More advanced: Reading data files and R markdown options

## Reading internet data into R

Suppose you have a larger data list in a file. You can read it from the folder where you save the Session (this is explained in the document Rintro.docx on LON CAPA). For our class I saved a data file on the STT Class pages. Let's read in the cereals data and the body fat data. The cereals data is an excel file (.csv extension) and the body fat data is a text file (.txt extension). They can be read into R directly with the URL as follows. The R command "head" will show you the top few lines, the command "tail" what show you the last few lines

cereal=read.csv("http://stt.msu.edu/Academics/ClassPages/uploads/FS14/200-1/Ch08\_Cereals.csv", header=T)  
  
head(cereal)

## name calories sugar  
## 1 100%\_Bran 70 6  
## 2 100%\_Natural\_Bran 120 8  
## 3 All-Bran 70 5  
## 4 All-Bran\_with\_Extra\_Fiber 50 0  
## 5 Almond\_Delight 110 8  
## 6 Apple\_Cinnamon\_Cheerios 110 10

body=read.table("http://stt.msu.edu/Academics/ClassPages/uploads/FS14/200-1/Ch08\_Body\_fatv2.txt", header=T, sep="")  
  
head(body)

## Waist Weight BodyFat  
## 1 32 175 6  
## 2 36 181 21  
## 3 38 200 15  
## 4 33 159 6  
## 5 39 196 22  
## 6 40 192 31

tail(body)

## Waist Weight BodyFat  
## 15 44 246 38  
## 16 33 160 10  
## 17 41 215 27  
## 18 34 159 12  
## 19 34 146 10  
## 20 44 219 28

In the line that reads in the data the "header=TRUE" or "header=T" says that there are variable names in the data file. For the text file (body fat) we need to say how the columns are separated, in this case with a space, sep="".

Have a look at the upper right window under the "Environment" tab. There are now "Data" (cereal and body) and Values (x). If you click on "body" it will give you a quick overview what is in the data set. In this case there are three variables, all integers (=discrete numerical data) and the first few observations of each variable. Under values you see the data list you named "x".

The data set "body" has three variables (stored in columns) and 20 observations each. You can access the first column, Waist measurements, as follows

# body[row number, column number]  
# if the row number is left empty, all rows are shown, anmely all 20 Waist measurments  
body[,1]

## [1] 32 36 38 33 39 40 41 35 38 38 33 40 36 32 44 33 41 34 34 44

# or using the variable name  
  
body$Waist

## [1] 32 36 38 33 39 40 41 35 38 38 33 40 36 32 44 33 41 34 34 44

If you wanted to see all measurements of the 4th individual you type

body[4,]

## Waist Weight BodyFat  
## 4 33 159 6

## Reading text files (data) into R

Suppose you have a dataset in a file (in your Stat folder!). You can read these all at once.

# Read the bodyfat data if you have such a file  
# If there are variable names, add "header=TRUE"  
# Indicate the type of separator (commas "'", space "", etc)  
body<- read.table("Ch08\_Body\_fat.txt", header=TRUE, sep="")

## Warning in file(file, "rt"): cannot open file 'Ch08\_Body\_fat.txt': No such  
## file or directory

## Error in file(file, "rt"): cannot open the connection

You may get error messages, if the datafile is not well organized.You may need to open the text file in a text editor and have a look (don't use Word for data files). Rstudio thinks the first line has 7 columns while the next line has only three columns. So open the text file, change variable names, make sure the name of each column is only one word, and save the file as "Ch08\_Body\_fatv2.txt". If you make changes to any dataset, leave the original and save the changes in another file.

# Read the bodyfat data (text file)  
# Be sure you changed "Body Fat" to one word "Bodyfat"  
body<- read.table("Ch08\_Body\_fatv2.txt", header=TRUE, sep="")

You can look at the top and bottom few lines

head(body)

## Waist Weight BodyFat  
## 1 32 175 6  
## 2 36 181 21  
## 3 38 200 15  
## 4 33 159 6  
## 5 39 196 22  
## 6 40 192 31

tail(body)

## Waist Weight BodyFat  
## 15 44 246 38  
## 16 33 160 10  
## 17 41 215 27  
## 18 34 159 12  
## 19 34 146 10  
## 20 44 219 28

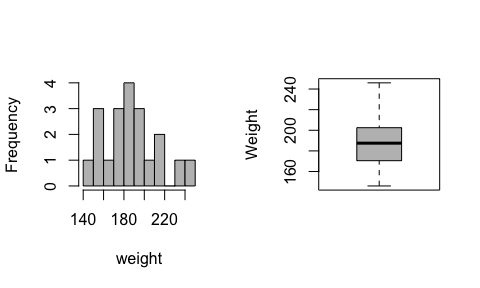
The structure in R called *data frame* is a collection of all variables. To extract values for the variable "Weight" we use the name of the dataset and the name of the Variable after a $ sign. Let's see the first three weight measurements from this dataset:

body$Weight[1:3]

## [1] 175 181 200

To see both a histogram and a boxplot in the same figure, we can tell R how to organize it: 1 row for 2 figures side-by-side. Try it different ways, too! There are also many ways to modify plots with labels of the axes, title, colors, lines, points, text, etc.

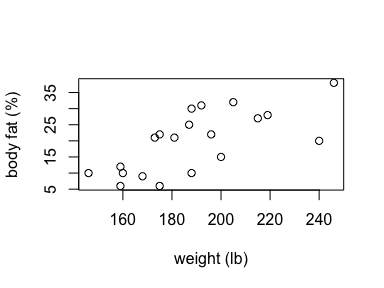
par(mfrow=c(1,2))  
hist(body$Weight, nclass=10, col="grey", main="", xlab="weight")  
boxplot(body$Weight, col="grey", ylab="Weight")



par(mfrow=c(1,1))

You can also embed plots, for example making a scatter plot of weight versus body fat. The plot function has the format "plot(x,y)" where x is the variable for the horizontal axis and y is the variable for the vertical axis.

plot(body$Weight, body$BodyFat, xlab="weight (lb)", ylab="body fat (%)")



## Reading Excel files into R

If you wanted to read an excel file instead of a textfile, you can use "read.csv", if it is a ".csv" file" or type the command install.packages(xlsx) and then do he following.

library(xlsx)

## Loading required package: rJava  
## Loading required package: xlsxjars

cereal<-read.xlsx("Ch08\_Cereals.xls", sheetName = "Sheet1")  
head(cereal)

## name calories sugar  
## 1 100%\_Bran 70 6  
## 2 100%\_Natural\_Bran 120 8  
## 3 All-Bran 70 5  
## 4 All-Bran\_with\_Extra\_Fiber 50 0  
## 5 Almond\_Delight 110 8  
## 6 Apple\_Cinnamon\_Cheerios 110 10

The correlation between sugar content and calories in a serving size of these cereal boxes is

cor(cereal$calories, cereal$sugar)

## [1] 0.5640922

## Things to know about the R chunk in R markdown files

* echo = FALSE in {r, echo=FALSE} will not show the R commands in the final document.
* include=FALSE will evaluate the R chunk, but no output will be shown in the document.
* Your document will stop compiling, if you have an error somehwere in the R code. It is recommended that you evaluate each R chunk separately to make sure it works. Also do a spell check before clicking on "Knit Word" or "Knit html" or "Knit pdf" for finalizing your document.
* You can use output from the R chunk in your text. For example, the correlation of calories and sugar content of cereal is 0.5640922. In order to do that you use a left apostrophe followed by "r" followed by a function or R variable followed by a left apostrophe `

## Other convenient features for R markdown documents

Here is a very neat feature of R markdown: you can add links to a text or to figures from the internet. For example, see [A short introduction to R] (<http://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf>) by Paul Torfs and Claudia Bauer. In order to do so you choose square brackets: [Here is the text for the link] followed by round brackets with the URL (<http://>...). If you want to insert a figure from the internet you do the same but with a ! in front of the square brackets, e.g. 