Importação e Exportação de dados

Cristiano de Carvalho Santos
professor.pacotes.estatisticos@gmail.com
Grupo Google: PacotesEstatisticos2016

Departamento de Estatística,
Universidade Federal de Minas Gerais (UFMG)

Importação de dados

- Muitas vezes os dados que iremos utilizar já foram digitados e armazenados num arquivo utilizando outro programa.
- Para importar dados externos quando eles estão em formato texto sem a necessidade de digitá-los novamente podemos considerar as funções do R:

$$scan(\ ),\quad read.table()\quad e\quad read.csv(\ ).$$

read.table

Principais argumentos:

- file: Nome do arquivo;
- header: um argumento lógico indicando se o arquivo contém os nomes das variáveis na primeira linha;
- sep: argumento que diz como os dados estão separados em cada linha. Se sep = " " (padrão) o separador é espaço em branco;
- dec: é o simbolo usado para separar decimal no banco de dados;

read.csv

- As funções read.csv(), read.csv2(), read.delim() e read.delim2() são praticamente iguais a read.table() porém com diferentes opções default.
- Em geral (mas não sempre) dados em formato csv usado no Brasil são lidos diretamente com read.csv2().

Instalação de pacotes

1. Menu Pacotes > instalar Pacotes > 0-Cloud > Escolher o pacote;
2. Menu Pacotes > carregar Pacote > Escolher o pacote;
3. Toda vez que o R for iniciado é necessário carregar os pacotes que serão utilizados;
4. É possível instalar e carregar pacotes através de comandos (ver script).

Arquivos em outros formatos

- É possível ler dados diretamente de outros formatos que não seja texto (ASCII).
- Isto em geral é mais eficiente e requer menos memória do que converter para formato texto.
- Muitas funções que permitem a importação de dados de outros programas são implementadas no pacote foreign.

Algumas destas funções são:

- read.dbf() para arquivos DBASE
- read.epiinfo() para arquivos .REC do Epi-Info
- read.mtp() para arquivos "Minitab Portable Worksheet"
- read.S() para arquivos do S-PLUS restore.data() para "dumps"do S-PLUS
- read.spss() para dados do SPSS
- read.dta() para dados do STATA
- read.xport() para ler do formato TRANSPORT do SAS
- read.ssd() pode escrever dados permanentes do SAS (.ssd ou .sas7bdat) no formato TRANSPORT
- O pacote Hmisc disponibiliza sas.get() que também requer o SAS no sistema

Para importar dados do MS-Excel para o R.

- A função read_excel do pacote readxl;
- A função read.xlsx do pacote xlsx;
- Um outro pacote capaz de ler dados diretamente de planilhas é o RODBC.

```
> require(RODBC)
> xlscon <- odbcConnectExcel("planilha.xls")
> dados1 <- sqlFetch(xlscon, "Planilha1")
> odbcClose(xlsConnect)
> head(dados1)
```

Exportação de dados

- write.table(mydata, "mydata.txt", sep="");
- A função WriteXLS do pacote WriteXLS;
- write.xlsx(mydata, "mydata.xlsx") do pacote xlsx;
- No pacote foreign:
  - SPSS: write.foreign(mydata, "mydata.txt", "mydata.sps", package="SPSS")
  - SAS: write.foreign(mydata, "mydata.txt", "mydata.sas", package="SAS")
  - STATA: write.dta(mydata, "mydata.dta")

Função attach

- A função attach é utilizada para fixar o data.frame a ser utilizado;
- Podemos fazer referência à variável apenas pelo seu nome, sem necessidade de usar o símbolo \$.
- Após terminar de trabalhar com o data.frame fixado utilize o comando detach (nome do dataset) para desabilitar o comando attach.

Análise descritiva - variáveis qualitativas

Construção de tabelas de frequências:

- As funções table e ftable podem ser usadas para construir tabelas de frequências;
- prop.table pode ser usada para calcular frequências relativas;
- round pode ser usada para determinar o número de casas decimais na exibição dos números;
- margin.table é utilizada para calcular as somas marginais em tabelas de frequências com mais de uma variável.

Construção de gráficos de frequências:

- barplot: função para gráfico de barras;
- pie: função para gráfico de pizza.

Principais parâmetros gráficos do R

- Função *par* define vários parâmetros da janela gráfica atual. Por exemplo, $par(mfrow = c(1, 2))$ permite colocar dois gráficos lado a lado.
- Cada tipo de gráfico permite configurar inúmeros parâmetros. Alguns deles são:
  - main: Título do gráfico;
  - sub: Subtítulo do gráfico;
  - xlab e ylab: nomes dos eixos;
  - col: relacionado a cor do gráfico;
  - cex.axis, cex.lab, cex.main, cex.sub: relativo ao tamanho da letra;
  - cex: relacionado a tamanho de pontos;
  - xlim e ylim: referente a escala que aparece nos eixos x e y;

Análise descritiva e exploratória de dados - variáveis quantitativas

Algumas funções usadas para a calcular medidas resumo de variáveis quantitativas são:

- mean, median, summary, quantile;
- sum, max, min;
- var, sd;

Gráficos

- Gráfico de dispersão;
- Histograma;
- Boxplot;
- Gráfico de ramo e folhas;
- Diagrama de pontos

Lista de Exercícios 2

- Forma de entrega: Mandar por email um arquivo ".txt" ou ".R" com os comandos utilizados na resolução da lista de exercícios.
- Salvar arquivo com nome Lista2-nomes dos autores-incompleta ou Lista1-nomes dos autores-final.

Exercícios:

1. Pesquise no help sobre a função *readLines* e use esta função para definir um vetor de caracteres com 5 nomes.

2. Pesquise como alterar um valor em um vetor. Por exemplo, suponha o vetor $(1, 2, \ldots, 100)$ e altere o valor da décima posição do vetor do número 10 para o número 57.

3. Escolha 3 bancos de dados entre os disponíveis no R em data() e que ainda não foram utilizados na disciplina. Salve estes bancos em arquivos ".txt", "csv" e "xls" (cada banco em um formato distinto). Importe estes dados novamente para o R, atribuindo-os aos objetos Dados1, Dados2 e Dados3, com as funções adequadas.

4. Faça os exercícios da página 54 do relatório técnico "BIOESTATÍSTICA BÁSICA USANDO O AMBIENTE COMPUTACIONAL R" disponível no site do departamento de estatística da UFMG.

5. Construa uma função para calcular os escores padronizados de uma variável.

6. Faça uma análise descritiva adequada do conjunto de dados women.

7. Alguns livros apresentam o histograma utilizando a frequência relativa no lugar da densidade. Descubra como fazer um histograma desta forma utilizando o pacote "histogram".

8. Pesquise na internet como fazer um gráfico de frequência acumulada.

9. Carregue o conjunto de dados USArrests com o comando data(USArrests). Examine a sua documentação com help(USArrests) e responda as perguntas a seguir.

   a) qual o número médio e mediano de cada um dos crimes?
   b) encontre a mediana e quartis para cada crime.
   c) encontre o número máximo e mínimo para cada crime.
   d) faça um gráfico adequado para o número de assassinatos (murder).
   e) faça um diagrama ramo-e-folhas para o número de estupros (rape).

f) verifique se há correlação entre os diferentes tipos de crime.
g) verifique se há correlação entre os crimes e a proporção de população urbana.
h) encontre os estados com maior e menor ocorrência de cada tipo de crime.
i) encontre os estados com maior e menor ocorrência per capta de cada tipo de crime.
j) encontre os estados com maior e menor ocorrência do total de crimes.