

Getting R

Download R from the R project website <http://www.r-project.org/> which requires a few clicks or directly from <http://cran.stat.ucla.edu/>. There are Windows, Mac, and Unix versions. These notes are for the Windows version. There will be minor differences for the other versions.

Useful reference manuals

- <http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf> is *SimpleR* by John Verzani.
- *R for Beginners* by Emanuel Paradis.

Where are the data?

- Some datasets are built into the basic package of R. For example there is a dataset of 150 observations of iris plants. If you know the name of the dataset (in this case `iris` you simply use the command

```
> data(iris)
> iris
   Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
1           5.1           3.5           1.4           0.2   setosa
2           4.9           3.0           1.4           0.2   setosa
3           4.7           3.2           1.3           0.2   setosa
4           4.6           3.1           1.5           0.2   setosa
```

which defines a `data.frame` of the same name. In this example we have shown the first four cases of this dataset.

- In some cases, the dataset is contained in a package that is not yet loaded into R. For example, the `alfalfa` dataset is in the `faraway` package.

```
> data(alfalfa)
Warning message:
data set 'alfalfa' not found in: data(alfalfa)
> library(faraway)
> data(alfalfa)
> alfalfa
   shade irrigation inoculum yield
1      1           1         A  33.8
2      1           2         B  33.7
3      1           3         D  30.4
4      1           4         C  32.7
```

In this example, you might have to download the `faraway` package from the R website. The command `data(package = .packages(all.available = TRUE))` will list all available datasets.

- In some cases, you may need to download the data from a file that is stored on your computer or on the web. It is important that R understand the format in which the data is stored. A very good format is CSV (comma separated values) since many programs read data in this format and Excel can easily create data in this format. Many such datasets can be found at <http://www.calvin.edu/stob/data/>. For example, team statistics for the recently completed baseball season can be found at <http://www.calvin.edu/stob/data/baseball2007.csv>. You can save that file to your own machine or read it directly from the web

```
> bball=read.csv('http://www.calvin.edu/~stob/data/baseball2007.csv')
> bball
      CLUB LEAGUE  BA  SLG  OBP  G  AB  R  H  TB X2B X3B HR
1    New York    A 0.290 0.463 0.366 162 5717 968 1656 2649 326 32 201
2    Detroit    A 0.287 0.458 0.345 162 5757 887 1652 2635 352 50 177
3    Seattle    A 0.287 0.425 0.337 162 5684 794 1629 2416 284 22 153
```

Printing Text

The easiest thing to do is to cut and paste the output from the **R** console to a wordprocessing program. For PCs, Notepad works fine.

Printing Graphics

You have a couple of choices.

- Make the graphics window active and then select **print** from the **file** menu.
- Choose **Save to File** from the **File** menu and save it in an appropriate format (PDF, JPG, etc.)

Data Frames

Probably the most important thing to remember about data structures in **R** is that, almost always, data are stored in **data.frames**. **data.frames** have rows and columns. The rows correspond to the individuals in the dataset. The columns correspond to the variables. The rows are often named and the columns are always named. Each individual column is a vector. The following code illustrates how individual rows and columns are accessed using the **bball** dataset loaded above.

```
> bball[1,] # gets the first individual, all variables
      CLUB LEAGUE  BA  SLG  OBP  G  AB  R  H  TB X2B X3B HR RBI SH SF
1 New York    A 0.29 0.463 0.366 162 5717 968 1656 2649 326 32 201 929 41 54
  HBP BB IBB  SO  SB CS GDP  LOB SHO  E  DP TP
1  78 637 32 991 123 40 138 1249 8 88 174 0
> bball$R # gets the vector of values in the variable R (runs)
[1] 968 887 794 822 867 756 811 782 718 816 706 753 741 693 860 804 735 810 892
[20] 725 752 790 783 724 801 723 673 683 741 712
> bball[,1] # gets the values of the first variable (team name)
[1] New York    Detroit    Seattle    Los Angeles  Boston
[6] Baltimore    Cleveland    Tampa Bay    Minnesota    Texas
[11] Kansas City  Toronto    Oakland    Chicago    Colorado
[16] New York    Los Angeles  Atlanta    Philadelphia  St. Louis
[21] Chicago    Florida    Cincinnati  Pittsburgh    Milwaukee
[26] Houston    Washington    San Francisco San Diego    Arizona
27 Levels: Arizona Atlanta Baltimore Boston Chicago Cincinnati ... Washington
>
```