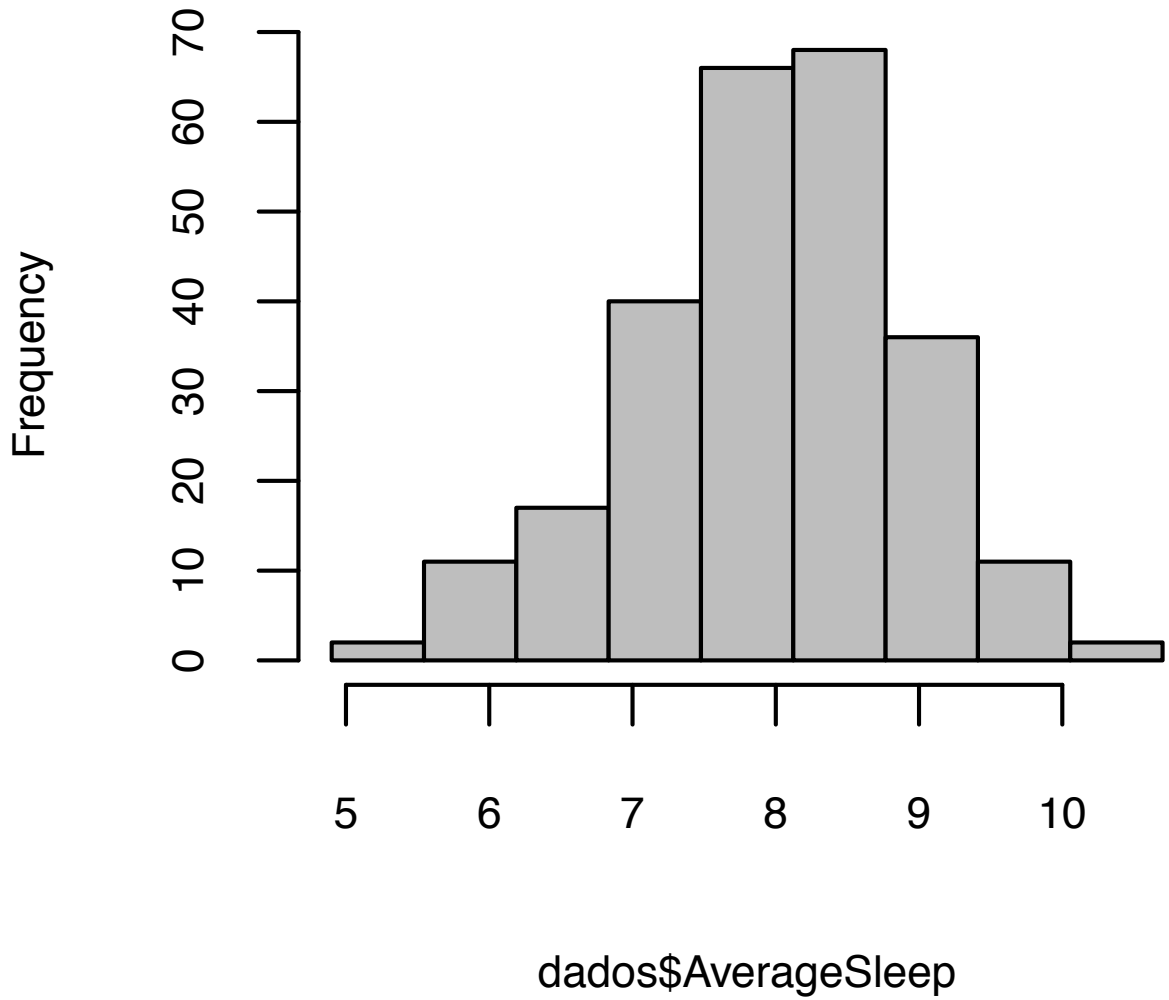
```
sum(AverageSleep.table)
```

```
[1] 253
```

```r
hist(dados$AverageSleep,main="",cex.axis=0.7,cex.lab=0.7,
     col="gray") # histograma
```



```r
hist(dados$AverageSleep,main="",cex.axis=0.7,cex.lab=0.7,
     breaks=seq(4.9,10.7,l=10),col="gray") # histograma
```



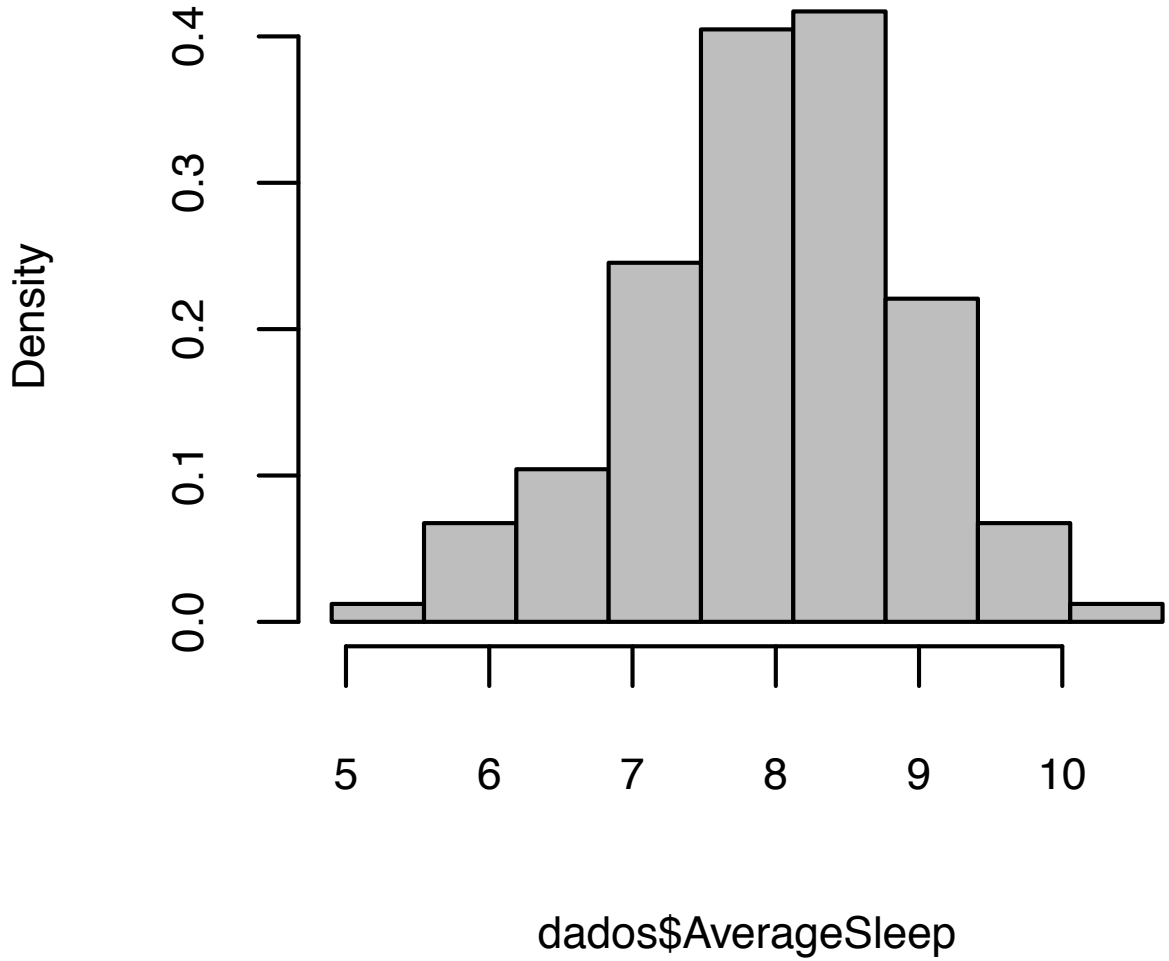
```
round(prop.table(AverageSleep.table),3)
```

```
 (4.8,5.47] (5.47,6.13]  (6.13,6.8]  (6.8,7.47] (7.47,8.13]  (8.13,8.8]
      0.008       0.036       0.071       0.162       0.261       0.277
 (8.8,9.47] (9.47,10.1] (10.1,10.8]
      0.134       0.043       0.008
```

```
sum(round(prop.table(AverageSleep.table),3))
```

```
[1] 1
```

```r
hist(dados$AverageSleep,main="",cex.axis=0.7,cex.lab=0.7,
     breaks=seq(4.9,10.7,l=10),col="gray",
     probability=T) # histograma
```



```
round(sort(dados$AverageSleep),1)
```

```
  [1]  5.0  5.0  5.6  5.8  5.8  5.9  6.0  6.0  6.0  6.1  6.1  6.1  6.1  6.2
 [15]  6.3  6.3  6.4  6.4  6.4  6.4  6.4  6.5  6.5  6.5  6.5  6.6  6.7  6.7
 [29]  6.8  6.8  6.9  6.9  6.9  6.9  6.9  7.0  7.0  7.0  7.0  7.0  7.0  7.0
 [43]  7.1  7.1  7.1  7.1  7.2  7.2  7.2  7.2  7.2  7.2  7.2  7.2  7.3  7.3
 [57]  7.3  7.3  7.3  7.3  7.4  7.4  7.4  7.4  7.4  7.4  7.4  7.4  7.5  7.5
 [71]  7.5  7.5  7.5  7.5  7.5  7.5  7.5  7.6  7.6  7.6  7.6  7.6  7.6  7.6
 [85]  7.6  7.6  7.6  7.6  7.7  7.7  7.7  7.7  7.7  7.7  7.8  7.8  7.8  7.8
 [99]  7.8  7.8  7.8  7.8  7.8  7.8  7.9  7.9  7.9  7.9  7.9  7.9  7.9  7.9
[113]  7.9  7.9  7.9  7.9  7.9  7.9  8.0  8.0  8.0  8.0  8.0  8.0  8.0  8.0
[127]  8.0  8.0  8.0  8.0  8.1  8.1  8.1  8.1  8.1  8.1  8.2  8.2  8.2  8.2
[141]  8.2  8.2  8.2  8.2  8.2  8.2  8.2  8.2  8.2  8.2  8.2  8.2  8.2  8.3
[155]  8.3  8.3  8.3  8.3  8.3  8.3  8.3  8.3  8.3  8.3  8.4  8.4  8.4  8.4
[169]  8.4  8.4  8.4  8.4  8.4  8.4  8.4  8.4  8.4  8.5  8.5  8.5  8.5  8.5
[183]  8.5  8.5  8.5  8.6  8.6  8.6  8.6  8.6  8.6  8.6  8.6  8.6  8.7  8.7
[197]  8.7  8.7  8.7  8.7  8.8  8.8  8.8  8.8  8.8  8.8  8.8  8.8  8.8  8.9
[211]  8.9  8.9  8.9  8.9  8.9  8.9  8.9  8.9  8.9  8.9  9.0  9.0  9.0  9.0
[225]  9.1  9.1  9.1  9.1  9.1  9.1  9.1  9.2  9.2  9.2  9.2  9.2  9.2  9.3
[239]  9.3  9.4  9.6  9.6  9.6  9.6  9.7  9.8  9.8  9.9  9.9 10.0 10.0 10.4
[253] 10.6
```

```r
stem(dados$AverageSleep) # ramo e folhas
```

```
  The decimal point is at the |

   4 |
   5 | 00
   5 | 6889
   6 | 000111123444444
   6 | 55556778899999
   7 | 000011111112222223333333344444444
   7 | 55555555666666666666777777888888889999999999999999
   8 | 000000000000111111222222222222222233333333333344444444444444
   8 | 5555555566666666677777788888888899999999999
   9 | 00001111111222223334
   9 | 666678999
  10 | 00
  10 | 56
```

```r
AverageSleep.media <- mean(dados$AverageSleep)
AverageSleep.media # média aritmética
```

```
[1] 7.965929
```

```r
AverageSleep.mediana <- median(dados$AverageSleep)
AverageSleep.mediana # mediana
```

```
[1] 8
```

```r
var(dados$AverageSleep) # variância
```

```
[1] 0.9309155
```

```r
AverageSleep.dp <- sd(dados$AverageSleep)
AverageSleep.dp # desvio padrão
```

```
[1] 0.9648396
```

```r
AverageSleep.amplitude <- diff(range(dados$AverageSleep))
AverageSleep.amplitude # amplitude
```

```
[1] 5.67
```

```r
AverageSleep.cv <- 100 * AverageSleep.dp/AverageSleep.media
AverageSleep.cv # coeficiente de variação
```

```
[1] 12.11208
```

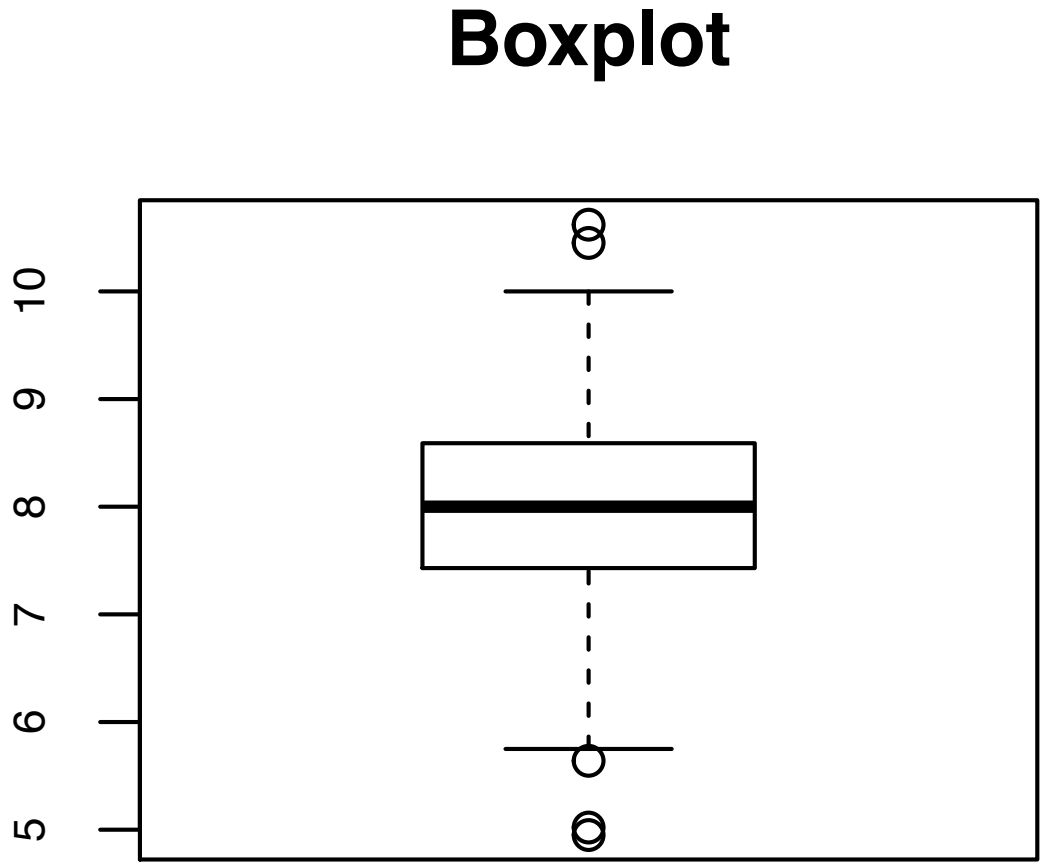
```r
AverageSleep.qt <- quantile(dados$AverageSleep)
AverageSleep.qt # quartis
```

```
   0%   25%   50%   75%  100%
 4.95  7.43  8.00  8.59 10.62
```

```r
AverageSleep.ai <- AverageSleep.qt[4] - AverageSleep.qt[2]
AverageSleep.ai # intervalo interquartil
```

```
 75%
1.16
```

```r
boxplot(dados$AverageSleep,cex.axis=0.7,cex.lab=0.7,
        main="Boxplot") # boxplot
```



```r
summary(dados$AverageSleep) # resumo dos dodos
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.950   7.430   8.000   7.966   8.590  10.620
```

```r
fivenum(dados$AverageSleep) # esquema do 5 números
```

```
[1]  4.95  7.43  8.00  8.59 10.62
```

Referências

- Bussab, W. O. & Morettin, P. A. (1987). *Estatística Básica. Atual Editora Ltda., São Paulo.*
- *http://www.leg.ufpr.br/~paulojus/embrapa/Rembrapa/Rembrapase8.html*
- *Magalhães, Marcos N.; Lima, Antonio Carlos P. (2010). Noções de probabilidade e estatística. São Paulo: Edusp, 2010.*
- *Wardrop, R. L. (1995). Statistics: Learning in the presence of variation.*