Análise de Dados com o Software R:
Métodos Estatísticos, Computacionais e Econométricos

Prof. Adriano Azevedo Filho (adrianoazevedofilho@gmail.com)

Preparação do Arquivo para Análise

```
Conteúdo do Módulo

  1 - Lendo o arquivo CSV BR denominado "ODB2013originalcorrigido.csv"
      que está no seu computador

  2 - Verificando se o arquivo foi lido corretamente

  3 - Tipos de valores e criação de vetores de interesse
      numeric, character, logic, NA, Inf, NaN, número:número, rep, seq

  4 - Modificando nomes e ordem das categorias (níveis) de variáveis
qualitativas

  5 - Modificando os nomes de níveis de variáveis com denominação
muito longa e redefinição de níveis

  6 - Filtros, seleções e estatísticas (medidas-resumo) incondicionais
e condicionais

  7 - Análise básica de frequências

  8 - Tratando a situação de mais de uma resposta por valor informado
no questionário

  9 - Salvando o data.frame modificado em arquivo no seu computador
      write.table, getwd, setwd
```

Os comandos também estão apresentados em arquivo texto (clique aqui para abrir)

1 - Lendo um arquivo CSV BR denominado "ODB2013originalcorrigido.csv" que está no seu computador

Baixe o arquivo seguindo as instruções dadas e salve numa pasta de fácil acesso no seu computador.

O nome do arquivo que usaremos é "ODB2013originalcorrigido.csv", se o arquivo csv (BR) estiver no seu próprio computador, encontre o nome do arquivo usando

```
rm(list=ls())   ## apaga (quase tudo) antes de uma nova análise (boa
prática)
nomearquivo<-file.choose()       #seleção de arquivo via menus
```

A função "file.choose" abre um menú que possibilita a escolha de um caminho que define um arquivo no seu computador.

Após a execução bem sucedida dessa linha, o conjunto de dados estará "contido" no objeto "alunos", um "data.frame".

```
o símbolo "<-" indica: guarde o objeto definido pelo
conteúdo da expressão da direita (no caso um "data.frame")
```

```
em uma região da memória do computador indentificada pelo
"identificador" especificado à esquerda ("alunos")
regras para identificadores: não exceda 20 caracteres
                             use só letras, números e
                             pontos, evite acentos
                             nota: isso é só uma sugestão,
                             há mais flexibilidade no R

ex: altura, peso.ind, pesoInd, renda.pessoal2, rendaPess
```

2 - Alternativa: Leitura do arquivo de site na internet (seu sistema deve permitir, caso contrário use o modo do tópico anterior)

Se não tiver restrições de acesso (ex. fora da rede corporativa), pode ler o arquivo de site na internet, usando:

```
caminho <- "http://s3.amazonaws.com/ihbs-
html/dados/ODB2013originalcorrigido.csv"
alunos <- read.csv2(caminho, fileEncoding = "latin1")
```

A opção fileEncoding="latin1" é em geral desnecessária em sistemas Windows configurado para o Brasil, mas pode ser importante para outros sistemas, que foram configurados para UTF-8 (não se preocupe se não entender isso).

2 - Verificando se o arquivo foi lido corretamente

É sempre oportuno verificar se o arquivo foi lido corretamente. Pode ter havido algum problema na conversão do arquivo original para o formato csv, ou problemas na opção de codificação. Algumas funções do R facilitam a verificação de potenciais problemas.

```
dim(alunos)  ## mostra número de linhas e colunas
## [1] 23 50
names(alunos)  ## mostra nomes das variáveis no data.frame
##  [1] "Indicação.de.data.e.hora"
##  [2] "Número"
##  [3] "Locais.principais.de.trabalho"
##  [4] "Sexo"
##  [5] "Data.de.nascimento"
##  [6] "Altura"
##  [7] "Peso"
##  [8] "Número.do.calçado.que.calça"
##  [9] "Circunferência.da.barriga..em.cm...na.altura.do.umbigo"
## [10] "Com.relação.ao.uso.das.mãos.você.é"
## [11] "Ensino.fundamental.e.médio..número.de.anos.em.escola.pública"
## [12] "Formação.acadêmica"
## [13] "Como.se.classificaria.como.aluno.a..na.graduação."
## [14] "Status.de.sua.formação.na.graduação"
## [15] "Tipo.de.escola.de.graduação.cursada"
## [16] "Estudos.de.pós.graduação"
## [17]
"Descreva.as.áreas.em.que.cursou.pós.e.local..caso.tenha.cursado."
## [18] "Como.avalia.sua.habilidade.de.comunicação."
## [19] "Como.avalia.a.sua.habilidade.com.métodos.quantitativos."
## [20] "Estado"
## [21] "Tipo.de.cidade.em.que.viveu.a.maior.parte.de.sua.vida"
## [22] "classe"
## [23] "Grau.de.instrução.maior.do.pai.ou.da.mãe"
## [24] "Conhecimento.de.Inglês...Leitura"
```

```
## [25]
"Conhecimento.de.Inglês...compreenção.auditiva.da.lingua.falada"
## [26] "Conhecimento.de.inglês...conversação"
## [27] "Conhecimento.de.inglês...habilidade.em.escrever"
## [28] "Outras.línguas.com.proficiência.elementar.ou.intermediária"
## [29] "Outras.línguas.com.proficiência.muito.boa.ou.excelente"
## [30] "Religião"
## [31]
"Se.respondeu..Outra...na..pergunta.anterior..especifique.qual.é"
## [32] "Tipo.de.música.preferida"
## [33]
"Indique.outras.músicas.que.gosta.caso.não.tenham.sido.especificadas.no.ítem.anterior"
## [34] "Hobbies.prediletos"
## [35]
"Indique.outros.Hobbies..caso.não.tenham.sido.descritos.na.pergunta.anterior"
## [36] "Fumo"
## [37]
"Consumo.de.bebida.alcoólica..indique.o.número.de.doses.consumidas..por.semana."
## [38] "Animal.de.estimação"
## [39] "Time.de.futebol.para.o.qual.torçe"
## [40] "Satisfação.pessoal.com.a.profissão.que.escolheu"
## [41] "Quando.ingressou.na.Odebrecht"
## [42] "Forma.de.ingresso.na.Odebrecht"
## [43] "Quantos.lançamentos.teve.desde.o.seu.ingresso.na.Odebrecht"
## [44] "Área.em.que.trabalha"
## [45]
"Indique.o.tempo.em.minutos.que.gasta.por.dia.para.ir.e.voltar.da.sua.casa.para.a.empresa"
## [46] "Qual.o.seu.custo.mensal.de.moradia.e.alimentação."
## [47]
"O.que.já.contribuiu.para.o.sucesso.da.empresa.em.que.trabalha"
## [48]
"Se.tiver..descreva.outras.habilidades.profissionais.importantes.que.têm.e.que..não.tenham.sido.abordadas.em.outras.questões"
## [49] "Quantos.livros.leu.em.2013"
## [50]
"Caso.tenha.lido.algum.livro.em.2013..indique.o.título.dos.livros.que.leu.em.2013..separados.por.vírgula"
alunos[1:3, 1:4]  ## mostra linhas 1 a 4 e colunas 1 a 5 do data.frame
##   Indicação.de.data.e.hora Número Locais.principais.de.trabalho
Sexo
## 1        7/11/2013 8:26:49  67788                       UCR, UAT
Masculino
## 2        7/11/2013 8:56:32  65790                            UCR
Feminino
## 3       7/16/2013 12:46:07  65788                            UCR
Masculino
```

3 - Tipos de valores e criação de vetores de interesse

Antes de prosseguirmos é importante explicar que o R considera 3 tipos de valores principais: numérico (“numeric”), que pode ser inteiro ou real, texto (“character”), lógic (“logic”). Variáveis que representam vetores contém coleções ordenadas de valores de um só tipo (listas, um outro tipo de objeto do R pode conter tipos diferentes). Textos envolvidos com aspas são frequentemente denominados “strings” na linguagem usada em computação.

Os vetores podem ser lidos externamente, em data.frames, ou definidos através do comando “c”, de “concatenate” como ilustrado a seguir:

```
x <- c(2, 4.2, 5, 4.88)  ## criando um vetor tipo numeric
x
## [1] 2.00 4.20 5.00 4.88
x <- c("a", "carro", "c2", "22")  ## criando um vetor com texto (tipo
character)
x
## [1] "a"     "carro" "c2"    "22"
x <- as.factor(c("a", "carro", "c2", "22"))  ## criando um fator (veja
a diferença)
x
## [1] a     carro c2    22
## Levels: 22 a c2 carro
x <- c(FALSE, TRUE, TRUE)  ## criando um vetor com valores lógicos
(TRUE e FALSE)
x
## [1] FALSE  TRUE  TRUE
```

O R também reconhece 3 valores especiais para indicar situações que ocorrem na prática da análise de dados, que são codificados com os símbolos

- NA - um valor que indica que o valor é inexistente (not available)
- Inf - infinito (resultado de uma operação como 1/0)
- NaN - resultado indefinido (resultado de uma operação como log(-1))

Alguns exemplos de operações em vetores envolvendo esses valores

```
x <- c(NA, 9, 16, -1)
x
## [1] NA  9 16 -1
sqrt(x)  # raiz
## Warning: NaNs produced
## [1]  NA   3   4 NaN
x/0
## [1]   NA  Inf  Inf -Inf
```

No caso da operação sqrt(x), o R também produziu um aviso (warning), para alertar o usuário sobre o resultado da operação que produziu um NaN.

Comandos adicionais para criação de vetores

Em muitas situações de análise, é necessário recorrer ao uso de certos vetores especiais realização de operações.

Vetores com sequências de números inteiros [número inteiro:número inteiro]

```
1:10  ## situação bem usual
##  [1]  1  2  3  4  5  6  7  8  9 10
-2:4  ## sequência iniciando com número negativo
## [1] -2 -1  0  1  2  3  4
10:1  ## sequência descrescente
##  [1] 10  9  8  7  6  5  4  3  2  1
```

Vetores com valores repetidos [rep(valor,repetições)]

```
rep(1, 5)
```

```
## [1] 1 1 1 1 1
rep("gato", 3)
## [1] "gato" "gato" "gato"
```

Vetores com valores reais com espaçamento constante [seq(início,fim,espaçamento)]

```
seq(1, 2, 0.2)
## [1] 1.0 1.2 1.4 1.6 1.8 2.0
seq(-2, 2, 0.5)
## [1] -2.0 -1.5 -1.0 -0.5  0.0  0.5  1.0  1.5  2.0
seq(3, 2, -0.1)
##  [1] 3.0 2.9 2.8 2.7 2.6 2.5 2.4 2.3 2.2 2.1 2.0
```

3 - Modificando nomes das variáveis no data.frame

Vamos modificar os nomes das variáveis para nomes mais compactos para facilitar o manuseio das variáveis nas análises. Além disso, como os nomes são usados como rótulos para identificação das variáveis nos resultados (tabelas, gráficos, etc.), nomes muito longos podem ser inconvenientes.

Antes de fazer a alteração, vamos armazenar os nomes originais num vetor:

```
nomeorig <- names(alunos)  # preservando nomes originais
```

Vamos agora fazer a alteração dos nomes:

```
novonome <- c("dh", "num", "loc", "sex", "dan", "alt", "pes", "cal",
"cir",
    "mao", "pub", "fac", "alu", "sta", "uni", "pg1", "pg2", "hco",
"hmq", "est",
    "cid", "cso", "ipa", "in1", "in2", "in3", "in4", "ol1", "ol2",
"rel1", "rel2",
    "mu1", "mu2", "ho1", "ho2", "fum", "alc", "ani", "tim", "sat",
"odi", "odf",
    "odl", "oda", "odt", "cus", "con", "out", "nlv", "liv")
names(alunos) <- novonome
names(alunos)  ## mostrando os novos nomes
##  [1] "dh"   "num"  "loc"  "sex"  "dan"  "alt"  "pes"  "cal"  "cir"
"mao"
## [11] "pub"  "fac"  "alu"  "sta"  "uni"  "pg1"  "pg2"  "hco"  "hmq"
"est"
## [21] "cid"  "cso"  "ipa"  "in1"  "in2"  "in3"  "in4"  "ol1"  "ol2"
"rel1"
## [31] "rel2" "mu1"  "mu2"  "ho1"  "ho2"  "fum"  "alc"  "ani"  "tim"
"sat"
## [41] "odi"  "odf"  "odl"  "oda"  "odt"  "cus"  "con"  "out"  "nlv"
"liv"
A função "c" do R cria um vetor, concatenando elementos separados por
vírgulas.
No caso, criou-se um vetor de textos ou "strings" na linguagem usada
em computação.
Cada "string" substituirá o "string" original que definia o nome da
variável,
na ordem mostrada quando da primeira execução de names(alunos)
```

4 - Modificando nomes e ordem das categorias (níveis) de variáveis qualitativas

Da mesma forma que fizemos para os nomes das variáveis, podemos também modificar os nomes ou identificadores que caracterizam as"categorias" ou “níveis” de variáveis qualitativas (também chamadas de fatores em estatística).

Para acesso a cada variável, precedemos o nome da variável com o nome do data.frame ao qual ela pertence, separando os dois nomes com o símbolo “$”. Podemos ver os valores da variável qualitativa alunos$sex (note a alteração do nome), usando:

```
alunos$sex
##  [1] Masculino Feminino  Masculino Feminino  Masculino Masculino
Masculino
##  [8] Masculino Feminino  Masculino Feminino  Feminino  Masculino
Feminino
## [15] Masculino Feminino  Feminino  Feminino  Masculino Masculino
Masculino
## [22] Masculino Masculino
## Levels: Feminino Masculino
```

Para alterar os nomes dessas categorias, que indicam o sexo do aluno, para “f” e “m”, letras minúsculas, inicialmente observe a ordem em que os níveis aparecem e faça as modificações desejadas como indicado abaixo:

```
levels(alunos$sex)  ## mostra os níveis ou categorias da variável sex
no data.frame alunos
## [1] "Feminino"  "Masculino"
levels(alunos$sex) <- c("f", "m")  # troca por identificadores mais
sintéticos
levels(alunos$sex)  ## mostrando os novos nomes, já alterados na
variável
## [1] "f" "m"
```

O R assume uma ordem para os níveis, a qual é a apresentada quando o comando levels é utilizado, como acabamos de fazer. Para mudar essa ordem, que é algo que pode ser interessante em algumas análises, podemos usar (nesse caso):

```
alunos$sex <- factor(alunos$sex, levels(alunos$sex)[c(2, 1)])  #
reordenação de níveis
levels(alunos$sex)  # note a reordenação abaixo
## [1] "m" "f"
```

O vetor c(2,1) mostra as novas posições para os níveis originais. O nível 1 vai para 2 e o nível 2 vai para 1. Podemos proceder de forma similar se existirem mais níveis.

Para retornar à forma anterior, usamos novamente

```
alunos$sex <- factor(alunos$sex, levels(alunos$sex)[c(2, 1)])
levels(alunos$sex)  # note o retorno à ordem anterior
## [1] "f" "m"
```

5 - Modificando os nomes de níveis de variáveis com denominação muito longa e redefinição de níveis

```
levels(alunos$loc)
## [1] "UAT"      "UCR"      "UCR, UAT"
levels(alunos$loc) <- c("at", "cr", "po")

levels(alunos$mao)
## [1] "Canhoto (usa a mão esquerda para escrever)"
```

```
## [2] "Destro (usa a mão direita para escrever)"
levels(alunos$mao) <- c("c", "d")

levels(alunos$fac)
##  [1] "Administração de Empresas"    "Engenharia Agronômica"
##  [3] "Engenharia Agrícola"          "Engenharia Ambiental"
##  [5] "Engenharia Eletrica"          "Engenharia Mecatrônica"
##  [7] "Engenharia Mecânica"          "Engenharia Produção Mecânica"
##  [9] "Engenharia Química"           "Engenharia de Alimentos"
## [11] "Engenharia de Automação"      "Engenharia de Meio Ambiente"
levels(alunos$fac) <- c("adm", "eagri", "eagro", "eamb", "eali",
"eauto", "emambi",
    "ee", "emec", "emeca", "eprod", "equi")


## Criando novas variáveis (tempg e cures) para facilitar a análise
dos
## cursos de pg
levels(alunos$pg1)
## [1] "Completei 1 ou mais cursos de especialização"
## [2] "Completei 1 ou mais cursos de especialização, Estou cursando a
pós em Engenharia de Segurança"
## [3] "Estou cursando a pós em Engenharia de Segurança"
## [4] "Estou cursando a pós em Engenharia de Segurança, Estou
cursando um ou mais cursos de especialização"
## [5] "Estou cursando um ou mais cursos de especialização"
## [6] "Nunca cursei Pós graduação"
## [7] "Tenho mestrado, Estou cursando a pós em Engenharia de
Segurança, Estou cursando um ou mais cursos de especialização"
alunos$tempg <- alunos$pg1
levels(alunos$tempg) <- c("esp", "esp", "cur", "cur", "cur", "nc",
"msc")
alunos$cures <- alunos$pg1
levels(alunos$cures) <- c("n", "s", "s", "s", "n", "n", "s")

# Criando variável alunos$itot com os pontos totais no inglês
alunos$itot <- alunos$in1 + alunos$in2 + alunos$in3 + alunos$in4

levels(alunos$ani)
## [1] "Não tenho animal de estimação" "Tenho um ou mais cachorros"
levels(alunos$ani) <- c("n", "s")

levels(alunos$tim)
##  [1] "Atlético Mineiro"             "Corinthians"
##  [3] "Cruzeiro"                     "E C Vitória"
##  [5] "Flamengo"                     "Goiás"
##  [7] "Grêmio"                       "Não me interesso por futebol"
##  [9] "Outro time"                   "Palmeiras"
## [11] "Santos"                       "São Paulo"
levels(alunos$tim) <- c("am", "co", "cr", "vi", "fl", "go", "gr",
"ni", "ot",
    "pa", "sa", "sp")

levels(alunos$odf)
## [1] "Jovem Parceiro" "Outras formas"
levels(alunos$odf) <- c("j", "o")

levels(alunos$oda)
## [1] "Agrícola (Operação)"
## [2] "Agrícola (Planejamento e/ou Controle)"
```

```
## [3] "Ambiente"
## [4] "Indústria (Planejamento e/ou Controle)"
## [5] "Manutenção Automotiva"
## [6] "Manutenção Automotiva, Manutenção Industrial"
## [7] "Manutenção Industrial"
## [8] "Parcerias e Fornecedores"
levels(alunos$oda) <- c("agop", "agpc", "ambi", "inpc", "mana",
"manai", "mani",
    "parf")
```

6 - Filtros, seleções e estatísticas (medidas-resumo) incondicionais e condicionais

Um recurso forte da linguagem do R envolve a facilidade de se observar, modificar e filtrar observações de variáveis a partir de critérios lógicos definidos, assim como possibilitar a obtenção de **estatísticas condicionais**. Alguns exemplos serão dados a seguir para ilustrar as possibilidades.

Observação e modificação

```
## Acesso a observação 3 da variável alunos$alt
alunos$alt[3]
## [1] 1.89
## Acesso às observações 2, 3 e 7
alunos$alt[c(2, 3, 7)]
## [1] 1.60 1.89 1.87
## Modificando valores de vetores
alt2 <- alunos$alt[c(2, 3, 7)]  ## criando uma réplica de alunos$alt
alt2[c(2, 3, 7)] <- c(1.5, 1.72, 1.8)  ## alterando as observações 2,
3 e 7
prop.table(table(alunos$sex))
##
##      f      m
## 0.3913 0.6087
```

Filtros

Quando usamos um vetor com valores lógicos como argumento para os índices de variáveis, extraímos os valores da variável que coincidem com o resultado lógico TRUE (Verdade), obtido pela aplicação do teste.

Se quisermos obter os valores de altura para as observações associadas a mulheres usaríamos:

```
## Observações de altura dos alunos do sexo feminino
alunos$alt[alunos$sex == "f"]
## [1] 1.60 1.65 1.69 1.64 1.60 1.70 1.58 1.64 1.58
```

Note que a avaliação de **alunos$sex=="f"** resultará em

```
## Observações de altura dos alunos do sexo feminino
alunos$sex == "f"
##  [1] FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE
TRUE
## [12]  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE
FALSE
## [23] FALSE
```

Somente os valores de **alunos$alt** correspondentes às posições que têm o resultado TRUE foram selecionadas no comando anterior.

Para obtermos as observações de altura correspondentes às mulheres (f) trabalhando em Alto Taquari (at) usaríamos

```
## Observações de altura dos alunos do sexo feminino
alunos$alt[alunos$sex == "f" & alunos$loc == "at"]  # (& corresponde
ao **e** lógico)
## [1] 1.65 1.69 1.58 1.64
```

Para observações de altura correspondentes às mulheres (f) ou pessoas com peso igual ou acima de 70 kg podemos usar

```
## Observações de altura dos alunos do sexo feminino
alunos$alt[alunos$sex == "f" | alunos$pes >= 70]  # (| corresponde ao
**ou** lógico)
##  [1] 1.71 1.60 1.89 1.65 1.83 1.87 1.75 1.69 1.74 1.64 1.60 1.72
1.70 1.80
## [15] 1.58 1.64 1.58 1.85 1.82 1.84 1.70
```

Alguns operadores lógicos usuais:

- == (igual exatamente)
- is.equal() (igual aproximadamente)
- > (maior), < (menor)
- >= (maior ou igual), <= (menor ou igual)
- <> (diferente)
- & (**e** lógico), | (**ou** lógico)
- parêntesis podem ser utilizados para deixar clara a prioridade das operações

Estatísticas elementares condicionais e incondicionais (variáveis quantitativas)

Podemos trabalhar com os resultados do filtro, que também será um vetor. O exemplo a seguir mostra o uso dos comandos **mean** (média), **sd** (desvio padrão), **median** (mediana), **max** (máximo) e **min** (mínimo), **which.max** (qual o índice do máximo valor) e **which.min** (qual é o índice do menor valor) a partir da utilização de filtros (em situações que denominamos de estatísticas condicionais)

```
## estatísticas elementares incondicionais
mean(alunos$alt)  # média
## [1] 1.718
median(alunos$alt)  # mediana
## [1] 1.71
sd(alunos$alt)  # desvio padrão
## [1] 0.0993
max(alunos$alt)  # máximo
## [1] 1.89
min(alunos$alt)  # mínimo
## [1] 1.58
## estatísticas elementares condicionais
mean(alunos$alt[alunos$sex == "f"])  # média da altura das mulheres
## [1] 1.631
mean(alunos$alt[alunos$sex == "m"])  # média da altura dos homens
## [1] 1.774
# algumas estatísticas do núm do sapato para pessoas com altura maior
que
# 1,6 trabalhando em Alto Taquari
median(alunos$cal[alunos$alt > 1.6 & alunos$loc == "at"])
## [1] 39.5
sd(alunos$cal[alunos$alt > 1.6 & alunos$loc == "at"])
## [1] 2.387
max(alunos$cal[alunos$alt > 1.6 & alunos$loc == "at"])
```

```
## [1] 43
min(alunos$cal[alunos$alt > 1.6 & alunos$loc == "at"])
## [1] 36
```

Para encontrarmos a altura da pessoa com maior peso podemos usar

```
alunos$alt[which.max(alunos$pes)]
## [1] 1.82
```

Explore os dados do **data.frame alunos** para se familiarizar com essas opções.

Uma opção interessante no R para estatísticas elementares condicionais, é o **tapply** ilustrado a seguir, com o cômputo da média de altura por sexo e da média de altura por sexo e local de trabalho.

```
tapply(alunos$alt, alunos$sex, mean)
##     f     m
## 1.631 1.774
tapply(alunos$alt, list(alunos$sex, alunos$loc), mean)
##     at    cr    po
## f 1.64 1.605 1.700
## m 1.76 1.792 1.765
```

7 - Análise básica de frequências - variáveis qualitativas

7.1 - Conceito básico de frequência absoluta e relativa

Para as definições a seguir considere:

- um conjunto de dados tem n observações
- x é uma variável qualitativa cujos valores podem assumir m_x categorias: c_1, c_2, $\ldots$, c_{m_x}

Com essa notação podemos mais formalmente definir as noções de frequência.

- $\mbox{frequência absoluta de } c_j\ \rightarrow\ \# (x == c_j),\ \ \ j=1,2,\ldots,m_x$

nesta última expressão, a notação $\#(x == c_j)$ indica o número de observações em que a variável categórica x apresenta a categoria c_j.

- $\mbox{frequência relativa de } c_j\ \rightarrow\ \displaystyle \frac{\# (x == c_j)}{n},\ \ \ j=1,2,\ldots,m_x$

A especificação da tabela de frequências exigirá a especificação de todas as m_x frequências para as m_x categorias disponíveis.

- nota: as frequências absolutas ou relativas, quando não envolvem qualquer condicionamento, também são chamadas de frequências (absolutas ou relativas) **incondicionais**.

Implementação de frequências absolutas e relativas no R: tabelas e gráficos

Suponha que deseja calcular as frequências de cada sexo nas observações do conjunto de dados (variável alunos$sex, com categorias “f” e “m”). Podemos usar os filtros e comandos como **length** (comprimento do vetor) e **sum** (soma) dos elementos (a soma de um vetor lógico assume o valor 1 para os resultados TRUE e 0 para os resultados FALSE)

```
n <- length(alunos$sex)  ## definindo o número de observações
```

```
sum(alunos$sex == "f")  ## frequência absoluta de mulheres
(possibilidade 1)
## [1] 9
length(alunos$sex[alunos$sex == "f"])  ## frequência absoluta de
mulheres (possibilidade 1)
## [1] 9
sum(alunos$sex == "f")/n  ## frequência relativa de mulheres
## [1] 0.3913
length(alunos$sex[alunos$sex == "m"])  ## frequência absoluta de
homens
## [1] 14
length(alunos$sex[alunos$sex == "m"])/n  ## frequência relativa de
mulheres
## [1] 0.6087
```

Comandos table e prop.table

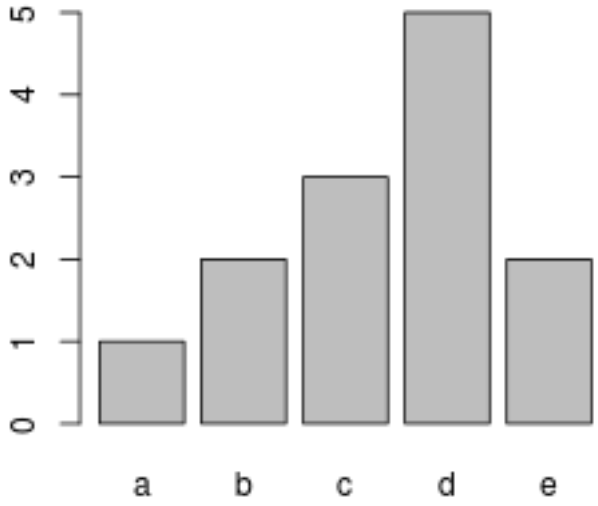
Há muitas formas de analisar frequências no R, além da forma mostrada nos parágrafos anteriores. Uma forma elementar mas prática utiliza os comandos **table** e **prop.table**. Veja alguns usos a seguir:

```
## Sexo dos alunos
table(alunos$sex)
##
##  f  m
##  9 14
prop.table(table(alunos$sex))
##
##      f      m
## 0.3913 0.6087
## Localização dos alunos
table(alunos$loc)
##
## at cr po
## 10 10  3
prop.table(table(alunos$loc))
##
##     at     cr     po
## 0.4348 0.4348 0.1304
```

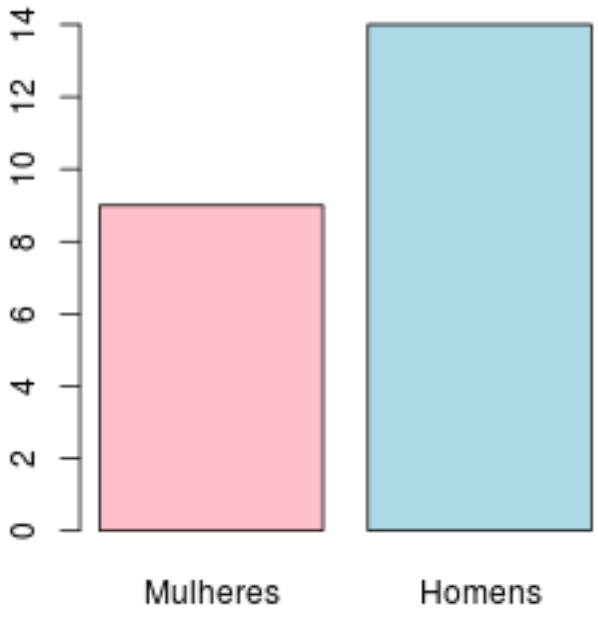
Visualização com gráficos de barra e gráficos tipo pizza

As frequências podem ser visualizadas graficamente, usando gráficos de barras elementares, que se aplicam à descrição de qualquer vetor de dados ou tabelas, como nos 2 exemplos abaixo:

```
x <- c(1, 2, 3, 5, 2)
ident <- c("a", "b", "c", "d", "e")
barplot(x, names.arg = ident)
```

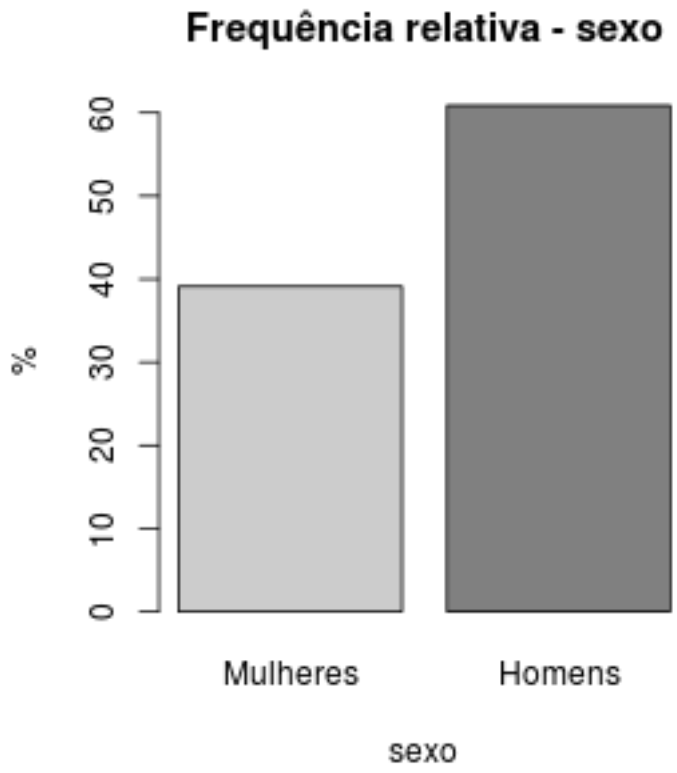



```r
ident <- c("Mulheres", "Homens")
barplot(table(alunos$sex), names.arg = ident, col = c("pink",
"lightblue"))
```



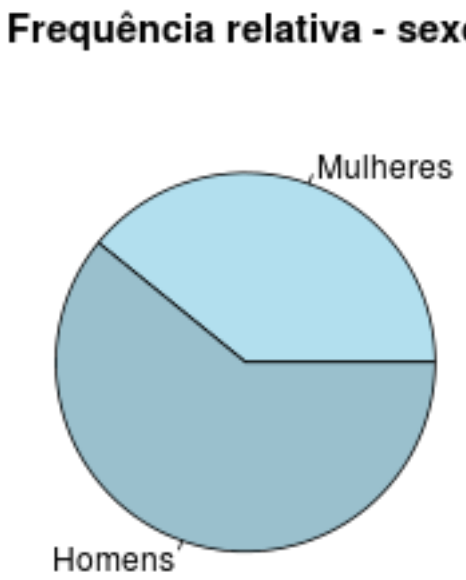
Podemos definir cores em tonalidades de cinza, usando a função **gray(x)** em que *x* é um valor entre 0 e 1 (0 é o preto e 1 é o branco). O próximo gráfico ilustra essa e outras opções:

```
ident <- c("Mulheres", "Homens")
barplot(prop.table(table(alunos$sex)) * 100, names.arg = ident, col =
c(gray(0.8),
    gray(0.5)))
title(main = "Frequência relativa - sexo", xlab = "sexo", ylab = "%")
```



É comum também a apresentação de dados de frequências em gráficos tipo **pizza**, mas esses são em geral não são recomendados por especialistas em visualização de dados, especialmente quando o número de níveis é muito grande (o gráfico de barras é mais recomendado para a visualização de diferenças).

```
ident <- c("Mulheres", "Homens")
pie(prop.table(table(alunos$sex)) * 100, label = ident, col =
c("lightblue2",
    "lightblue3"))
title(main = "Frequência relativa - sexo")
```



Veja o site 1 e o site 2 para mais detalhes sobre a definição de cores.

7.3 Frequências conjuntas

Conceito de frequência conjunta (para 2 variáveis)

Para implementação do conceito, considere que:

- conjunto de dados tem \(n \) observações
- \(x \) é uma variável cujos valores podem assumir \(m_x \) categorias: \(c_1 \), \(c_2 \), \(\ldots \), \(c_{m_x} \)
- \(y \) é uma variável cujos valores podem assumir \(m_y \) categorias: \(k_1 \), \(k_2 \), \(\ldots \), \(k_{m_y} \)

O conceito de frequências conjuntas será definido para 2 variáveis mas pode ser expandido para o caso de mais de 2 variáveis. No caso de duas variáveis temos

- \(\mbox{frequência conjunta absoluta de } c_j\ \&\ k_t\ \rightarrow\ \displaystyle \# (x\ ==\ c_j\ \&\ y\ ==\ k_t) \)
- \(\mbox{frequência conjunta relativa de } c_j\ \&\ k_t\ \rightarrow\ \displaystyle \frac{\# (x\ ==\ c_j\ \&\ y\ ==\ k_t)}{n} \)

Implementação do conceito de frequência conjunta

Poderíamos usar os filtros, para obter, por exemplo, as frequências conjuntas absolutas (e dai as relativas), associadas às variáveis **alunos$sex** e **alunos$loc** através de

```
sum(alunos$sex == "f" & alunos$loc == "at")
## [1] 4
sum(alunos$sex == "m" & alunos$loc == "at")
```

```
## [1] 6
sum(alunos$sex == "f" & alunos$loc == "cr")
## [1] 4
sum(alunos$sex == "m" & alunos$loc == "cr")
## [1] 6
sum(alunos$sex == "f" & alunos$loc == "po")
## [1] 1
sum(alunos$sex == "m" & alunos$loc == "po")
## [1] 2
```

Ao dividirmos os valores obtidos por **n<-length(alunos$sex)** podemos obter as frequências **relativas conjuntas**.

A obtenção das frequências conjuntas exigirá a especificação das expressões para todas as possibilidades (produto cartesiano) de $j=1,2,\ldots,m_x$, e $t=1,2,\ldots,m_y$, ou seja, $m_x \times m_y$ frequências. Uma alternativa mais fácil que a anterior pode utilizar as funções **table** e **prop.table**, a qual é demonstrada a seguir.

```
table(alunos$sex, alunos$loc)
##
##     at cr po
##   f  4  4  1
##   m  6  6  2
prop.table(table(alunos$sex, alunos$loc))
##
##          at      cr      po
##   f 0.17391 0.17391 0.04348
##   m 0.26087 0.26087 0.08696
```

Nota: Podemos alterar o número de digitos significativos apresentado (para 3 por exemplo), usando a opção

```
oldoptions <- options()  # preservando as opções existentes
options(digits = 3)
```

Não há restrição com relação ao número de variáveis em tabelas de frequência conjunta. O próximo exemplo mostra as frequências conjuntas de todas as possibilidades envolvendo 3 variáveis:

Situações envolvendo 3 variáveis (conjunta)

```
prop.table(table(alunos$odf, alunos$loc, alunos$sex))
## , ,  = f
##
##
##         at     cr     po
##   j 0.1739 0.0870 0.0435
##   o 0.0000 0.0870 0.0000
##
## , ,  = m
##
##
##         at     cr     po
##   j 0.1304 0.1304 0.0000
##   o 0.1304 0.1304 0.0870
##
```

7.4 Frequências condicionais

Conceito

Para formalização do conceito, considere que:

- conjunto de dados tem n observações
- x é uma variável cujos valores podem assumir m_x categorias: c_1, c_2, $\ldots$, c_{m_x}
- y é uma variável cujos valores podem assumir m_y categorias: k_1, k_2, $\ldots$, k_{m_y}

Caso de 2 variáveis com 1 delas condicionando (situação mais trivial):

- $\mbox{frequência condicional absoluta de } c_j \mid k_t \rightarrow \displaystyle \# (x == c_j \& y == k_t)$

O conceito de frequência condicional relativa é definido por

- $\mbox{frequência condicional relativa de } c_j \mid k_t \rightarrow \displaystyle \frac{\# (x == c_j \& y == k_t)}{\# (y == k_t)}$

Implementação elementar de frequência condicional usando os comandos table e prop.table

As frequências condicionais podem se obtidas pelo uso de filtros. Por exemplo, a frequência de homens (h) condicional ao local de trabalho ser Alto Taquari (at) será dada por:

Frequências condicionais (alunos$loc|alunos$sex):

```
length(alunos$sex[alunos$sex == "h" & alunos$loc ==
"at"])/length(alunos$sex[alunos$loc ==
    "at"])
## [1] 0
```

ou, equivalentemente,

```
sum(alunos$sex == "h" & alunos$loc == "at")/sum(alunos$loc == "at")
## [1] 0
```

É mais conveniente, contudo utilizar os comandos **table** e **prop.table** para essa finalidade (o índice no comando indica a variável condicionadora).

Frequências condicionais (alunos$sex|alunos$loc):

```
prop.table(table(alunos$sex, alunos$loc), 2)
##
##        at    cr    po
##   f 0.400 0.400 0.333
##   m 0.600 0.600 0.667
```

Para inverter o condicionamento, usamos: (note que o valor do indice corresponde à variável que condiciona)

Frequências condicionais (alunos$loc|alunos$sex):

```
prop.table(table(alunos$sex, alunos$loc), 1)
##
##        at    cr    po
##   f 0.444 0.444 0.111
##   m 0.429 0.429 0.143
```

Situações envolvendo 3 variáveis (condicional em sexo)

```
prop.table(table(alunos$odf, alunos$loc, alunos$sex), 3)
## , ,  = f
##
##
##        at    cr    po
##   j 0.444 0.222 0.111
##   o 0.000 0.222 0.000
##
## , ,  = m
##
##
##        at    cr    po
##   j 0.214 0.214 0.000
##   o 0.214 0.214 0.143
##
```

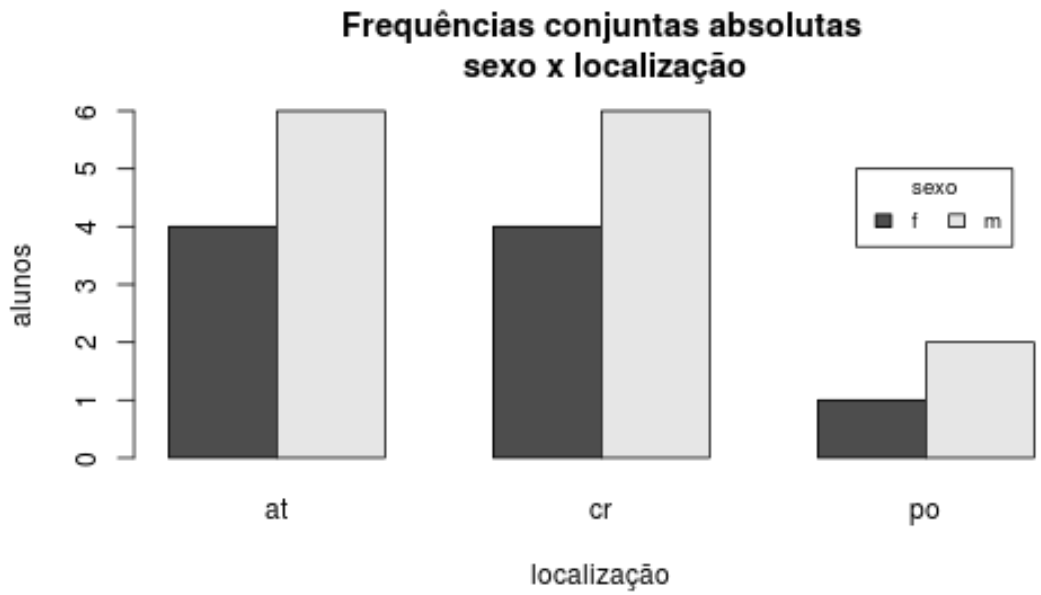
Situações envolvendo 3 variáveis (condicional em sexo e local de trabalho)

```
prop.table(table(alunos$odf, alunos$loc, alunos$sex), c(2, 3))
## , ,  = f
##
##
##      at  cr  po
##   j 1.0 0.5 1.0
##   o 0.0 0.5 0.0
##
## , ,  = m
##
##
##      at  cr  po
##   j 0.5 0.5 0.0
##   o 0.5 0.5 1.0
##
```

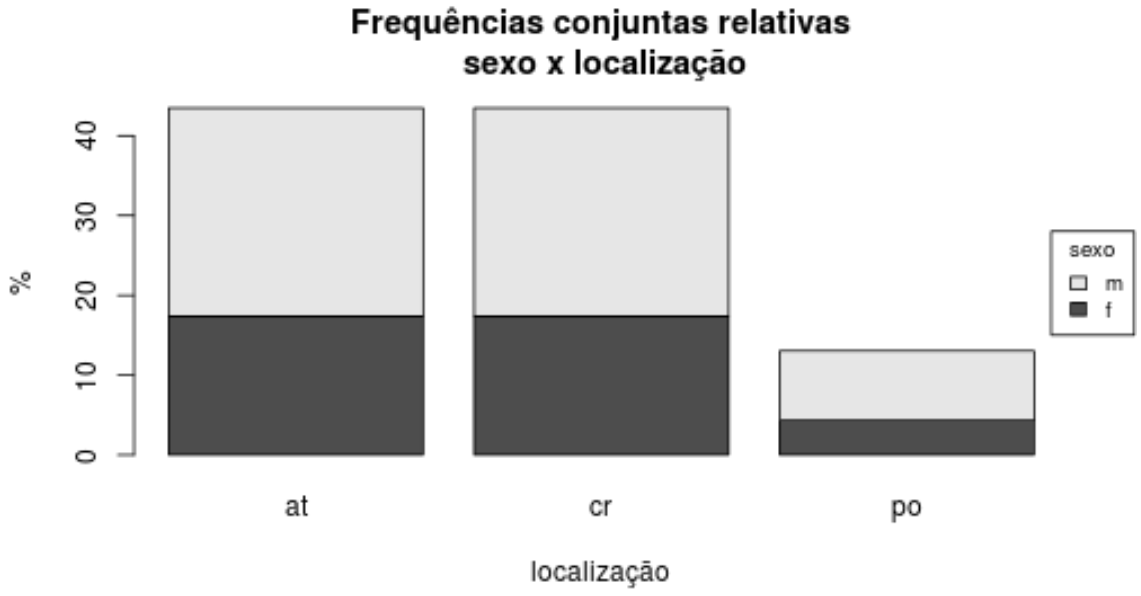
7.5 Gráficos para frequências conjuntas e condicionais

Há muitos recursos importantes (packages específicos, como o vcd por exemplo) especializados na análise e visualização gráfica de frequências conjuntas e condicionais. Neste tópico, mostraremos alguns recursos elementares fundamentados nas funções básicas e um breve exemplo do uso do package vcd (que deve ser instalado no seu computador) e uso elementar de funções básicas.

```
## gráfico de barras justapostas (segunda variável no eixo x) -
Frequência
## conjunta absoluta
barplot(table(alunos$sex, alunos$loc), beside = TRUE, legend.text =
TRUE, args.legend = list(x = 8.8,
    y = 5, title = "sexo", horiz = TRUE, cex = 0.8))
title("Frequências conjuntas absolutas\n sexo x localização", xlab =
"localização",
    ylab = "alunos")
```

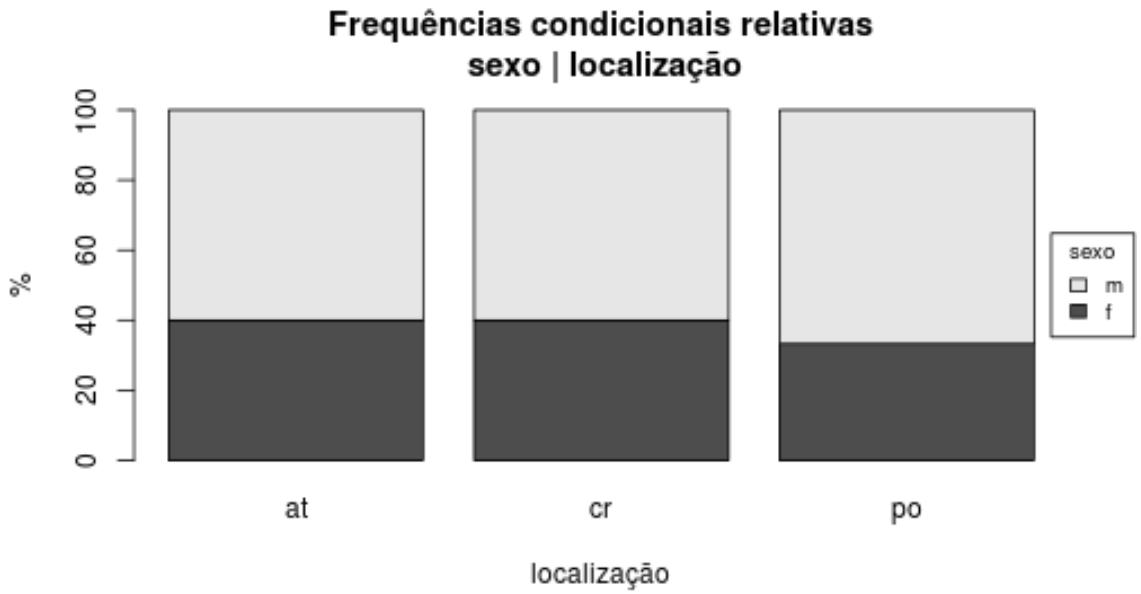



```
## gráfico de barras empilhadas (segunda variável no eixo x) -
Frequência
## conjunta relativa
barplot(prop.table(table(alunos$sex, alunos$loc)) * 100, legend.text =
TRUE,
    xpd = TRUE, args.legend = list(x = "right", title = "sexo", horiz
= FALSE,
        inset = -0.07, cex = 0.8))
title("Frequências conjuntas relativas\n sexo x localização", xlab =
"localização",
    ylab = "%")
```



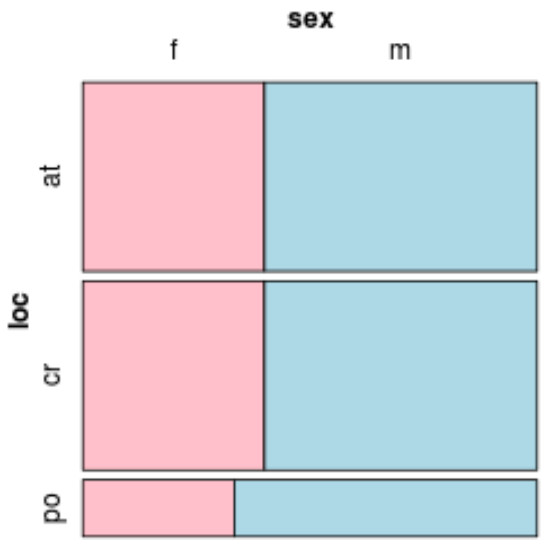

```
## Gráfico de frequência condicional (alunos$sex|alunos$loc)
barplot(prop.table(table(alunos$sex, alunos$loc), 2) * 100,
legend.text = TRUE,
```

```
    xpd = TRUE, ylim = c(0, 100), args.legend = list(x = "right",
title = "sexo",
        horiz = FALSE, inset = -0.07, cex = 0.8))
title("Frequências condicionais relativas\n sexo | localização", xlab
= "localização",
    ylab = "%")
```

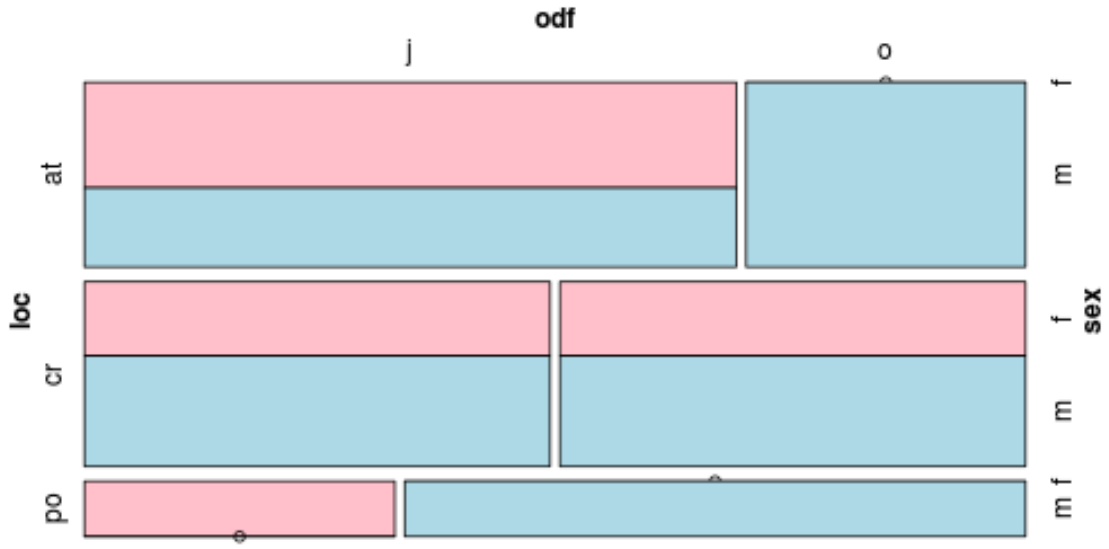


Os próximos exemplos usam o gráfico tipo Mosaico do package **vcd** que deve estar instalado para que possa ser executado. Nos gráficos as regiões são proporcionais ao número de pessoas em cada categoria.

```
## carregamento do package vcd (deve ser instalado antes)
require(vcd)
## gráfico tipo mosaico - frequências condicionais sexo | localização
mosaic(sex ~ loc, data = alunos, highlighting_fill = c("pink",
"lightblue"))
```

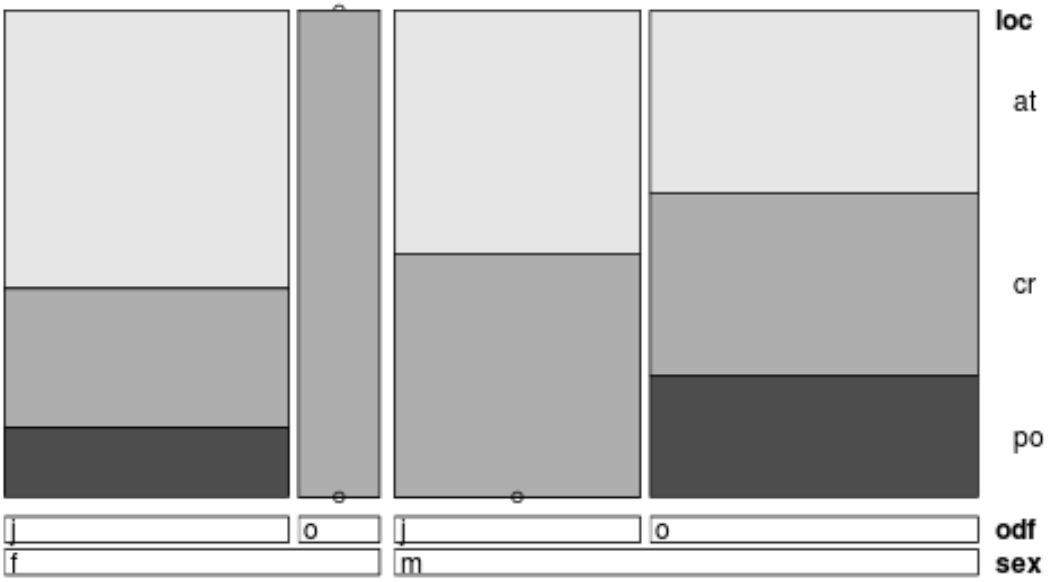



```
## gráfico tipo mosaico - frequências condicionais sexo | localização,
## tipo de ingresso
mosaic(sex ~ loc + odf, data = alunos, highlighting_fill = c("pink",
"lightblue"))
```



Também do package vcd há o gráfico tipo **doubledecker** que é útil para apresentar dados de frequências condicionais, exemplificado a seguir. Da mesma forma que no caso anterior, as regiões são proporcionais ao número de pessoas em cada categoria.

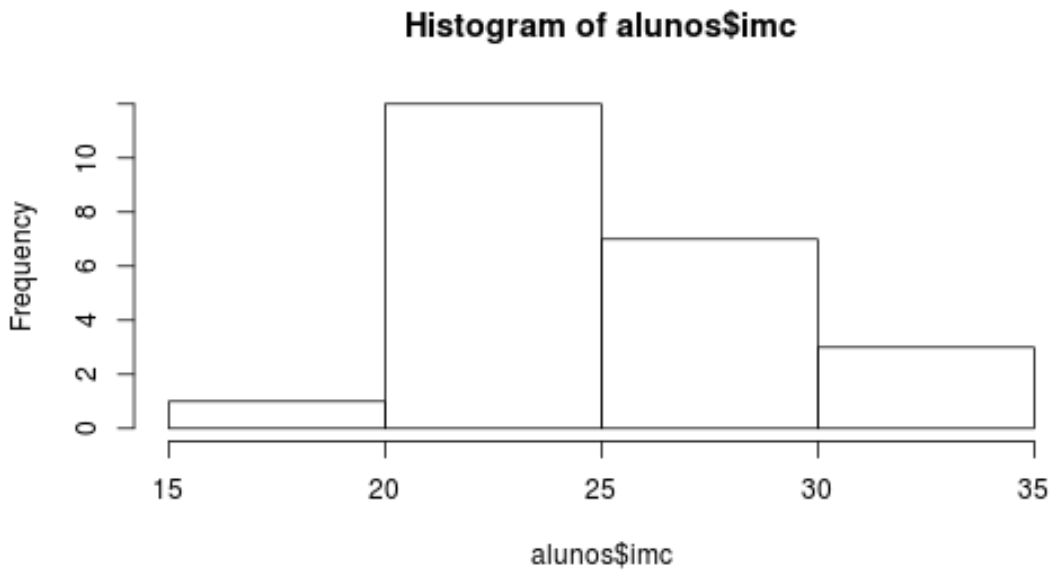
```
## carregamento do package vcd (deve ser instalado antes)
require(vcd)
## gráfico tipo doubledecker, frequência condicional sexo |
localização,
## tipo de ingresso
doubledecker(loc ~ sex + odf, data = alunos)
```



9 - Visualização de variáveis quantitativas com histogramas e outros mecanismos

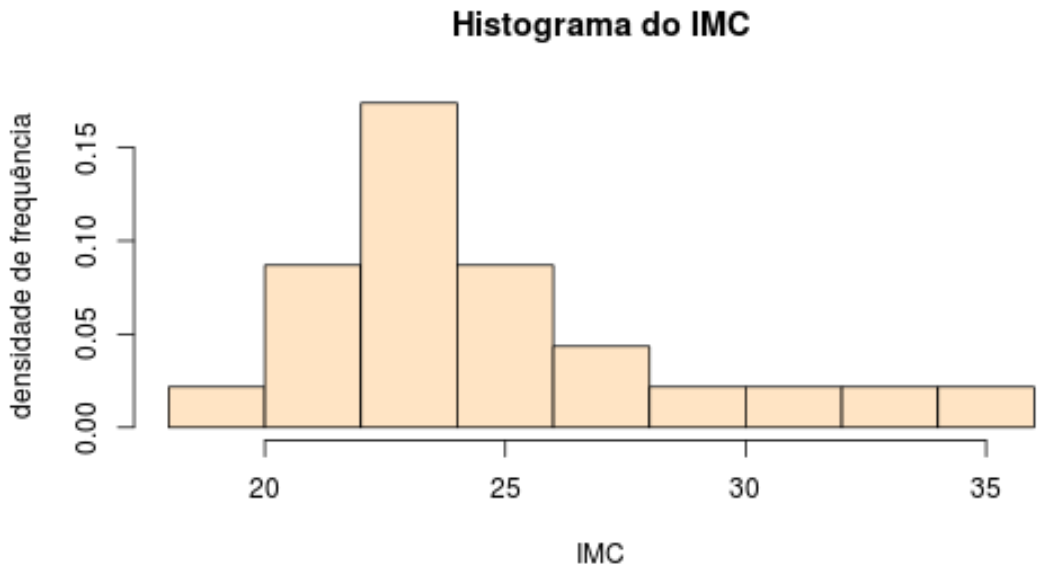
A visualização de variáveis contínuas através de gráficos pode considerar muitos conceitos diferentes. Dois deles são bem fundamentais: o histograma e o boxplot (diagrama de Tuckey) os quais serão ilustrados a seguir na descrição de variáveis quantitativas do levantamento. Para tanto examinaremos o índice de massa corporal dos alunos (que denominaremos **imc**). Caso seja definido, o argumento breaks tenta especificar o número de categorias que o histograma irá considerar.

```
alunos$imc <- alunos$pes/alunos$alt^2
## histograma do peso dos alunos - básico
hist(alunos$imc, breaks = 5)
```



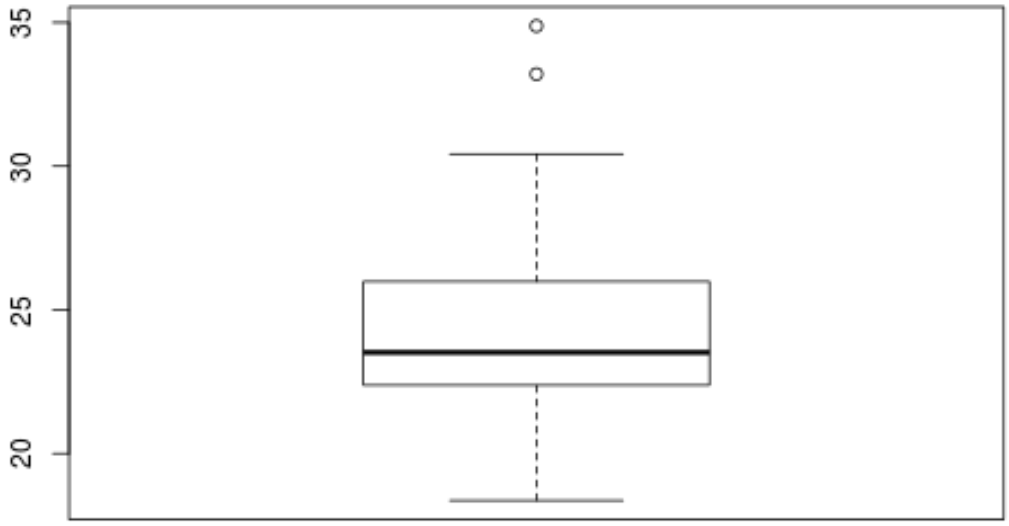

```
## a próxima implementação incorpora algumas opções específicas e
deixa o
```

```r
## número de categorias para o R especificar
hist(alunos$imc, xlab = "IMC", ylab = "densidade de frequência", main
= "Histograma do IMC",
    col = "bisque", freq = FALSE)
```

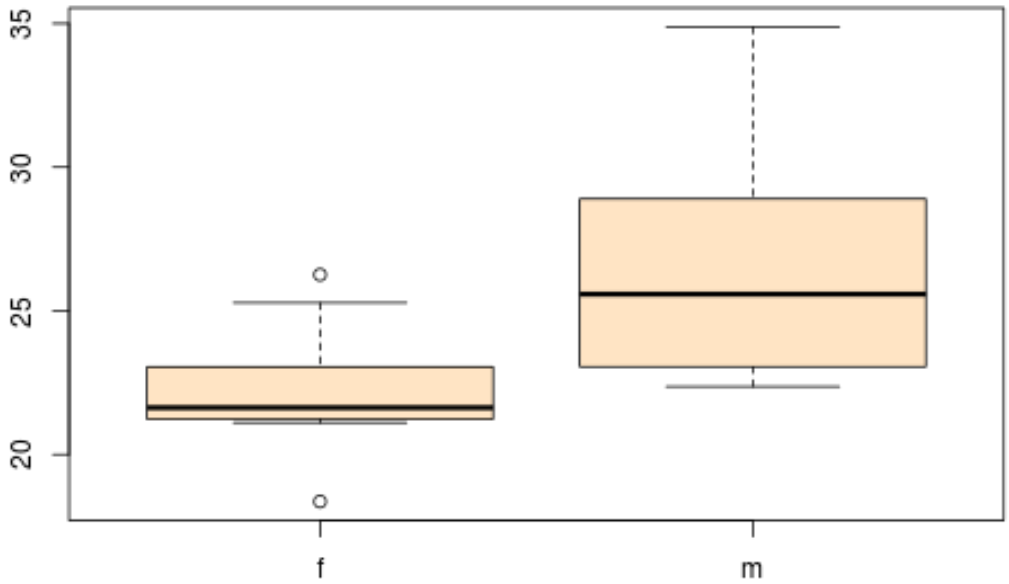


Uma outra opção é o boxplot, que mostra o máximo, mínimo, a mediana e os quartis 25% e 75%. Pode ser um gráfico incondicional ou condicional.

```r
boxplot(alunos$imc)
```



```
boxplot(alunos$imc ~ alunos$sex, col = "bisque")
```



9 - Tratando a situação de mais de uma resposta por valor informado no questionário

Em muitos casos, há mais de uma informação indicada por resposta (ex. religião, hobbies, livros). Como tratar essa situação. Há várias formas. Uma delas, a mais simples é simplesmente organizar a informação em uma lista que registra a ocorrência de cada caso. Assim poderiamos saber, pelo menos, o total de pessoas que informou uma dada possibilidade. Para fazermos essa conversão, usaremos a função abaixo, que converte um vetor com as respostas separadas por um dado separador, em um vetor com as respostas já separadas.

```
abrestring <- function(mvec, sep) {
    n <- length(mvec)
    nvec <- list()
    for (i in 1:n) {
        nvec <- list(nvec, strsplit(as.character(mvec[i]), sep))
    }
    nvec <- unlist(nvec)
    return(nvec)
}
```

Usaremos essa função a seguir para identificar as respostas associadas à informação de hobbies.

```
relhobbies <- abrestring(alunos$ho1, ", ")
table(relhobbies)
## relhobbies
##                                  Academia
##                                         6
## Acessar internet para buscar conhecimento
##                                        15
```

```
##                                    Corrida
##                                          3
##                                    Leitura
##                                          9
##                                     Outras
##                                          9
##                                Ouvir música
##                                         10
##                             Reuniões sociais
##                                          6
##                                Tocar musica
##                                          4
```

Para religião temos:

```
totrelig <- as.factor(abrestring(alunos$rel1, ", "))
levels(totrelig) <- c("at", "ca", "es", "ev", "ad", "ou", "pr")
table(totrelig)
## totrelig
## at ca es ev ad ou pr
##  1 11  4  4  3  1  1
```

No caso dos títulos dos livros podemos fazer:

```
livvec <- as.factor(abrestring(alunos$liv, ";"))
levels(livvec)
##  [1] " Cinquenta Tons de Cinza mais Escuro"
##  [2] " Diablo III"
##  [3] "100 anos de solidão"
##  [4] "A Cabana"
##  [5] "A Cilada"
##  [6] "A Cómedia Trágica ou a Tragédia Cômica de Mr. Punch"
##  [7] "A Travessia"
##  [8] "A cabana"
##  [9] "ABC da adubação"
## [10] "As Esganadas"
## [11] "Boas práticas"
## [12] "Cana de açúcar"
## [13] "Cinquenta Tons de Cinza"
## [14] "Cinquenta Tons de Liberdade"
## [15] "Como conviver com os Outros"
## [16] "Conhecer Jesus é Tudo"
## [17] "Educação pelo trabalho"
## [18] "Einstein por Ele Mesmo"
## [19] "Equador"
## [20] "Filosofia Sentimental, ensaios de lucidez"
## [21] "Fortaleza Digital"
## [22] "Lugar Nenhum"
## [23] "Mandela"
## [24] "Manutenção Mecânica"
## [25] "Meninas Normais Vão ao Shopping: Meninas Iradas Vão à Bolsa"
## [26] "O Contador de Lágrimas"
## [27] "O Que Steve Jobs Faria"
## [28] "O Vendedor de Sonhos, a revolução dos anônimos"
## [29] "O livro de Ouro da Mitologia"
## [30] "O Último Reino"
## [31] "Pai Rico Pai Pobre"
## [32] "Sementeira de Luz"
## [33] "Senhores da Escuridão"
```

```
## [34] "Sobreviver, Crescer e Perpetuar vol.1"
## [35] "Sobreviver, crescer e perpetuar"
## [36] "The Diary of a Wimpy Kid (vol 2,3,4)"
## [37] "Viva para Contar"
```

Observe que há muitas correções a fazer (espaços, títulos mal-padronizados, letras com caixa diferente). A correção pode se processar pela substituição dos valores originais por valores padronizados como operacionalizado a seguir:

```
levels(livvec)[c(1, 2, 4, 8, 34, 35)] <- c("Cinquenta Tons de Cinza
mais Escuro",
    "Diablo III", "A Cabana", "A Cabana", "Sobreviver, Crescer e
Perpetuar",
    "Sobreviver, Crescer e Perpetuar")
```

Verifique se agora tudo está correto, usando:

```
levels(livvec)
##  [1] "Cinquenta Tons de Cinza mais Escuro"
##  [2] "Diablo III"
##  [3] "100 anos de solidão"
##  [4] "A Cabana"
##  [5] "A Cilada"
##  [6] "A Cómedia Trágica ou a Tragédia Cômica de Mr. Punch"
##  [7] "A Travessia"
##  [8] "ABC da adubação"
##  [9] "As Esganadas"
## [10] "Boas práticas"
## [11] "Cana de açúcar"
## [12] "Cinquenta Tons de Cinza"
## [13] "Cinquenta Tons de Liberdade"
## [14] "Como conviver com os Outros"
## [15] "Conhecer Jesus é Tudo"
## [16] "Educação pelo trabalho"
## [17] "Einstein por Ele Mesmo"
## [18] "Equador"
## [19] "Filosofia Sentimental, ensaios de lucidez"
## [20] "Fortaleza Digital"
## [21] "Lugar Nenhum"
## [22] "Mandela"
## [23] "Manutenção Mecânica"
## [24] "Meninas Normais Vão ao Shopping: Meninas Iradas Vão à Bolsa"
## [25] "O Contador de Lágrimas"
## [26] "O Que Steve Jobs Faria"
## [27] "O Vendedor de Sonhos, a revolução dos anônimos"
## [28] "O livro de Ouro da Mitologia"
## [29] "O Último Reino"
## [30] "Pai Rico Pai Pobre"
## [31] "Sementeira de Luz"
## [32] "Senhores da Escuridão"
## [33] "Sobreviver, Crescer e Perpetuar"
## [34] "The Diary of a Wimpy Kid (vol 2,3,4)"
## [35] "Viva para Contar"
```

Podemos mostrar as frequências absolutas, já ordenadas em ordem decrescente, usando:

```
sort(table(livvec), decreasing = TRUE)
## livvec
##                                                  A Cabana
```

```
##                                                                  2
##                              Sobreviver, Crescer e Perpetuar
##                                                                  2
##                           Cinquenta Tons de Cinza mais Escuro
##                                                                  1
##                                                      Diablo III
##                                                                  1
##                                               100 anos de solidão
##                                                                  1
##                                                        A Cilada
##                                                                  1
##          A Cómedia Trágica ou a Tragédia Cômica de Mr. Punch
##                                                                  1
##                                                     A Travessia
##                                                                  1
##                                                ABC da adubação
##                                                                  1
##                                                   As Esganadas
##                                                                  1
##                                                  Boas práticas
##                                                                  1
##                                                Cana de açúcar
##                                                                  1
##                                       Cinquenta Tons de Cinza
##                                                                  1
##                                    Cinquenta Tons de Liberdade
##                                                                  1
##                                    Como conviver com os Outros
##                                                                  1
##                                          Conhecer Jesus é Tudo
##                                                                  1
##                                       Educação pelo trabalho
##                                                                  1
##                                       Einstein por Ele Mesmo
##                                                                  1
##                                                         Equador
##                                                                  1
##                     Filosofia Sentimental, ensaios de lucidez
##                                                                  1
##                                             Fortaleza Digital
##                                                                  1
##                                                  Lugar Nenhum
##                                                                  1
##                                                       Mandela
##                                                                  1
##                                             Manutenção Mecânica
##                                                                  1
## Meninas Normais Vão ao Shopping: Meninas Iradas Vão à Bolsa
##                                                                  1
##                                       O Contador de Lágrimas
##                                                                  1
##                                       O Que Steve Jobs Faria
##                                                                  1
##               O Vendedor de Sonhos, a revolução dos anônimos
##                                                                  1
##                                    O livro de Ouro da Mitologia
##                                                                  1
##                                                 O Último Reino
##                                                                  1
##                                             Pai Rico Pai Pobre
```

```
##                                                          1
##                                          Sementeira de Luz
##                                                          1
##                                       Senhores da Escuridão
##                                                          1
##                          The Diary of a Wimpy Kid (vol 2,3,4)
##                                                          1
##                                           Viva para Contar
##                                                          1
```

9 - Salvando o data.frame modificado em arquivo no seu computador

Se quiser salvar o data.frame alunos em arquivo, já com as modificações no seu computador, use

```
write.table(alunos, "ODB2013originalmodificado02.csv", sep = ";", dec
= ",",
    row.names = FALSE)
```

O data.frame alunos será salvo no arquivo “ODB2013originalmodificado02.csv”, com as opções corretas que definem um arquivo csv no formato BR desejado: “;” separando valores, “,” separando decimais, com nomes das variáveis na primeira linha do arquivo, e sem nomes identificando linhas.

Onde será salvo o arquivo? Se o caminho não for especificado junto com o nome, o arquivo será salvo no caminho padrão definido na instalação do R. Como saber esse caminho?

```
getwd()
## [1] "C:/caminhodefinido"
```

Esse resultado mostraria que o caminho padrão é “C:/caminhodefinido”. No seu caso poderá obter um valor diferente, definido durante a instalação e/ou configuração do R.

Se quiser alterar esse caminho para um caminho “c:/caminhodesejado” use

```
setwd("c:/caminhodesejado")  ## defina o caminho desejado (mude a
letra do drive se quiser)
```

Após a execução dessa última linha o R passará a usar “c:/caminhodesejado” como o caminho padrão