Introdução ao R

Anderson Castro Soares de Oliveira

Estatística Descritiva

- Para iniciar uma análise descritiva é importante verificar o tipo de variáveis do conjunto de dados.

Estatística Descritiva

- Para iniciar uma análise descritiva é importante verificar o tipo de variáveis do conjunto de dados.
- `summary` para uma analise preliminar, retorna:

Estatística Descritiva

- Para iniciar uma análise descritiva é importante verificar o tipo de variáveis do conjunto de dados.
- `summary` para uma analise preliminar, retorna:
  - Variáveis qualitativas - frequências absolutas;

Estatística Descritiva

- Para iniciar uma análise descritiva é importante verificar o tipo de variáveis do conjunto de dados.
- `summary` para uma analise preliminar, retorna:
  - Variáveis qualitativas - frequências absolutas;
  - Variáveis quantitativas - menor valor, primeiro quartil, mediana, média, terceiro quartil e maior valor.

Tabelas

- Variáveis Qualitativas -`table`

Tabelas

- Variáveis Qualitativas -`table`
- Variáveis Quantitativas Discretas -`table`

Tabelas

- Variáveis Qualitativas -`table`
- Variáveis Quantitativas Discretas -`table`
- Variáveis Quantitativas Contínuas - `fdt` pacote `fdth`

```
fdt(x, k, start, end, h, breaks=c("Sturges",
 "Scott", "FD"))
```

em que:

Tabelas

- Variáveis Qualitativas -`table`
- Variáveis Quantitativas Discretas -`table`
- Variáveis Quantitativas Contínuas - `fdt` pacote `fdth`

```
fdt(x, k, start, end, h, breaks=c("Sturges",
 "Scott", "FD"))
```

em que:

  - x - são os dados que devem agrupados;

Tabelas

- Variáveis Qualitativas -`table`
- Variáveis Quantitativas Discretas -`table`
- Variáveis Quantitativas Contínuas - `fdt` pacote `fdth`

```
fdt(x, k, start, end, h, breaks=c("Sturges",
 "Scott", "FD"))
```

em que:

  - x - são os dados que devem agrupados;
  - k - numero de intervalos de classe;

Tabelas

- Variáveis Qualitativas -`table`
- Variáveis Quantitativas Discretas -`table`
- Variáveis Quantitativas Contínuas - `fdt` pacote `fdth`

```
fdt(x, k, start, end, h, breaks=c("Sturges",
 "Scott", "FD"))
```

em que:

  - x - são os dados que devem agrupados;
  - k - numero de intervalos de classe;
  - start - limite inferior da primeira classe;

Tabelas

- Variáveis Qualitativas -`table`
- Variáveis Quantitativas Discretas -`table`
- Variáveis Quantitativas Contínuas - `fdt` pacote `fdth`

```
fdt(x, k, start, end, h, breaks=c("Sturges",
 "Scott", "FD"))
```

em que:

  - x - são os dados que devem agrupados;
  - k - numero de intervalos de classe;
  - start - limite inferior da primeira classe;
  - end - limite superior da ultima classe;

Tabelas

- Variáveis Qualitativas -`table`
- Variáveis Quantitativas Discretas -`table`
- Variáveis Quantitativas Contínuas - `fdt` pacote `fdth`

```
fdt(x, k, start, end, h, breaks=c("Sturges",
 "Scott", "FD"))
```

em que:

  - x - são os dados que devem agrupados;
  - k - numero de intervalos de classe;
  - start - limite inferior da primeira classe;
  - end - limite superior da ultima classe;
  - h - amplitude da classe;

Tabelas

- Variáveis Qualitativas -`table`
- Variáveis Quantitativas Discretas -`table`
- Variáveis Quantitativas Contínuas - `fdt` pacote `fdth`

```
fdt(x, k, start, end, h, breaks=c("Sturges",
 "Scott", "FD"))
```

em que:

  - x - são os dados que devem agrupados;
  - k - numero de intervalos de classe;
  - start - limite inferior da primeira classe;
  - end - limite superior da ultima classe;
  - h - amplitude da classe;
  - breaks - caso não seja definido `k`, `start`, `end` e `h`, pode-se utilizar um método para isso.

Tabelas

```
##Distribuição de frequência para variável Grau de Instrução.
> fa=table(GI) ##frequencia absoluta
> fa
GI
fundamental       medio    superior
          7          14           8
> fr=fa/sum(fa) ##frequencia relativa
> fr
GI
fundamental       medio    superior
  0.2413793   0.4827586   0.2758621
> fp=100*fr ##fp=frequencia percentual
> fp
GI
fundamental       medio    superior
   24.13793    48.27586    27.58621
> dist=cbind(fa,fr,fp) ##distribuição de frequências
> dist
             fa        fr       fp
fundamental   7 0.2413793 24.13793
medio        14 0.4827586 48.27586
superior      8 0.2758621 27.58621
```

Tabelas

```
> fa=table(SEXO,AT) ##frequencia absoluta
> fa
    AT
SEXO Nao Sim
   F   9   8
   M   9   3
> fr=fa/sum(fa) ##frequencia relativ
> fr
    AT
SEXO       Nao       Sim
   F 0.3103448 0.2758621
   M 0.3103448 0.1034483
> fp=100*fr ##fp=frequencia percentual
> fp
    AT
SEXO      Nao      Sim
   F 31.03448 27.58621
   M 31.03448 10.34483
```

Tabelas

```
> fa=table(NFD) ##frequencia absoluta
> fa
NFD
 2  3  4  5
10  8  6  5
> fr=fa/sum(fa) ##frequencia relativa
> fr
NFD
        2         3         4         5
0.3448276 0.2758621 0.2068966 0.1724138
> fp=100*fr ##fp=frequencia percentual
> fp
NFD
       2        3        4        5
34.48276 27.58621 20.68966 17.24138
> dist=cbind(fa,fr,fp) ##distribuição de frequências
> dist
  fa        fr       fp
2 10 0.3448276 34.48276
3  8 0.2758621 27.58621
4  6 0.2068966 20.68966
5  5 0.1724138 17.24138
```

Tabelas

```
> require(fdth)
> x=fdt(IDADE)
> x
 Class limits  f   rf rf(%) cf  cf(%)
  [19.8,25.1) 10 0.34 34.48 10  34.48
  [25.1,30.4)  8 0.28 27.59 18  62.07
  [30.4,35.7)  2 0.07  6.90 20  68.97
  [35.7,40.9)  3 0.10 10.34 23  79.31
  [40.9,46.2)  2 0.07  6.90 25  86.21
  [46.2,51.5)  4 0.14 13.79 29 100.00
> x1=fdt(IDADE,k=4)
> x1
 Class limits  f   rf rf(%) cf  cf(%)
  [19.8,27.7) 14 0.48 48.28 14  48.28
  [27.7,35.7)  6 0.21 20.69 20  68.97
  [35.7,43.6)  5 0.17 17.24 25  86.21
  [43.6,51.5)  4 0.14 13.79 29 100.00
```

Tabelas

```
> x2=fdt(IDADE,start=20,end=50,h=5)
> x2
 Class limits f   rf rf(%) cf cf(%)
      [20,25) 9 0.31 31.03  9 31.03
      [25,30) 8 0.28 27.59 17 58.62
      [30,35) 1 0.03  3.45 18 62.07
      [35,40) 5 0.17 17.24 23 79.31
      [40,45) 2 0.07  6.90 25 86.21
      [45,50) 3 0.10 10.34 28 96.55
> x3=fdt(IDADE,breaks="Scott")
> x3
 Class limits  f   rf rf(%) cf  cf(%)
  [19.8,30.4) 18 0.62 62.07 18  62.07
  [30.4,40.9)  5 0.17 17.24 23  79.31
  [40.9,51.5)  6 0.21 20.69 29 100.00
```

Gráficos

- `plot`

Gráficos

- `plot`
  - Variáveis Qualitativas -gráfico de barras

Gráficos

- `plot`
  - Variáveis Qualitativas -gráfico de barras
  - Variáveis Quantitativas - gráfico de dispersão

Gráficos

- `plot`
  - Variáveis Qualitativas -gráfico de barras
  - Variáveis Quantitativas - gráfico de dispersão
- `barplot` - gráfico de barras

Gráficos

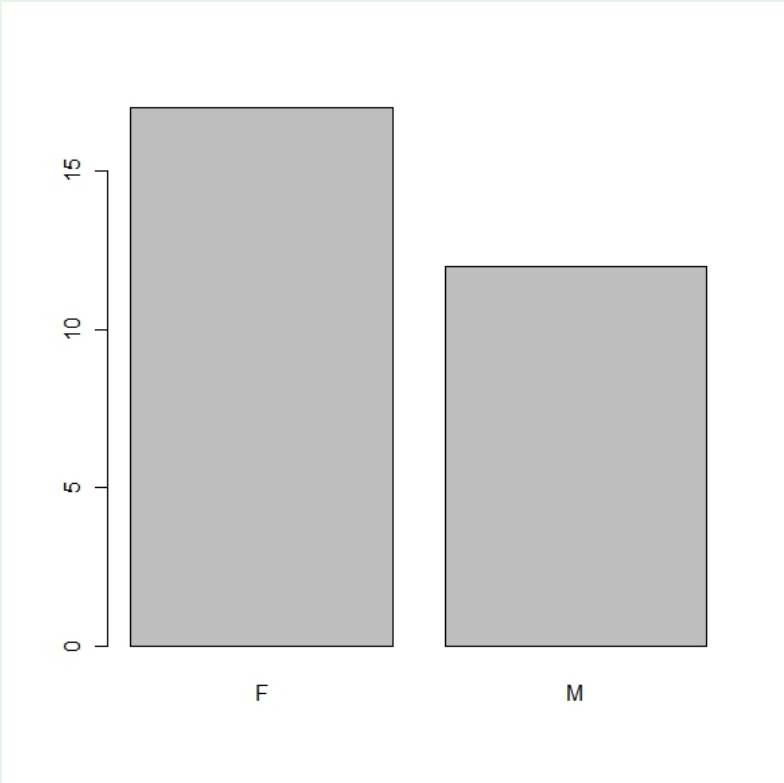
- `plot`
  - Variáveis Qualitativas -gráfico de barras
  - Variáveis Quantitativas - gráfico de dispersão
- `barplot` - gráfico de barras
- `pie` - gráfico de pizza

Gráficos

- `plot`
  - Variáveis Qualitativas -gráfico de barras
  - Variáveis Quantitativas - gráfico de dispersão
- `barplot` - gráfico de barras
- `pie` - gráfico de pizza
- `hist` - histograma.

Gráficos

```
##Grafico de barras
> fa=table(SEXO) #frequência absoluta
> barplot(fa)    #plota gráfico.
> plot(SEXO)   ##plota gráfico
```



Gráficos

- `main` para adicionar titulo;

Gráficos

- `main` para adicionar titulo;
- `xlab` titulo para o eixo x;

Gráficos

- `main` para adicionar titulo;
- `xlab` titulo para o eixo x;
- `ylab` titulo para o eixo y;

Gráficos

- `main` para adicionar titulo;
- `xlab` titulo para o eixo x;
- `ylab` titulo para o eixo y;
- `xlim` delimita os valores de x;

Gráficos

- `main` para adicionar titulo;
- `xlab` titulo para o eixo x;
- `ylab` titulo para o eixo y;
- `xlim` delimita os valores de x;
- `ylim` delimita os valores de y;

Gráficos

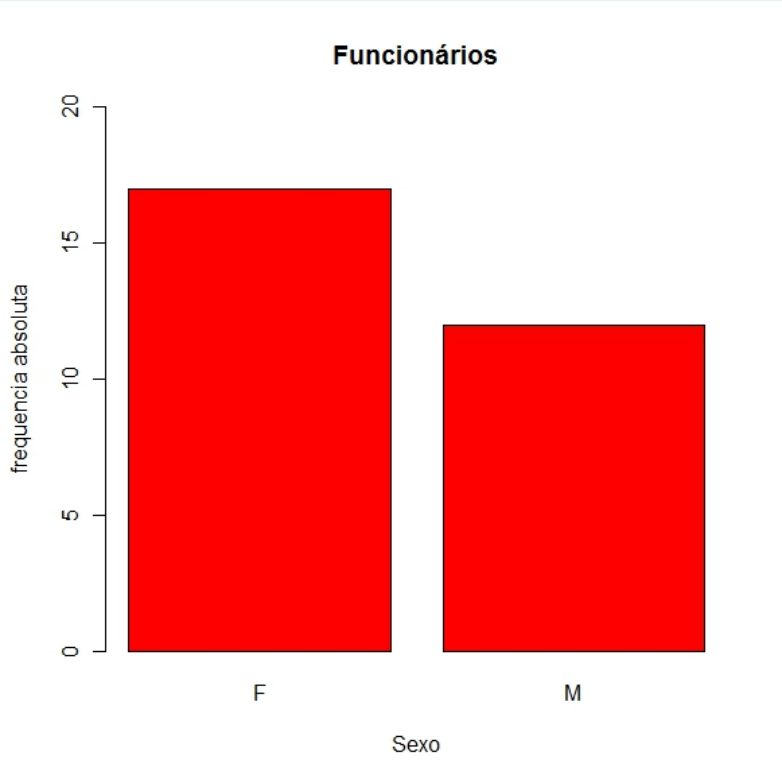
- `main` para adicionar titulo;
- `xlab` titulo para o eixo x;
- `ylab` titulo para o eixo y;
- `xlim` delimita os valores de x;
- `ylim` delimita os valores de y;
- `col` para definir cor;

Gráficos

- `main` para adicionar titulo;
- `xlab` titulo para o eixo x;
- `ylab` titulo para o eixo y;
- `xlim` delimita os valores de x;
- `ylim` delimita os valores de y;
- `col` para definir cor;
- `horiz` para definir se as barras são horizontais ou verticais.

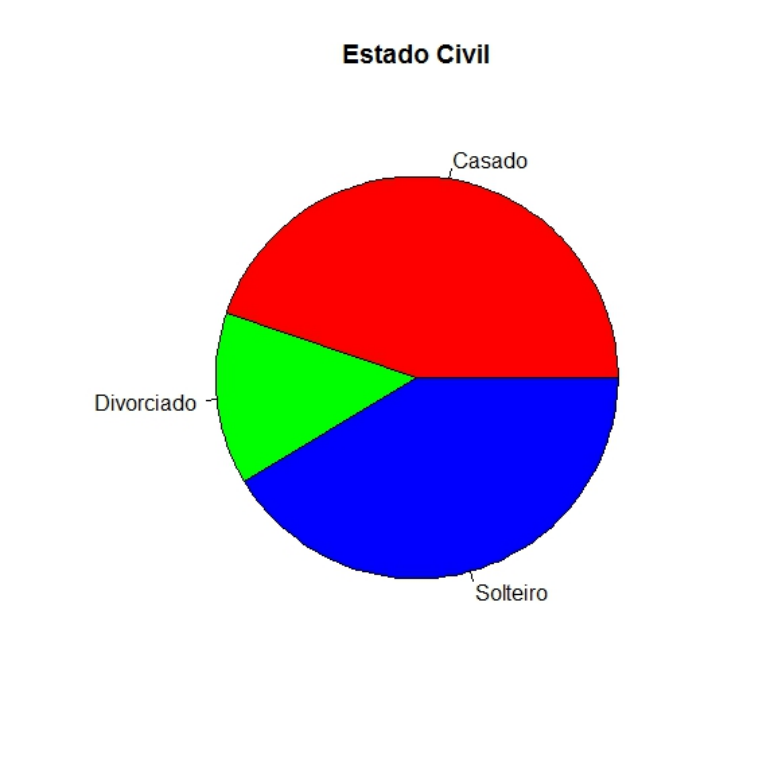
Gráficos

```
> plot(SEXO,main="Funcionários",xlab="Sexo",
+ ylab="frequencia absoluta",ylim=c(0,20),col="red"
```



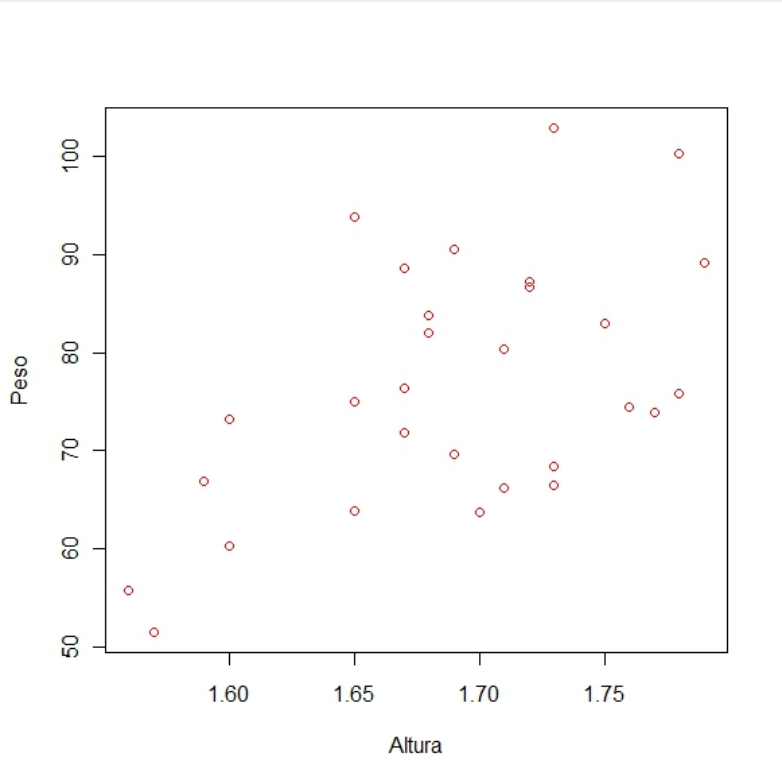
Gráficos

```
> fa=table(EC)
> fr=fa/sum(fa)
> n=length(fr) ##contar quantos classes
> pie(fr,main="Estado Civil",col=rainbow(n))
```



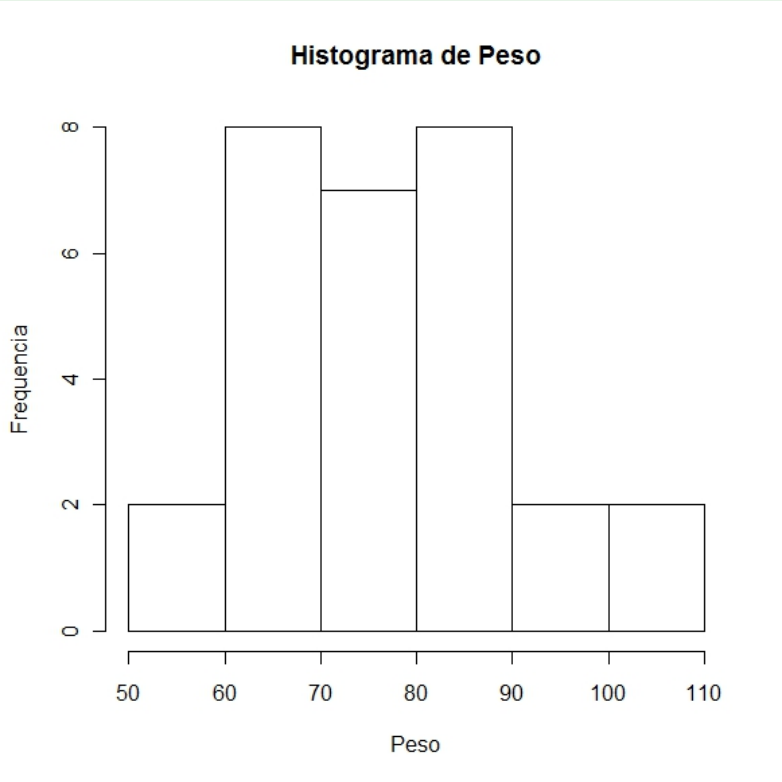
Gráficos

```
> plot(A,P,xlab="Altura",ylab="Peso",col="red")
```



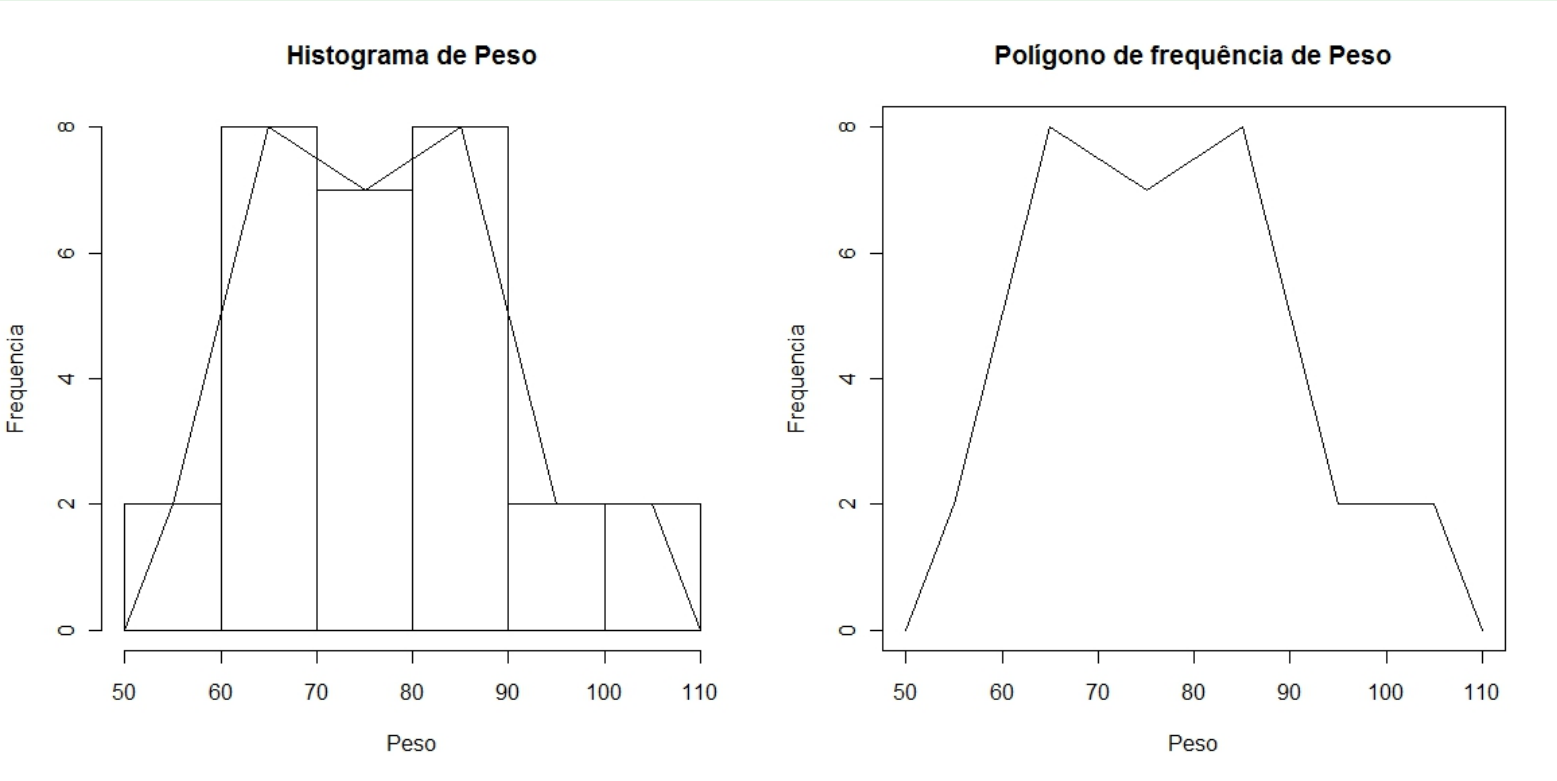
Gráficos

```
#histogram para variável peso.
hist(P,main="Histograma de Peso",xlab="Peso",ylab="
```



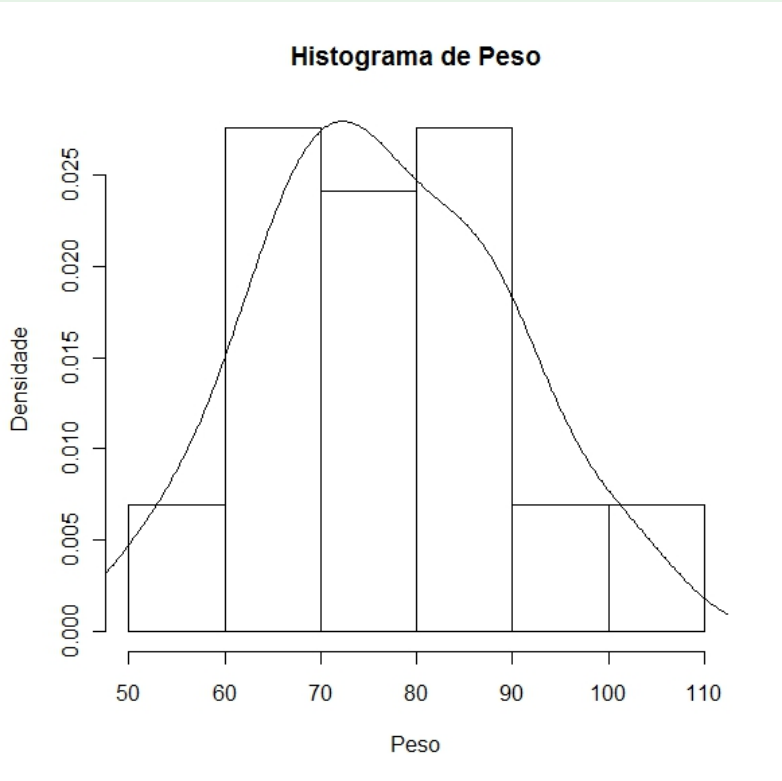
Gráficos

```
h=hist(P,main="Histograma de Peso",xlab="Peso",ylab
##polígono de frequência com histograma
lines(c(min(h$breaks), h$mids, max(h$breaks)), c(0,
 ##polígono de frequência
> plot(c(min(h$breaks), h$mids, max(h$breaks)), c(0
+ type = "l",main="Polígono de frequência de Peso",
```



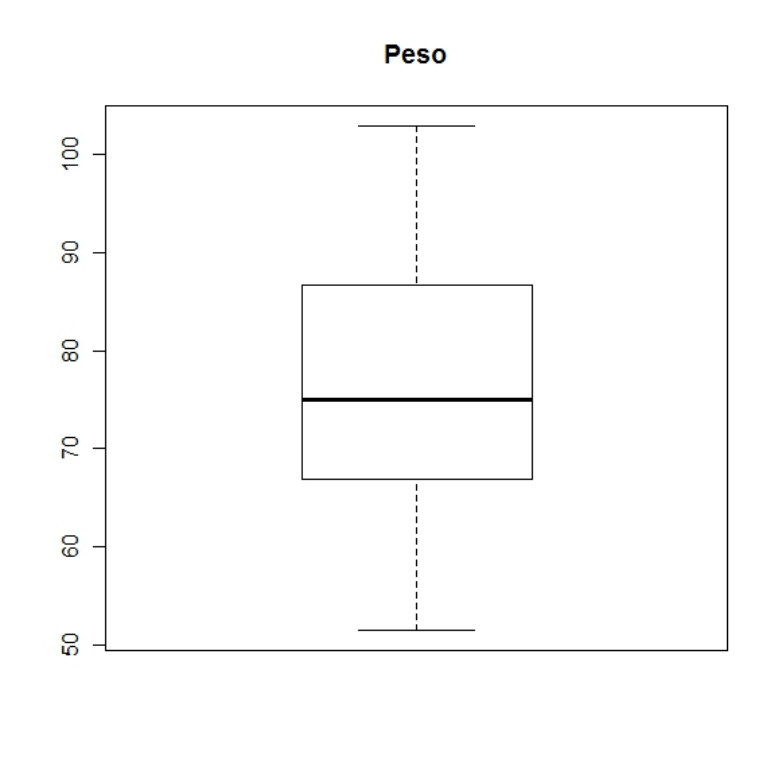
Gráficos

```
##histograma pela densidade de frequência
hist(P,freq=FALSE)
h1=density(P) ##obtendo a densidade dos dados
lines(h1) ##adicinar um gráfico de linhas
```



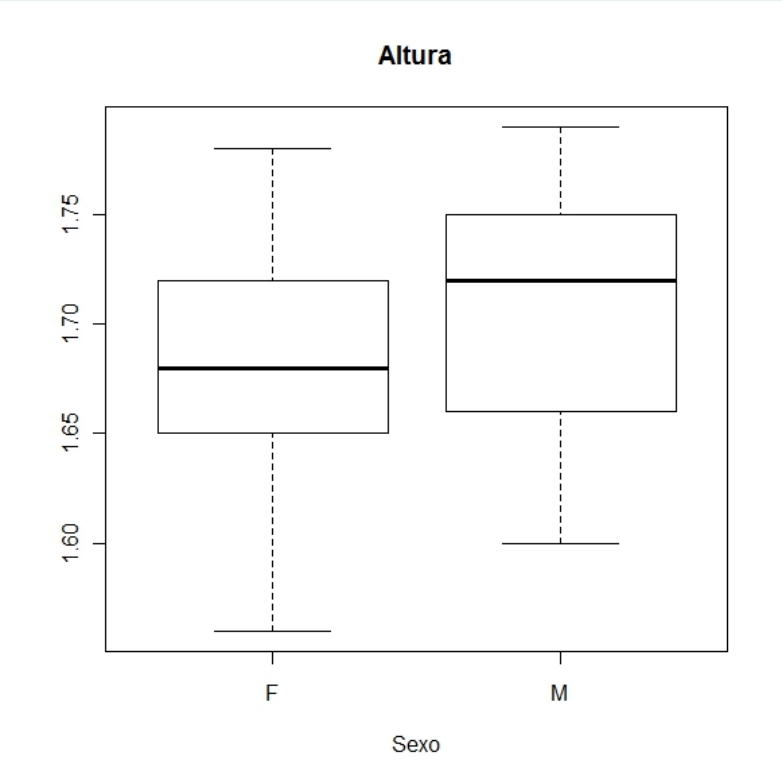
Gráficos

```
#Boxplot para variável peso.
> boxplot(P,main="Peso")
```



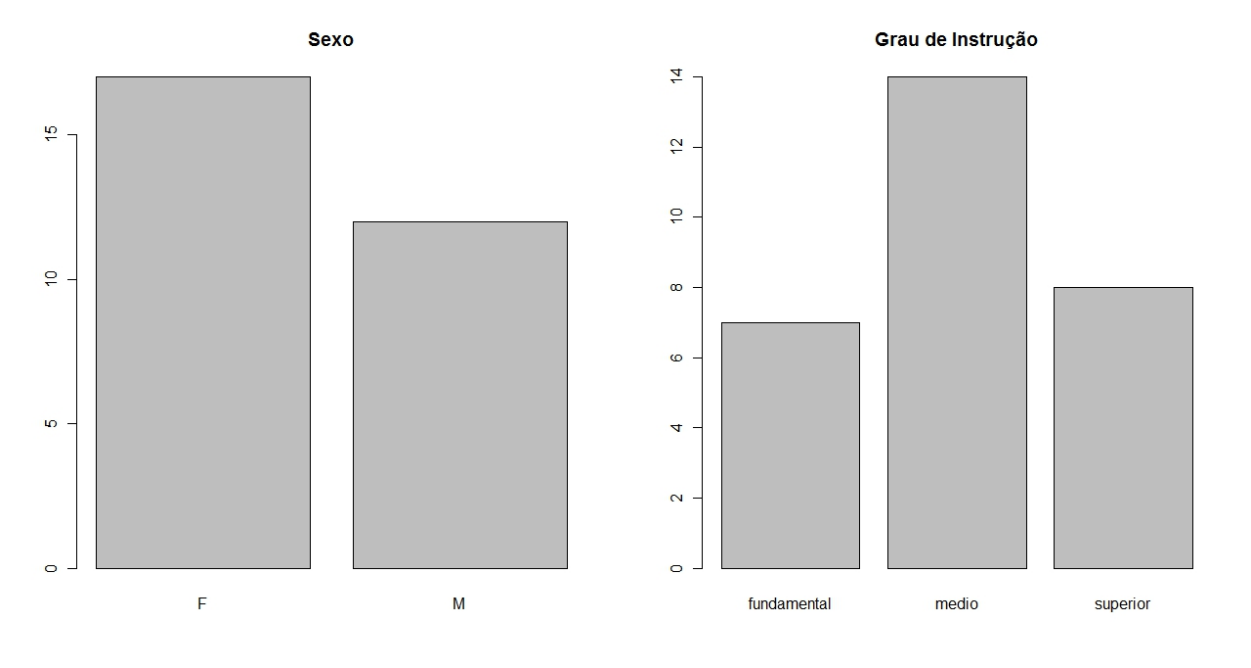
Gráficos

```
#Boxplot para variável altura dentro de cada sexo.
> boxplot(A~SEXO,main="Altura",xlab="Sexo")
```



Dividir Janela de Gráficos

```
> par(mfrow=c(1,2)) ##uma linha e duas colunas
> plot(SEXO,main="Sexo")
> plot(GI,main="Grau de Instrução")
```



Medidas de Posição

- média - `mean`

Medidas de Posição

- média - `mean`
- mediana - `median`

Medidas de Posição

- média - `mean`
- mediana - `median`
- moda - pacote `modeest`.

Medidas de Posição

- média - `mean`
- mediana - `median`
- moda - pacote `modeest`.
  - Variável discreta - `mfv`

Medidas de Posição

- média - `mean`
- mediana - `median`
- moda - pacote `modeest`.
  - Variável discreta - `mfv`
  - Variável contínua por diferentes métodos - `mlv`

Medidas de Posição

- média - `mean`
- mediana - `median`
- moda - pacote `modeest`.
  - Variável discreta - `mfv`
  - Variável contínua por diferentes métodos - `mlv`
- Para obter medidas por grupos - `tapply`.

Medidas de Posição

```
> ##Média de Peso
> mean(P)
[1] 76.59655
> ##Média de Altura por Sexo
> tapply(A, SEXO, mean)
       F        M
1.678824 1.705000
> ##Mediana de Peso
> median(P)
[1] 75
> ##Mediana de Altura por Sexo
> tapply(A, SEXO, median)
   F    M
1.68 1.72
```

Medidas de Posição

```
> require(modeest)
> ##Moda de Idade
> mfv(IDADE)
[1] 27
> ##Moda de Idade por Sexo
> tapply(IDADE, SEXO,mfv)
$F
[1] 20 27
$M
[1] 22 39 49
> ##moda de Peso
> mlv(P)
Mode (most likely value): 69.71333
Bickel's modal skewness: 0.3103448
Call: mlv.default(x = P)
Warning message:
In mlv.default(P) :
  argument 'method' is missing. Data are supposed t
```

Medidas de Dispersão

- Amplitude - `range` combinado com `diff`.

Medidas de Dispersão

- Amplitude - `range` combinado com `diff`.
- Variância - `var`

Medidas de Dispersão

- Amplitude - `range` combinado com `diff`.
- Variância - `var`
- Desvio padrão - `sd`.

Medidas de Dispersão

- Amplitude - `range` combinado com `diff`.
- Variância - `var`
- Desvio padrão - `sd`.
- moda - pacote `modeest`.

Medidas de Dispersão

```
> ##Amplitude de Altura
> range(A)
[1] 1.56 1.79
f(range(A))
[1] 0.23
> X=tapply(A, SEXO,range)
> X
$F
[1] 1.56 1.78

$M
[1] 1.60 1.79

> diff(X$F) ##Ampliture Altura Mulheres
[1] 0.22
> diff(X$M) ##Ampliture Altura Homens
[1] 0.19
```

Medidas de Dispersão

```
> ##Variância de Altura
> var(A)
[1] 0.004010591
> ##Variância de Altura por Sexo
> tapply(A, SEXO,var)
          F           M
0.003886029 0.004118182
> ##Desvio Padrão de Altura
> sd(A)
[1] 0.06332923
> ##Desvio Padrão de Altura por Sexo
> tapply(A, SEXO,sd)
         F          M
0.06233803 0.06417306
```