

Introdução ao R Commander

Notas de aula

Felipe Rafael Ribeiro Melo
Departamento de Métodos Quantitativos
Universidade Federal do Estado do Rio de Janeiro

Rio de Janeiro, Brasil

Fevereiro de 2018

Sumário

1	Introdução	4
1.1	<i>Download</i> e instalação do <i>software</i> R	4
1.2	Instalando o pacote Rcmdr	5
1.3	Carregando o pacote Rcmdr	6
2	Conjuntos de dados	7
2.1	Criando um conjunto de dados	8
2.2	Importando conjunto de dados com formatos <i>.xls</i> ou <i>.xlsx</i>	11
2.3	Importando conjunto de dados com formato <i>.txt</i> ou <i>.csv</i>	12
2.4	Exportando um conjunto de dados do R Commander	13
2.5	Conjuntos de dados nativos do R Commander	14
3	O menu <i>Dados</i>	14
3.1	Salvando conjuntos de dados	15
3.2	Carregando conjuntos de dados (no formato <i>.RData</i>)	15
3.3	Agrupar dois conjuntos de dados	15
3.4	Verificando os nomes das variáveis no conjunto de dados ativo	16
3.5	Definir nomes dos casos	16
3.6	Ordenar linhas do conjunto de dados	17
3.7	Modificação de variáveis no conjunto de dados	19
3.7.1	Converter variável numérica para variável qualitativa	19
3.7.2	Reordenar níveis dos fatores	20
3.7.3	Agrupar em classes uma variável numérica	21
3.7.4	Computar nova variável	22
3.7.5	Renomear e apagar variáveis	22
3.7.6	Recodificar variáveis	23
4	O menu <i>Gráficos</i>	25
4.1	Gradiente de cores	25
4.2	Gráfico de setores (gráfico de pizza)	26
4.3	Gráfico de barras	27
4.4	Gráfico de barras múltiplas	30

4.5	Gráfico de pontos	33
4.6	Gráfico de hastes (<i>Plot discrete numeric variable</i>)	35
4.7	Histograma	37
4.8	Boxplot	39
4.9	Diagrama de dispersão	40
4.10	Salvando gráficos	43
5	O menu <i>Estatísticas</i>	43
5.1	Resumos numéricos de todas as variáveis	44
5.2	Resumo numérico de uma variável (quantitativa)	44
5.3	Contando dados faltantes	45
5.4	Distribuições de frequências	46
5.5	Matriz de correlação	47
5.6	Tabela de contingência	47
	Referências bibliográficas	50

1 Introdução

Neste curso, usaremos o *software* estatístico R. Inicialmente, tal *software* apresenta uma interface sob a qual é necessário escrever linhas de comando para a geração ou importação de dados, bem como para a aplicação de qualquer ferramenta estatística. Todavia, o R possui uma grande quantidade de **pacotes**, os quais fornecem funcionalidades específicas. Usaremos neste curso um pacote chamado Rcmdr (uma abreviação de R Commander), que fornece ao usuário do R uma interface mais “amigável”, com menus que propiciam várias funcionalidades do R sem a necessidade de escrever linhas de comando, tais como:

- digitação de conjuntos de dados;
- importação de conjuntos de dados construídos em outros *softwares* (Excel, Bloco de notas, *etc.*);
- análises de dados por meio de gráficos;
- análises de dados por meio de tabelas.

A seguir, uma breve explicação de como baixar e instalar o *software* R, e como instalar e carregar o pacote Rcmdr.¹

1.1 Download e instalação do *software* R

O *software* R é gratuito e pode ser obtido em <http://www.r-project.org>. Acesse essa página no seu navegador de preferência e siga os seguintes passos:

1. Clique em **CRAN** (lado esquerdo da tela do seu navegador);
2. Opte por um dos *mirrors* disponíveis (aconselhável optar um *mirror* de instituição brasileira);
3. Escolha o seu sistema operacional (*Linux*, *Mac OS X* ou *Windows*) no campo *Download and Install R* (os passos seguintes derivam da escolha do sistema operacional *Windows*);

¹Este material foi escrito utilizando o *software* R versão 3.4.3 e pacote Rcmdr versão 2.4-2.

4. Clique em **base**;
5. Por fim, clique no *link* destacado (algo do tipo **Download R 3.4.3 for Windows**) para baixar o programa.

Uma vez realizado o *download*, o processo de instalação do programa é simples, bastando apenas dar um duplo clique no arquivo executável baixado e clicar sempre em *Avançar*. Agora, o R está instalado e pronto para ser utilizado.

Observação 1 *Por padrão, as versões mais recentes do R instalam simultaneamente uma versão 32 bits (R i386) e uma versão 64 bits (R x64). Neste curso, vamos utilizar sempre a versão 32 bits (R i386).*



1.2 Instalando o pacote Rcmdr

Ao abrirmos o R, visualizamos uma janela grande (chamada *R Gui*) contendo uma janela chamada *R Console*. Nesta última, vemos uma série de créditos na cor azul e, em seguida, o sinal `>` na cor vermelha. Este símbolo é chamado **prompt de comando**, e significa que o R está apto a receber um **comando** nesta linha. Ou seja, na interface apresentada, é necessário escrever linhas de comando para a realização de qualquer operação. Entretanto, o R possui uma grande quantidade de **pacotes**, os quais fornecem funcionalidades específicas.

Usaremos neste curso um pacote chamado **Rcmdr** (uma forma abreviada de **R Commander**), que fornece ao usuário do R **uma interface mais “amigável”**, com menus que propiciam várias funcionalidades do R *sem a necessidade de escrever linhas de comando*.

Para a instalação do pacote *Rcmdr*, vá na barra de ferramentas (parte superior da janela *R Gui*), selecione *Packages* (ou *Pacotes*) e, em seguida, *Install packages* (ou *Instalar pacotes*). Uma janela chamada *CRAN Mirror* abrirá, e nela, selecione o local de

preferência (sugestões: *0-Cloud[https]* ou *Brazil(RJ)*). Dê OK e uma lista com todos os pacotes disponíveis para instalação será aberta.


Os pacotes estão listados em ordem alfabética. Procure o pacote *Rcmdr* e clique nele² (não dê OK ainda). Para instalar junto a ele todos os seus *plugins*, mantenha a tecla  apertada e, com a seta  do seu teclado, marque todos os pacotes que começam com a expressão “Rcmdr”. Agora sim, com todos estes pacotes (*Rcmdr* até *RcmdrPlugin.UCA*) já marcados com fundo azul, clique em OK para instalá-los. Após isto, caso sejam abertas duas pequenas janelas uma após a outra com opções *Sim* e *Não*, selecione *Sim* em ambas. A conclusão do processo de instalação se dá quando o *prompt* de comando (sinal `>` na cor vermelha) surgir novamente no canto esquerdo inferior da janela *R Console*.

Observação 2 *É necessária conexão com a internet para a instalação de pacotes do R.*

1.3 Carregando o pacote Rcmdr

Uma vez **instalado** o pacote Rcmdr no seu computador, não será mais necessário repetir o processo descrito na Subseção 1.2. Quando quisermos trabalhar com o R Commander, será necessário apenas **carregá-lo**. Para isto, digite no R (janela *R Console*) o comando

```
require(Rcmdr)
```

e aperte  (note que apenas uma das letras no comando acima é maiúscula). Feito isto, uma nova janela será aberta: a janela *R Commander* (Figura 1), na qual vamos trabalhar.

Observação 3 *Na primeira vez que o R Commander for carregado, pode surgir uma pequena janela com opções Sim e Não. Selecione Sim e, na janela seguinte, clique em OK.*

Note que a janela *R Commander* possui três janelas: *R Script*, *Output* e *Mensagens*. Mais a frente, entenderemos melhor o que cada uma destas janelas significa. Neste

²Caso o pacote *Rcmdr* não esteja nesta lista, clique em *Cancel* e escolha outro *CRAN Mirror* no menu *Packages* → *Set CRAN mirror* antes de acessar novamente o menu *Packages* → *Install Packages*.

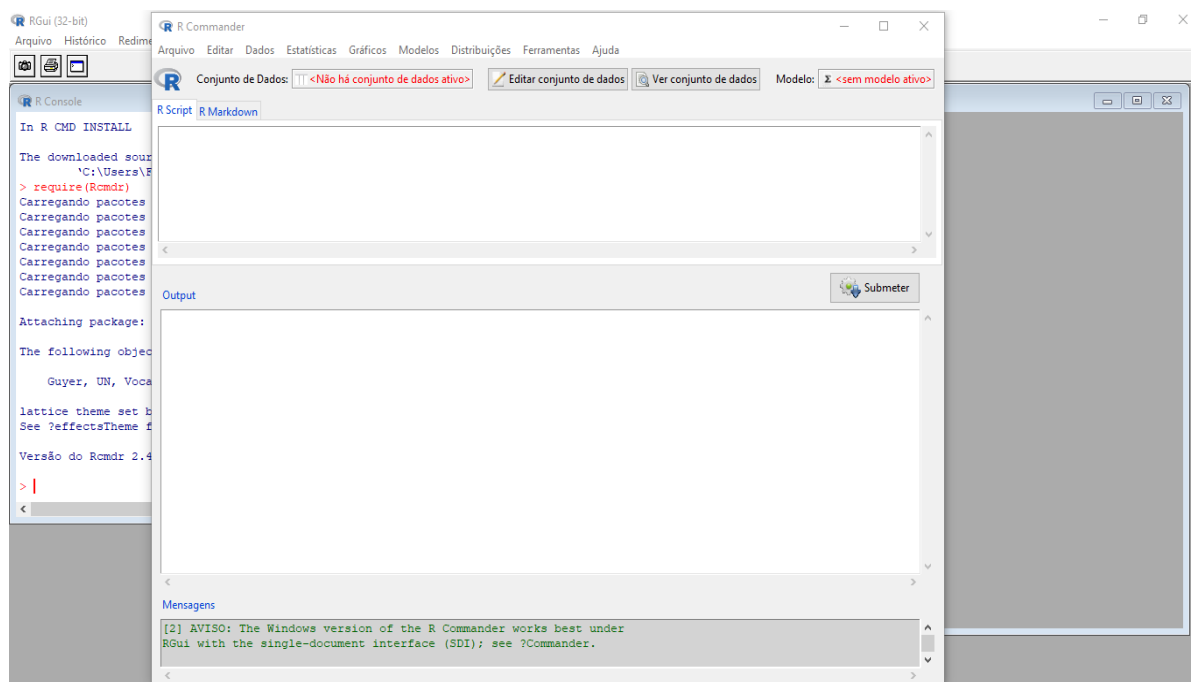


Figura 1: Apresentando o R Commander.

material, serão detalhadas e explicadas apenas algumas das muitas funcionalidades possíveis de implementar com o R Commander.

2 Conjuntos de dados

Já temos um *software* estatístico capaz de fazer análises estatísticas, bem como um pacote que torna viável executar tais análises para quem desconhece as linhas de comando do R. Interessa-nos agora aplicar as análises desejadas a um **conjunto de dados** de interesse.

Em geral, conjuntos de dados não estão em formatos que o R reconhece imediatamente. É comum, por exemplo, um conjunto de dados ser escrito numa planilha Excel (formato *.xls* ou *.xlsx*), ou até mesmo no Bloco de Notas (formato *.txt*). Para que o R e, conseqüentemente, o R Commander consigam “entender” um conjunto de dados criado fora deles, é necessário *importar* este arquivo (com o conjunto de dados) com o R Commander. Aqui, trataremos apenas da importação de conjuntos de dados em arquivos nos formatos: *.xls*, *.xlsx*, *.txt* e *.csv*. Mas primeiramente, vamos ver como criar

um conjunto de dados diretamente no R Commander via digitação.

2.1 Criando um conjunto de dados

A título de ilustração, vamos escrever este pequeno conjunto de dados abaixo, o qual chamaremos de *Turma_1*.

Aluno	Sexo	Nota	Faltas
André	M	8.5	0
Carla	F	7.0	3
Fernando	M	4.5	4
Larissa	F	9.0	2

Na barra de ferramentas da janela R Commander, vá em *Dados → Novo conjunto de dados*. Defina um nome para o conjunto de dados que será criado, sem utilizar espaço (no nosso exemplo, vamos usar o nome *Turma_1*). Ao dar OK, uma pequena janela chamada *Editor de dados* (ou *Data Editor*) será aberta por cima da janela *R Commander*.

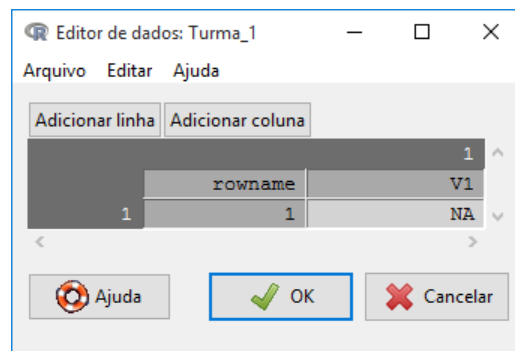



Figura 2: Planilha para criação de conjunto de dados no R Commander.

A cada clique no botão *Adicionar linha*, uma linha será adicionada, e a cada clique no botão *Adicionar coluna*, uma coluna será adicionada. Lembrando que o conjunto de dados que iremos digitar possui quatro colunas (*Aluno*, *Sexo*, *Nota* e *Faltas*) e quatro observações (alunos). Ignore a coluna *rowname* e faça o seguinte:

- clique no botão *Adicionar coluna* três vezes para a planilha ficar com quatro colunas (nomeadas genericamente *V1*, *V2*, *V3* e *V4*);

- clique no botão *Adicionar linha* três vezes para a planilha ficar com quatro linhas;
- clique na célula onde está escrito *V1* e escreva *Aluno*;
- clique na célula onde está escrito *V2* e escreva *Sexo*;
- clique na célula onde está escrito *V3* e escreva *Nota*;
- clique na célula onde está escrito *V4* e escreva *Faltas*.

Agora preencha a planilha com nome, sexo, nota e faltas de cada aluno nas células das respectivas colunas (não se preocupe em apagar os NA's; basta “escrever por cima”). Não use  para mudar de célula! Use o *mouse* ou as setas do teclado. Ainda, ao digitar as notas, use ponto ao invés de vírgula como separador de casa decimal (o R e seus pacotes sempre trabalham com ponto como separador de casa decimal). Ao terminar, clique em OK.

Observação 4 *Para apagar linhas adicionadas acidentalmente: clique com o botão esquerdo do mouse em qualquer célula dessa linha; clique com o botão direito do mouse; e escolha a opção Delete current row (ou Apagar linha corrente). Procedimento análogo para apagar colunas adicionadas acidentalmente, porém clicando em Delete current column (ou Apagar coluna corrente).*

Observe a Figura 3. Note que, no canto superior esquerdo, está escrito *Turma_1* na cor azul (ao lado de “Conjunto de dados:”), que é o conjunto de dados que acabamos de criar no R Commander. Isto quer dizer que este é, atualmente, o **conjunto de dados ativo**, ou seja, o conjunto de dados selecionado para a análise de dados. Ao lado direito do nome do conjunto de dados ativo, temos dois botões úteis:

- **Editar conjunto de dados:** Uma janela com o conjunto de dados ativo é aberta para a realização de modificações em qualquer célula desejada. Feitas as modificações, não se esqueça de clicar em OK.
- **Ver conjunto de dados.** Uma janela com o conjunto de dados ativo é aberta, porém apenas para a visualização (ao fechá-la, o conjunto de dados não será perdido). Observe a Figura 4. A coluna *rowname* (que ignoramos na digitação) é, na verdade, a coluna com fundo cinza à esquerda do conjunto de dados, com os números inteiros de 1 até 4 indicando 1ª linha, 2ª linha, 3ª linha e 4ª linha.

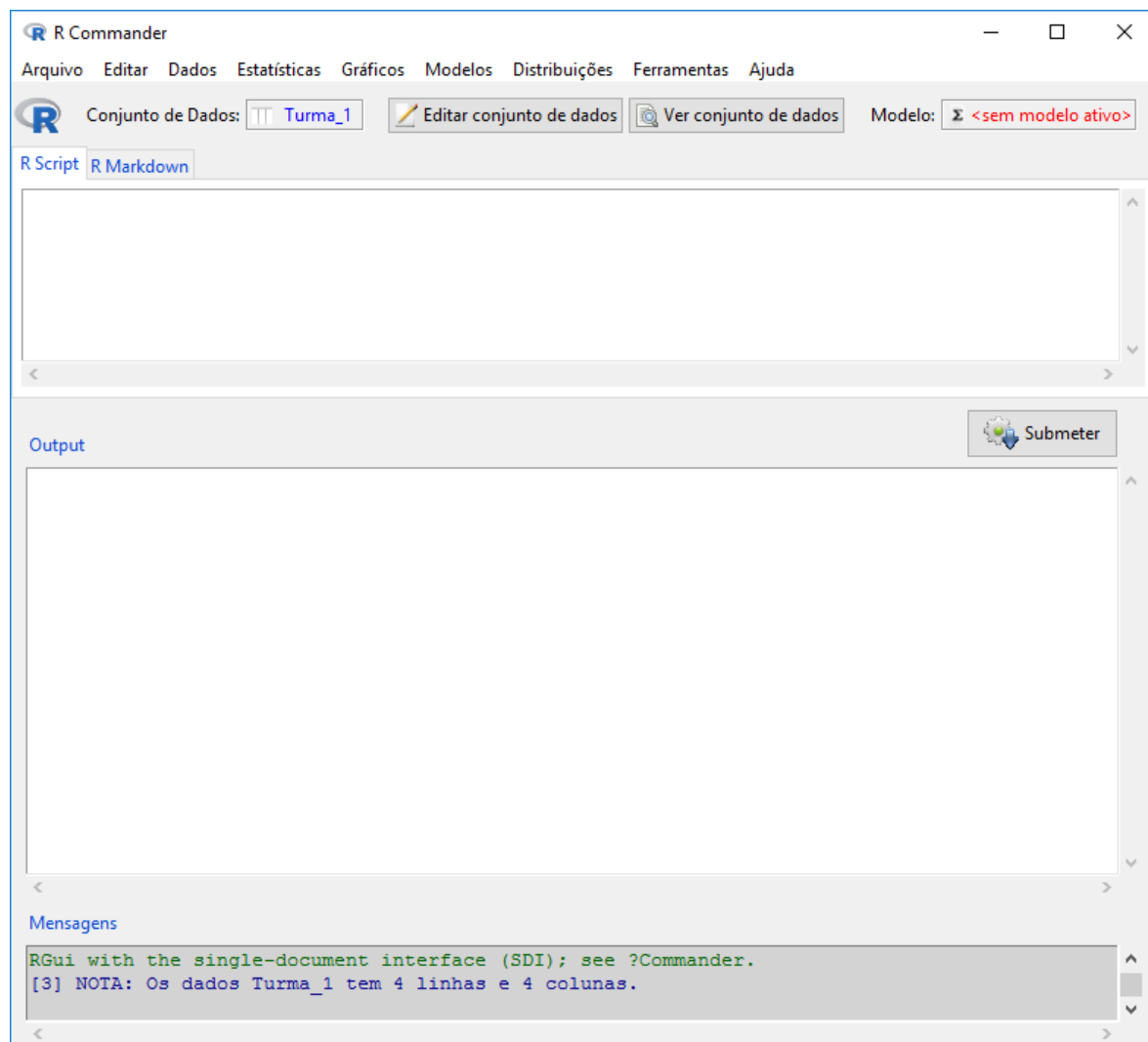




Figura 3: Apresentando as janelas *R Script*, *Output* e *Mensagens*.

	Aluno	Sexo	Nota	Faltas
1	André	M	8.5	0
2	Carla	F	7.0	3
3	Fernando	M	4.5	4
4	Larissa	F	9.0	2

Figura 4: Visualizando conjunto de dados no R Commander.

Ainda visualizando a Figura 3, estamos em posição de discutir as janelas *R Script*, *Output* e *Mensagens* da janela *R Commander*.

- **R Script:** Nesta janela são impressas as linhas de comando que foram executadas. Isso mesmo, linhas de comando! Quase tudo que é feito por meio de menus no R Commander (como, por exemplo, importação ou edição de um conjunto de dados) gera na janela *R Script* a respectiva linha de comando que é utilizada no R para obter a mesma saída. A última linha de comando nesta janela se refere ao último comando executado, a penúltima linha de comando se refere ao penúltimo comando executado, e assim sucessivamente.
- **Output:** Esta é a janela de saída, na qual são exibidos os resultados de alguns comandos executados.
- **Mensagens:** Informações que o sistema julga relevante compartilhar com o usuário, incluindo alertas.

Observação 5 *Uma linha de comando na janela R Script sempre começa colada à margem esquerda. Quando uma linha na janela R Script não começa colada à margem esquerda, isto significa que ela é continuação da linha imediatamente acima, e não uma nova linha de comando. Ainda, é possível executar qualquer linha de comando exibida na janela R Script: basta selecionar toda a linha de comando desejada com o mouse e clicar em Submeter (ou utilizar o atalho do teclado  + ).*

2.2 Importando conjunto de dados com formatos *.xls* ou *.xlsx*

É comum trabalharmos com conjuntos de dados digitados em uma planilha Excel (formato *.xls* ou *.xlsx*). Para que o R consiga “entender” este conjunto de dados, é necessário importá-lo com o R Commander. Para tal, façamos o seguinte:

1. Na barra de ferramentas da janela *R Commander*, vá em *Dados → Importar arquivos de dados → do arquivo Excel*.
2. Dê um nome para o seu conjunto de dados (sem utilizar espaço). No campo *Símbolo p/ dados faltantes*, é aconselhável substituir *<casela vazia>* por *NA*. Dê OK.
3. Procure o arquivo e dê duplo clique nele.
4. Selecione o nome da planilha na qual está o conjunto de dados. Dê OK.

Exercício 1 *Importe o conjunto de dados Turma2.xls e dê a ele o nome Turma_2.*

Observação 6 *Feito o Exercício 1, note que temos dois conjuntos de dados carregados: Turma_1 e Turma_2, este último tomando a posição de conjunto de dados ativo. Para alternar entre conjuntos de dados já carregados, basta clicar em cima do nome do conjunto de dados em azul (logo abaixo da barra de ferramentas) e escolher o conjunto com o qual se deseja trabalhar (ou seja, qual conjunto de dados desejamos tornar o conjunto de dados ativo).*

Observação 7 *A edição de dados também pode ser feita em conjuntos de dados importados (e não apenas com conjuntos criados diretamente no R Commander), bem como a sua visualização.*

2.3 Importando conjunto de dados com formato *.txt* ou *.csv*

Não é raro encontrarmos conjuntos de dados salvos no formato *.txt* ou no formato *.csv*. Apesar de programas como o MS Excel terem suporte para abrir arquivos com estas extensões, a forma de importação para o R é diferente daquela explicada na Subseção 2.2. Segue o passo a passo.

1. Na barra de ferramentas da janela *R Commander*, vá em *Dados → Importar arquivos de dados → de arquivo texto, clipboard ou URL*. Uma janela tal qual a Figura 5 será aberta.
2. Dê um nome para o seu conjunto de dados (não utilize espaço).
3. Nome das variáveis no arquivo: deixe marcado (recomendado).
4. Símbolo para dados faltantes: mantenha NA (recomendado).
5. Localização do Arquivo de dados: mantenha em *Sistema de Arquivos Local*.
6. Para preencher corretamente os campos *Separador de campos* e *Separador de decimais*, abra o arquivo que contém o conjunto de dados com o Bloco de Notas e verifique qual o símbolo utilizado para separar os campos e qual símbolo utilizado para separador de casa decimal. Após a verificação, feche-o.

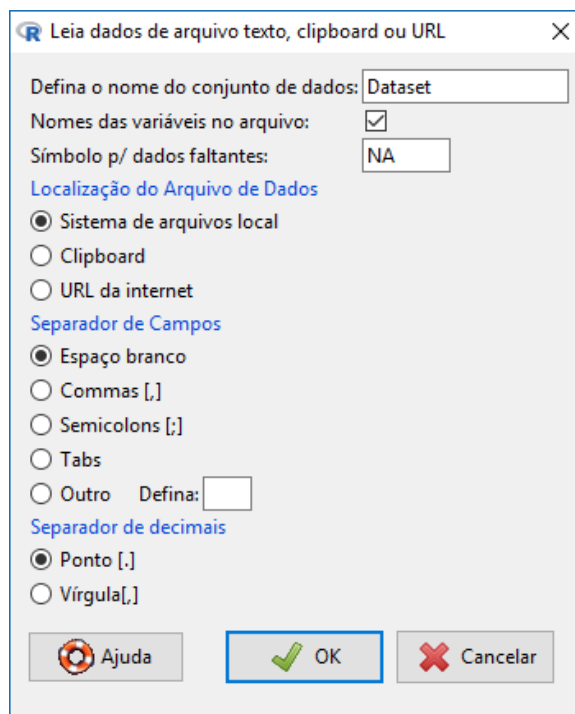


Figura 5: Importação de conjunto de dados em formato *.txt* ou *.csv*.

7. Dê OK e procure o arquivo para concluir a importação.

Exercício 2 *Importe o conjunto de dados Turma3.txt e dê a ele o nome Turma_3. Neste arquivo em particular, é utilizado ponto e vírgula (Semicolons) como separador de campos, e vírgula como separador de casas decimais.*

2.4 Exportando um conjunto de dados do R Commander

Assim como a importação de conjuntos de dados, o R Commander também possibilita a exportação de um conjunto de dados (para o formato *.txt*). O caminho é *Dados* → *Conjuntos de dados ativo* → *Exportar conjunto de dados ativo*. Feito isto, uma janela tal qual a Figura 6 será aberta. A primeira caixinha refere-se à inclusão do nome das variáveis (colunas) no arquivo exportado, e a segunda caixinha refere-se à inclusão do nome das linhas (coluna cinza à esquerda do conjunto de dados) no arquivo exportado. A terceira caixinha menciona a inclusão de aspas em torno dos atributos das variáveis qualitativas (o que pode ser útil ou não, a depender dos nossos propósitos).

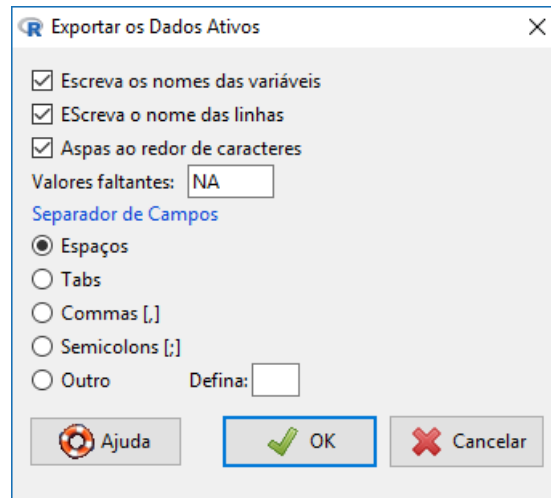


Figura 6: Exportando conjunto de dados.

Exercício 3 *Selecione o conjunto de dados Turma_1 como conjunto de dados ativo e exporte-o, deixando marcada apenas a caixinha referente à inclusão do nome das variáveis. Sugestão: em Separador de Campos, selecione Semicolons.*

2.5 Conjuntos de dados nativos do R Commander

Existem alguns conjuntos de dados que são disponibilizados ao carregarmos o pacote Rcmdr. Para acessá-los, vá em *Dados → Conjuntos de dados em pacotes → Ler dados de pacote “attachado”*. Como exercício, acesse este menu, dê um duplo clique no pacote *car* e depois um duplo clique no conjunto de dados *Davis*. Após carregá-lo (clitando em OK)³, visualize este conjunto de dados clicando em *Ver conjunto de dados*.

3 O menu *Dados*

Como visto na seção anterior, utilizamos o menu *Dados* para a criação, importação e exportação de conjuntos de dados. Nesta seção, veremos outros comandos interessantes deste menu.


³Antes de clicar em OK, pode ser interessante clicar na caixa *Ajuda no conjunto de dados selecionado*, pois dessa forma será aberta (na janela do navegador de internet padrão) uma breve explicação do conjunto de dados selecionado, incluindo o significado de suas variáveis, as quais são comumente escritas de forma abreviada.

3.1 Salvando conjuntos de dados

Para evitar a necessidade de criar ou importar novamente um conjunto de dados, podemos salvá-lo no formato que o R reconhece (*.RData*). Primeiramente, clique no nome do conjunto de dados em azul (abaixo da barra de ferramentas) e escolha como conjunto de dados ativo aquele que se deseja salvar. Feito isto, vá em *Dados → Conjunto de dados ativo → Salvar conjunto de dados ativo*.

3.2 Carregando conjuntos de dados (no formato *.RData*)

Para carregar um conjunto de dados salvo no formato *.RData*, vá em *Dados → Carregar conjunto de dados ativo*.

Observação 8 *Caso não estejamos com o R aberto e o conjunto de dados salvo (no formato *.RData*) estiver iconizado por uma letra R, é possível carregá-lo dando um duplo clique neste arquivo salvo. Dessa forma, o R se abrirá, e basta carregar o R Commander (digitando `require(Rcmdr)` e apertando  logo após). O R Commander expressará, na cor vermelha, que “Não há conjunto de dados ativo”, porém basta clicar sobre esta expressão em vermelho para eleger o conjunto de dados carregado como o conjunto de dados ativo.”⁴*

3.3 Agrupar dois conjuntos de dados

O R Commander possibilita que agrupemos dois conjuntos de dados por meio do menu *Dados → Merge de conjunto de dados*. Esse agrupamento pode ser por linhas (*Merge de linhas*) ou por colunas (*Merge de colunas*). O agrupamento por linhas é indicado quando temos dois conjuntos de dados com as mesmas variáveis dispostas na mesma ordem, tal como nos conjuntos de dados *Turma_1*, *Turma_2* e *Turma_3*. Já o agrupamento por colunas é indicado quando as observações são as mesmas (dispostas na mesma ordem) em dois conjuntos de dados, porém em cada um deles são avaliadas variáveis diferentes.

⁴Ao carregar conjuntos de dados desta forma alternativa, é provável que o R e o R Commander abertos sejam da versão de 64-bits (*R x64 3.4.3*), e não a versão que estamos usando neste material, de 32-bits (*R i386 3.4.3*).

Exercício 4 *Agrupe o conjunto de dados `Turma_2` com o conjunto de dados `Turma_3` (sugestão de nome: `Turmas_2_e_3`). Logo após, clique em `Ver conjunto de dados` para visualizar o conjunto de dados gerado por este agrupamento.*

3.4 Verificando os nomes das variáveis no conjunto de dados ativo

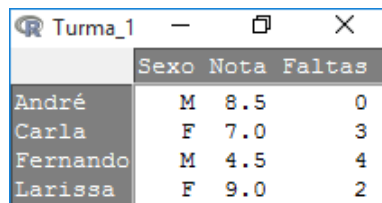
Note que ao criarmos, importarmos ou carregarmos um conjunto de dados, a janela *Mensagens* (localizada no rodapé da janela *R Commander*) relata o número de linhas e de colunas deste conjunto de dados. No formato “padrão” de conjunto de dados, cada coluna representa uma variável⁵. Para verificar o nome dessas variáveis (ou seja, o nome atribuído a cada coluna), vá em *Dados → Conjunto de dados ativo → Variáveis no conjunto de dados ativo*. O nome das variáveis aparecerá na janela *Output* do R Commander (na cor azul).

3.5 Definir nomes dos casos

Note que, nos conjuntos de dados que dispomos (*Turma_1*, *Turma_2*, *Turma_3* e *Turmas_2_e_3*), a primeira coluna não representa uma variável, mas sim uma identificação das observações (os alunos, identificados pelo primeiro nome) em cada linha. Todavia, o R Commander entende que toda coluna é uma variável (o que, a rigor, não chega a ser um grande problema). Uma forma da coluna *Aluno* (em qualquer um destes conjuntos de dados) ser de fato compreendida pelo R Commander como identificação das observações em cada linha é acessar o menu *Dados → Conjunto de dados ativo → Definir nomes dos casos*. Feito isto (por exemplo, com *Turma_1* como conjunto de dados ativo), escolha *Aluno* e dê OK. Clique agora no botão *Ver conjunto de dados* para verificar a diferença: os nomes dos alunos foram movidos para a coluna com fundo cinza à esquerda do conjunto de dados, e agora é vista apenas como identificadora de cada linha, ou seja, de cada aluno (compare as Figuras 4 e 7).

Exercício 5 *Repita o procedimento acima com os outros três conjuntos de dados.*

⁵A primeira coluna de um conjunto de dados pode estar associada simplesmente à identificação de cada observação, não se tratando portanto de uma variável propriamente dita.



	Sexo	Nota	Faltas
André	M	8.5	0
Carla	F	7.0	3
Fernando	M	4.5	4
Larissa	F	9.0	2

Figura 7: Definindo a coluna *Aluno* como nomes dos casos no conjunto de dados *Turma_1*.

3.6 Ordenar linhas do conjunto de dados

Em alguns conjuntos de dados, pode ser interessante ordenar as linhas em ordem crescente ou decrescente de acordo com uma coluna (variável) de referência. Por exemplo, podemos ordenar as linhas de qualquer um dos conjuntos de dados que digitamos/importamos anteriormente conforme a nota do(a) aluno(a) (da maior nota para a menor nota, ou vice-versa). Para isto, primeiramente defina o conjunto de dados de interesse como conjunto de dados ativo (a título de ilustração, escolha o conjunto de dados *Turma_1*) e, logo após, acesse o menu *Dados → Conjunto de dados ativo → Sort active dataset*. Uma janela tal qual a Figura 8 será aberta. Clique em *Nota* e marque a opção *Decreasing* para que as linhas sejam organizadas por ordem decrescente de nota. Ainda, no campo *Nome para o novo conjunto de dados*, é possível digitar outro nome para este “novo” conjunto de dados com a ordenação desejada, de forma que a ordem original do conjunto de dados *Turma_1* não seja alterada. Por comodidade, não altere o campo *Nome para o novo conjunto de dados* e clique em OK. Uma nova janela surgirá, perguntando se desejamos sobrescrever o conjunto de dados original (clique em *Sim*). Logo após, clique no botão *Ver conjunto de dados* e note que as linhas de *Turma_1* estão ordenadas pelas notas (da maior para a menor), e não mais na ordem alfabética do nome dos alunos, conforme digitamos na Subseção 2.1.

No caso de duas notas semelhantes (o que não ocorre no conjunto de dados *Turma_1*, mas ocorre no conjunto de dados *Turmas_2_e_3*), poderíamos estar interessados em usar o número de faltas como um “critério de desempate” para a ordenação. Isto também é possível por meio do mesmo menu acima: note que, na janela exibida na Figura 8, é possível escolher uma *ou mais* variáveis. No caso da escolha de pelo menos duas variáveis, uma nova janela será exibida após clicarmos em OK, para definir qual será o

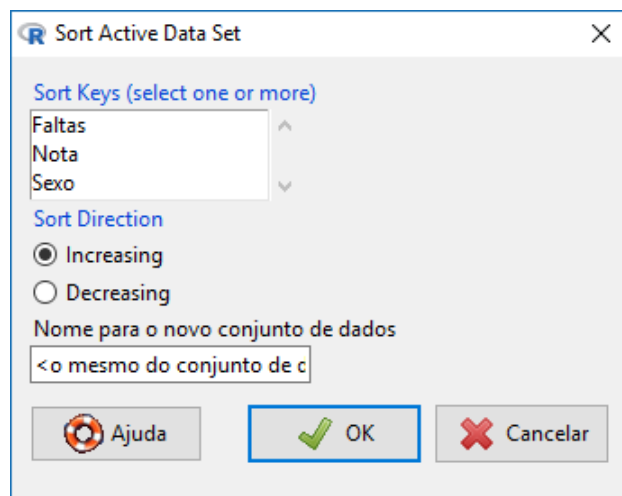


Figura 8: Ordenando linhas do Conjunto de dados *Turma_1* pela nota dos alunos.

primeiro critério de ordenação (indicado pelo número 1), qual será o segundo critério de ordenação (indicado pelo número 2), e assim sucessivamente.

Para ilustrar a situação acima, selecione o conjunto de dados *Turmas_2_e_3* como conjunto de dados ativo e acesse o menu *Dados* → *Conjunto de dados ativo* → *Sort active dataset*. Marque as variáveis *Nota* e *Faltas*, bem como a opção *Decreasing* e dê OK (novamente clique em *Sim* na janela com o aviso de sobrescrever conjunto de dados). Na janela seguinte, digite 1 para *Nota* e 2 para *Faltas*, como mostra a Figura 9, e dê OK. Clicando em *Ver conjunto de dados*, note que há dois alunos com nota 6.5, porém primeiramente é listado(a) o(a) aluno(a) com maior número de faltas (lembre que foi marcada a opção *Decreasing*, ou seja, ordem decrescente).

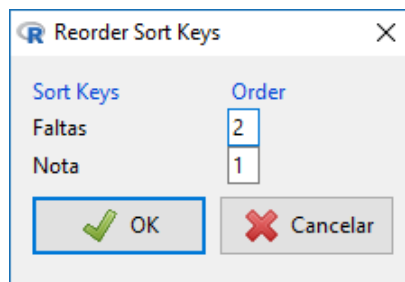


Figura 9: Ordenando linhas do conjunto de dados *Turmas_2_e_3* pela nota dos alunos e, em seguida, pelo número de faltas dos alunos.

Observação 9 *O ponto negativo deste tipo de “edição” é o fato de que todos os critérios são ordenados na mesma direção (todos crescentes ou todos decrescentes). Há uma forma de contornar isto por meio de edição da linha de comando gerada por esta ordenação, entretanto não entraremos no mérito de edição de linhas de comando neste material.*

3.7 Modificação de variáveis no conjunto de dados

Esta subseção é dedicada às principais funcionalidades das muitas fornecidas pelo menu *Dados → Modificação de variáveis no conjunto de dados*. Primeiramente, **vamos importar o conjunto de dados *Bussab.xlsx*** tal como ensinado na Seção 2.2. Este conjunto de dados traz 36 observações (indivíduos) e registra, para cada um destes indivíduos: seu grau de instrução; sua renda; se é ou não casado; idade em anos completos; região de procedência (capital, interior ou outro); e o número de filhos (dos indivíduos casados).

3.7.1 Converter variável numérica para variável qualitativa

No conjunto de dados *Bussab*, note que as variáveis *Casado* e *Instr* estão representadas por rótulos em forma de números, apesar da natureza qualitativa de ambas. Para que estas variáveis sejam vistas pelo R Commander como variáveis qualitativas, é necessário converter estes rótulos numéricos para **níveis**, ou seja, transformar uma variável (vista pelo R como) quantitativa em qualitativa.

Vá em *Dados → Modificação de variáveis no conjunto de dados → Converter variável numérica para fator*. Selecione a variável de interesse (por exemplo, a variável *Casado* do conjunto de dados *Bussab*). Em *Níveis dos fatores*, selecione *Defina nome dos níveis*, e dê um nome para esta nova variável (sugestão: *Casado_fator*, ver Figura 10). Ao clicar em OK, uma janela se abrirá para que seja escrito o nome do nível (atributo) para cada rótulo numérico. Neste caso, escreva: *Não* ao lado de 0; e *Sim* ao lado de 1.

Após concluir esta conversão, clique em *Ver conjunto de dados* para visualizar a nova variável criada.

Observação 10 *É possível manter os rótulos como números (optando por Use números em vez de Defina nome dos níveis no campo Níveis dos fatores). A princípio, podemos pensar que isto não fará diferença. Mas, a rigor, a variável convertida passa a se*

comportar como variável qualitativa (e não mais como quantitativa), apesar dos rótulos numéricos.

Observação 11 Neste processo de conversão, podemos manter o nome da variável convertida como o mesmo da variável original. Entretanto, a “nova” variável criada irá sobrescrever a variável original, e desta forma **não conseguiremos recuperar a variável na sua forma original** (a não ser que o conjunto de dados seja novamente importado ou carregado).

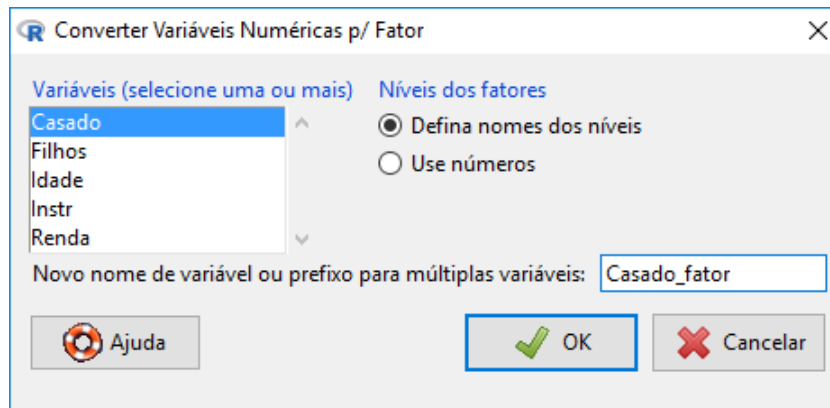


Figura 10: Convertendo a variável de rótulos numéricos *Casado* para a variável qualitativa *Casado_fator*.

Exercício 6 Crie uma nova variável no conjunto de dados Bussab chamada Instrução que converta os rótulos numéricos 0, 1 e 2 da variável Instr em seus respectivos significados (Fundamental, Médio e Superior).

3.7.2 Reordenar níveis dos fatores

Quando o R lida com variáveis qualitativas, a ordenação dos níveis (atributos) é feita em ordem alfabética. Por exemplo, um gráfico de barras da variável *Casado_fator* terá a barra associada à frequência da resposta “Não” à esquerda da barra associada à frequência da resposta “Sim”. Porém, do ponto de vista estético, é mais interessante que este gráfico trace a barra da resposta “Sim” à esquerda da barra da resposta “Não”.

No caso da variável *Instrução* do Exercício 6, temos um exemplo de variável qualitativa ordinal, cujos níveis (em ordem crescente) são: Fundamental, Médio e

Superior. Coincidentemente, essa ordenação segue a ordem alfabética. Mas poderia não ser assim (por exemplo, se tivéssemos o nível “Pós-graduação”).

Para fazer a reordenação dos níveis de uma variável qualitativa (nominal ou ordinal), vá em *Dados → Modificação de variáveis no conjunto de dados → Reordenar níveis dos fatores*. Faça esta reordenação com a variável *Casado_fator*, selecionando-a e clicando em OK. Como o campo *Nome do fator* não foi alterado, a “nova” coluna com a ordenação desejada sobrescreverá a coluna com a ordenação original (em ordem alfabética), portanto clique em *Sim* ao ser alertado quanto a isto. Em sequência, digite 1 para “Sim” e 2 para “Não” no campo *Nova ordem*. Aparentemente, nada mudou, porém, a partir dessa modificação, a ordem dos atributos de *Casado_fator* está configurada como “Sim” para primeiro atributo e “Não” para segundo atributo em qualquer análise que for feita para esta variável.

3.7.3 Agrupar em classes uma variável numérica

Quando lidamos com uma variável quantitativa contínua (ou uma variável quantitativa discreta com muitos valores distintos), pode ser de interesse resumir este conjunto de valores em **classes**. Para fazer isto pelo do R Commander, vá em *Dados → Modificação de variáveis no conjunto de dados → Agrupar em classes uma variável numérica (para criar fator)*. Uma janela conforme exibida na Figura 11 será aberta (lembrando que estamos utilizando *Bussab* como conjunto de dados ativo).

A título de ilustração, acesse o menu acima destacado e escolha a variável *Renda*. Dê um nome para a “variável agrupada” que será criada no campo *Novo nome de variável* (sugestão: *Renda_classes*). Defina o número de classes (sugestão: 4 classes) e opte pela forma com a qual cada classe será rotulada em *Nome dos níveis*. Esses rótulos podem ser nomes, números ou o próprio intervalo (sugestão inicial: marcar *Intervalos*). No campo *Método usado para definição de classes*, mantenha em *Classes de mesma largura*. Concluído este agrupamento por classes, clique em *Ver conjunto de dados* para visualizar esta nova variável criada.

Exercício 7 Repita o agrupamento da variável *Renda* feito acima, porém selecionando Definir nomes no campo Nomes dos níveis (pode fazer “por cima” da variável anterior, isto é, continuar usando o nome *Renda_classes*). Sugestões de nomes para as classes: Baixa, Média, Alta e Muito alta.

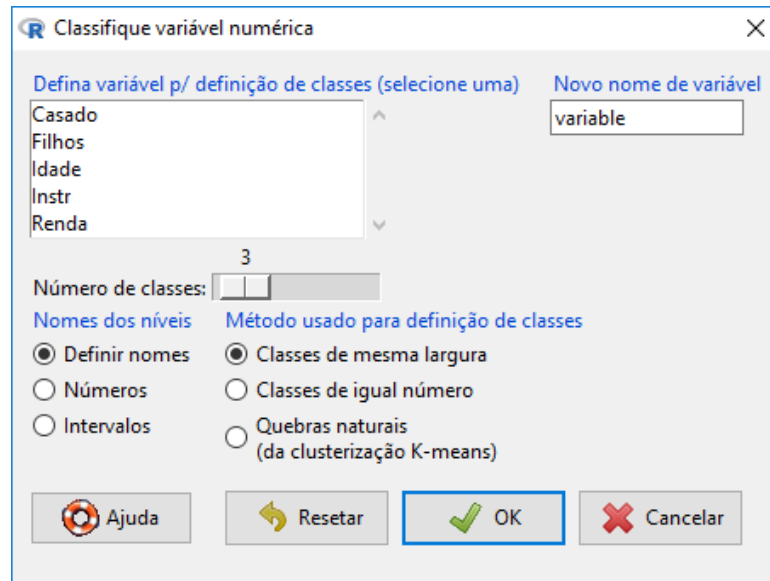


Figura 11: Agrupando uma variável numérica em classes.


3.7.4 Computar nova variável

No conjunto de dados *Bussab*, note que a variável *Renda* está expressa em salários mínimos. Uma vez que se conheça o valor de um salário mínimo (digamos, R\$954), também pode ser interessante avaliar a renda em reais, isto é,

$$\text{Renda em reais} = \text{Renda em salários mínimos} \times 954.$$

O R Commander possibilita fazer funções de variáveis, tal como acima. Vá em *Dados* → *Modificação de variáveis no conjunto de dados* → *Computar nova variável*. Preencha a janela *Compute Nova Variável* conforme a Figura 12 (duplo clique no nome da variável faz surgir seu nome no campo *Expressão p/ computar*). Use asterisco para sinal de multiplicação. Após dar OK, clique em *Ver conjunto de dados* para visualizar as rendas em reais (parece que a sugestão de nomes para as classes no Exercício 7 não está muito realista).

3.7.5 Renomear e apagar variáveis

Para **renomear** uma variável do conjunto de dados ativo, vá em *Dados* → *Modificação de variáveis no conjunto de dados* → *Renomear variáveis*. É possível selecionar mais de uma variável de uma só vez (basta segurar a tecla  e clicar nas variáveis desejadas).

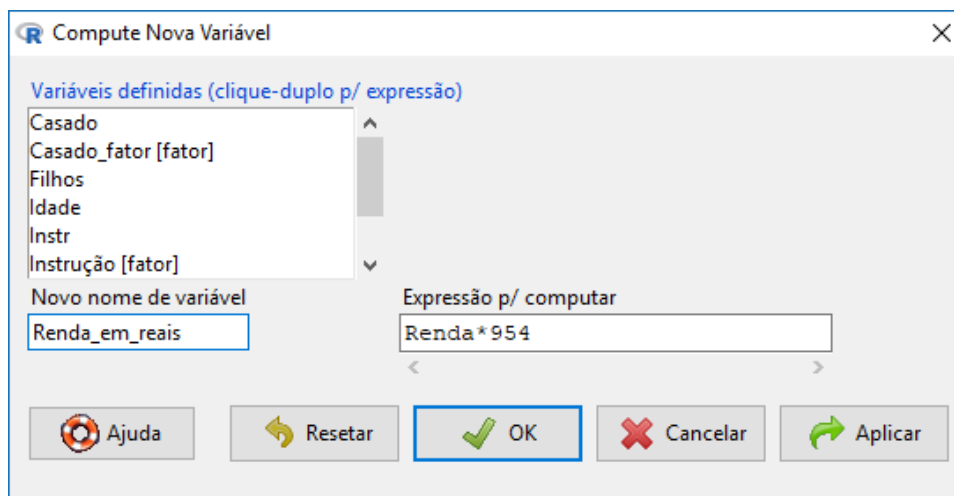


Figura 12: Computando nova variável: Renda (em R\$).

Para **apagar** uma variável (ou seja, eliminar uma coluna) do conjunto de dados ativo, vá em *Dados → Modificação de variáveis no conjunto de dados → Apagar variáveis de um conjunto de dados*. É possível selecionar mais de uma variável de uma só vez (basta segurar a tecla **Ctrl** e clicar nas variáveis desejadas).

Exercício 8 *Apague a variável “equivocada” Renda_classes do conjunto de dados Bussab.*

3.7.6 Recodificar variáveis

A recodificação de uma variável consiste na criação de outra variável (outra coluna) no conjunto de dados, que é feita baseada nos resultados da variável a ser recodificada. Nas Subseções 3.7.1 até 3.7.4, foram realizados casos particulares de recodificação de variáveis. Entretanto, o menu

Dados → Modificação de variáveis no conjunto de dados → Recodificar variáveis

permite uma forma geral de recodificação. Suponha que queiramos incluir, no conjunto de dados *Bussab*, uma variável (qualitativa) chamada *Superior_completo*, com atributos *Sim* (se o indivíduo tem ensino superior) e *Não* (se o indivíduo não tem ensino superior). Para inserir esta variável, acesse o menu de recodificação de variáveis, preencha a janela que será aberta conforme a Figura 13 e dê OK. As aspas em volta de cada nome faz com

que o R de fato entenda-os como rótulos (atributos). Clique em *Ver conjunto de dados* para visualizar a nova variável criada.

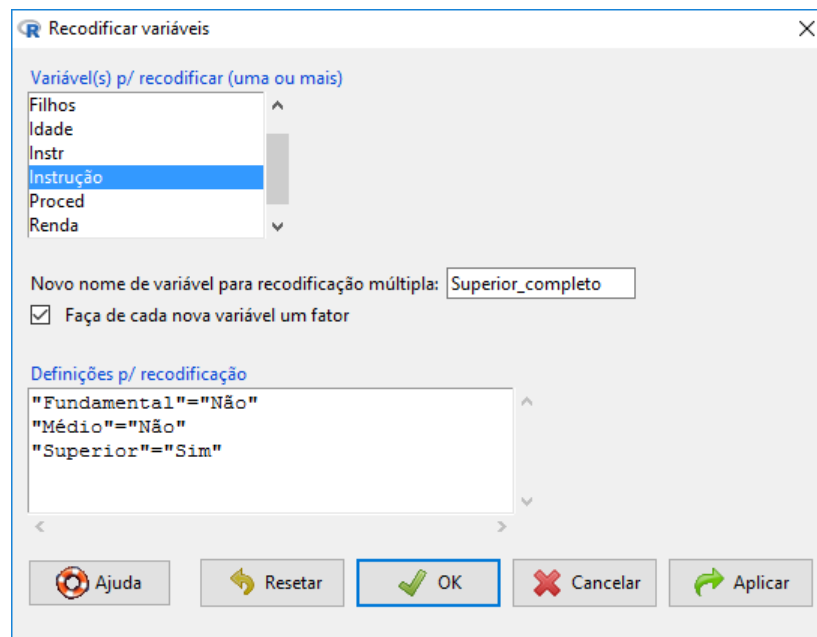


Figura 13: Criando a variável *Superior_completo* no conjunto de dados *Bussab* por meio de recodificação da variável *Instrução*.

O menu de recodificação de variáveis não se restringe apenas à variáveis qualitativas. Também podemos recodificar variáveis quantitativas, bem como ter uma variável quantitativa como resultado de recodificação. A título de ilustração: selecione o conjunto de dados *Turmas_2_e_3* como conjunto de dados ativo, acesse o menu de recodificação de variáveis e preencha a janela que será aberta conforme a Figura 14. Os termos *lo* e *hi* significam *low* (inferior, abaixo de) e *high* (superior, acima de). Após clicar em OK, clique na caixa *Ver conjunto de dados* para visualizar a nova variável criada.

Observação 12 Quando o resultado da recodificação for de natureza quantitativa (ou seja, quando no campo “Definições p/ recodificação” são postos números ao invés de palavras à direita dos sinais de igual), é **imprescindível desmarcar a caixinha “Faça de cada nova variável um fator”**. Além disso, tanto entradas numéricas como saídas numéricas não devem ser postas entre aspas, diferentemente do que ocorre com atributos de variáveis qualitativas.

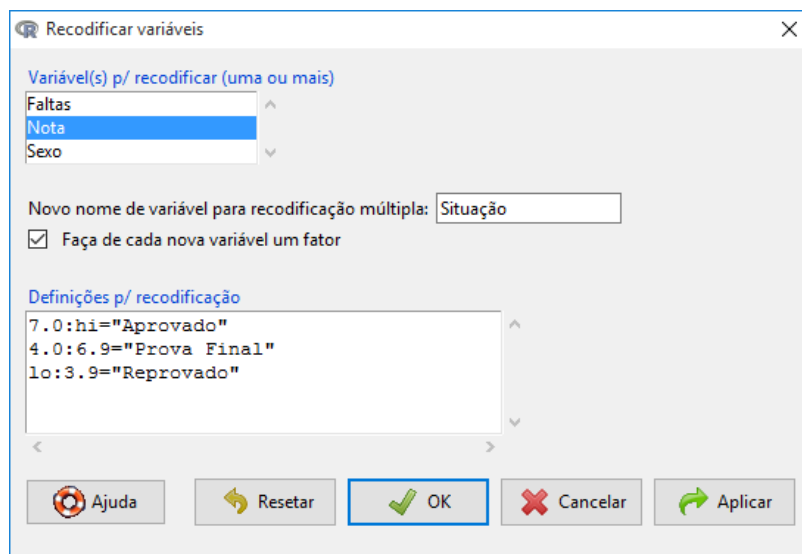


Figura 14: Criando a variável *Situação* no conjunto de dados *Turmas_2_e_3* por meio de recodificação da variável *Notas*.

4 O menu *Gráficos*

Nesta seção, veremos como gerar gráficos de variáveis do conjunto de dados ativo. De posse desses gráficos, é possível fazer (ou começar a fazer) uma análise estatística de variáveis de interesse. Outros tipos de resultados (como tabelas de distribuição de frequências, tabelas de contingência e medidas resumo) serão abordados na Seção 5.

Falaremos aqui apenas dos gráficos mais interessantes para este curso, e **adotaremos o conjunto de dados *Bussab* como conjunto de dados ativo no decorrer de toda esta seção**. É importante ressaltar que **os gráficos feitos pelo R Commander surgirão na janela *R Gui* (ao lado da janela *R Console*, em uma janela chamada *R Graphics*)**. Além disto, gráficos gerados em sequência sobrescrevem os anteriores. Ou seja, caso executemos dois gráficos em sequência, o segundo apagará o primeiro. Portanto, é importante **salvar** gráficos para não perdê-los. Tal procedimento é simples e será detalhado na Subseção 4.10.

4.1 Gradiente de cores

Antes de tratarmos dos gráficos propriamente ditos, vamos começar explorando o menu *Gráficos* → *Gradiente de cores (color palette)*. Ao clicar sobre qualquer um dos

retângulos coloridos da janela oriunda deste menu (Figura 15), é possível configurar um nova cor para tal retângulo, alterando assim a paleta de cores original ao clicar em OK.

Mas para o que isto serve?

Em alguns gráficos apresentados nesta subseção, é possível optar que os mesmos sejam coloridos segundo esta paleta de cores (a primeira cor da paleta, originalmente preta, não será considerada). Ou seja, em gráficos com apenas uma cor (como gráficos de barras simples), ele será colorido conforme a cor definida para o segundo retângulo. Já em gráficos com duas cores (como em gráficos de pizza para uma variáveis com apenas dois atributos), serão usadas a segunda e a terceira cor da paleta, e assim sucessivamente.

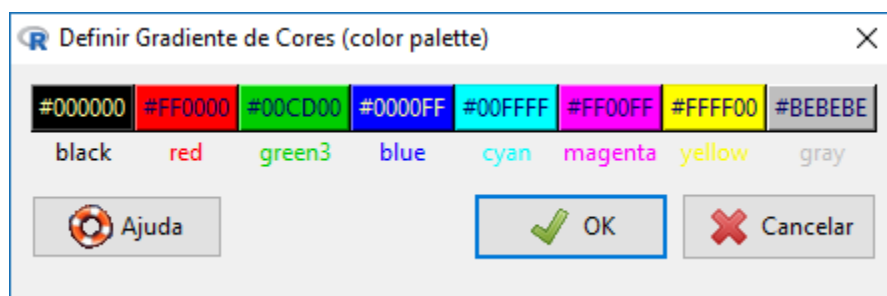


Figura 15: Gradiente de cores original do R Commander.

4.2 Gráfico de setores (gráfico de pizza)

Uma representação gráfica frequentemente utilizada para variáveis qualitativas (sobretudo qualitativas nominais) é o gráfico de setores, popularmente conhecido como gráfico de pizza. Para gerá-lo no R Commander, o caminho é *Gráficos → Gráfico de Pizza*. *Com o conjunto de dados Bussab selecionado como conjunto de dados ativo*, acesse o menu acima. A janela de construção do gráfico de pizza é simples: além da escolha da variável para a qual será feito o gráfico, é possível editar o rótulo do eixo-x, o rótulo do eixo-y e o título do gráfico. Ainda, é possível optar pela paleta de cores debatida na Subseção 4.1 selecionando *From color palette* no campo *Color selection*. Para gráficos de pizza, em particular, os nomes dos eixos devem ficar em branco (isto ocorre mantendo a expressão *<auto>* em ambos). O gráfico na Figura 16 foi obtido digitando

Indivíduos casados\nConjunto de dados Bussab

no lugar de *<auto>* no campo de *Título do gráfico* (a expressão *\n* implica quebra de linha), e mantendo o campo *Color selection* em *Default* (o R Commander usa, por padrão, uma paleta de cores especial para gráficos de pizza, oriunda do pacote *colorspace*).

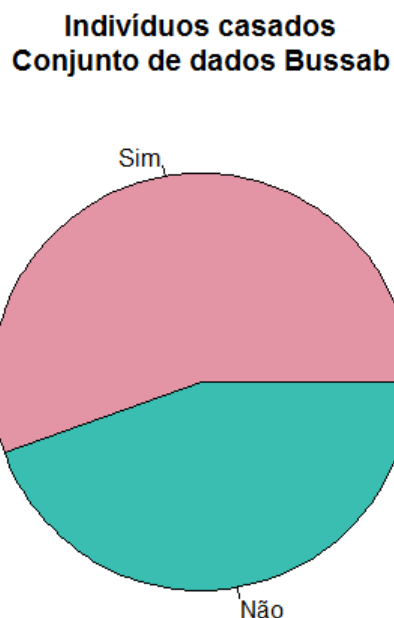


Figura 16: Gráfico de setores da variável *Casado_fator*.

Observação 13 Ao repetir a geração do gráfico acima, porém optando por *From color palette* no campo *Color selection*, note que a cor associada ao “Sim” é definida como a cor do segundo retângulo da paleta de cores do R Commander, e a cor associada ao “Não” é definida como a cor do terceiro retângulo desta mesma paleta. Lembrando que, nesta variável em particular, Sim vem “antes” de Não em virtude da reordenação realizada na Subseção 3.7.2.

4.3 Gráfico de barras

O gráfico de barras é útil para variáveis qualitativas (sobretudo qualitativas ordinais), mas também pode ser interessante para variáveis quantitativas discretas, desde que tenhamos poucos valores distintos desta variável no conjunto de dados em questão.

O caminho para fazer este tipo de gráfico é *Gráficos → Gráfico de Barras*. A janela de construção possui duas abas: *Dados* e *Opções*. Na aba *Dados*, selecione a variável *Proced* (região de procedência). Na aba *Opções*, apenas os campos *Legendas*, *Escala do eixo* e *Color selection* são de interesse para gráficos de barras simples. Preenchendo esta aba conforme a Figura 17 e clicando em OK, o gráfico de barras gerado segue na Figura 18. Ainda, note que o campo *rótulo do eixo-x* na Figura 17 está propositalmente em branco, pois caso contrário teríamos uma repetição desnecessária de informação (o título do gráfico já informa que a variável em questão é a região de procedência).

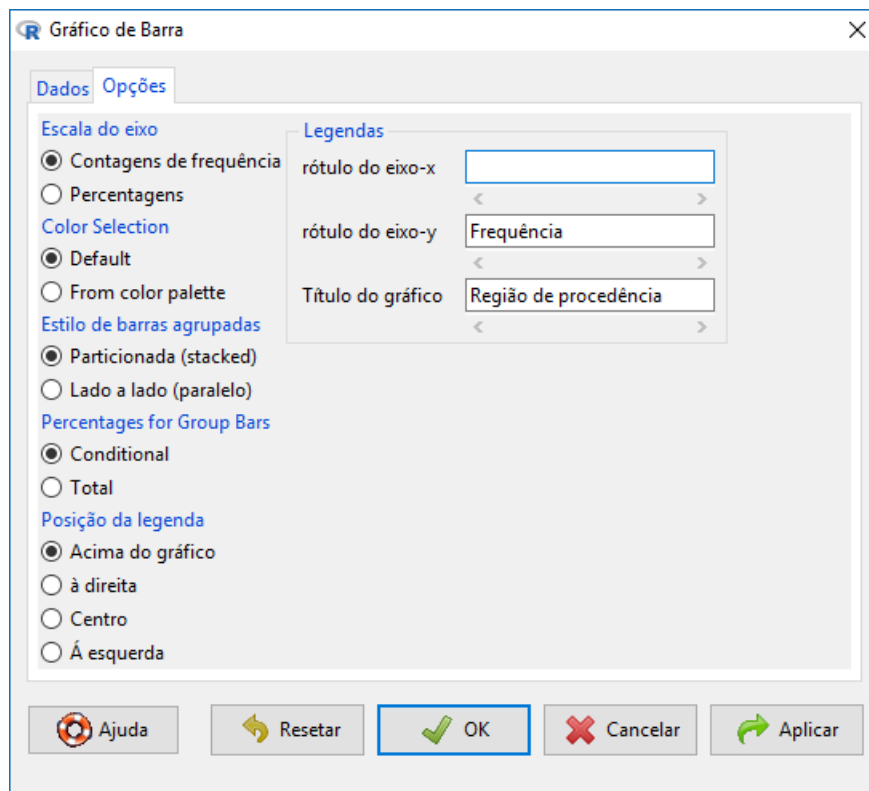


Figura 17: Construindo gráfico de barras para a variável *Proced*.

Observação 14 O campo *Color selection* é análogo ao que já foi apresentado para gráficos de pizza, porém todas as barras serão coloridas da mesma cor – a cor do segundo retângulo da paleta de cores. Já o campo *Escala do eixo* permite gerar gráficos de barras com porcentagens no lugar das contagens de frequência, e caso seja feita esta alteração, não faz sentido digitarmos *Frequência* para o rótulo do eixo-y; devemos usar, no seu lugar, termos como *Porcentagem*, *Percentual*, etc.

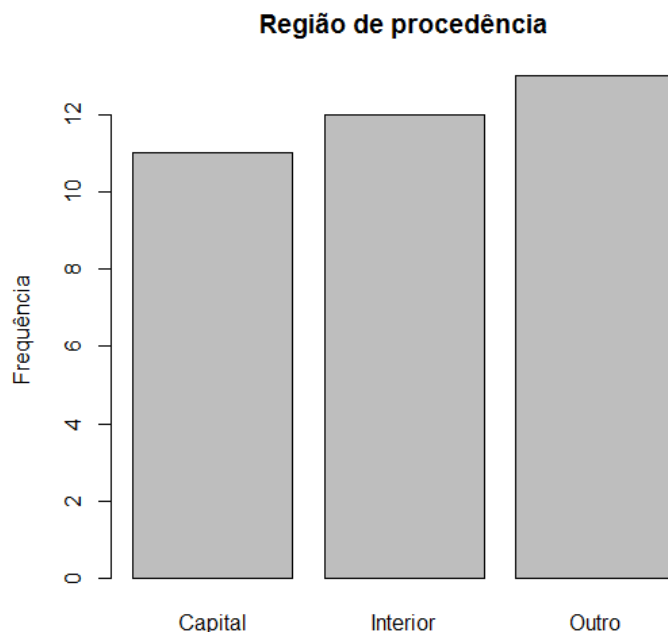


Figura 18: Gráfico de barras da variável *Proced.*

Note que o R Commander permite o traçado de gráfico de barras apenas para variáveis qualitativas. Se tivermos, no conjunto de dados de interesse, uma variável quantitativa discreta com poucos valores distintos e desejarmos fazer um gráfico de barras para ela, é necessário converter seus valores numéricos para níveis, tal como mostrado na Subseção 3.7.1 (mais precisamente, na Observação 10). Por exemplo, para gerarmos um gráfico de barras do número de filhos, é necessário criar uma nova variável no conjunto de dados *Bussab* por meio do menu de conversão de variáveis numéricas para fator, optando pela variável *Filhos* e, por comodidade, em *Use números* no campo *Níveis dos fatores* (sugestão de nome para a nova variável: *Filhos_fator*). O gráfico em questão segue na Figura 19. Entretanto, em alguns casos, esta representação pode ser “visualmente traiçoeira”, pois os números são tratados como meros rótulos. Perceba que, neste conjunto de dados em particular, nenhum indivíduo possui exatamente 4 filhos, porém a distância entre as barras de 3 filhos e de 5 filhos não difere da distância entre os outros pares de barras vizinhas.

Alternativas mais adequadas – e menos trabalhosas – ao gráfico de barras para variáveis quantitativas discretas são o **gráfico de pontos** e o **gráfico de hastes**, que serão discutidos nas Subseções 4.5 e 4.6.

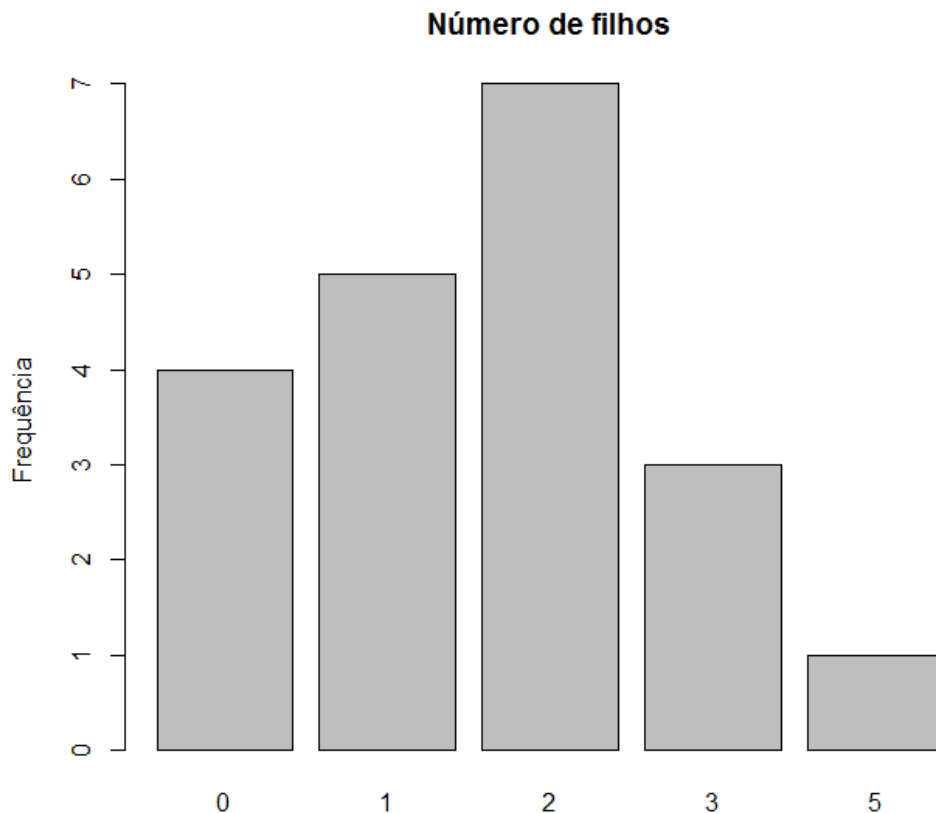


Figura 19: Gráfico de barras da variável *Filhos_fator*.

4.4 Gráfico de barras múltiplas

Além de gráficos de barras simples, também é possível gerar gráficos de barras múltiplas no menu *Gráficos → Gráfico de Barras*. Por exemplo, para “quebrarmos” cada barra do gráfico da Figura 18 (correspondentes a cada região de procedência) conforme o nível de instrução, retorne ao menu de construção de gráficos de barras, escolha a variável *Proced* e, logo abaixo, clique em *Gráfico por grupos*. Escolha a variável *Instrução* como variável de grupo e dê OK. Por fim, preencha a aba *Opções* da janela de construção do gráfico conforme a Figura 20 e dê OK para gerar o gráfico de barras múltiplas exibido na Figura 21. Note que a legenda está à esquerda, conforme definido no campo *Posição da legenda* na janela de construção do gráfico. Porém, para não correr o risco da legenda sobrepor algumas barras de um gráfico de barras múltiplas, é recomendável selecionar a opção *Acima do gráfico* no campo *Posição da legenda*.

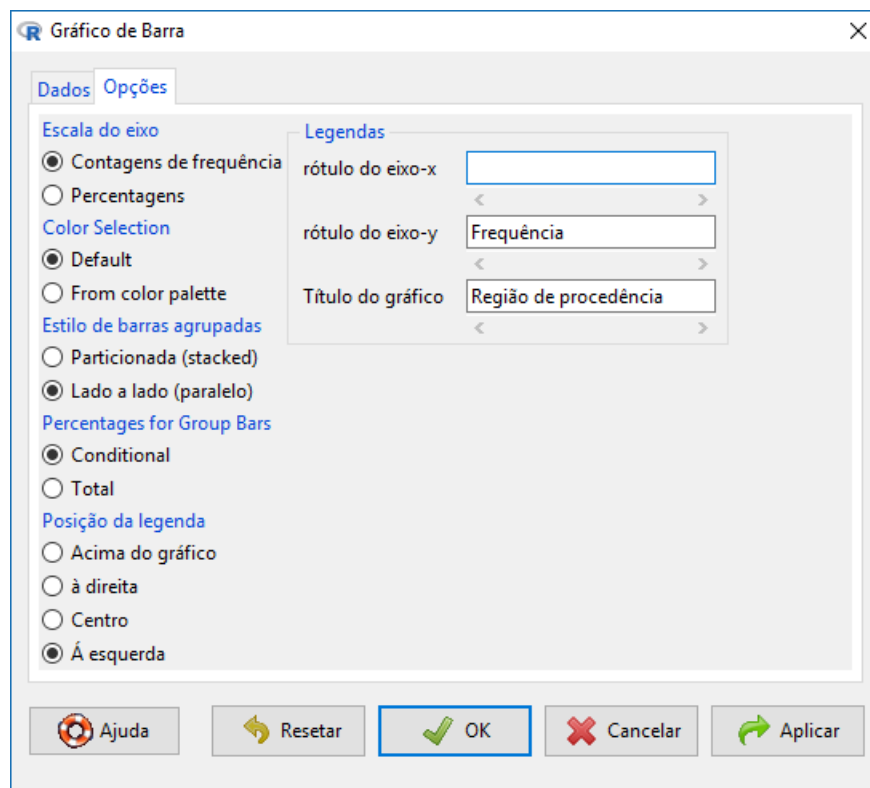


Figura 20: Construindo gráfico de barras múltiplas com as variáveis *Proced* e *Instrução*.

Observação 15 *O R Commander trata o gráfico que fizemos acima como um gráfico de barras (múltiplas) da variável *Proced* (com a variável *Instrução* escolhida como variável de grupo). Entretanto, pode ser mais fácil entender o gráfico de barras múltiplas na Figura 21 como três gráficos de barras (simples) da variável *Instrução*: o primeiro (à esquerda) considera apenas os indivíduos cuja região de procedência é *Capital*; o segundo (no centro) considera apenas os indivíduos cuja região de procedência é *Interior*; e o terceiro (à direita) considera apenas os indivíduos cuja região de procedência é *Outro*. Tal raciocínio pode ser estendido para qualquer gráfico de barras múltiplas.*

Um gráfico de barras múltiplas permite visualizar graficamente a distribuição de frequências de uma variável qualitativa em diferentes grupos (fornecidos por outra variável qualitativa). No gráfico da Figura 21, é possível visualizar graficamente a distribuição de frequências do nível de instrução dos indivíduos, separadamente em cada região de procedência da qual estes indivíduos provém. Entretanto, o número de indivíduos procedentes de cada região (*Capital*, *Interior* e *Outro*) é diferente, e assim torna-se

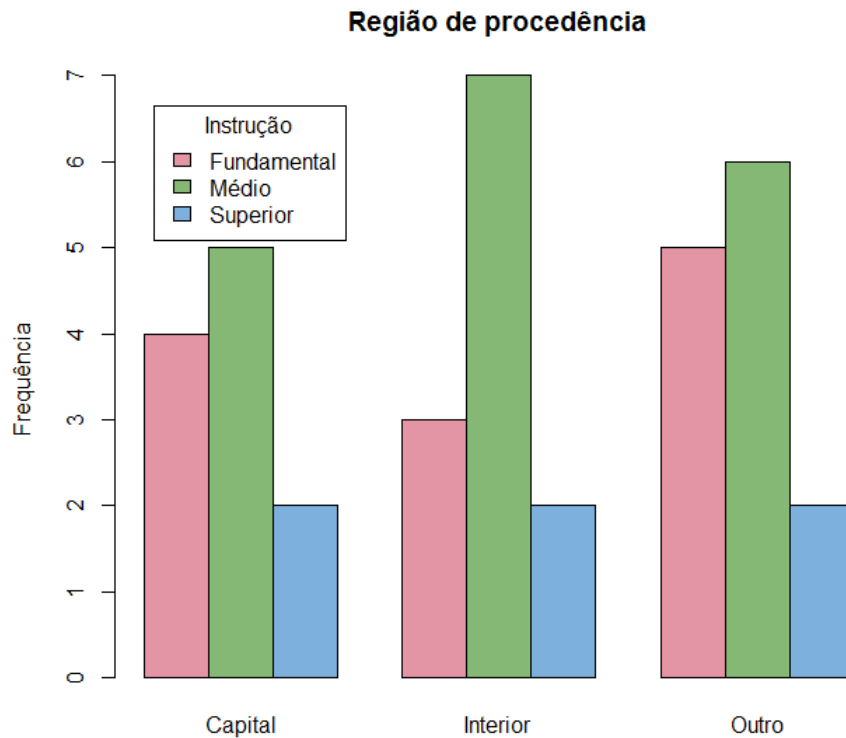


Figura 21: Gráfico de barras múltiplas com as variáveis *Proced* e *Instrução*.

razoável tomar o perfil percentual do nível de instrução em cada região de procedência quando o objetivo é comparar como se distribui o nível de instrução nas diferentes regiões de procedência dentro do conjunto de dados *Bussab*. Para realizar tal gráfico com porcentagens no lugar das frequências, tomando por base novamente a Figura 20, basta selecionar *Porcentagens* no campo *Escala do eixo* (e para que o gráfico faça sentido, digite *Porcentagem* no lugar de *Frequência* no campo *rótulo do eixo-y*). O gráfico exibido na Figura 22 surge como fruto das alterações acima e também das alterações nos campos *Color selection* e *Posição da legenda* (de *À esquerda* para *Acima do gráfico*). Ainda, é importante mencionar que o campo *Percentages for Group Bars* não deve ser alterado quando desejamos verificar o perfil percentual da variável de grupo em cada atributo da outra variável escolhida, pois caso contrário, a altura de cada barra seria o percentual total do respectivo cruzamento de nível de instrução com região de procedência.

Observação 16 Para gerar um gráfico de barras simples após gerar gráficos de barras múltiplas, é necessário clicar no botão *Resetar* (ou *Reset*) na janela de construção de gráficos de barras.

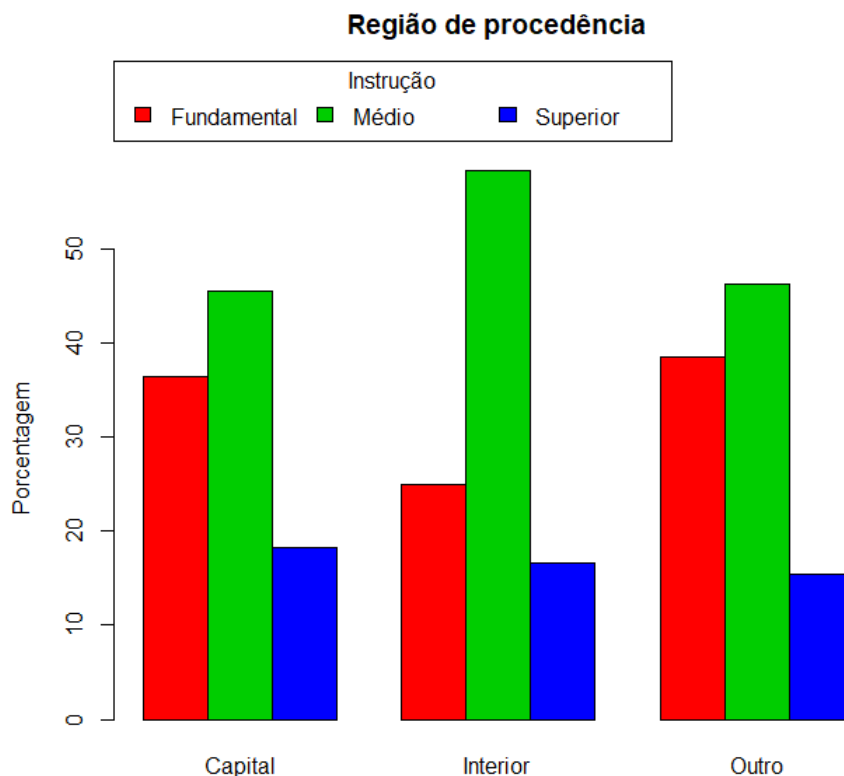
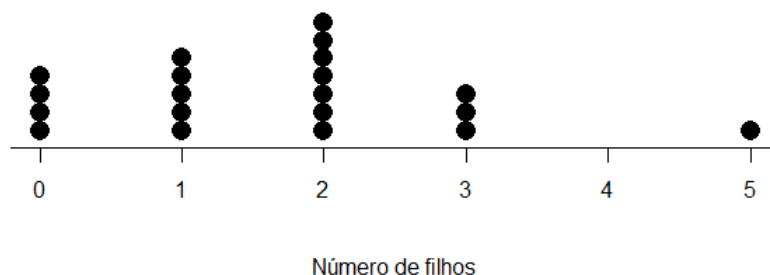


Figura 22: Gráfico de barras múltiplas com o perfil percentual do nível de instrução em cada região de procedência.

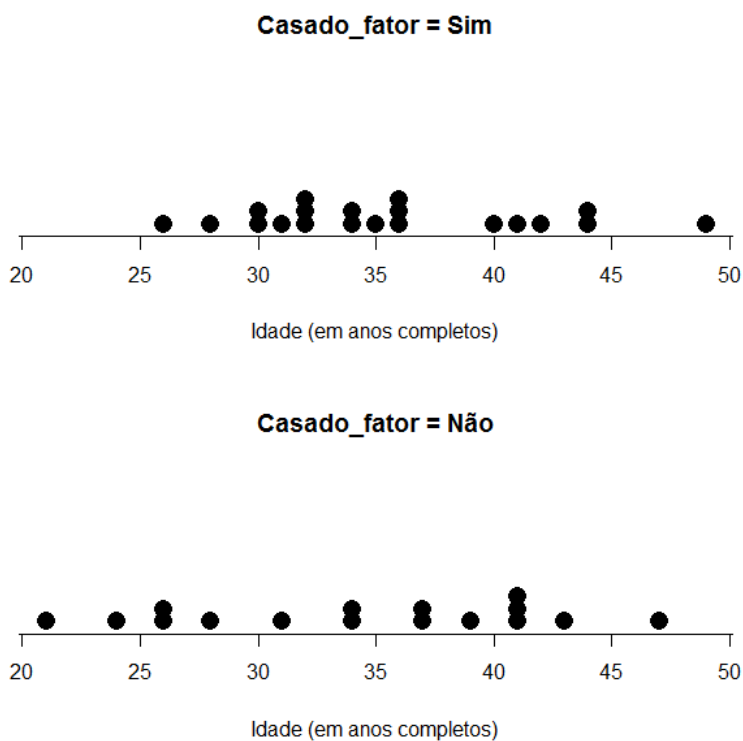
4.5 Gráfico de pontos

Conforme comentado na Subseção 4.3, o gráfico de pontos é uma alternativa ao gráfico de barras para variáveis quantitativas discretas, e está disponível no menu *Gráficos → Gráfico de pontos*. A variável de interesse deve ser selecionada na aba *Dados* da janela de construção deste gráfico (ignore por enquanto a caixa *Gráfico por grupos*). Já na aba *Opções*, é possível editar o rótulo do eixo-x (nome desejado para a variável no gráfico). A Figura 23 exhibe o gráfico de pontos para a variável *Filhos* do conjunto de dados *Bussab* (no campo *rótulo do eixo-x*, foi escrito *Número de filhos*). Diferentemente da representação gráfica na Figura 19, a natureza numérica da variável foi respeitada.

É possível gerar gráficos de pontos (de uma mesma variável quantitativa) de diferentes grupos do conjunto de dados de acordo com alguma variável qualitativa deste mesmo conjunto de dados. Isto é interessante para compararmos o comportamento de uma mesma variável em diferentes grupos dentro do conjunto de dados sob análise. Por

Figura 23: Gráfico de pontos da variável *Filhos*.

exemplo, podemos gerar um gráfico de pontos das idades dos indivíduos casados e um gráfico de pontos das idades dos indivíduos que não são casados. Para tal, acesse o menu de construção do gráfico de pontos, escolha a variável *Idade* e clique na caixa *Gráfico por grupos*. No campo *Variável de grupo*, selecione *Casado_fator* e dê OK nesta janela e na janela de construção do gráfico de pontos. O gráfico gerado segue na Figura 24 (no campo *rótulo do eixo-x*, foi escrito *Idade (em anos completos)*).

Figura 24: Gráficos de pontos da variável *Idade* para os casados e para os não casados.

Observação 17 Para gerar um gráfico de pontos sem divisão por grupos após gerar gráficos de pontos por grupos, é necessário clicar no botão *Resetar* (ou *Reset*) na janela de construção deste gráfico.

4.6 Gráfico de hastes (*Plot discrete numeric variable*)

Alternativa mais elegante ao gráfico de pontos, o gráfico de hastes assemelha-se ao gráfico de barras, porém com hastes no lugar das barras. No R Commander, é possível gerar este tipo de gráfico no menu *Gráficos → Plot discrete numeric variable*. A variável de interesse deve ser selecionada na aba *Dados* da janela de construção deste gráfico (ignore por enquanto a caixa *Gráfico por grupos*). Já na aba *Opções*, é possível editar o rótulo do eixo-x, o rótulo do eixo-y e o título do gráfico, além do campo *Escala do eixo*, com opções de frequência ou de porcentagem para a altura das hastes. Assim como no gráfico de pontos, a natureza numérica da variável quantitativa no gráfico de hastes também é respeitada.

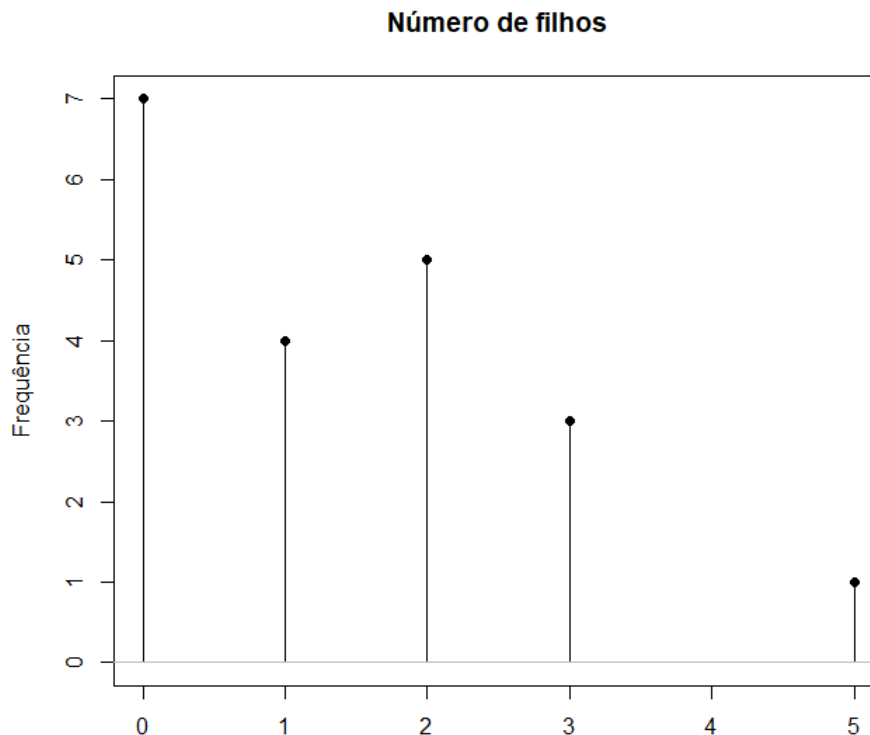


Figura 25: Gráfico de hastes da variável *Filhos*.

Ainda seguindo analogia ao gráfico de pontos, é possível gerar gráficos de hastes (de uma mesma variável quantitativa) de diferentes grupos do conjunto de dados de acordo com alguma variável qualitativa do mesmo conjunto de dados. Como este tipo de gráfico permite trabalhar com escala percentual, ilustremos aqui a geração de um gráfico de hastes das idades dos indivíduos casados e um gráfico de hastes das idades dos indivíduos que não são casados, ambos na escala percentual. Para tal, acesse o menu de construção do gráfico de hastes, escolha a variável *Idade* e clique na caixa *Gráfico por grupos*. No campo *Variável de grupo*, selecione *Casado_fator* e dê OK nesta janela. Já na aba *Opções*, selecione *Percentagens* no campo *Escala do eixo* e atente ao rótulo do eixo-y, caso tenha editado-o anteriormente. O gráfico gerado segue na Figura 26: em *rótulo do eixo-y*, foi escrito *Porcentagem*, e no rótulo do eixo-x, *Idade (em anos completos)*.

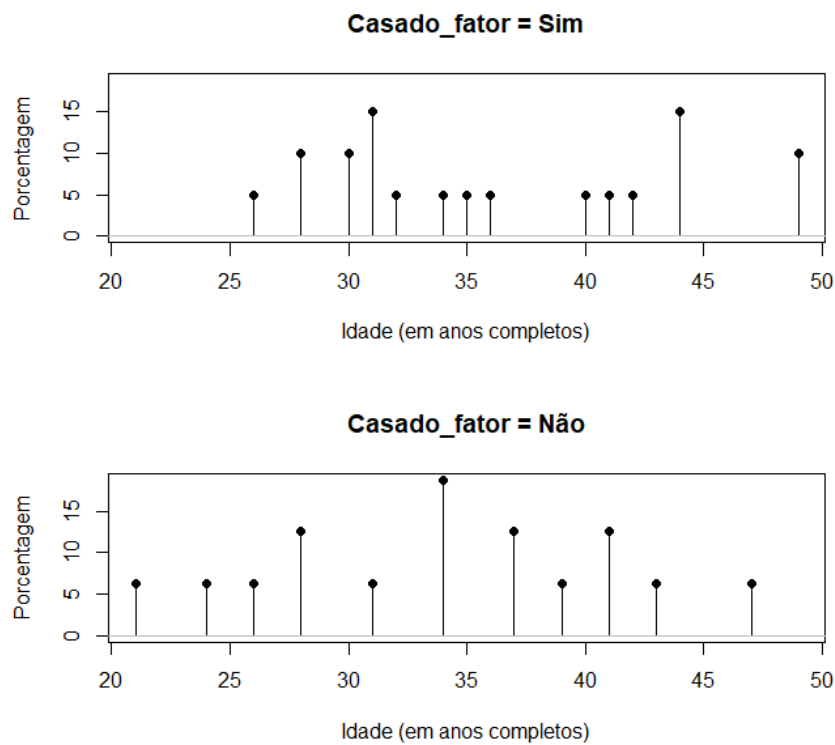


Figura 26: Gráficos de hastes da variável *Idade* para os casados e para os não casados.

Observação 18 Para gerar um gráfico de hastes sem divisão por grupos após gerar gráficos de hastes por grupos, é necessário clicar no botão *Resetar* (ou *Reset*) na janela de construção deste gráfico.

4.7 Histograma

O histograma é um dos gráficos mais utilizados para verificar o comportamento de variáveis quantitativas (sobretudo quantitativas contínuas). Ao acessarmos o menu *Gráficos* → *Histograma*, somos direcionados a uma janela com duas abas (*Dados* e *Opções*), tal qual as janelas de construção de gráficos de barras, pontos e hastes.

Na aba *Dados*, escolha a variável cujo comportamento queremos verificar por meio de um histograma (ignore por enquanto a caixa *Gráfico por grupos*). Na aba *Opções*, podemos editar: o rótulo do eixo-x, o rótulo do eixo-y e o título do gráfico; o número de classes; e a escala de altura das barras (*Contagens de frequência*, *Percentagens* ou *Densidades*). Como exercício, reproduza o histograma da Figura 27. Mantenha o número de classes em *<auto>*, e lembre que a expressão $\backslash n$ implica quebra de linha.

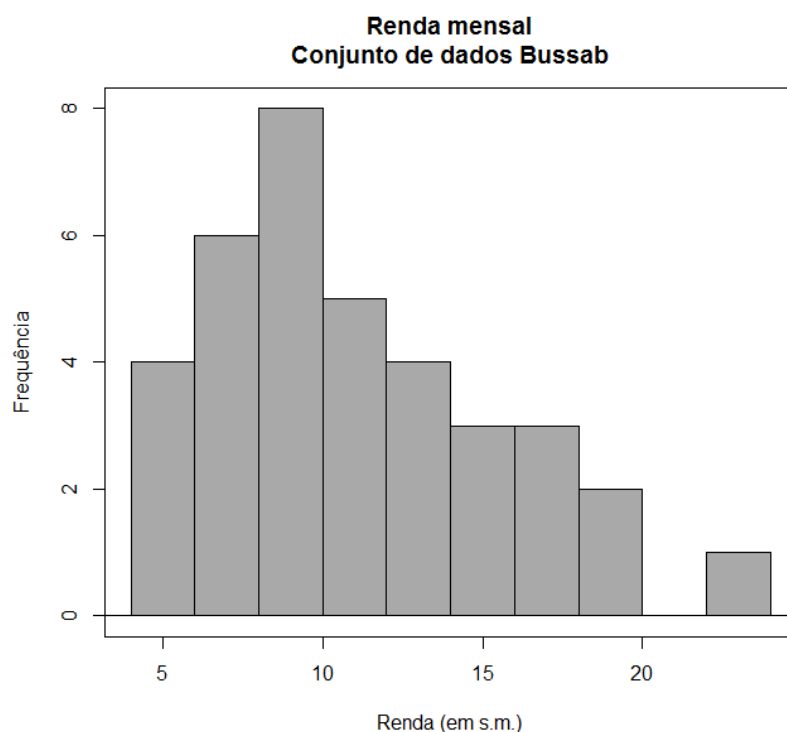


Figura 27: Histograma da variável *Renda*.

Na construção de histogramas pelo R Commander, é possível gerar histogramas (de uma variável quantitativa) de diferentes grupos do conjunto de dados de acordo com alguma variável qualitativa deste mesmo conjunto de dados, o que nos permite comparar o comportamento de uma mesma variável quantitativa em diferentes grupos dentro do

conjunto de dados sob análise. Por exemplo, podemos gerar um histograma da renda dos indivíduos casados e um histograma da renda dos indivíduos que não são casados, preferencialmente em escala percentual, pois a frequência de indivíduos que são casados e a frequência de indivíduos que não são casados não são iguais. Para tal, acesse o menu de construção de histogramas, escolha a variável *Renda* e clique na caixa *Gráfico por grupos*. No campo *Variável de grupo*, selecione *Casado_fator* e dê OK nesta janela. Na aba *Opções*, selecione *Percentagens* no campo *Escala do eixo* e atente ao rótulo do eixo-y, caso tenha editado-o anteriormente. Clique em OK para obter o gráfico exibido na Figura 28 (o título foi deixado em branco em virtude do espaço reduzido na janela *R Graphics*).

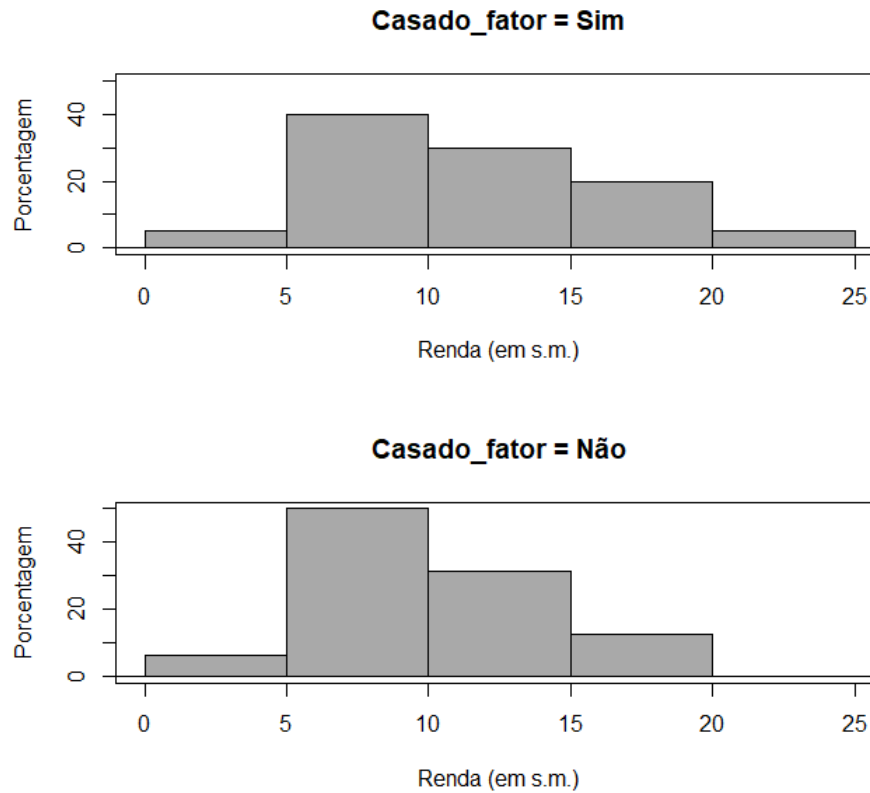


Figura 28: Histogramas da variável *Renda* para os casados e para os não casados.

Observação 19 Para fazer um histograma sem divisão por grupos (como o exibido na Figura 27) após fazer histogramas por grupos (como os exibidos na Figura 28), é necessário clicar no botão *Resetar* (ou *Reset*) na janela de construção de histogramas.

4.8 Boxplot

Além do histograma, outro gráfico bastante utilizado para analisar graficamente uma variável quantitativa é o boxplot. Para obtê-lo, acesse o menu *Gráficos → Boxplot*. A janela de construção do boxplot se assemelha bastante à janela de construção do histograma. A diferença é que, no lugar do número de classes e da escala de altura das barras, temos o campo *Identificar “outliers”*, que remete à identificação de cada observação que é considerada valor discrepante na variável em análise. Todavia, tal identificação pode deixar o gráfico visualmente poluído, pois os números (ou nomes) de identificação das observações podem ficar “um por cima do outro” quando ocorrem valores discrepantes muito próximos. Para um gráfico mais “limpo”, marque a opção *Não* em *Identificar “outliers”*, ou a opção *Com o mouse*. Neste último caso, os valores discrepantes são identificados quando clicamos sobre eles no gráfico (clicar com o botão direito sobre a janela *R Graphics* abre a opção de encerramento da marcação dos valores discrepantes).

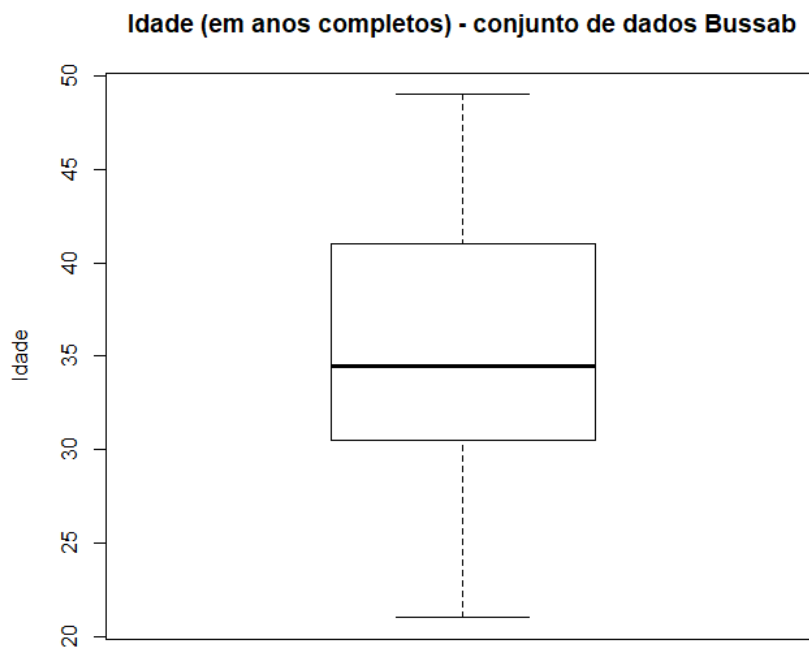


Figura 29: Boxplot da variável *Idade*.

Podemos fazer boxplots por grupos no R Commander, tal como fora feito para histogramas, gráficos de pontos e gráficos de hastes. O procedimento é análogo aos

apresentados anteriormente: clique na caixa *Gráfico por grupos* no menu de construção do gráfico, selecione a variável de grupo e clique em OK. Na Figura 30, seguem boxplots da variável *Renda* do conjunto de dados *Bussab*, agrupados pelo nível de instrução (variável *Instrução* escolhida como variável de grupo).

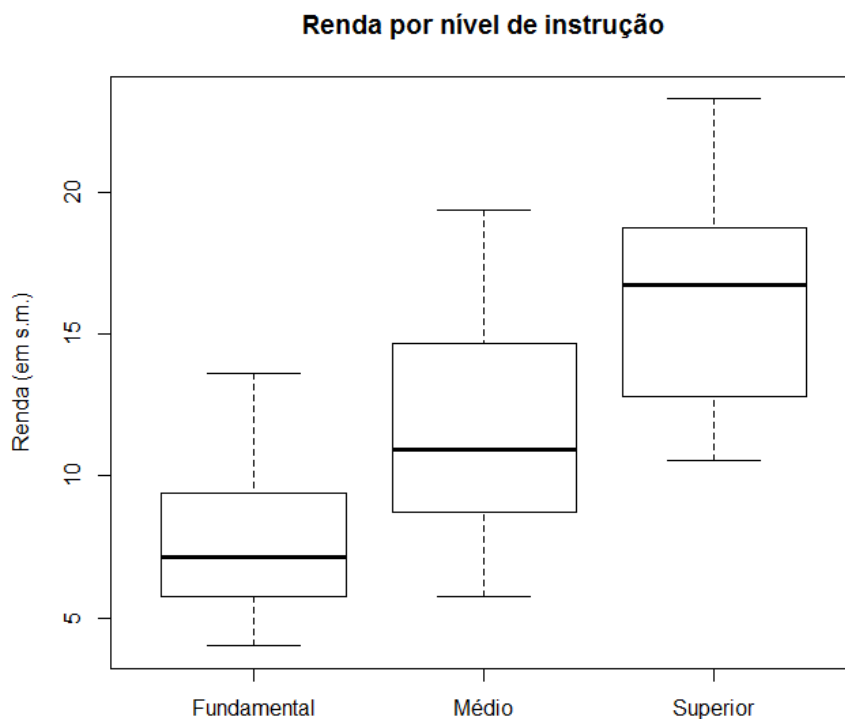


Figura 30: Boxplots da variável *Renda* agrupados pelo nível de instrução.

Observação 20 Para gerar um boxplot simples (sem divisão por grupos) após ter feito boxplots por grupos, clique em *Resetar* (ou *Reset*) na janela de construção de boxplots.

4.9 Diagrama de dispersão

O diagrama de dispersão é um gráfico que permite visualizar associação (ou falta de associação) entre duas variáveis quantitativas em um conjunto de dados e pode ser gerado por meio do menu *Gráficos* → *Diagrama de Dispersão*. Por exemplo, para visualizarmos como estão associadas as idades e as rendas dos indivíduos do conjunto de dados *Bussab*, acesse este menu e, na aba *Dados*, escolha *Idade* para variável-x e *Renda* para variável-y. Para evitar excesso de informações em um gráfico que veremos pela primeira vez,

preencha a aba *Opções* tal qual mostra a Figura 31 e dê OK. A Figura 32 mostra o diagrama de dispersão obtido, onde a linha verde representa a reta de regressão (linha de quadrados mínimos).

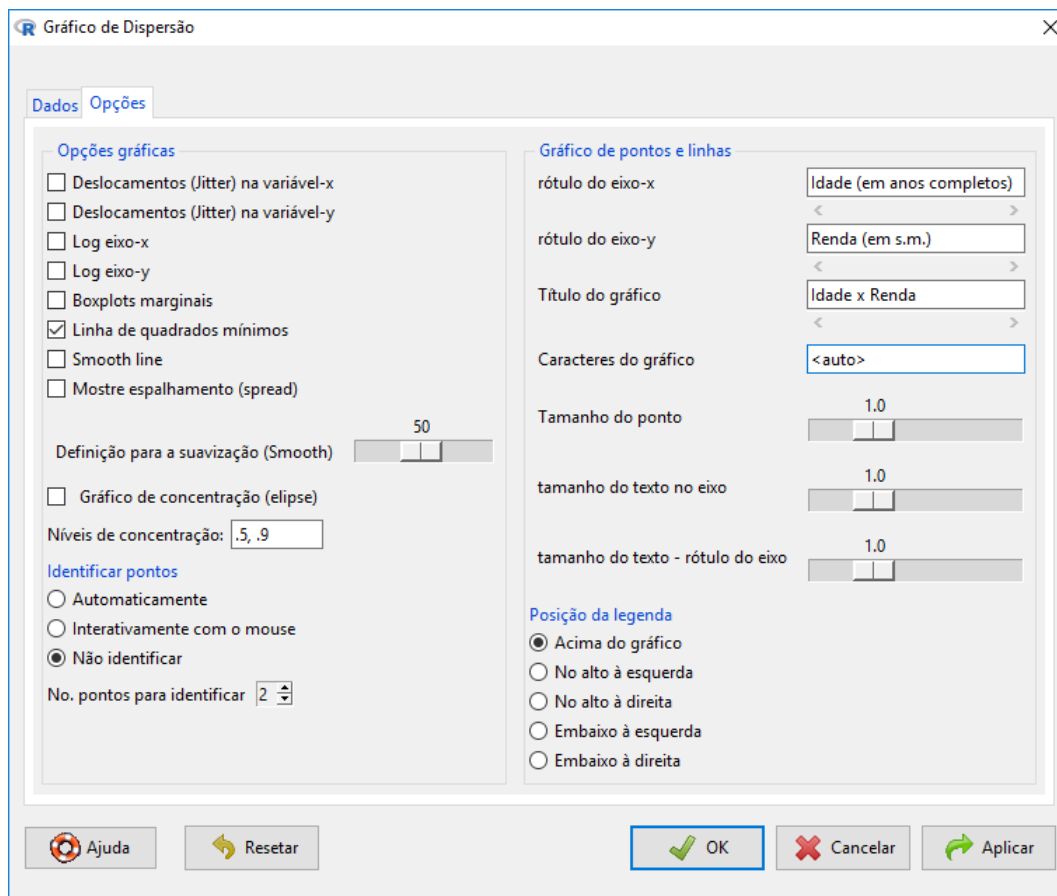


Figura 31: Construindo diagrama de dispersão.

Após obter este gráfico, acesse novamente a janela de construção de diagramas de dispersão, mas agora clique na caixa *Gráfico por grupos*, selecione *Casado_fator* como variável de grupo e dê OK⁶. Isto gera dois diagramas de dispersão sobrepostos (Figura 33): um diagrama de dispersão das variáveis *Idade* e *Renda* dos indivíduos casados; e um diagrama de dispersão entre as mesmas variáveis, porém apenas entre os indivíduos que não são casados. As cores associadas a cada diagrama de dispersão baseiam-se na paleta apresentada na Subseção 4.1, considerando inclusive a cor do primeiro retângulo.

⁶Se a caixinha *Gráfico de linhas por grupo* (na janela de escolha da variável de grupo) for desmarcada, o gráfico exibirá a reta de regressão de todas as observações ao invés das retas de regressão de cada grupo.

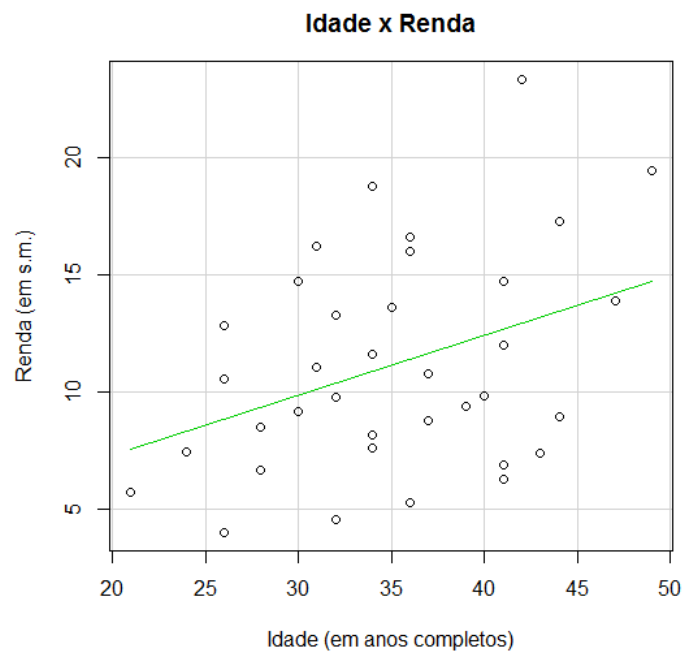


Figura 32: Diagrama de dispersão entre as variáveis *Idade* e *Renda*.

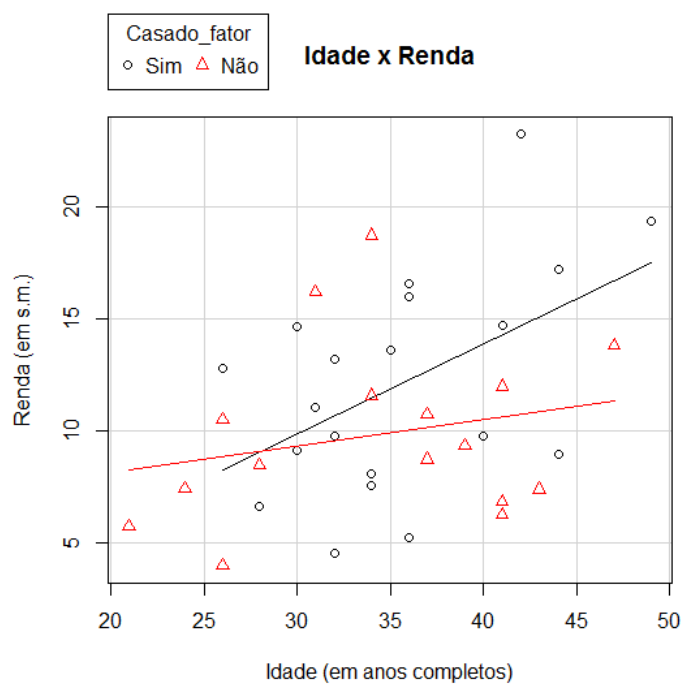


Figura 33: Diagramas de dispersão entre as variáveis *Idade* e *Renda* dos indivíduos casados (cor preta) e dos indivíduos não casados (cor vermelha).

Observação 21 Para construir um diagrama de dispersão sem divisão por grupos (como o exibido na Figura 32) após executar diagramas de dispersão por grupos (conforme exposto na Figura 33), é necessário clicar no botão *Resetar* (ou *Reset*) na janela de construção de diagramas de dispersão.

Observação 22 Na construção de diagramas de dispersão (Figura 31), podemos escrever um número inteiro de 1 até 25 no lugar de *<auto>* no campo *Caracteres* do gráfico. Cada número retorna um símbolo diferente para os pontos do diagrama de dispersão. O número 19, por exemplo, resulta em círculos preenchidos. No caso de diagramas de dispersão por grupos, é necessário escrever um número para cada atributo da variável de grupo, separados por vírgulas.

4.10 Salvando gráficos

Para **salvar um gráfico pela janela R Gui**, clique na janela do gráfico (*R Graphics*), dimensione-a na forma que desejar e vá em *Arquivo → Salvar como* (ou *File → Save as*). Escolha o formato do arquivo de imagem (sugestão: PNG).

Para **salvar um gráfico pelo R Commander**, existem dois caminhos: salvar como figura (*Gráficos → Salvar gráfico em arquivo → como Bitmap*) ou como PDF (*Gráficos → Salvar gráfico em arquivo → como PDF/Postscript/EPS*).

A desvantagem em salvar figuras pelo R Commander é que não conseguimos visualizar de antemão a dimensão da figura que estamos salvando. Já pelo *R Gui*, podemos visualizar como ficará o gráfico antes de salvá-lo, ao manusear diretamente a janela *R Graphics*.

5 O menu *Estatísticas*

Em geral, a análise de gráficos consiste apenas de parte de uma análise estatística. Nesta seção, veremos alguns comandos do menu *Estatísticas*, que tornarão nossas análises mais completas. **Os resultados destes comandos serão exibidos na janela *Output* do R Commander, na cor azul.**

Apresentaremos aqui apenas os principais comandos dos dois primeiros submenus do menu *Estatísticas*: *Resumos* e *Tabelas de Contingência*.

5.1 Resumos numéricos de todas as variáveis

Este comando gera um resumo numérico de cada variável do conjunto de dados ativo. Para variáveis qualitativas, é gerada a distribuição de frequência dos seus atributos (níveis). E para variáveis quantitativas, é feito, de fato, um resumo numérico (isto é, um conjunto de medidas descritivas) contendo: menor valor, 1º quartil, mediana, média, 3º quartil e maior valor. Também são computados os dados faltantes (NA's) em cada variável, caso existam. O caminho para gerar estes resumos é *Estatísticas* → *Resumos* → *Conjunto de dados ativo*.

5.2 Resumo numérico de uma variável (quantitativa)

Para gerar um resumo numérico de uma variável quantitativa em particular, vá em *Estatísticas* → *Resumos* → *Resumos numéricos*. Na aba *Dados*, escolha a(s) variável(is) de interesse. Na aba *Estatísticas*, é possível escolher quais medidas descritivas serão geradas pelo comando, entre elas: média, mediana, 1º quartil, 3º quartil, desvio padrão, distância interquartil e coeficiente de variação.

Deixando a aba *Estatísticas* na sua forma original, o resumo numérico da variável *Renda* (do conjunto de dados *Bussab*) segue na Figura 34, no qual temos: média (*mean*), desvio padrão (*sd*, de *standard deviation*), distância interquartil (*IQR*), menor valor (0%), 1º quartil (25%), mediana (50%), 3º quartil (75%), maior valor (100%) e total de observações (*n*). Quando a variável apresenta dados faltantes, também é exibido no resumo numérico quantos deles temos, isto é, o total de NA's para a variável em questão no conjunto de dados ativo.

mean	sd	IQR	0%	25%	50%	75%	100%	n
11.12222	4.587458	6.5075	4	7.5525	10.165	14.06	23.3	36

Figura 34: Resumo numérico da variável *Renda* do conjunto de dados *Bussab*.

Note que é possível fazer resumos numéricos por grupos ao clicar na caixa *Resuma por grupos*, de forma análoga a feita com gráficos de pontos, gráficos de hastes, histogramas, boxplots e diagramas de dispersão ao clicar na caixa *Gráfico por grupos*. Na Figura 35, segue o resumo numérico da variável *Renda*, com a variável *Instrução* escolhida como

variável de grupo, o que gera três resumos numéricos de rendas (em salários mínimos) dispostos em três linhas. A primeira linha corresponde ao resumo numérico das rendas dos indivíduos do conjunto de dados *Bussab* cujo grau de instrução é ensino fundamental. Analogamente, a segunda e a terceira linha correspondem ao resumo numérico das rendas dos indivíduos com ensino médio e com ensino superior como nível de instrução, respectivamente.

	mean	sd	IQR	0%	25%	50%	75%	100%	data:n
Fundamental	7.815833	2.910765	3.1550	4.00	6.0075	7.125	9.1625	13.6	12
Médio	11.542222	3.723802	5.6425	5.73	8.8375	10.910	14.4800	19.4	18
Superior	16.475000	4.502438	4.7300	10.53	13.6475	16.740	18.3775	23.3	6

Figura 35: Resumos numéricos por nível de instrução da variável *Renda* do conjunto de dados *Bussab*.

Observação 23 Para gerar um resumo numérico simples (sem divisão por grupos) após ter sido feito um resumo numérico por grupos, clique em *Resetar* (ou *Reset*) ao acessar o menu de geração de resumos numéricos.

Observação 24 Na janela de geração de resumos numéricos, note que há uma caixinha (por padrão, desmarcada) chamada *Binned Frequency Counts* na aba *Estatísticas*. Marcando esta caixinha, o resumo numérico gerado vem acompanhado de uma distribuição de frequências por classes da variável escolhida, o que pode ser muito útil para verificar exatamente as alturas das barras do histograma desta mesma variável, desde que as classes sejam as mesmas. Como exercício, refaça o resumo numérico da variável *Renda* (sem divisão por grupos), porém marcando a caixinha *Binned Frequency Counts*, e compare a frequência de cada classe com a altura de cada barra do histograma na Figura 27.

5.3 Contando dados faltantes

Apesar dos resumos numéricos debatidos na subseções anteriores computarem dados faltantes (comumente representados por *NA*), há um menu no R Commander específico para tal: *Estatísticas* → *Resumos* → *Contar observações faltantes*. Ao acessar este menu, o R Commander exibe (na janela *Output*) o total de dados faltantes de todas as variáveis

do conjunto de dados ativo. Em particular, no conjunto de dados *Bussab*, originalmente apenas a variável *Filhos* apresenta NA's, pois na coleta de dados que gerou este conjunto de dados, a informação do número de filhos foi restrita aos indivíduos casados.

Observação 25 *Há situações nas quais pode ser interessante remover linha(s) do conjunto de dados na(s) qual(is) uma ou mais variáveis apresentem dados faltantes. Isto pode ser feito rapidamente no R Commander por meio do menu [Dados → Conjunto de dados ativo → Remover observações com dados faltantes](#). Por segurança, é possível dar um outro nome para o conjunto de dados sem as “linhas indesejáveis”, de forma a preservar o conjunto de dados com todas as observações.*

5.4 Distribuições de frequências

Para obtermos a distribuição de frequências de uma variável no R Commander, basta acessarmos o menu [Estatísticas → Resumos → Distribuições de frequência](#), escolhermos a(s) variável(is) de interesse e clicarmos em OK. Este comando produz duas tabelas de distribuição de frequências para cada variável escolhida: uma com contagens (frequências absolutas) e outra com percentuais (frequências relativas $\times 100$).

counts:		
Instrução		
Fundamental	Médio	Superior
12	18	6
percentages:		
Instrução		
Fundamental	Médio	Superior
33.33	50.00	16.67


Figura 36: Distribuição de frequências da variável *Instrução* do conjunto de dados *Bussab*.

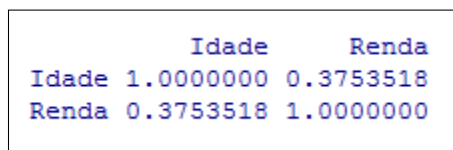
Note que as variáveis disponíveis para escolha são apenas as variáveis qualitativas⁷ do conjunto de dados ativo. Porém, quando temos uma variável quantitativa discreta com poucos valores distintos, também é interessante obter sua tabela de distribuição de frequências. Isto pode ser contornado convertendo esta variável quantitativa em rótulos

⁷A rigor, variáveis que o R Commander “enxerga” como qualitativas, isto é, variáveis não-numéricas.

numéricos (ver Seção 3.7.1), criando assim uma nova coluna no conjunto de dados, a qual será interpretada pelo R Commander como uma variável qualitativa e, portanto, passível da aplicação de distribuição de frequências.

5.5 Matriz de correlação

Na Seção 4.9, mostramos como obter um diagrama de dispersão (com a reta de regressão sobreposta ao gráfico). Quando fazemos esta representação gráfica, é importante também relatar o coeficiente de correlação entre as variáveis analisadas. Por meio da matriz de correlação, podemos verificar o(s) coeficiente(s) de correlação de um ou mais pares de variáveis quantitativas do conjunto de dados ativo. Tal matriz é obtida no R Commander no menu *Estatísticas* → *Resumos* → *Matriz de correlação*. Uma vez acessado este menu, escolha duas ou mais variáveis (segurando a tecla  enquanto as seleciona) e não altere o restante da janela. Podemos verificar na Figura 37 que o coeficiente de correlação (de Pearson) entre as variáveis *Idade* e *Renda* do conjunto de dados *Bussab* é aproximadamente 0,375, que configura uma correlação positiva moderada.



	Idade	Renda
Idade	1.0000000	0.3753518
Renda	0.3753518	1.0000000

Figura 37: Matriz de correlação com as variáveis *Idade* e *Renda* do conjunto de dados *Bussab*.

5.6 Tabela de contingência

Para um conjunto de dados com pelo menos duas variáveis qualitativas, é possível construir tabelas de contingência (tabelas de dupla entrada). Para tal, o caminho é *Estatísticas* → *Tabelas de Contingência* → *Tabela de dupla entrada*.

Na janela de construção da tabela de contingência, selecione as duas variáveis de interesse na aba *Dados* (a ordem de escolha não é relevante). Na aba *Estatísticas*, temos as opções de fazer testes de hipóteses baseados na tabela de contingência (o único que já está marcado por padrão é o teste de independência de qui-quadrado). Deixando esta

aba do jeito que está e escolhendo as variáveis *Casado_fator* e *Instrução* (do conjunto de dados *Bussab*) para comporem a tabela de contingência, obtemos uma saída conforme expressa na Figura 38: acima de *Pearson Chi-squared test*, encontra-se de fato a tabela de contingência solicitada, e abaixo de *Pearson Chi-squared test* é listado o resultado do teste de independência de qui-quadrado (*p-value* expressa o p-valor do teste).

```
Frequency table:
      Instrução
Casado_fator Fundamental Médio Superior
      Sim           6      11       3
      Não           6       7       3

      Pearson's Chi-squared test

data:  .Table
X-squared = 0.45, df = 2, p-value = 0.7985
```

Figura 38: Tabela de contingência com as variáveis *Casado_fator* e *Instrução* do conjunto de dados *Bussab*, seguida do teste qui-quadrado de Pearson (teste de independência de qui-quadrado) entre estas variáveis.

Além do campo *Testes de hipóteses*, a aba *Estatísticas* do menu de construção de tabelas de dupla entrada apresenta o campo *Computar Percentagens*. A opção padrão *Sem percentual* expressa apenas a tabela de dupla entrada exibida na Figura 38. As outras três opções deste campo (*Percentual nas linhas*, *Percentual nas colunas* e *Percentagens do total*) exibem, além da tabela de dupla entrada, uma tabela com a mesma estrutura, cujas entradas são todas na escala percentual (exceto na linha ou coluna *Count*), conforme as descrições abaixo.

- **Percentual nas linhas:** Apresenta o perfil percentual da variável escolhida como variável coluna em cada atributo da variável escolhida como variável linha, como ilustra a Figura 39. As entradas na coluna *Count* consistem do total de observações que satisfazem o atributo da respectiva linha.
- **Percentual nas colunas:** Apresenta o perfil percentual da variável escolhida como variável linha em cada atributo da variável escolhida como variável coluna, como

ilustra a Figura 40. As entradas na linha *Count* consistem do total de observações que satisfazem o atributo da respectiva coluna.

- **Percentagens do total:** Apresenta o percentual de cada cruzamento de atributos das duas variáveis em relação ao total de observações (ver Figura 41).

```
Row percentages:
      Instrução
Casado_fator Fundamental Médio Superior Total Count
      Sim      30.0  55.0    15.0 100.0    20
      Não      37.5  43.8    18.8 100.1    16
```

Figura 39: Perfil percentual da variável *Instrução* (escolhida como variável coluna) em cada atributo da variável *Casado_fator* (escolhida como variável linha).

```
Column percentages:
      Instrução
Casado_fator Fundamental Médio Superior
      Sim           50  61.1      50
      Não           50  38.9      50
      Total        100 100.0     100
      Count         12  18.0       6
```

Figura 40: Perfil percentual da variável *Casado_fator* (escolhida como variável linha) em cada atributo da variável *Instrução* (escolhida como variável coluna).

```
Total percentages:
      Fundamental Médio Superior Total
Sim      16.7  30.6      8.3  55.6
Não      16.7  19.4      8.3  44.4
Total    33.3  50.0     16.7 100.0
```

Figura 41: Percentual de cada cruzamento dos atributos das variáveis *Casado_fator* e *Instrução* em relação ao total de observações.

Observação 26 A coluna Total da tabela de percentuais nas linhas e a linha Total da tabela de percentuais nas colunas representam as somas das percentagens na respectiva linha ou coluna. Matematicamente, estas somas sempre resultam em 100%. Todavia, alguns dos percentuais vistos nas três figuras acima estão aproximados por apenas uma casa decimal, e isto pode gerar uma soma equivocada. Por exemplo, na Figura 39, a entrada na coluna Total (linha Não) retorna 100,1%. Este percentual é, na verdade, uma soma de percentuais que passaram por tal aproximação de apenas uma casa decimal: as entradas 43,8% e 18,8% são aproximações de, respectivamente, $7/16 = 43,75\%$ e $3/16 = 18,75\%$. De fato, temos $37,5\% + 43,75\% + 18,75\% = 100\%$.

Referências bibliográficas

- Bussab, W. & Morettin, P. (2010). **Estatística Básica**. 6ª edição. São Paulo: Editora Saraiva.
- Fox, J. & Bouchet-Valat, M. (2017). *Getting started with the R Commander - Version 2.4-0*. Disponível em: <https://socialsciences.mcmaster.ca/jfox/Misc/Rcmdr/Getting-Started-with-the-Rcmdr.pdf>
- Fox, J. (2005). *The R Commander: A Basic-Statistics Graphical User Interface to R*. Disponível em: <https://www.jstatsoft.org/article/view/v014i09/v14i09.pdf>
- Mignozzetti, U.G. (2009). *Introdução ao R Commander*. Disponível em: <http://www.nadd.prp.usp.br/cis/arqs/aprcmdr.pdf>