Noções de Probabilidade e Estatística - Resolução Exercícios Pares

Gledson Luiz Picharski

July 26, 2007

Capitulo 1

Seção 1.1

Exercício 2

a) As crianças do estado de São Paulo são a população de interesse, estão fazendo parte da amostra 200 mães de recém nascidos, não é interessante utilizar esta amostra, pois não é representativa, pode ser que algumas mães estejam no primeiro filho e apenas um posto de saúde não representa o estado todo.

b) A população é o sangue do paciente, a amostra é um pouco deste sangue, como o sangue é homogêneo então esta é uma amostra representativa e podemos tirar conclusões sobre todo o sangue do paciente.

c) A população de interesse são os telespectadores de um programa de TV, a amostra são os 563 indivíduos que foram entrevistados por telefone com relação ao canal em eu estavam sintonizados.Não seria valido inferir atravez desta amostra, pois apenas um seleto grupo esta participando da pesquisa, como a pesquisa é por telefone, pode ser que telespectadores não tenham telefone, ou não quiseram atender, ou então não quiseram atender.

d) Os eleitores brasileiros formam a população, a amostra são as 122 pessoas entrevistadas em Brasília, a amostra não é representativa, para saber a intenção de voto dos brasileiros, precisariamos pesquisar com um número bem maior de pessoas e distribuido entre vários estados, apenas um estado não representa o país todo.

Seção 1.2

Exercício 2

```
> Fisioterapia <- c(7, 8, 5, 6, 4, 5, 7, 7, 6, 8, 6, 5, 5, 4, 5)
> Sequelas <- factor(c(1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0,
+     0), label = c("N", "S"), lev = 0:1)
> Cirurgia <- factor(c(3, 2, 3, 2, 2, 1, 3, 2, 1, 2, 1, 1, 2, 2,
+     3), label = c("B", "M", "A"), lev = 1:3, ord = T)
> dados <- data.frame(Fisioterapia, Sequelas, Cirurgia)
> rm(Fisioterapia, Sequelas, Cirurgia)
> head(dados)

  Fisioterapia Sequelas Cirurgia
1            7        S        A
2            8        S        M
3            5        N        A
4            6        N        M
5            4        N        M
6            5        S        B
```

a) Fisioterapia é quantitativa discreta, Sequelas é qualitativa nominal e Cirurgia é qualitativa ordinal.

b) As Figuras 1, 2 e 3 mostram os gráficos de cada uma das variaveis e as tabelas de frequência são geradas pelos comandos a seguir.

```
> tb1 <- with(dados, table(Fisioterapia))
> tb1

Fisioterapia
4 5 6 7 8
2 5 3 3 2

> barplot(tb1)
```

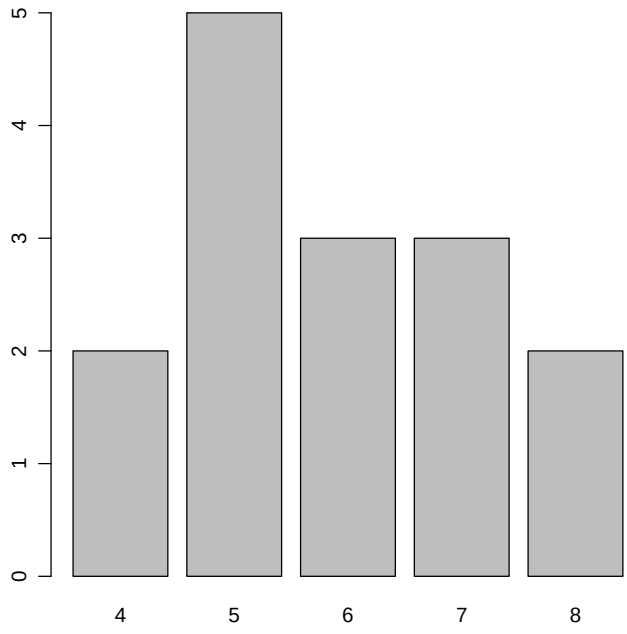

Figure 1: barplot de Fisioterapia

```
> seque.tb <- table(dados$Sequelas)
> seque.tb
> pie(seque.tb)
```

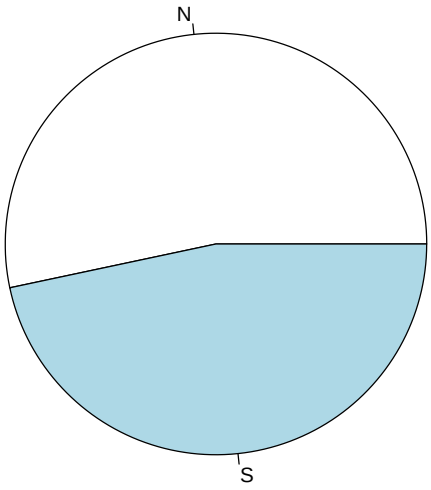

Figure 2: Gráfico sobre Sequelas

```
> cir.tb <- table(dados$Cirurgia)
> cir.tb
> barplot(cir.tb)
```

c) Nota-se que o tempo de fisioterapia é menor nos pacientes sem sequelas, observamos isso na Figura 4

```
> fisio.tb <- table(with(dados, Fisioterapia[Sequelas == "N"]))
> fisio.tb
> barplot(fisio.tb)
```

Exercício 4

Para gerar os dados em classes percebi 3 possibilidades, aqui está resolvido pela que considerei mais simples, as duas outras maneiras estrão no final do capitulo.

```
> freqs <- c(14, 28, 27, 11, 4)
> dados <- rep(0:4 * 2 + 1, freqs)
> dados.tb <- table(cut(dados, seq(0, 10, l = 6)))
> dados.tb

 (0,2]  (2,4]  (4,6]  (6,8] (8,10]
    14     28     27     11      4
```

a) A Figura 5 representa o histograma das notas.

```
> hist(dados, breaks = 0:5 * 2, main = "", xlab = "")
```

b)Primeiramente monto uma tabela de frequência acumulada, descubro o percentual que tirou acima de 4 e acima de 6 então trato a média 5 linearmente e encontro o percentual de aprovados.

```
> freqAc <- cumsum(prop.table(freqs))
> result <- 1 - (freqAc[2] + freqAc[3])/2
> result

[1] 0.3392857
```

Encontramos então aprovação de aproximadamente 0.339.

No histograma da Figura 6 está representado o percentual de aprovados.

```
> cir.tb <- table(dados$Cirurgia)
> cir.tb
> barplot(cir.tb)
```

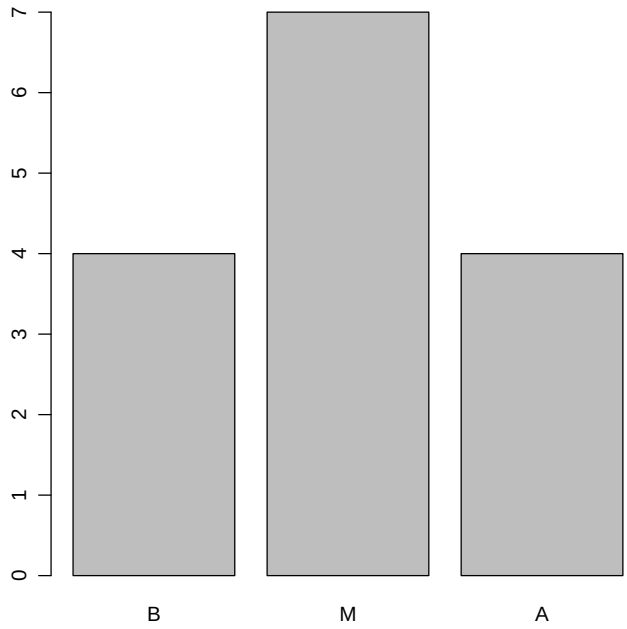

Figure 3: Barplot da tabela de cirurgias

```
> hist(dados, breaks = 0:5 * 2, main = "", xlab = "")
> rect(5, 0, 6, dados.tb[3], col = "gray")
> rect(6, 0, 8, dados.tb[4], col = "gray")
> rect(8, 0, 10, dados.tb[5], col = "gray")
> legend("topright", c("reprovados", "aprovados"), fill = c("white",
+     "gray"))
```

c)A Figura 7 representa o boxplot das notas.

```
> boxplot(dados)
```

```
> fisio.tb <- table(with(dados, Fisioterapia[Sequelas == "N"]))
> fisio.tb
> barplot(fisio.tb)
```

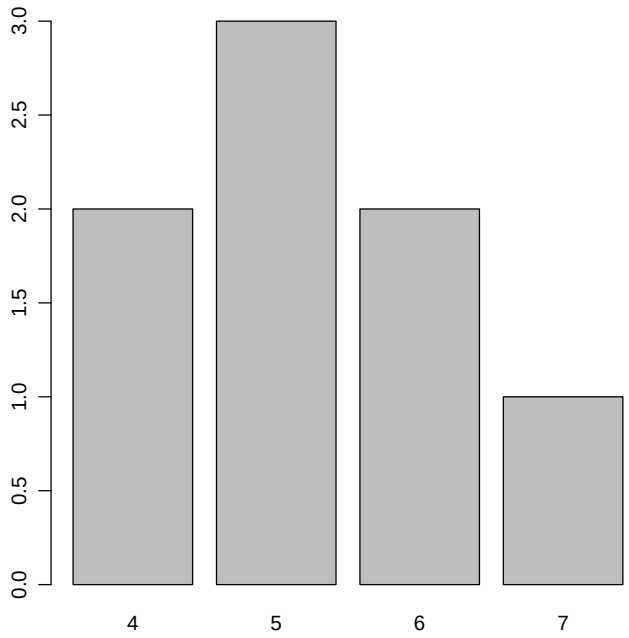

Figure 4: barplot para a variável Fisioterapia.

Seção 1.3

Exercício 2

A tabela da página 7 do livro foi obtida no endereço `http://www.ime.usp.br/~noproest`. É possivel notar que grande parte dos estudantes teêm entre 17 e 18 anos e a quantidade de pessoas é menor quanto maiores são as idades. Percebe-se que aparecem mais pessoas conforme maior a altura até 1,70, alturas entre 1,7 e 1,85 têm frequencia de aproximadamente 4 pessoas a cada 5cm. Nota-se que a maioria das pessoas têm peso entre 50 e 60Kg. Percebe-se ainda que grande parte das pessoas teêm 1 ou 2 filhos. Estes dados estão representados na Figura **??**

```
> tab1.1 <- read.xls("questionario.xls",head=T)

> par(mfrow = c(2, 2))
> with(tab1.1, hist(Idade))
> with(tab1.1, hist(Alt))
> with(tab1.1, hist(Peso))
> with(tab1.1, hist(Filhos))
> par(mfrow = c(1, 1))
```

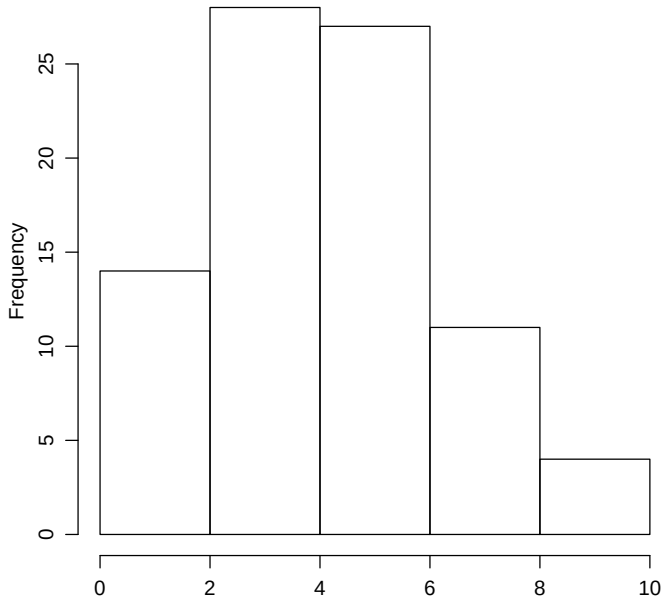

Figure 5: Histograma das notas

Seção 1.4

Exercício 2

Para poder representar os histogramas fiz uma soposição dos dados. Na Figura 9 é possivel verificar que o salário na empresa A está distribuido de forma mais uniforme, o que indica que ela deve pagar mais para pessoas em cargos intermediarios, a empresa B mostra ter um salário inicial um pouco maior e também poucas pessoas ganham mais do que na A, então se eu fosse ser contratado para um auto cargo escolheria A e se fosse para cargos intermediários, que são a maioria ds cargos, escolheria B.

```
> A <- rep(1:9 * 5 + 2.5, c(50, 95, 100, 96, 100, 97, 101, 98,
+     50))
> B <- rep(1:6 * 10 + 5, c(99, 100, 45, 15, 3, 2))

> hist(A, main = "Empresa A")

> hist(B, breaks = 1:7 * 10, main = "Empresa B")
```

Exercício 4

As idades são apresentadas a seguir.

```
> idade <- c(rep(22:28, c(4, 2, 4, 2, 4, 1, 1)), 35, 40)
> idade

 [1] 22 22 22 22 23 23 24 24 24 24 25 25 26 26 26 26 27 28 35 40
```

a)

```
> table(idade)

idade
22 23 24 25 26 27 28 35 40
 4  2  4  2  4  1  1  1  1

> range(idade)

[1] 22 40

> nclass.Sturges(idade)
```

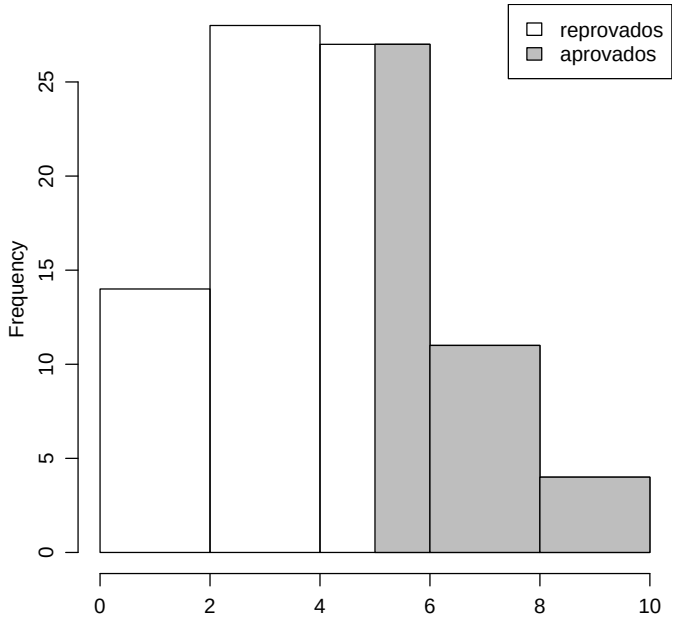

Figure 6: Representação das notas, salientando aprovados.

```
[1] 6

> idade.class <- ordered(cut(idade, seq(21.5, 41.5, 5)))
> idade.class

 [1] (21.5,26.5] (21.5,26.5] (21.5,26.5] (21.5,26.5] (21.5,26.5] (21.5,26.5]
 [7] (21.5,26.5] (21.5,26.5] (21.5,26.5] (21.5,26.5] (21.5,26.5] (21.5,26.5]
[13] (21.5,26.5] (21.5,26.5] (21.5,26.5] (21.5,26.5] (26.5,31.5] (26.5,31.5]
[19] (31.5,36.5] (36.5,41.5]
Levels: (21.5,26.5] < (26.5,31.5] < (31.5,36.5] < (36.5,41.5]

> idade.class.tb <- table(idade.class)
> idade.class.tb

idade.class
(21.5,26.5] (26.5,31.5] (31.5,36.5] (36.5,41.5]
         16           2           1           1
```

b) Usando o box-plot apresentado na Figura 10, percebemos que 35 e 40 são dados atípicos. Na tabela de frequência a seguir é possivel perceber que todas as classes estão com alguma informação, sendo assim o resultado fica melhor resumido, além disso, na tabela de frequência do item a ocorreu um acumulo de frequência nos menores valores e nesta todos os dados são representativos da grande maioria.

```
> boxplot(idade)

> idade1 <- idade[idade < 35]
> idade1

 [1] 22 22 22 22 23 23 24 24 24 24 25 25 26 26 26 26 27 28

> table(idade1)

idade1
22 23 24 25 26 27 28
 4  2  4  2  4  1  1

> range(idade1)

[1] 22 28
```

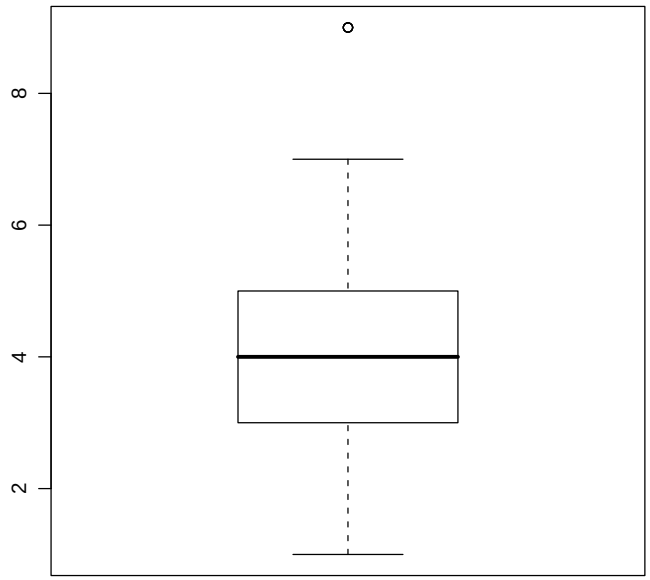

Figure 7: Boxplot das notas.

```
> nclass.Sturges(idade1)

[1] 6

> idade1.class <- ordered(cut(idade1, seq(20.5, 28.5, 4)))
> idade1.class

 [1] (20.5,24.5] (20.5,24.5] (20.5,24.5] (20.5,24.5] (20.5,24.5] (20.5,24.5]
 [7] (20.5,24.5] (20.5,24.5] (20.5,24.5] (20.5,24.5] (24.5,28.5] (24.5,28.5]
[13] (24.5,28.5] (24.5,28.5] (24.5,28.5] (24.5,28.5] (24.5,28.5] (24.5,28.5]
Levels: (20.5,24.5] < (24.5,28.5]

> idade1.class.tb <- table(idade1.class)
> idade1.class.tb

idade1.class
(20.5,24.5] (24.5,28.5]
         10           8
```

Exercício 6

```
> crian <- c(rep(1:5, c(3, 4, 7, 5, 6)), 6, 6, 10, 11)
> crian

 [1]  1  1  1  2  2  2  2  3  3  3  3  3  3  3  4  4  4  4  4  5  5  5  5  5  5
[26]  6  6 10 11
```

a) A tabela de frequência é apresentada a seguir

```
> table(crian)

crian
 1  2  3  4  5  6 10 11
 3  4  7  5  6  2  1  1
```

b) A representação gráfica é mostrada através do box-plot da Figura 11

```
> boxplot(crian)
```

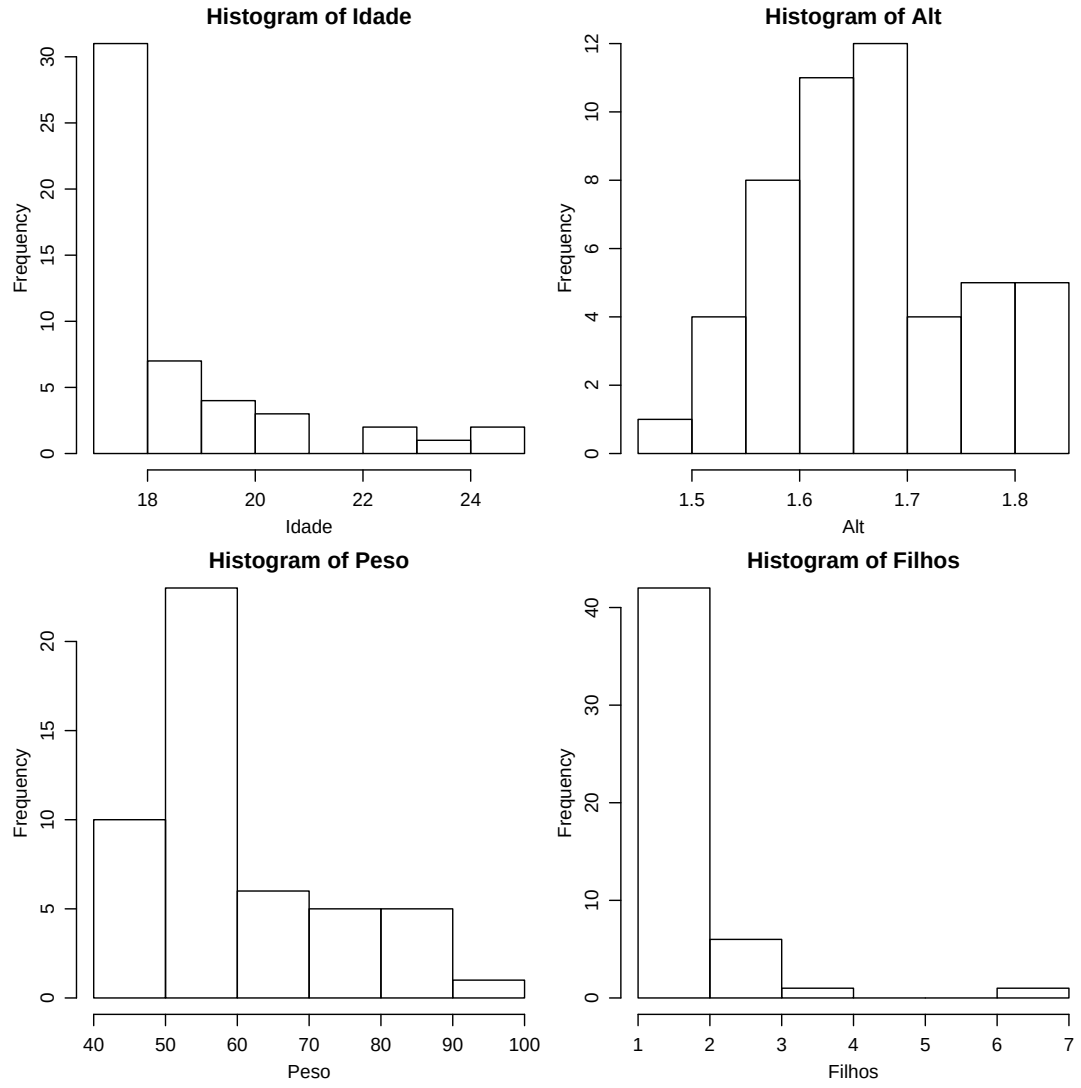

Figure 8:

c) Percebemos que 10 e 11 são valores discrepantes, eles podem ser retirados da amostra, afim de analisar melhor os dados, pois esses valores influênciam as medidas resumo o que interferiria em qualquer tomada de decisão, nota-se por exemplo a diferença entre a média considerando ou não estes valores.

```
> mean(crian)

[1] 3.965517

> mean(crian[crian < 10])

[1] 3.481481
```

Exercício 8

```
> freq <- c(46, 57, 21, 15, 4)
> n.esc <- rep(c(1, 2, 3, 4, 5), freq)
> table(n.esc)

n.esc
 1  2  3  4  5
46 57 21 15  4
```

a) Percebe-se fazendo uma simples operação que em torno de 68% dos alunos cursaram em mais de uma escola.

```
> 1 - cumsum(prop.table(table(n.esc)))[1]
```

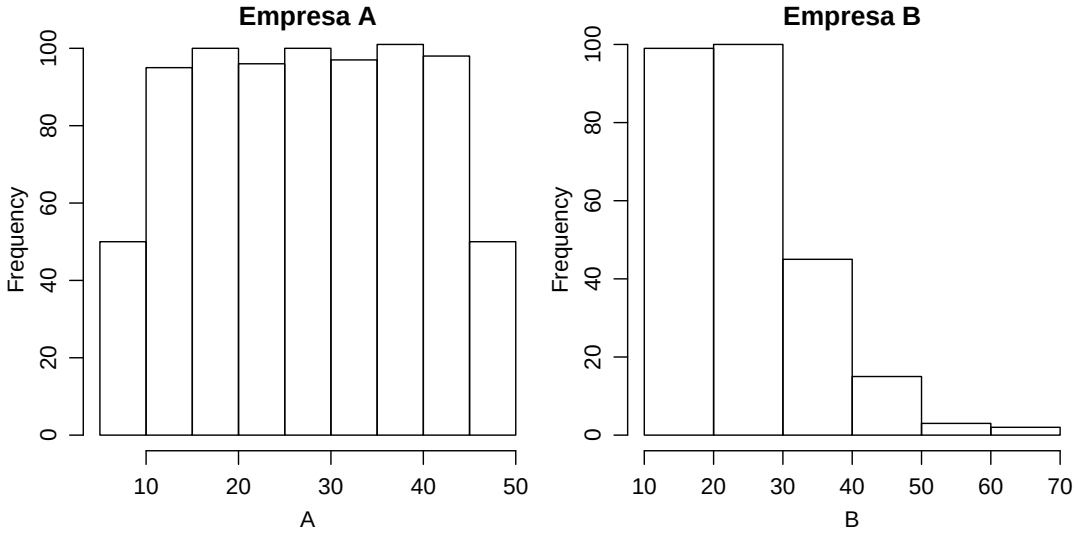

Figure 9: Comparação entre empresas

```
        1
0.6783217
```

b) o gráfico de barras é apontado na Figura 12

```
> barplot(table(n.esc))
```

c) A tabela de frequência é obtida a seguir.

```
> n.esc[n.esc > 2] <- "alta"
> n.esc[n.esc <= 2] <- "baixa"
> table(n.esc)

n.esc
 alta baixa
   40   103
```

Exercício 10

```
> temp <- c(1.1, 1.2, 1.7, 1.5, 0.9, 1.3, 1.4, 1.6, 1.7, 1.6, 1,
+     0.8, 1.5, 1.3, 1.7, 1.6, 1.4, 1.2, 1.2, 1, 0.9, 1.8, 1.7,
+     1.5, 1.3, 1.5)
```

a) A tabela de frequência vem a seguir.

```
> table(temp)

temp
0.8 0.9   1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8
  1   2   2   1   3   3   2   4   3   4   1
```

b) Podemos observar a tabela de frequência por classes a seguir.

```
> range(temp)

[1] 0.8 1.8

> table(ordered(cut(temp, seq(0.8, 1.8, by = 0.2), include.lowest = T)))

  [0.8,1]   (1,1.2] (1.2,1.4] (1.4,1.6] (1.6,1.8]
        5         4         5         7         5
```

c) Percebemos que no item b fica mais fácil de visualizar os dados por eles estarem mais resumidos.

```
> boxplot(idade)
```

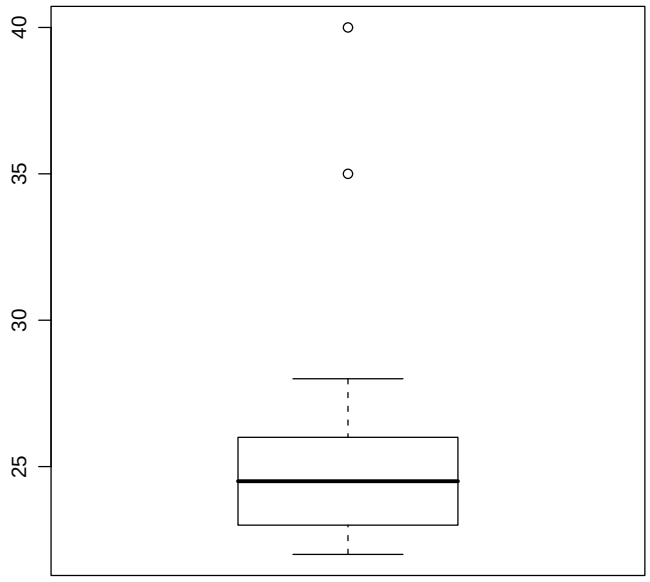

Figure 10: Boxplot representando outliers

d) se tivessemos estas 1000 medidas no digitadas no computador poderiamos ultilizar várias outras jeitos de resumir os dados, entre eles agrupar em poucas classes poderia ser uma solução, mas tentar visualizar todos os 1000 dados não parece ser uma boa alternativa.

Exercício 12

```
> hem <- c(11.1, 12.2, 11.7, 12.5, 13.9, 12.3, 14.4, 13.6, 12.7,
+     12.6, 11.3, 11.7, 12.6, 13.4, 15.2, 13.2, 13, 16.9, 15.8,
+     14.7, 13.5, 12.7, 12.3, 13.5, 15.4, 16.3, 15.2, 12.3, 13.7,
+     14.1)
```

a) Separando os dados em classes de tamanho 1, obtemos 5 classes

```
> range(hem)

[1] 11.1 16.9

> table(ordered(cut(hem, 11:17)))

(11,12] (12,13] (13,14] (14,15] (15,16] (16,17]
      4      10       7       3       4       2
```

b) O Histograma é representado na Figura 13

```
> hist(hem, main = "")
```

c) Atravéz do comando a seguir, podemos verificar a mediana o terceiro quartil e outras medidas resumo.

```
> summary(hem)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  11.10   12.35   13.30   13.46   14.32   16.90
```

d) Obtemos a tabela de acompanhamento médico substituindo os valores numéricos, pelos caracteres sim e não de acordo com a situação, como percebe-se nos comandos a seguir.

```
> hem[hem < 12 | hem > 16] <- "sim"
> hem[hem != "sim"] <- "nao"
> table(hem)
```

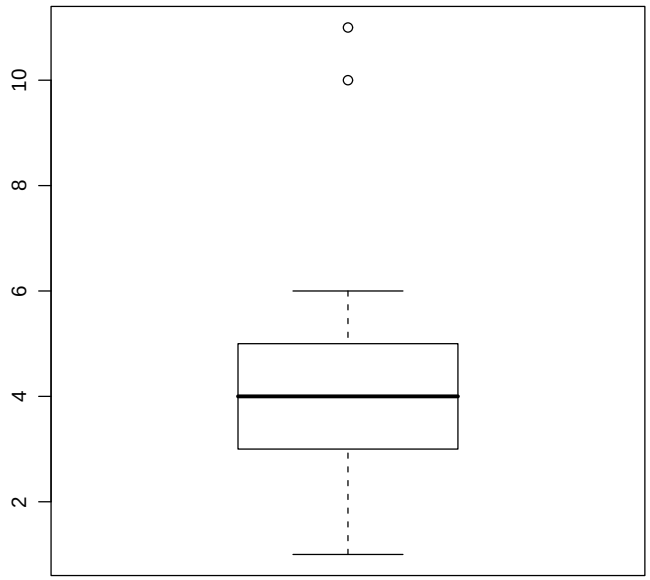

Figure 11: Número de crianças nas familias que utilizam o posto

```
hem
nao sim
 24   6
```

Exercício 14

```
> gols <- c(32, 42, 73, 35, 79, 57, 37, 52, 35, 25, 55, 70, 42,
+     41, 63, 66, 74, 29, 47, 53)
```

a) A variável é quantitativa ordinal, não parece interessante construir uma tabela de frequência com os valores dados, pois eles estariam pouco resumidos e seria quase a mesma coisa que olhar para os valores originais.

b) A tabela de frequência iniciando em 20 e de comprimento 10 é obtida a seguir.

```
> range(gols)

[1] 25 79

> table(ordered(cut(gols, 2:8 * 10)))

(20,30] (30,40] (40,50] (50,60] (60,70] (70,80]
      2       4       4       4       3       3
```

c) O Histograma é obtido na Figura 14

```
> hist(gols)
```

d) Observamos na Figura 15 que Sexprlength(gols[gols > 38])/length(gols) dos times marcaram mais que 38 gols

```
> por <- length(gols[gols > 38])/length(gols)
> por

[1] 0.7

> hist(gols, main = "")
> rect(c(38, 4:7 * 10), rep(0, 5), 4:8 * 10, c(4, 4, 4, 3, 3),
+     col = "gray")
> legend("topright", c("> 38 gols", "< 38 gols"), fill = c("gray",
+     "white"))
```

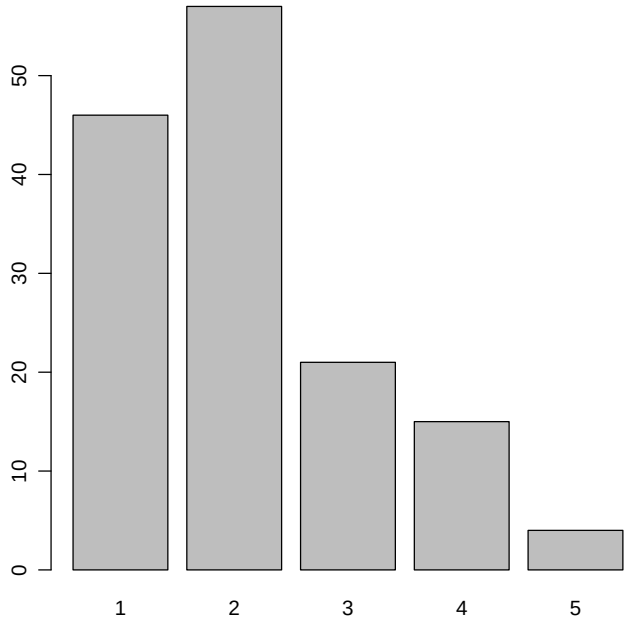

Figure 12: Número de Escolas cursadas pelos alunos.

Exercício 16

Primeiramente suponho que os dados são o ponto médio de cada classe, e em seguida monto as classes através desses valores, é uma suposição bastante válida para as análises a seguir.

```
> freq <- c(8, 20, 42, 18, 10, 2)
> pm <- c(67.5, 0:4 * 5 + 77.5)
> sgerm <- rep(pm, freq)
> sgerm.tb <- table(ordered(cut(sgerm, c(60, seq(75, 100, by = 5)))))
> sgerm.tb
```

```
 (60,75]  (75,80]  (80,85]  (85,90]  (90,95] (95,100]
       8       20       42       18       10        2
```

a) Os dados são continuos, mas as classes podem ser tratadas como dados discretos e podemos melhor visualizar os dados em um gráfico de barras, mostrado pela Figura 16.

```
> barplot(sgerm.tb)
```

b) O Box-plot é mostrado na Figura **??**

```
> boxplot(sgerm)
```

c) Para verificar se a afirmação do fabricante é rasoavel poderia ser feito um teste de hipoteses, mas isso é visto apenas no capitulo 8, então intuitivamente podemos perceber que em média a germinação é bastante próxima da afirmada pelo fabricante.

```
> mean(sgerm)
```

```
[1] 82.5
```

```
> rm("freq", "pm", "sgerm", "sgerm.tb")
```

Exercício 18

```
> esp <- rep(1:5, c(210, 342, 109, 91, 35))
> nesp <- rep(1:5, c(106, 222, 338, 292, 164))
```

a) Os histogramas são mostrados na Figura 18

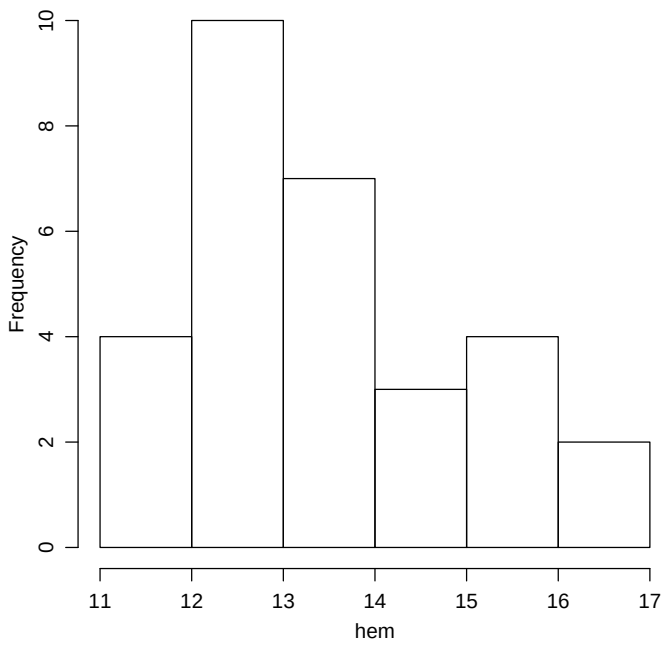

Figure 13: Histograma sobre a Taxa de Hemoglobina

```
> par(mfrow = c(1, 2))
> hist(esp, breaks = 0:5, freq = F, main = "especializados")
> hist(nesp, breaks = 0:5, freq = F, main = "não especializados")
> par(mfrow = c(1, 1))
```

b) Podemos observar o diagrama de barras na Figura **??**

```
> barplot(table(c(nesp, esp)), main = "")
```

c) Percebemos, através do item a, que os trabalhadores especializados trocam menos de emprego do que os não especializados, isso está no fato de termos uma quantidade maior de especializados com menor variação de empregos.

Exercício 20

O número de acerto em cada disciplina de cada aluno é apresentado a seguir.

```
> Port <- c(35, 35, 34, 32, 31, 30, 26, 26, 24, 23, 23, 12, 11,
+     20, 17, 12, 14, 20, 8, 10)
> Mat <- c(31, 29, 27, 28, 28, 26, 30, 28, 25, 23, 21, 32, 31,
+     20, 21, 25, 20, 13, 23, 20)
> notas <- t(matrix(c(Port, Mat), ncol = 2, nrow = 20))
> dimnames(notas) <- list(c("Port", "Mat"), 1:20)
> notas

      1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
Port 35 35 34 32 31 30 26 26 24 23 23 12 11 20 17 12 14 20  8 10
Mat  31 29 27 28 28 26 30 28 25 23 21 32 31 20 21 25 20 13 23 20
```

a) Por termos poucos dados, parece ser de fácil visualisação, por isso não é necessário que os dados sejam separados em classes(mas poderiam), as tabelas são apresentadas a seguir.

```
> table(Port)

Port
 8 10 11 12 14 17 20 23 24 26 30 31 32 34 35
 1  1  1  2  1  1  2  2  1  2  1  1  1  1  2

> table(Mat)
```

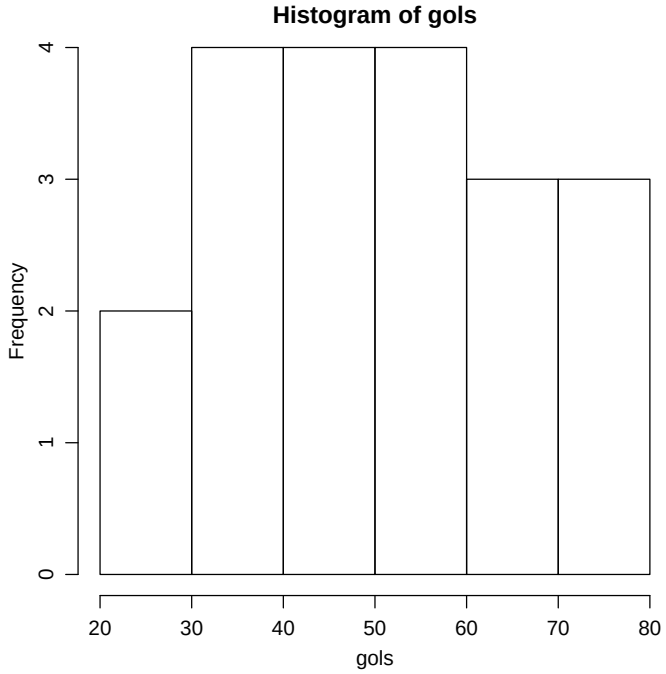

Figure 14: Taxa de Hemoglobina

```
Mat
13 20 21 23 25 26 27 28 29 30 31 32
 1  3  2  2  2  1  1  3  1  1  2  1
```

b) A Figura 20 representa as tabelas obtidas no item a.

```
> par(mfrow = c(1, 2))
> hist(Port, freq = T)
> hist(Mat, freq = T)
> par(mfrow = c(1, 1))
```

c) O total de pontos de cada aluno, pode ser obtido com a soma entre as duas linhas, que representam as disciplinas, da matriz gerada anteriormente, talvez fosse interessante colocarmos os dados em classes, mas por termos pouca informação não considero necessário.

```
> table(notas[1, ] + notas[2, ])

30 31 33 34 37 38 40 42 44 46 49 54 56 59 60 61 64 66
 1  1  1  1  1  1  1  1  2  1  1  1  2  1  1  1  1  1
```

d) Nota-se, nos histogramas do item b, que poucos alunos tiraram notas mais altas em matemática, o que demonstra que eles se sairam melhor em português.

Exercício 22

a) Pelo box-plot apresentado no livro, encontramos medianas de aproximadamente 6,7, 9,5 e 8 para as váriaveis A, B e C

b) O intervalo interquartil pode ser obtido observando o grafico do livro. Entre os pacientes submetidos a cada uma das três técnicas, seu valor é de aproximadamente 2, 0,5 e 1,5 dias para as técnicas A, B e C.

c) O tempo de recuperação para a técnica A é entre 4,8 e 8,7 dias, para a B é entre 8,7 e 10 dias e para a C é entre 6 e 9,5 dias, está variação é dada pelas caracteristicas de cada técnica.

d) Escolheria a técnica A, pois tem um possui um grupo razoavel com menor tempo de recuperação do que as outras técnicas.

Exercício 24

Os arquivos do livro são encontrados em `http://www.ime.usp.br/~noproest`, de onde podemos cancer.xls.

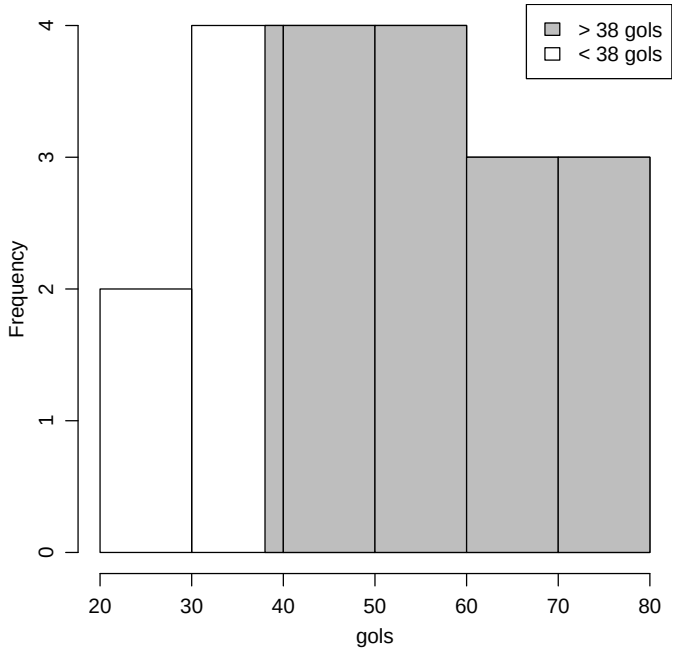

Figure 15: Número de gols.

```
> cancer <- read.xls('cancer.xls',head=T)
> #poderia ser...
> #cancer <- read.xls("http://www.ime.usp.br/~noproest/dados/cancer.xls",head=T)
> #cancer <- read.table("http://www.ime.usp.br/~noproest/dados/cancer.txt",head=T)
> head(cancer) # mostra as linhas iniciais do arquivo

  Ident Grupo Idade  AKP   P  LDH ALB  N  GL
1     1     1    71  8.0 3.2  7.8  62  6 113
2     2     1    66 10.5 5.1 50.1  57  9  93
3     3     1    83  8.5 3.3 15.3  53 21 109
4     4     1    52 12.8 3.2 18.8  45 14  91
5     5     1    61  7.4 4.3 12.9  69 19  78
6     6     1    54  8.1 2.7 15.9  57 10 122

> attach(cancer)
```

a) O Grupo é uma variavel qualitativa nominal, GL é quantitativa continua e Idade é uma quantitativa continua. As Figuras 21, 22 e 23, mostram os histogramas das três variaveis.

```
> table(Grupo)
> hist(Grupo, breaks = 0:4, main = "")

> range(Idade)
> table(ordered(cut(Idade, 0:5 * 20 + 5)))
> hist(Idade, breaks = 0:5 * 20 + 5, main = "")

> range(GL)
> table(ordered(cut(GL, 0:5 * 60, include.lowest = T)))
> hist(GL, breaks = 0:5 * 60, main = "")
```

b) Pela Figura 24, podemos perceber que o grupo com falso-positivos é um pouco mais jovem do que o outro, pois temos um maior quantidade de pessoas mais novas nesse grupo.

```
> range(Idade[Grupo == 1])

[1]  18 101

> range(Idade[Grupo == 4])
```

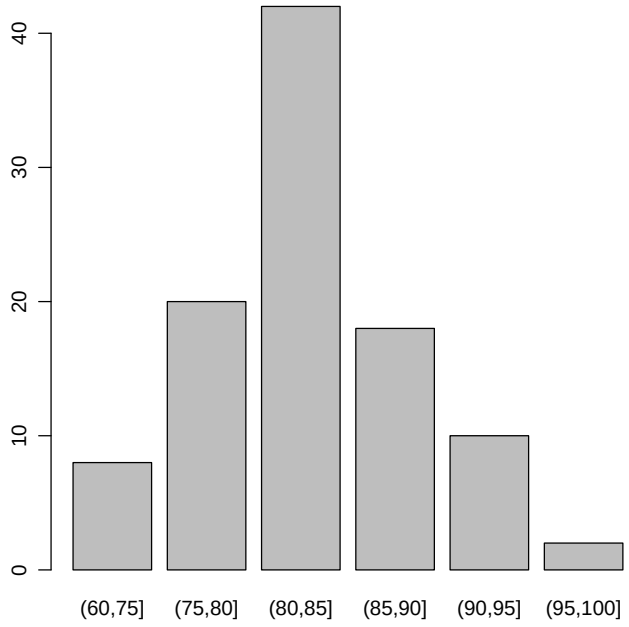

Figure 16: Índice de germinação de sementes de milho do fabricante.

```
[1]  9 88
```

```
> table(ordered(cut(Idade[Grupo == 1], 0:9 * 10 + 15)))
```

```
 (15,25]  (25,35]  (35,45]  (45,55]  (55,65]  (65,75]  (75,85]  (85,95]
       5        7        8        9       13        9        2        2
(95,105]
       1
```

```
> table(ordered(cut(Idade[Grupo == 4], 0:8 * 10 + 5)))
```

```
 (5,15] (15,25] (25,35] (35,45] (45,55] (55,65] (65,75] (75,85]
      1       9       3       7      18      11      10       5
```

```
> par(mfrow = c(1, 2))
> hist(Idade[Grupo == 1], freq = F, main = "falso-negativo")
> hist(Idade[Grupo == 4], freq = F, main = "falso-positivo")
> par(mfrow = c(1, 1))
```

```
> detach(cancer)
```

```
> rm("cancer")
```

Exercício 26

```
> se <- read.xls("aeusp.xls", head = T)
> head(se)
```

```
  Num    Comun Sexo Idade Ecivil X.Reproce X.Temposp X.Resid Trab Ttrab X.Itrab
1   1 JdRaposo    2     4      4  Nordeste        21       9    3    NA       20
2   2 JdRaposo    2     1      1   Sudeste        24       9    1     1       14
3   3 JdRaposo    2     2      1  Nordeste        31       3    1     1       14
4   4 JdRaposo    1     2      2  Nordeste        10       3    1     4       10
5   5 JdRaposo    2     4      2  Nordeste        31       6    1     1       11
6   6 JdRaposo    2     4      2   Sudeste        24       4    2    NA       15
  X.Renda X.Acompu X.Serief
1       1        2        1
2       2        2        7
3       5        2        7
```

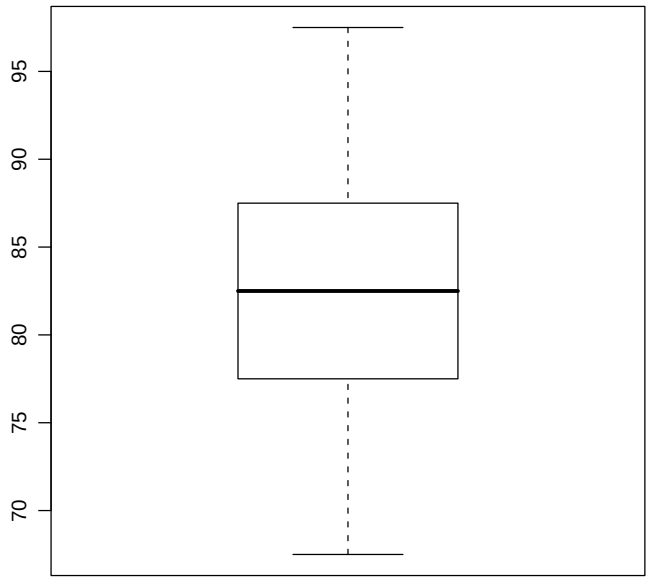

Figure 17: Box-plot sobre as germinações.

```
4       5       2       11
5       6       1        4
6       4       2        4
```

```
> attach(se)
```

a)

Classificação das veriaveis:

quantitativas continuas:Tempo de residencia em SP e Idade que começou a trabalhar.

quantitativas discretas:Faixa de Idade, Número de residencias e faixa da renda familiar.

qualitativas nominais:Comunidade, sexo, estado civil, região de procedência, trabalho, tipo de trabalho e acesso ao computador.

qualitativas ordinais:Série em que parou de estudar.

A seguir faço o teste para verificar se todos os dados apresentados, são possiveis, para os dados não coerentes substituo por NA. Existem outras atitudes que poderiam ser tomadas conforme o caso, os testes de verificação também poderiam ser de várias formas, poderiamos por exemplo tentar perceber se o dado está errado por erro de digitação, ou por que a resposta do individuo foi incoerente, ou pelo pesquisador não ter colotado os dados de forma correta, mas aqui vou assumir que seja o suficiente substituir por NA.

```
> with(se, Sexo[Sexo != 1 & Sexo != 2] <- NA)
> with(se, Idade[Idade < 1 | Idade > 4] <- NA)
> with(se, Ecivil[Ecivil < 1 | Ecivil > 5] <- NA)
> with(se, X.Temposp[X.Temposp[Idade == 1] > 25] <- NA)
> with(se, X.Temposp[X.Temposp[Idade == 2] > 35] <- NA)
> with(se, X.Temposp[X.Temposp[Idade == 3] > 45] <- NA)
> with(se, X.Temposp[X.Temposp[Idade == 4] > Inf] <- NA)
> with(se, Idade[X.Temposp == NA] <- NA)
> with(se, Trab[Trab < 1 | Trab > 3] <- NA)
> with(se, Ttrab[Ttrab < 1 | Ttrab > 5] <- NA)
> with(se, X.Renda[X.Renda < 1 | X.Renda > 6] <- NA)
> with(se, X.Acompu[X.Acompu < 1 | X.Acompu > 2] <- NA)
> with(se, X.Serief[X.Serief < 1 | X.Serief > 12] <- NA)
```

As váriaveis em branco podem aparecer por que o item não foi respondido pelo morador.

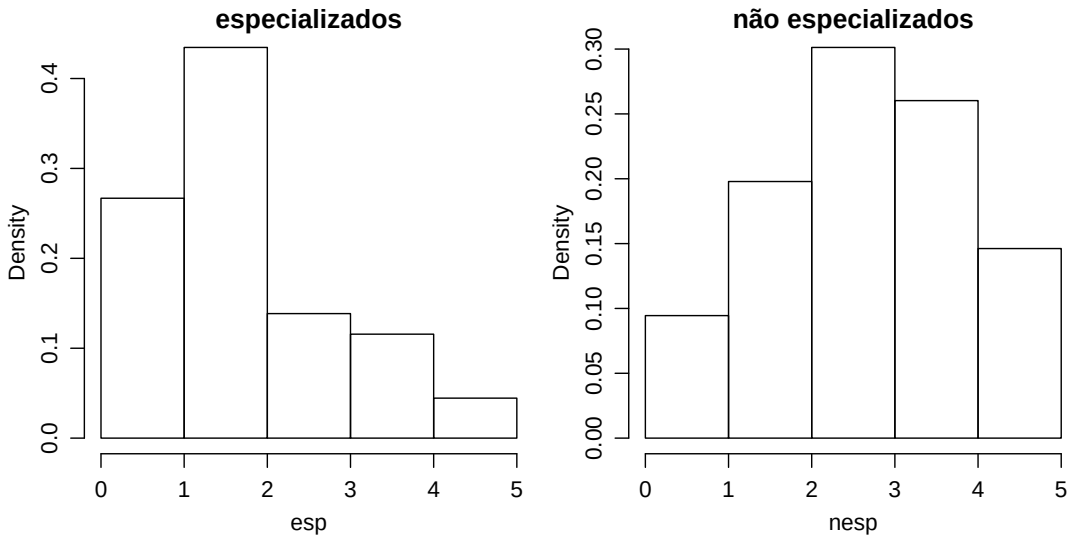

Figure 18: Rotatividade de mão de obra na industria.

b) Pelos histogramas apresentados na Figura é possivel perceber que o Jardim d'Abril tem uma renda um pouco menor

```
> ren.c <- X.Renda[Comun == "Cohab"]
> ren.j <- X.Renda[Comun == "JddAbril"]
> table(ren.c)

ren.c
 1  2  3  4  5  6
 3  7  9 36 17 14

> table(ren.j)

ren.j
 1  2  3  4  5  6
 5 16 10 14  4  1

> par(mfrow = c(1, 2))
> hist(ren.c, breaks = 0:6, main = "Cohab", freq = F)
> hist(ren.j, breaks = 0:6, main = "Jardim d'Abril", freq = F)
> par(mfrow = c(1, 1))
```

c) Podemos verificar na Figura que o tempo de residencia em SP independe do tipo de trabalho, pois o tipo 1 e 4 acumulam aproximadamente o mesma quantia de pessoas com o limite de idade próximo e são bem distintos.

```
> par(mfrow = c(1, 2))
> boxplot(X.Temposp ~ Ttrab)
> stripchart(X.Temposp ~ Ttrab, vertical = TRUE)
> par(mfrow = c(1, 1))
```

d) O boxplot está representado na Figura 27.

```
> boxplot(X.Itrab, main = "")


> detach(se)

> rm("se", "ren.c", "ren.j")
```

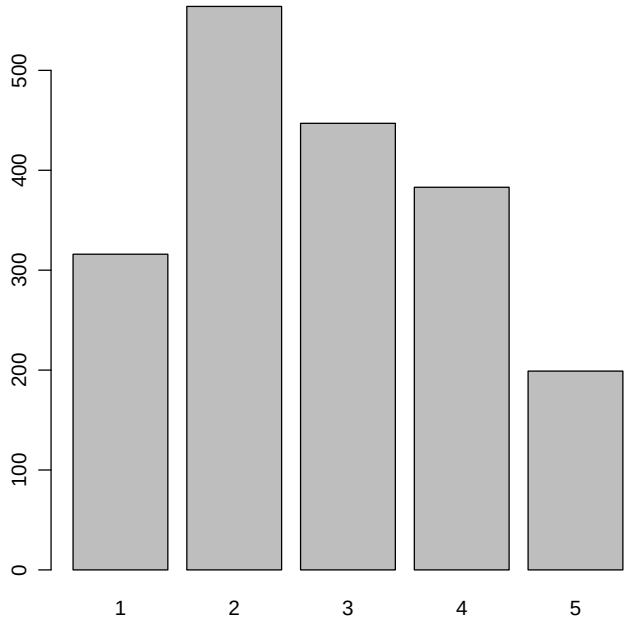

Figure 19: Rotatividade de mão de obra na industria.

Existem muitas soluções para um mesmo exercício, em alguns casos acho interessante fazer uma demonstração de mais casos possíveis.

No caso em que recebemos uma tabela de classe e precisamos tratar dos dados, percebo 3 soluções, uma delas seria pegar o ponto médio de cada classe e gerar ele conforme a frequência que a classe a parece, foi essa a possição que assumi e está resolvido em exercícios como o 4 da seção 1.2, outras soluções seriam pegar números espaçados igualmente dentro de cada classe, ou então pegar números aleatórios dentro de cada classe, vou usar o exercício citado para fazer isto.

1.2.4

Solução 2

Dentro da primeira classe tem 14 números igualmente espaçados, e assi ocorre para cada classe com a quantidade de números relativos a sua frequência.

```
> freq <- c(14, 28, 27, 11, 4)
> x <- paste("a", 1:5, sep = "")
> for (i in 1:5) (assign(x[i], seq((0:4 * 2.001)[i], (1:5 * 2)[i],
+     l = freq[i])))
> y <- matrix(unlist(sapply(x, get)))
> table(ordered(cut(y, seq(0, 10, by = 2), include.lowest = T)))

 [0,2]  (2,4]  (4,6]  (6,8] (8,10]
    14     28     27     11      4

> hist(y, breaks = 0:5 * 2)
```

Solução 3

Muito semelhante a anterior, mas agora os números foram gerados de forma aleatória dentro de cada classe.

```
> freq <- c(14, 28, 27, 11, 4)
> x <- paste("a", 1:5, sep = "")
> for (i in 1:5) (assign(x[i], runif(freq[i], (0:4 * 2.001)[i],
+     (1:5 * 2)[i])))
> y <- matrix(unlist(sapply(x, get)))
> table(ordered(cut(y, seq(0, 10, by = 2), include.lowest = T)))

 [0,2]  (2,4]  (4,6]  (6,8] (8,10]
    14     28     27     11      4

> hist(y, breaks = 0:5 * 2)
```

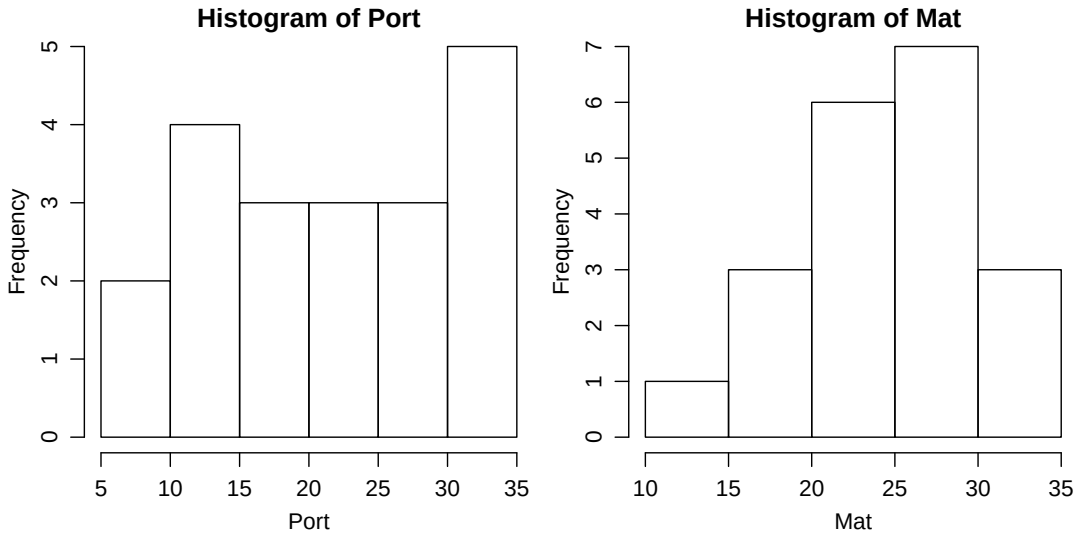

Figure 20: Comparação de Notas.

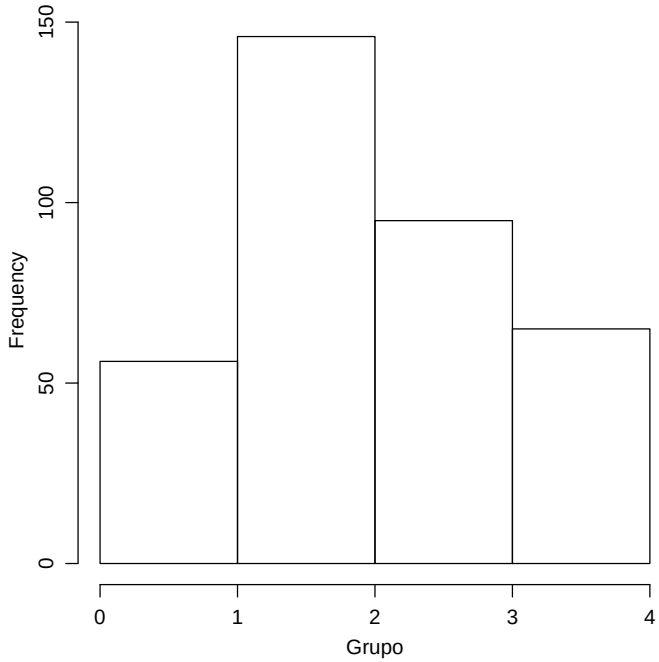

Figure 21: Grupos de diagnóstico.

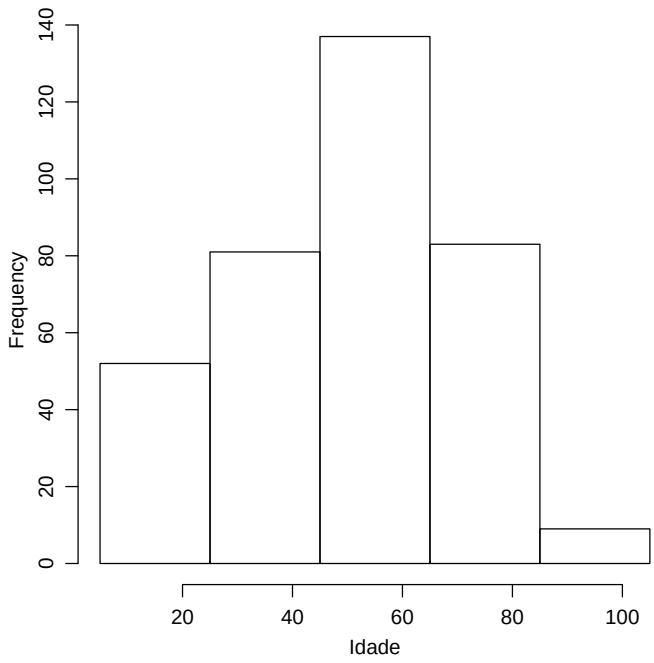

Figure 22: Idades dos Pacientes.

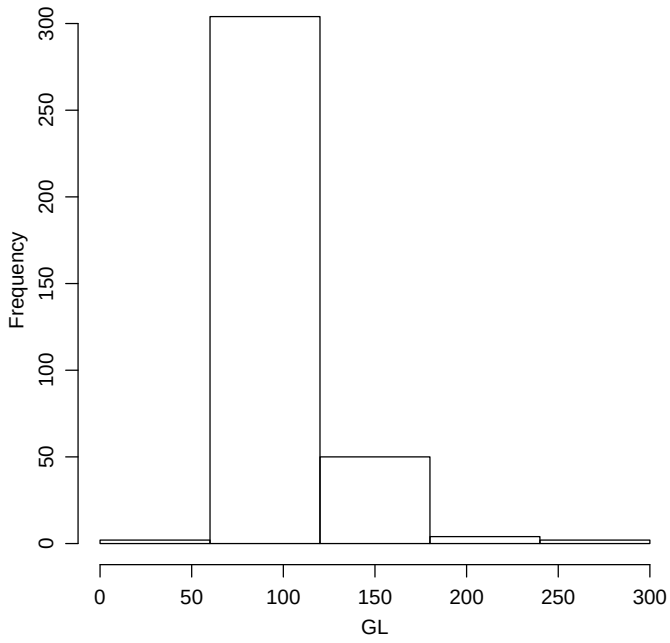

Figure 23: Glicose no sangue dos paciêntes.

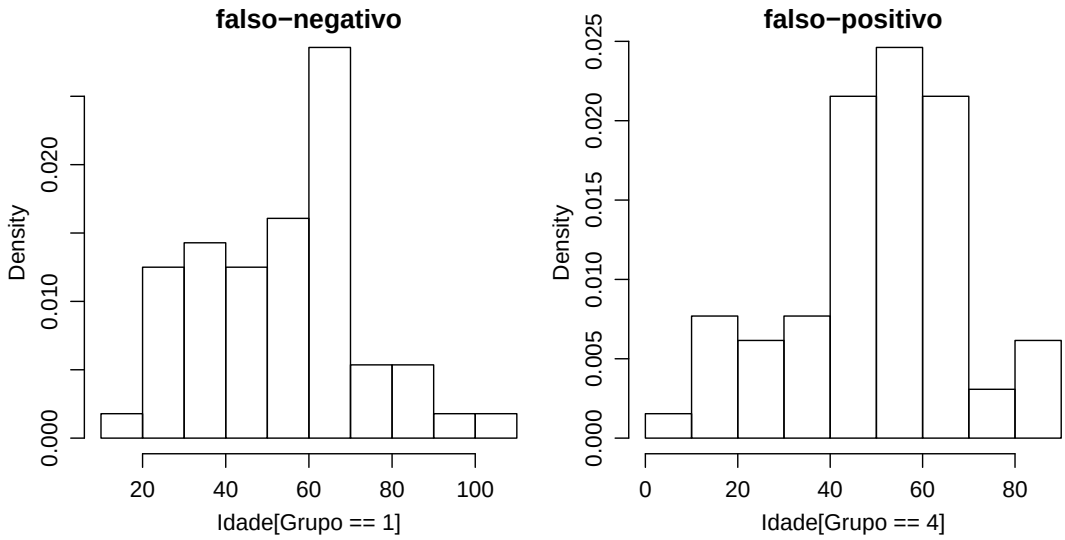

Figure 24: Comparativode idade entre falso-negativo e falso-positivo

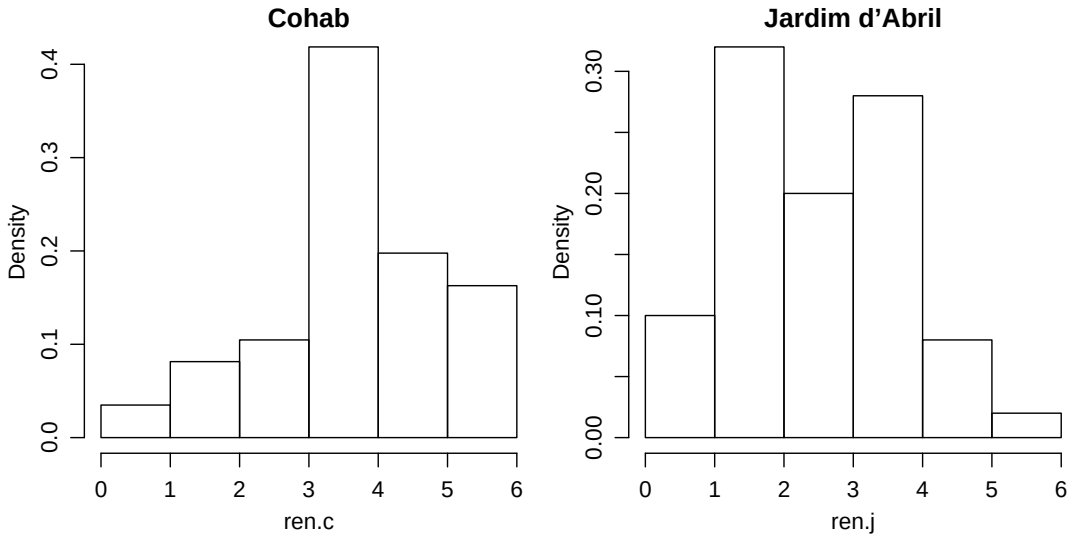

Figure 25: Comparativo de renda entre os dois bairros.

É importante observar que as duas primeiras soluções fornecem as mesmas médias que o livro ensina calcular, já a solução 3 a média pode se distinguir, pois os dados supostos foram gereados aleatóriamente dentro de cada classe.

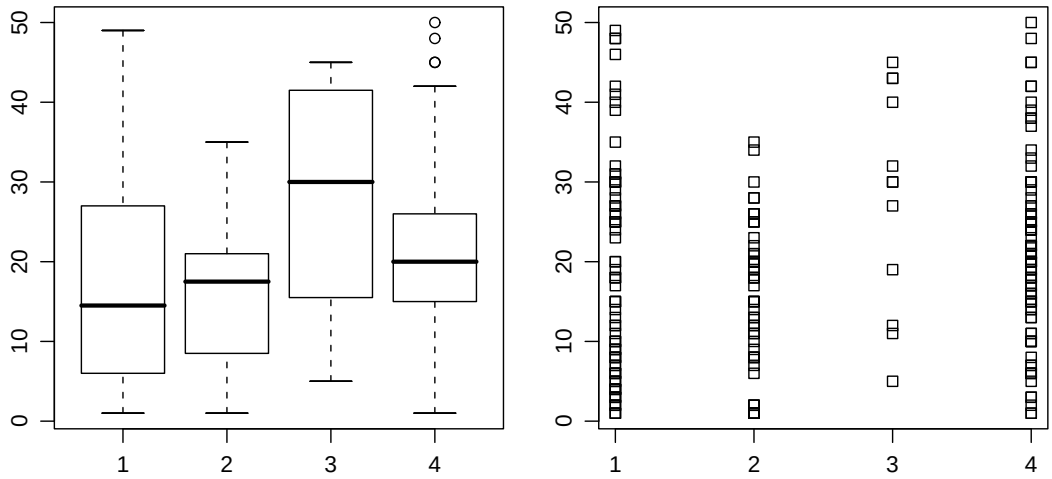

Figure 26: Camparação entre tempo em SP e tipo de trabalho.

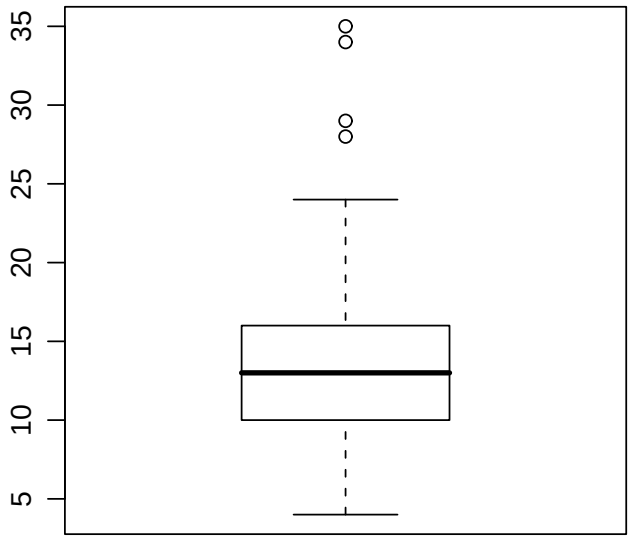

Figure 27: Idade em que começou a trabalhar.