

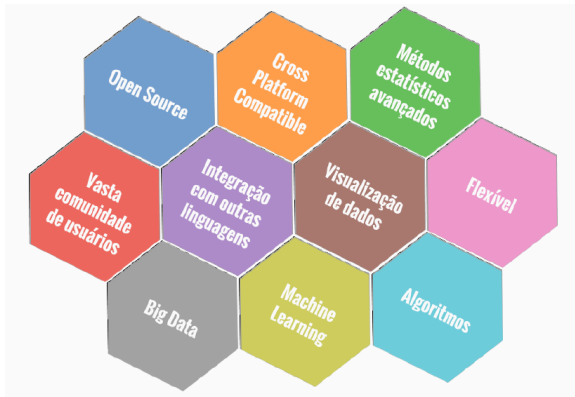
Aula 04 – Introdução à Estatística Descritiva – Programando em R

Luciana Rocha Pedro

GCC 1518 – Estatística e Probabilidade – CEFET Maracanã

26 de março de 2018

Linguagem R



Quem Usa R



Comandos Importantes

| Função | Descrição |
|------------------------------|---|
| <code>rm(x)</code> | deletar variável x |
| <code>rm(x, y)</code> | deletar as variáveis x e y |
| <code>rm(list = ls())</code> | deletar todas as variáveis (limpar a workspace) |
| <code>q()</code> | sair do R com a opção de salvar o workspace em um arquivo ("Name.RData") e o histórico de comandos em outro arquivo ("Name.RHistory") |
| <code>ctrl + L</code> | no teclado, pressione "ctrl+L" para limpar a tela |
| <code>esc</code> | cancelar a linha do cursor |
| seta para cima | acessar últimos comandos |
| seta para baixo | limpar a linha do cursor |
| ?comando | informações sobre um comando específico |

Tabela 01

Tabela 01 – Informações sobre os alunos da disciplina Inferência Estatística do curso de Estatística da UEM – 21/03/2005.

| Num | Sexo | Id | Altura | Peso | Est.Civil | N.Ir. | Transp. | Procedência | Trabalho | Inform | Disc. |
|-----|------|----|--------|------|-----------|-------|----------|-------------|-----------|----------|-------|
| 1 | F | 20 | 1,60 | 58 | Solteiro | 1 | Próprio | Maringá | Não rel. | TV | 2 |
| 2 | F | 26 | 1,65 | 59 | Solteiro | 2 | Coletivo | Fora do PR | Não trab. | Revista | 0 |
| 3 | F | 18 | 1,64 | 55 | Solteiro | 2 | Próprio | Maringá | Não trab. | TV | 0 |
| 4 | F | 25 | 1,73 | 60 | Solteiro | 2 | Coletivo | Outro no PR | Não rel. | TV | 2 |
| 5 | M | 35 | 1,76 | 83 | Casado | 6 | Coletivo | Outro no PR | Não rel. | TV | 2 |
| 6 | F | 20 | 1,62 | 58 | Solteiro | 2 | Coletivo | Outro no PR | Não rel. | Rádio | 5 |
| 7 | F | 29 | 1,72 | 70 | Solteiro | 3 | Coletivo | Maringá | Não trab. | TV | 0 |
| 8 | M | 23 | 1,71 | 62 | Separado | 2 | Próprio | Outro no PR | Não rel. | Internet | 2 |
| 9 | F | 20 | 1,63 | 63 | Solteiro | 2 | Próprio | Maringá | Não trab. | TV | 1 |
| 10 | M | 20 | 1,79 | 75 | Solteiro | 2 | Próprio | Fora do PR | Não trab. | Internet | 2 |
| 11 | M | 20 | 1,82 | 66 | Solteiro | 1 | Próprio | Fora do PR | Não trab. | TV | 2 |
| 12 | F | 30 | 1,68 | 46 | Solteiro | 3 | Próprio | Outro no PR | Parc.rel. | TV | 4 |
| 13 | F | 18 | 1,69 | 64 | Solteiro | 1 | Próprio | Maringá | Parc.rel. | TV | 0 |
| 14 | M | 37 | 1,82 | 80 | Casado | 2 | Próprio | Maringá | Não rel. | TV | 3 |
| 15 | M | 25 | 1,83 | 62 | Solteiro | 1 | Próprio | Outro no PR | Não rel. | TV | 2 |
| 16 | F | 20 | 1,63 | 68 | Solteiro | 2 | Coletivo | Maringá | Não trab. | TV | 2 |
| 17 | M | 21 | 1,71 | 80 | Solteiro | 2 | Coletivo | Maringá | Não rel. | Internet | 0 |
| 18 | M | 25 | 1,80 | 82 | Casado | 1 | Próprio | Outro no PR | Não rel. | Internet | 3 |
| 19 | F | 24 | 1,62 | 55 | Solteiro | 2 | Próprio | Maringá | Não trab. | Jornal | 2 |
| 20 | M | 19 | 1,74 | 58 | Solteiro | 2 | Próprio | Maringá | Com.rel. | TV | 3 |
| 21 | F | 21 | 1,55 | 65 | Solteiro | 1 | Próprio | Maringá | Não trab. | TV | 1 |
| 22 | M | 22 | 1,73 | 62 | Solteiro | 0 | Próprio | Maringá | Não trab. | Jornal | 4 |

Tabela Simples

Tabela 02 – Número de alunos matriculados na disciplina Probabilidade I do curso de Estatística da Universidade Estadual de Maringá.

| Ano | Nº de Alunos |
|------|--------------|
| 2000 | 40 |
| 2001 | 59 |
| 2002 | 63 |
| 2003 | 69 |
| 2004 | 71 |

Fonte: DES/UEM.
Nota: Os números de 2003 e 2004 correspondem a duas turmas.

```

> rm(list=ls())
> ano = c(2000,2001,2002,2003,2004)
> alunos = c(40, 59, 63, 69, 71)
> ano
[1] 2000 2001 2002 2003 2004
> alunos
[1] 40 59 63 69 71
> tab2 = cbind(ano, alunos)

```

```

> tab2
      ano alunos
[1,] 2000     40
[2,] 2001     59
[3,] 2002     63
[4,] 2003     69
[5,] 2004     71

```

Tabela Simples

Tabela 03 – Município de procedência dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

| Município de Procedência | Nº de Alunos |
|--------------------------|--------------|
| Maringá | 12 |
| Outro no Paraná | 7 |
| Fora do Paraná | 3 |
| Total | 22 |

Fonte: Tabela 01.

```
dados = read.table("D:\\Dropbox\\Estatistica\\Aulas\\Tabela 1.csv", header = TRUE, sep = ",")
> names(dados)
[1] "Num"      "Sexo"      "Id"        "Altura"    "Peso"      "Est.Civil"  "N.Ir."     "Transp."
[9] "Procedencia" "Trabalho"  "Inform"    "Disc."
> attach(dados)
> Procedencia
[1] Maringa      Fora do PR   Maringa      Outro no PR  Outro no PR  Outro no PR  Maringa      Outro no PR
[9] Maringa      Fora do PR   Fora do PR   Outro no PR  Maringa      Maringa      Outro no PR   Maringa
[17] Maringa      Outro no PR  Maringa      Maringa      Maringa      Maringa
Levels: Fora do PR Maringa Outro no PR
> summary(Procedencia)
Fora do PR      Maringa      Outro no PR
          3          12          7
```

Tabela Simples

Tabela 03 – Município de procedência dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

| Município de Procedência | Nº de Alunos |
|--------------------------|--------------|
| Maringá | 12 |
| Outro no Paraná | 7 |
| Fora do Paraná | 3 |
| Total | 22 |

Fonte: Tabela 01.

```
> x = table(Procedencia)
> x
Procedencia
Fora do PR      Maringa      Outro no PR
           3           12           7
> x = sort(x, decreasing = TRUE)
> x
Procedencia
Maringa      Outro no PR      Fora do PR
           12           7           3
> tab3 = c(x, sum(x))
> tab3
Maringa      Outro no PR      Fora do PR
           12           7           3           22
> names(tab3)[4]=" Total"
> tab3
Maringa      Outro no PR      Fora do PR      Total
           12           7           3           22
```


Frequências Relativas em Percentual

Tabela 04 – Município de procedência dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

| Município de Procedência | Nº de Alunos | Percentual |
|--------------------------|--------------|------------|
| Maringá | 12 | 55 |
| Outro no Paraná | 7 | 32 |
| Fora do Paraná | 3 | 13 |
| Total | 22 | 100 |

Fonte: Tabela 01.

```
> x
Procedencia
  Maringa  Outro no PR  Fora do PR
      12      7      3
> p-fi = 100*prop.table(x)
> p-fi
Procedencia
  Maringa  Outro no PR  Fora do PR
  54.54545  31.81818  13.63636
> tab41 = c(x,sum(x))
> names(tab41)[4]=" Total"
> tab42 = c(p-fi,sum(p-fi))
> tab4 = cbind(tab41,tab42)
> tab4
```

| | tab41 | tab42 |
|-------------|-------|-----------|
| Maringa | 12 | 54.54545 |
| Outro no PR | 7 | 31.81818 |
| Fora do PR | 3 | 13.63636 |
| Total | 22 | 100.00000 |

```
> tab4 = round(tab4)
> tab4
```

| | tab41 | tab42 |
|-------------|-------|-------|
| Maringa | 12 | 55 |
| Outro no PR | 7 | 32 |
| Fora do PR | 3 | 14 |
| Total | 22 | 100 |

Tabela Simples

Tabela 05 – Meio de transporte mais utilizado pelos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

| Meio de transporte | Nº de Alunos |
|--------------------|--------------|
| Coletivo | 7 |
| Próprio | 15 |
| Total | 22 |

Fonte: Tabela 01.

```
> tab5 = table(Transp.)
> tab5
Transp.
Coletivo  Proprio
      7      15
> tab5 = c(tab5, sum(tab5))
> tab5
Coletivo  Proprio
      7      15      22
> names(tab5)[3]=" Total"
> tab5
Coletivo  Proprio      Total
      7      15      22
```

Tabela Cruzada

Tabela 06 – Meio de transporte mais utilizado segundo o sexo dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

| Meio de transporte | Sexo | | Total |
|--------------------|-----------|----------|-------|
| | Masculino | Feminino | |
| Coletivo | 2 | 5 | 7 |
| Próprio | 8 | 7 | 15 |
| Total | 10 | 12 | 22 |

Fonte: Tabela 01.

```
> tab6 = table(Transp.,Sexo)
```

```
> tab6
```

| | Sexo | |
|----------|------|---|
| Transp. | F | M |
| Coletivo | 5 | 2 |
| Próprio | 7 | 8 |

```
> tab6 = addmargins(tab6)
```

```
> tab6
```

| | Sexo | | |
|----------|------|----|-----|
| Transp. | F | M | Sum |
| Coletivo | 5 | 2 | 7 |
| Próprio | 7 | 8 | 15 |
| Sum | 12 | 10 | 22 |

Distribuição de Freqüências Completa

Tabela 08 – Número de irmãos dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

| Número de irmãos (x_i) | F_i | f_i % | F_{ai} | f_{ai} % |
|----------------------------|-------|---------|----------|------------|
| 0 | 1 | 4,55 | 1 | 4,55 |
| 1 | 6 | 27,26 | 7 | 31,81 |
| 2 | 12 | 54,55 | 19 | 86,36 |
| 3 | 2 | 9,09 | 21 | 95,45 |
| 6 | 1 | 4,55 | 22 | 100,00 |
| Total | 22 | 100,00 | | |

Fonte: Tabela 01.

```
> ni = N.Ir.
> Freq = table(ni)
> Relativa = 100*prop.table(Freq)
> Acumulada = cumsum(Freq)
> AcumuladaRelativa = cumsum(Relativa)
> tab8 = cbind(Freq, Relativa, Acumulada, AcumuladaRelativa)
> tab8 = round(tab8, digits = 2)
```

```
> tab8
      Freq Relativa Acumulada AcumuladaRelativa
0      1     4.55         1         4.55
1      6    27.27         7        31.82
2     12    54.55        19        86.36
3      2     9.09        21        95.45
6      1     4.55        22       100.00
```

Distribuição de Frequência Pontual

Tabela 09 – Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

| Idade | F_i |
|-------|-------|
| 18 | 2 |
| 19 | 1 |
| 20 | 6 |
| 21 | 2 |
| 22 | 1 |
| 23 | 1 |
| 24 | 1 |
| 25 | 3 |
| 26 | 1 |
| 29 | 1 |
| 30 | 1 |
| 35 | 1 |
| 37 | 1 |
| Total | 22 |

Fonte: Tabela 01.

```
> Id
[1] 20 26 18 25 35 20 29 23 20 20 20 30 18 37 25 20 21 25 24 19 21 22
> tab9 = table(Id)
> tab9
Id
18 19 20 21 22 23 24 25 26 29 30 35 37
 2  1  6  2  1  1  1  3  1  1  1  1  1
```

Distribuição de Frequência em Classes

- ▶ $AT = 37 - 18 = 19$ anos;
- ▶ $k = \sqrt{22} = 4,69 \cong 5$ classes;
- ▶ $h = \frac{19}{5} = 3,8 \cong 4$ anos.

| | | |
|---|--|--|
| <pre>> xmax = max(Id) > xmin = min(Id) > AT = xmax - xmin > AT [1] 19</pre> | <pre>> n = length(Id) > k = sqrt(n) > k [1] 4.690416 > k = round(k) > k [1] 5</pre> | <pre>> h = AT/k > h [1] 3.8 > h = round(h) > h [1] 4</pre> |
|---|--|--|

Distribuição de Frequência em Classes

Tabela 10 – Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

| Idade | x_i | F_i | $f_i\%$ | F_{a_i} | $f_{a_i}\%$ |
|----------|-------|-------|---------|-----------|-------------|
| 18 ---22 | 20 | 11 | 50,00 | 11 | 50,00 |
| 22 ---26 | 24 | 6 | 27,27 | 17 | 77,27 |
| 26 ---30 | 28 | 2 | 9,09 | 19 | 86,36 |
| 30 ---34 | 32 | 1 | 4,55 | 20 | 90,91 |
| 34 ---38 | 36 | 2 | 9,09 | 22 | 100,00 |
| Total | - | 22 | 100,00 | - | - |

Fonte: Tabela 01.

```
> Id
[1] 20 26 18 25 35 20 29 23 20 20 20 30 18 37 25 20 21 25 24 19 21 22
> xmax = xmax + 1
> Fi = table(cut(Id, breaks = seq(xmin, xmax, h), right = FALSE))
> xi = seq(xmin+h/2, xmax, h)
> fi = 100*prop.table(Fi)
> Fai = cumsum(Fi)
> fai = cumsum(fi)
> tab10 = cbind(xi, Fi, fi, Fai, fai)

> tab10 = round(tab10, digits = 2)
> tab10
      xi Fi    fi Fai    fai
[18,22) 20 11 50.00  11 50.00
[22,26) 24  6 27.27  17 77.27
[26,30) 28  2  9.09  19 86.36
[30,34) 32  1  4.55  20 90.91
[34,38) 36  2  9.09  22 100.00
```

Gráfico de Barras

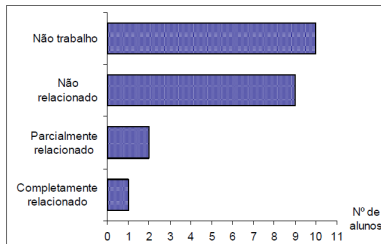


Figura 02 - Relação trabalho e curso dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM, 21/03/2005.

Fonte: Tabela 01.

```
> fig2 = table(Trabalho)
> fig2
Trabalho
Com.rel.  Nao rel.  Nao trab.  Parc.rel.
      1         9        10         2
> barplot(fig2)
> barplot(sort(fig2), horiz = TRUE)
> barplot(sort(fig2), horiz = TRUE, xlab = "Numero de alunos")
```

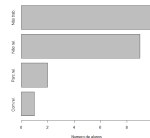


Gráfico de Colunas

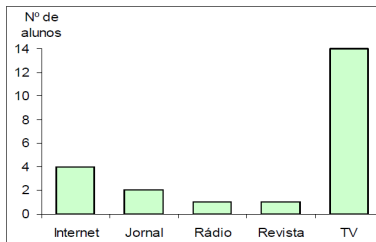
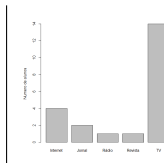


Figura 03 - Meios de informação utilizados pelos alunos da disciplina Inferência Estatística, curso de Estatística da UEM, 21/03/2005.

Fonte: Tabela 01.

```
> fig3 = table(Inform)
> barplot(fig3, ylab = "Numero de alunos")
```



Gráficos Comparativos

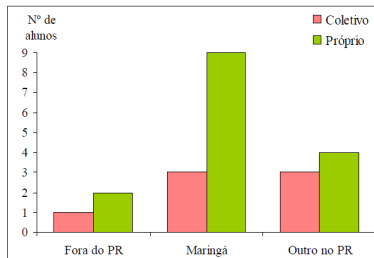


Figura 04 - Município de procedência segundo o tipo de transporte utilizado pelos alunos da disciplina Inferência Estatística do curso de Estatística da UEM, 21/03/2005.

Fonte: Tabela 01.

```
> fig4 = table(Transp., Procedencia)
> barplot(fig4, beside = TRUE)
> barplot(fig4, beside = TRUE, legend = rownames(fig4))
> barplot(fig4, beside = TRUE, legend = rownames(fig4), col = c("pink", "green"))
```

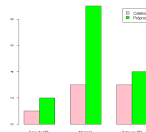


Gráfico de Setores

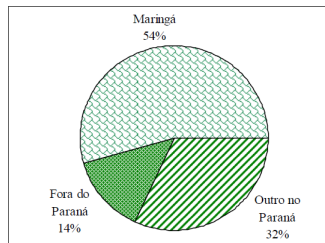


Figura 05 - Município de procedência dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM, 21/03/2005.

Fonte: Tabela 01.

```
> fig5 = table(Procedencia)
> fig5 = round(100*prop.table(fig5))
> lab = paste(names(fig5), round(fig5))
> lab = paste(lab, "%", sep = "")
> pie(fig5, labels = lab)
```



Gráfico de Linhas

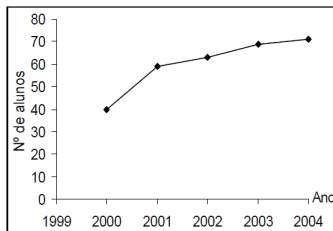


Figura 06 - Número de matrículas anuais na disciplina Probabilidade do curso de Estatística da UEM, 21/03/2005.

Fonte: Tabela 01.

```
> ano = c(2000, 2001, 2002, 2003, 2004)
> alunos = c(40, 59, 63, 69, 71)
> plot(ano, alunos, ylim = c(0, 80), xlim = c(1999, 2004))
> lines(ano, alunos)
```

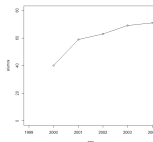


Gráfico de Bastões

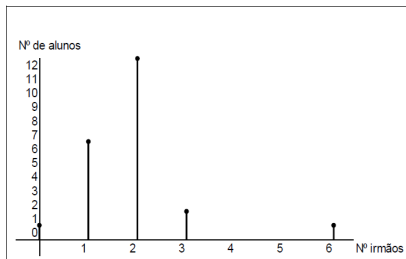


Figura 07 – Número de irmãos dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM, 21/03/2005.

Fonte: Tabela 01.

```
> N.Ir.
[1] 1 2 2 2 6 2 3 2 2 2 1 3 1 2 1 2 2 2 1 2 1 0
> fig7 = table(N.Ir.)
> fig7
N.Ir.
 0  1  2  3  6
 1  6 12  2  1
> plot(fig7, type = "h")
> par(new = TRUE)
> plot(fig7, type = "p")
```

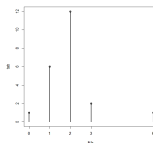


Gráfico de Freqüência Acumulada

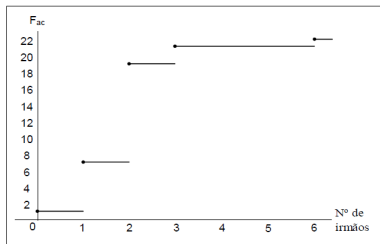
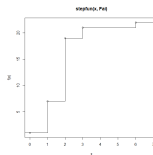


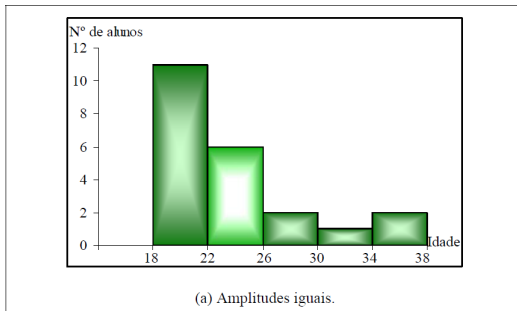
Figura 08 - Número acumulado de irmãos dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM. 21/03/2005.

Fonte: Tabela 01.

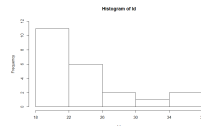
```
> y = c(0, 1, 2, 3, 6)
> yai = c(1, 1, 7, 19, 21, 22)
> fig8 = stepfun(y, yai)
> plot(fig8)
> plot(fig8, xlim = c(0, 7))
```



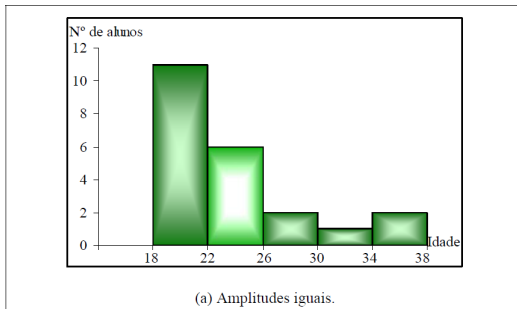
Histograma



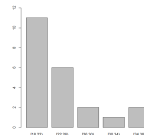
```
> hist(Id, breaks = seq(xmin-h, xmax, h), right = FALSE, xaxt = "n", ylim = c(0, 12))
> axis(1, seq(xmin-h, xmax, h))
```



Histograma



```
> fig9 = table(cut(Id, breaks = seq(xmin, xmax, h), right = FALSE))
> fig9
[18,22) [22,26) [26,30) [30,34) [34,38)
      11         6         2         1         2
> barplot(fig9)
> barplot(fig9, ylim = c(0, 12))
```



Polígono de Freqüência

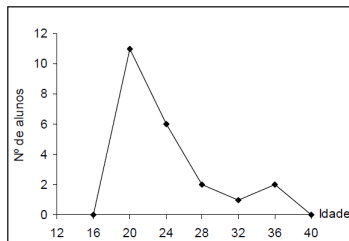


Figura 10 - Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM, 21/03/2005.

Fonte: Tabela 01.

```
> Fi = table(cut(Id, breaks = seq(xmin, xmax, h), right = FALSE))
> xi = seq(xmin-h/2, xmax+h/2, h)
> xi
[1] 16 20 24 28 32 36 40
> Fi0 = c(0, Fi, 0)
> Fi0
      [18,22) [22,26) [26,30) [30,34) [34,38)
      0      11      6      2      1      2      0
> plot(xi, Fi0, ylim = c(0, 12), xaxt = "n")
> lines(xi, Fi0)
> axis(1, xi)
```

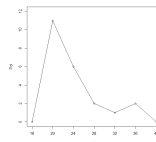


Gráfico de Freqüência Acumulada ou Ogiva

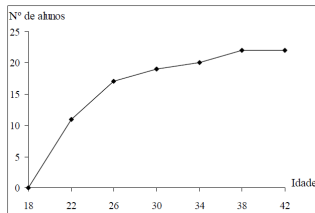


Figura 11 - Idade acumulada dos alunos da disciplina Inferência Estatística do curso de Estatística da UEM, 21/03/2005.

Fonte: Tabela 01.

```
> x = seq(xmin, xmax+h, h)
> x
[1] 18 22 26 30 34 38 42
> Fi0
      [18,22) [22,26) [26,30) [30,34) [34,38)
      0       11      6       2       1       2
> fi0 = cumsum(Fi0)
> fi0
      [18,22) [22,26) [26,30) [30,34) [34,38)
      0       11      17      19      20      22
> plot(x, fi0, xaxt = "n", ylim = c(0, 25))
> lines(x, fi0)
> axis(1, x)
```

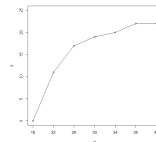


Diagrama Ramo-e-Folhas

| Ramo | Folha | Frequência |
|---------------------------|-------------------------------|------------|
| 1 | 8 8 9 | 3 |
| 2 | 0 0 0 0 0 0 1 1 2 3 4 5 5 6 9 | 16 |
| 3 | 5 7 | 2 |
| (a) Sem divisão de ramos. | | |
| Ramo | Folha | Frequência |
| 1 | 8 8 9 | 3 |
| 2 | 0 0 0 0 0 0 1 1 2 3 4 | 12 |
| 2* | 5 5 6 9 | 4 |
| 3 | 5 7 | 2 |
| (b) Com divisão de ramos. | | |

Figura 12 - Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

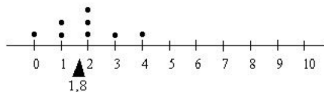
Fonte: Tabela 01.

```
> stem(Id)
```

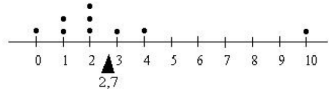
```
The decimal point is 1 digit(s) to the right of the |
```

```
1 | 889
2 | 00000011234
2 | 55569
3 | 0
3 | 57
```

Média



```
> vet = c(0, 1, 1, 2, 2, 2, 3, 4)
> vet
[1] 0 1 1 2 2 2 3 4
> mean(vet)
[1] 1.875
```



```
> vet2 = c(vet, 10)
> vet2
[1] 0 1 1 2 2 2 3 4 10
> mean(vet2)
[1] 2.777778
```

Média

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} = \frac{20 + 26 + 18 + \dots + 21 + 22}{22} = \frac{518}{22} = 23,5 \text{ anos.}$$

```
> Id
[1] 20 26 18 25 35 20 29 23 20 20 20 30 18 37 25 20 21 25 24 19 21 22
> mean(Id)
[1] 23.54545
```

Média

$$\bar{X} = \frac{\sum_{i=1}^k x_i F_i}{n} = \frac{20 \cdot 11 + 24 \cdot 6 + \dots + 36 \cdot 2}{22} = \frac{524}{22} = 23,8 \text{ anos.}$$

```
> Fi
[18,22) [22,26) [26,30) [30,34) [34,38)
      11         6         2         1         2
> xi = seq(xmin+h/2, xmax, h)
> xi
[1] 20 24 28 32 36
> n
[1] 22
> m = sum(Fi*xi)/n
> m
[1] 23.81818
```

Moda

A **moda** (M_o) é o valor que apresenta a maior frequência da variável entre os valores observados.

```
> moda = function(x) {
+   t = table(x)
+   return(as.numeric(names(t)[t == max(t)]))
+ }
```

```
> moda = function(x) {
+   t = table(x)
+   return(as.numeric(names(t)[t == max(t)]))
+ }
> moda(Id)
[1] 20
```

Moda

$$\begin{aligned}
 M_o &= l_i + \frac{h(F_i - F_{i-1})}{(F_i - F_{i-1}) + (F_i - F_{i+1})} \\
 &= 18 + \frac{4 \cdot (11 - 0)}{(11 - 0) + (11 - 6)} = 18 + \frac{44}{16} = 18 + 2,75 = 20,75 \text{ anos.}
 \end{aligned}$$

```

> modaclass = function(li, h, F1, F2, F3) {
  modaclass = li + h*(F2 - F1)/((F2 - F1) + F2 - F3)
  modaclass
}

```

```

> modaclass = function(li, h, F1, F2, F3) {
+   modaclass = li + h*(F2 - F1)/((F2 - F1) + F2 - F3)
+   modaclass
+ }
> modaclass(x[1],h,0,Fi[1],Fi[2])
[18,22)
20.75

```


Mediana

| X | x ₁ | x ₂ | x ₃ | x ₄ | | x ₅ | x ₆ | x ₇ | x ₈ |
|-------------------|----------------|----------------|----------------|----------------|-----------------------|-------------------|----------------|----------------|----------------|
| Valor observado | 0 | 1 | 1 | 2 | $\frac{x_4 + x_5}{2}$ | 2 | 2 | 3 | 4 |
| ← 4 observações → | | | | Md=2 | | ← 4 observações → | | | |

```
> vet = c(0, 1, 1, 2, 2, 2, 2, 3, 4)
[1] 0 1 1 2 2 2 3 4
> median(vet)
[1] 2
```

Mediana

| X | X ₁ | X ₂ | X ₃ | X ₄ | X ₅ | X ₆ | X ₇ | X ₈ | X ₉ |
|-----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Valor observado | 0 | 1 | 1 | 1 | 2 | 2 | 3 | 4 | 10 |

\longleftrightarrow 4 observações \longrightarrow Md=2 \longleftrightarrow 4 observações \longrightarrow

```

> vet2
[1] 0 1 1 2 2 2 3 4 10
> median(vet2)
[1] 2

```

Mediana

18, 18, 19, 20, 20, 20, 20, 20, 20, 20, 21, 21, 22, 23, 24, 25, 25, 25, 26, 29, 30, 35, 37

$$M_d = x_{11} + 0,5 \cdot (x_{12} - x_{11}) = 21 + 0,5 \cdot (22 - 21) = 21,5 \text{ anos.}$$

```
> Id
[1] 20 26 18 25 35 20 29 23 20 20 20 30 18 37 25 20 21 25 24 19 21 22
> sort(Id)
[1] 18 18 19 20 20 20 20 20 20 21 21 22 23 24 25 25 25 26 29 30 35 37
> median(Id)
[1] 21.5
>
```

Mediana

$$\begin{aligned}
 M_d &= l_i + \frac{h(p - F_{a_{i-1}})}{F_i} \\
 &= 18 + \frac{4 \cdot (11 - 0)}{11} = 18 + \frac{44}{11} = 18 + 4 = 22 \text{ anos.}
 \end{aligned}$$

```

> medianaclass = function(li, h, p, Fa1, F2) {
+   medianaclass = li + h*(p - Fa1)/F2
+   medianaclass
+ }

> medianaclass(x[1], h, Fi[1], 0, Fi[1])
[18,22)
22
    
```

Medidas Separatrizes

18, 18, 19, 20, 20, 20, 20, 20, 20, 20, 21, 21, 22, 23, 24, 25, 25, 25, 26, 29, 30, 35, 37

- ▶ o terceiro quartil,
- ▶ o quadragésimo percentil,

```
> summary(Id)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00  20.00   21.50   23.55   25.00   37.00
> quantile(Id, seq(0.1, 0.9, 0.1))
 10%  20%  30%  40%  50%  60%  70%  80%  90%
19.1 20.0 20.0 20.4 21.5 23.6 25.0 25.8 29.9
> quantile(Id, c(0.4))
 40%
20.4
```

Medidas Separatrizes em Classes

Tabela 10 – Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá, 21/03/2005.

| Idade | x_i | F_i | $f_i\%$ | F_{a_i} | $f_{a_i}\%$ |
|----------|-------|-------|---------|-----------|-------------|
| 18 ---22 | 20 | 11 | 50,00 | 11 | 50,00 |
| 22 ---26 | 24 | 6 | 27,27 | 17 | 77,27 |
| 26 ---30 | 28 | 2 | 9,09 | 19 | 86,36 |
| 30 ---34 | 32 | 1 | 4,55 | 20 | 90,91 |
| 34 ---38 | 36 | 2 | 9,09 | 22 | 100,00 |
| Total | - | 22 | 100,00 | - | - |

Fonte: Tabela 01.

```
> x
[1] 18 22 26 30 34 38
> Fi
[18,22) [22,26) [26,30) [30,34) [34,38)
      11         6         2         1         2
> Fai
[18,22) [22,26) [26,30) [30,34) [34,38)
      11        17        19        20        22
> n
[1] 22
> h
[1] 4
```

Medidas Separatrizes em Classes

- ▶ o primeiro quartil,
- ▶ o terceiro quartil,
- ▶ o sétimo decil,
- ▶ o nonagésimo percentil.

```
> k = 1
> p = n/4*k
> p
[1] 5.5
> i = 1
> Q1 = medianaclass(x[i], h, p, 0, Fi[i])
> Q1
[18,22)
      20
```

```
> k = 3
> p = n/4*k
> p
[1] 16.5
> i = 2
> Q3 = medianaclass(x[i], h, p, Fai[i-1], Fi[i])
> Q3
[18,22)
25.66667
```

```
> k = 7
> p = n/10*k
> p
[1] 15.4
> i = 2
> D7 = medianaclass(x[i], h, p, Fai[i-1], Fi[i])
> D7
[18,22)
24.93333
```

```
> k = 90
> p = n/100*k
> p
[1] 19.8
> i = 4
> P90 = medianaclass(x[i], h, p, Fai[i-1], Fi[i])
> P90
[26,30)
      33.2
```

Amplitude

$$AT = 37 - 18 = 19 \text{ anos}$$

```
> xmax = max(Id)
> xmin = min(Id)
> AT = xmax - xmin
> AT
[1] 19
```


Amplitude Interquartílica

$$dq = 25,67 - 20 = 5,67 \text{ anos.}$$

```
> dq = Q3 - Q1
> dq
[1] 5.666667
```

Amplitude Semi-interquartílica

$$dq_m = 2,84 \text{ anos.}$$

```
> dqm = dq/2
> dqm
[1] 2.833333
```

Variância e Desvio Padrão

$$\text{Variância: } s^2 = \frac{(20 - 23,5)^2 + \dots + (36 - 23,5)^2}{22 - 1} = 27,11 \text{ anos.}$$

$$\text{Desvio padrão: } s = \sqrt{27,11} = 5,21 \text{ anos.}$$

```
> variancia = function(n, m, x) {
+   soma = 0
+   for (i in 1:n){
+     soma = soma + (x[i]-m)^2
+   }
+   soma
+   variancia = soma/(n-1)
+   variancia
+ }

variancia = function(n, m, x) {
+   soma = 0
+   for (i in 1:n){
+     soma = soma + (x[i]-m)^2
+   }
+   soma
+   variancia = soma/(n-1)
+   variancia
+ }

> x = Id
> variancia(n, m, x)
[1] 27.19481

> var(Id)
[1] 27.11688
> sqrt(var(Id))
[1] 5.207387
> sd(Id)
[1] 5.207387
```

Variância e Desvio Padrão – Classes

$$\text{Variância: } s^2 = \frac{(20 - 23,8)^2 \cdot 11 + \dots + (36 - 23,8)^2 \cdot 2}{22 - 1} = 26,63 \text{ anos.}$$

```
> varianciaclass = function(k, n, m, x, F) {
  soma = 0
  for (i in 1:k){
    soma = soma + (x[i]-m)^2*F[i]
  }
  soma
  varianciaclass = soma/(n-1)
  varianciaclass
}
```

```
> varianciaclass = function(k, n, m, x, F) {
+   soma = 0
+   for (i in 1:k){
+     soma = soma + (x[i]-m)^2*F[i]
+   }
+   soma
+   varianciaclass = soma/(n-1)
+   varianciaclass
+ }
```

Variância e Desvio Padrão – Classes

Desvio padrão: $s = \sqrt{23,63} = 5,16$ anos.

```
> n = length(Id)
> n
[1] 22
> k = round(sqrt(n))
> k
[1] 5
> m
[1] 23.81818
> xi = seq(xmin+h/2, xmax, h)
> xi
[1] 20 24 28 32 36
> Fi
[18,22) [22,26) [26,30) [30,34) [34,38)
      11         6         2         1         2
```

```
> s2 = varianciaclass (k, n, m, xi, Fi)
> s2
[18,22)
26.63203
> s = sqrt(s2)
> s
[18,22)
5.160623
```

Coeficiente de Variação

$$CV = \frac{\sigma}{\mu} \cdot 100 = \frac{5,16}{23,8} \cdot 100 = 21,68\%.$$

```
> s
[1] 5.160657
> m
[1] 23.81818
> CV = 100*s/m
> CV
[18,22)
21.66674
```

Medida de Assimetria

$$A_s = \frac{\mu - M_o}{\sigma} = \frac{23,8 - 20,75}{5.16} = 0,89.$$

```
> x[1]
[1] 18
> h
[1] 4
> Fi
[18,22) [22,26) [26,30) [30,34) [34,38)
      11         6         2         1         2

> mo = modaclass(x[1],h,0,Fi[1],Fi[2])
> mo
[18,22)
      20.75
> As = (m - mo)/s
> As
[18,22)
      0.594537
```

Medida de Curtose

$$K = \frac{(Q_3 - Q_1)}{2 \cdot (P_{90} - P_{10})} = \frac{(25,67 - 20)}{2(33,2 - 18,8)} = 0,1969.$$

| | | |
|--|---|--|
| <pre>> Q3 [18,22) 25.66667 > Q1 [18,22) 20 > P90 [26,30) 33.2</pre> | <pre>> k = 10 > p = n/100*k > p [1] 2.2 > i = 1</pre> | <pre>> P10 = medianaclasses(x[i], h, p, 0, Fi[i]) > P10 [18,22) 18.8 > K = (Q3 - Q1)/(2*(P90 - P10)) > K [1] 0.1967593</pre> |
|--|---|--|

BoxPlot

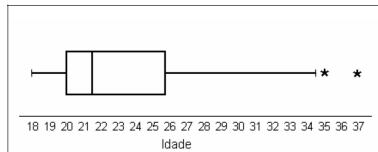
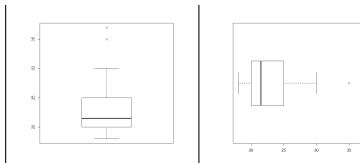


Figura 17 - Idade dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá.

```
> boxplot(Id)
> boxplot(Id, horizontal = TRUE)
```



BoxPlot

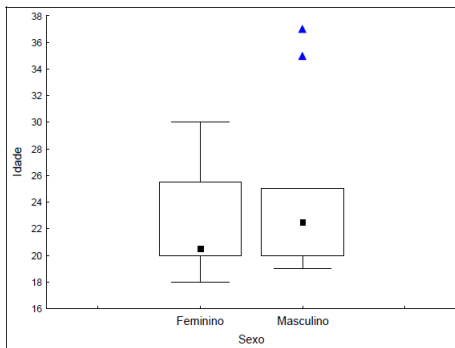


Figura 18 – Box plot da idade segundo o sexo dos alunos da disciplina Inferência Estatística do curso de Estatística da Universidade Estadual de Maringá.

BoxPlot

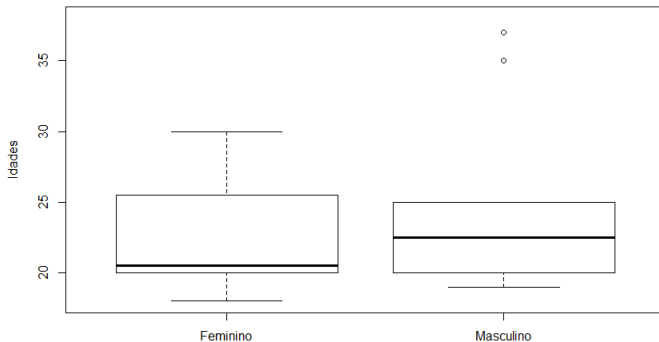
```

> tab = cbind(Id, Sexo)
> tab
      Id Sexo
[1,] 20    1
[2,] 26    1
[3,] 18    1
[4,] 25    1
[5,] 35    2
[6,] 20    1
[7,] 29    1
[8,] 23    2
[9,] 20    1
[10,] 20   2
[11,] 20   2
[12,] 30    1
[13,] 18    1
[14,] 37    2
[15,] 25    2
[16,] 20    1
[17,] 21    2
[18,] 25    2
[19,] 24    1
[20,] 19    2
[21,] 21    1
[22,] 22    2

> tabf = tab[which(tab[,2] == 1),]
> tabf
      Id Sexo
[1,] 20    1
[2,] 26    1
[3,] 18    1
[4,] 25    1
[5,] 20    1
[6,] 29    1
[7,] 20    1
[8,] 30    1
[9,] 18    1
[10,] 20    1
[11,] 24    1
[12,] 21    1
> tabm = tab[which(tab[,2] == 2),]
> tabm
      Id Sexo
[1,] 35    2
[2,] 23    2
[3,] 20    2
[4,] 20    2
[5,] 37    2
[6,] 25    2
[7,] 21    2
[8,] 25    2
[9,] 19    2
[10,] 22    2
> boxplot(tabf[,1], tabm[,1], ylab = "Idades", ylim = c(xmin, xmax))
> axis(1, at = 1:2, labels = c("Feminino", "Masculino"))

```

BoxPlot



Exemplo

Um investigador deseja estudar a possível relação entre o **Salario** (em mil reais) e o tempo de **Experiencia** (em anos completos) no cargo de gerente de agências bancárias de uma grande empresa utilizando 27 pares de observações.

| Experiencia | Salario |
|-------------|---------|
| 0 | 1.9307 |
| 17 | 3.1769 |
| 8 | 2.2769 |
| 15 | 3.1307 |
| 9 | 2.7769 |
| 15 | 3.0923 |
| 8 | 2.6538 |
| 5 | 2.2230 |
| 13 | 2.8538 |
| 20 | 3.2307 |
| 11 | 2.8230 |
| 1 | 1.9076 |
| 6 | 2.5384 |
| 7 | 2.5692 |

| Experiência | Salário |
|-------------|---------|
| 23 | 4.2230 |
| 20 | 4.0923 |
| 18 | 3.6000 |
| 27 | 4.7076 |
| 11 | 3.1461 |
| 10 | 2.9923 |
| 29 | 4.7461 |
| 23 | 4.1153 |
| 4 | 2.3615 |
| 22 | 4.0923 |
| 25 | 4.5076 |
| 9 | 2.9076 |
| 25 | 4.4846 |

Estatísticas Descritivas

Uma maneira fácil de obter algumas estatísticas descritivas das variáveis em estudo é através do comando *summary()*, que retorna as estatísticas mínimo, quartis, média e máximo.

Para medir a variabilidade, utilizamos as funções *var()* e *sd()* para obtermos a variância e o desvio padrão.

```
> dados = read.table("D:\\Dropbox\\Estatistica\\Aulas\\BD.csv", sep = ',', header = TRUE)

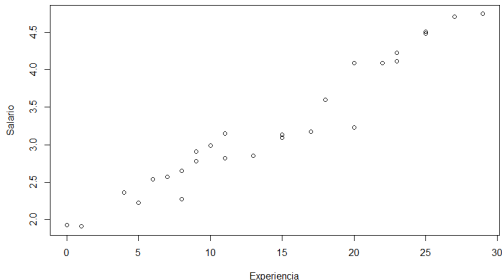
> names(dados)
[1] "Experiencia" "Salario"
> attach(dados)
> summary(dados)
  Experiencia      Salario 
Min.   : 0.00   Min.   :1.908 
1st Qu.: 8.00   1st Qu.:2.611 
Median :13.00   Median :3.092 
Mean   :14.11   Mean   :3.228 
3rd Qu.:21.00   3rd Qu.:4.092 
Max.   :29.00   Max.   :4.746 

> summary(Salario)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
1.908  2.611   3.092   3.228  4.092   4.746 
> var(Salario)
[1] 0.7366968
> summary(Experiencia)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.00   8.00   13.00   14.11  21.00  29.00 
> var(Experiencia)
[1] 68.33333
```

Diagrama de Dispersão

Para verificar a existência de alguma relação entre **Salário** e **Experiência**, podemos construir um diagrama de dispersão e calcular o coeficiente de correlação linear de pearson entre as variáveis utilizando a função *cor*.

```
> plot(Experiencia,Salario)
> cor(Experiencia,Salario)
[1] 0.9704137
```



Regressão Linear Simples

Sejam X e Y , respectivamente, as variáveis Experiência (explicativa) e Salário (resposta). Propõe-se um modelo de regressão linear de primeira ordem, dado pela equação: $y = \alpha + \beta x$, em que α e β são parâmetros desconhecidos. Para ajustar um modelo de regressão linear no R utiliza-se a função *lm*.

O R retorna o valor dos coeficientes α e β estimados via Método de Mínimos Quadrados. Logo, a equação da reta ajustada é dada por $y = 1,81 + 0,10x$.

```
> ajuste = lm(Salario ~ Experiencia)
> ajuste

Call:
lm(formula = Salario ~ Experiencia)

Coefficients:
(Intercept)  Experiencia
    1.8063      0.1008
```


Diagrama de Dispersão

Para esboçar a reta ajustada no diagrama de dispersão, utilize a função *abline*.

```
> plot(Experiencia,Salario)  
> abline(lm(Salario ~ Experiencia))
```

