Data Analysis with R Laboratories – Sets of Exercises, with R Code

John Maindonald

July 31, 2007

Laboratory exercises make extensive use of datasets from the *DAAG* and *DAAGxtras* packages. Make sure that they are installed.

For background to laboratory exercises I – II, see the document: *The R System – An Introduction and Overview.*

Contents

I R Basics 3

1 Data Input 3

2 Data Input from a Web Page 3

3 The `paste()` Function; Further Details 3

4 Missing Values 4

5 Subsets of Dataframes 4

6 Scatterplots 4

7 Factors 6

8 Stripcharts (base graphics) and Stripplots (*lattice*) 7

9 Tabulation 7

10 Sorting 8

11 Use of a For Loop 8

12 A Function 8

II Practice with R 9

1 Information about the Columns of Data Frames 9

2 Data Input 9

3 A Tabulation Exercise 9

4 A For Loop 10

5 Data Exploration – Distributions of Data Values 10

6 Random Samples 11

7 Smooth Curves 12

8 Information on Workspace Objects 12

9 Different Ways to Do a Calculation – Timings 13

10 Functions – Making Sense of the Code 14

11 *Use of `sapply()` to Give Multiple Graphs 15

12 *The Internals of R – Functions are Pervasive 15

III Informal and Formal Data Exploration 17

1 Informal Data Exploration 17

2 Bootstrap sampling 19

IV Models and Model Accuracy 21

1 Straight Line Regression 21

2 Analysis Using *randomForest* 25

Part I
R Basics

1 Data Input

Exercise 1
Ensure that the *DAAGxtras* package is installed. From the R command line, attach it, and type `dataFile(c("molclock1", "molclock2"))`, thus:

```
> library(DAAGxtras)
> dataFile(c("molclock1", "molclock2"))  # NB dataFile, not datafile
```

This places the files **molclock1.txt** and **molclock2.txt** in the working directory. Use `file.show()` to examine the contents of each of these files. Alternatively, you can use R's script editor (under Windows, go to File | Open script...), or use another editor such as the Windows tinn-R editor that is designed to interface to R.

Use `read.table()` to read each of them into R. Check carefully whether you need `header=TRUE`. Then display the data frame and check that the data have been input correctly.

Note: If the file is located in a directory other than the working directory a fully specified file name, including the path, is necessary. For example, to input a file **travelbooks.txt** that has been placed in the directory **c:/datafiles/**, type

```
> travelbooks <- read.table("c:/datafiles/travelbooks.txt")
```

For input to R functions, forward slashes replace backslashes.

2 Data Input from a Web Page

Exercise 2
Files can be read directly from a web page, providing that there is a live internet connection, Here are examples:

```
> webfolder <- "http://www.maths.anu.edu.au/~johnm/datasets/text/"
> webpage <- paste(webfolder, "molclock.txt", sep="")
> molclock <- read.table(url(webpage))
```

Use this approach to input the file **travelbooks.txt** that is available from this same web page.

3 The `paste()` Function; Further Details

Exercise 3
Here are further examples that illustrate the use of `paste()`:

```
> paste("Leo", "the", "lion")
> paste("a", "b")
> paste("a", "b", sep="")
> paste(1:5)
> paste(1:5, collapse="")
```

4 Missing Values

Exercise 4
The following counts, for each species, the number of missing values for the column `root` of the data frame `rainforest` (*DAAG*):

```
> library(DAAG)
> with(rainforest, table(complete.cases(root), species))
```

For each species, how many rows are "complete", i.e., have no values that are missing?

Exercise 5
For each column of the data frame `Pima.tr2` (*MASS*), determine the number of missing values.

5 Subsets of Dataframes

Exercise 6
Use `head()` to check the names of the columns, and the first few rows of data, in the data frame `rainforest` (*DAAG*). Use `table(rainforest$species)` to check the names and numbers of each species that are present in the data. The following extracts the rows for the species *Acmena smithii*

```
> library(DAAG)
> Acmena <- subset(rainforest, species=="Acmena smithii")
```

The following extracts the rows for the species `Acacia mabellae` and `Acmena smithii`

```
> AcSpecies <- subset(rainforest, species %in% c("Acacia mabellae",
+                                                "Acmena smithii"))
```

Now extract the rows for all species except `C. fraseri`.

Exercise 7
Extract the following subsets from the data frame `ais` (*DAAG*):

(a) Extract the data for the rowers.

(b) Extract the data for the rowers, the netballers and the tennis players.

(c) Extract the data for the female basketabllers and rowers.

6 Scatterplots

Exercise 8
Using the Acmena data from the data frame `rainforest`, plot `wood` (wood biomass) vs `dbh` (diameter at breast height), trying both untransformed scales and logarithmic scales. Here is suitable code:

Exercise 8, continued

```
> Acmena <- subset(rainforest, species=="Acmena smithii")
> plot(wood ~ dbh, data=Acmena)
> plot(wood ~ dbh, data=Acmena, log="xy")
```

Use of the argument `log="xy"` gives logarithmic scales on both the x and y axes. For purposes of adding additional features to the plot, note that logarithms to base 10 are used.
For the second plot, add a fitted line, thus:

```
> plot(wood~dbh, data=Acmena, log="xy")
> ## The lm() command will fit a line; more details later
> ## abline() then plots this line.
> Acmena10.lm <- lm(log10(wood) ~ log10(dbh), data=Acmena)
> abline(Acmena10.lm)

> ## Now print the coefficents, for a log10 scale
> coef(Acmena10.lm)
> ## For comparison, print the coefficients for a natural log scale
> Acmena.lm <- lm(log(wood) ~ log(dbh), data=Acmena)
> coef(Acmena.lm)
```

Write down the equation that gives the fitted relationship between `wood` and `dbh`.

Exercise 9
The `orings` data frame gives data on the damage that had occurred in US space shuttle launches prior to the disastrous Challenger launch of January 28, 1986. Only the observations in rows 1, 2, 4, 11, 13, and 18 were included in the pre-launch charts used in deciding whether to proceed with the launch. Add a new column to the data frame that identifies rows that were included in the pre-launch charts. Now make three plots of `Total` incidents against `Temperature`:

(a) Plot only the rows that were included in the pre-launch charts.

(b) Plot all rows.

(c) Plot all rows, using different symbols or colors to indicate whether or not points were included in the pre-launch charts.

Comment, for each of the first two graphs, whether and open or closed symbol is preferable. For the third graph, comment on the your reasons for choice of symbols.

Use the following to identify rows that hold the data that were presented in the pre-launch charts:

```
> orings$Included <- logical(23)  # orings has 23 rows
> orings$Included[c(1,2,4,11,13,18)] <- TRUE
```

The construct `logical(23)` creates a vector of length 23 in which all values are `FALSE`. The following are two possibilities for the third plot; can you improve on these choices of symbols and/or colors?

```
> plot(Total ~ Temperature, data=orings, pch=orings$included+1)
> plot(Total ~ Temperature, data=orings, col=orings$Included+1)
```

Exercise 10
Using the data frame `oddbooks`, use graphs to investigate the relationships between:

(a) weight and volume;

(b) density and volume;

(c) density and page area.

7 Factors

Exercise 11
Investigate the use of the functions `as.character()` and `unclass()` with a factor argument. Comment on their use in the following code:

```
> par(mfrow=c(1,2), pty="s")
> plot(weight ~ volume, pch=unclass(cover), data=allbacks)
> plot(weight ~ volume, data=allbacks, type="n")
> with(allbacks, text(weight ~ volume, labels=as.character(cover)))
> par(mfrow=c(1,1))
```

[The setting `mfrow=c(1,2)` gives side by side plots. The setting `pty="s"` gives a square plotting region.]

Exercise 12
Run the following code:

```
> gender <- factor(c(rep("female", 91), rep("male", 92)))
> table(gender)
> gender <- factor(gender, levels=c("male", "female"))
> table(gender)
> gender <- factor(gender, levels=c("Male", "female")) # Note the mistake
>                               # The level was "male", not "Male"
> table(gender)
> rm(gender)                    # Remove gender
```

Explain the output from the final `table(gender)`.
The output is

```
gender
female   male
    91     92

> table(gender)
> gender <- factor(gender, levels=c("Male", "female")) # Note the mistake
>                               # The level was "male", not "Male"
> table(gender)
> rm(gender)                    # Remove gender
```

8 Stripcharts (base graphics) and Stripplots (*lattice*)

Exercise 13
Look up the help for the lattice function `dotplot()`.
Compare the following:

```
> ## First, use the regular graphics function stripchart()
> with(ant111b, stripchart(harvwt ~ site))
> ## Next, use lattice graphics
> library(lattice)
> stripplot(site ~ harvwt, data=ant111b)
> ## Next, use lattice graphics, but switch the x and y axes
> library(lattice)
> stripplot(site ~ harvwt, data=ant111b)
```

Note the differences in syntax between the two graphics systems.

Exercise 14
Check the class of each of the columns of the data frame `cabbages` (*MASS*). Do side by side plots of `HeadWt` against `Date`, for each of the levels of `Cult`.

```
> stripplot(Date ~ HeadWt | Cult, data=cabbages)
```

As the lattice graphics `stripplot()` function allows you to do a lot more than `stripchart()`, and as the lattice syntax is highly consistent across the different *lattice* functions, it seems best to use `stripplot()`.

Exercise 15
In the data frame `nsw74psid3`, use `stripplot()` to compare, between levels of `trt`, the continuous variables `age`, `educ`, `re74` and `re75`
It is possible to generate all the plots at once, side by side. A simplified version of the plot is:

```
> stripplot(trt ~ age + educ, data=nsw74psid1, outer=T, scale="free")
```

What are the respective effects of `scale = "free"`, and `outer = TRUE`? (Try leaving these at their defaults.)

9 Tabulation

Exercise 16
In the data set `nsw74psid3`, compare for each of the two levels of `trt`:

(a) the relative numbers of `black`;

(b) the relative numbers of hispanics (`hisp`);

(c) the relative numbers of married (`marr`).

10 Sorting

Exercise 17
Sort the rows in the data frame `Acmena` in order of increasing values of `dbh`.
[Hint: Use the function `order()`, applied to `age` to determine the order of row numbers required to sort rows in increasing order of age. Reorder rows of `Acmena` to appear in this order.]

```
> Acmena <- subset(rainforest, species=="Acmena smithii")
> ord <- order(Acmena$dbh)
> acm <- Acmena[ord, ]
```

Sort the row names of `possumsites` (*DAAG*) into alphanumeric order. Reorder the rows of `possumsites` in order of the row names.

11 Use of a For Loop

Exercise 18

(a) Create a `for` loop that, given a numeric vector, prints out one number per line, with its square and cube alongside.

(b) Look up `help(while)`. Show how to use a `while` loop to achieve the same result.

(c) Show how to achieve the same result without the use of an explicit loop.

12 A Function

Exercise 19
The following function calculates the mean and standard deviation of a numeric vector.

```
> meanANDsd <- function(x){
+     av <- mean(x)
+     sdev <- sd(x)
+     c(mean=av, sd = sdev) # The function returns this vector
+ }
```

Modify the function so that: (a) the default is to use `rnorm()` to generate 20 random normal numbers, and return the standard deviation; (b) if there are missing values, the mean and standard deviation are calculated for the remaining values.

Part II
Practice with R

1 Information about the Columns of Data Frames

Exercise 1
Functions that may be used to get information about data frames include `str()`, `dim()`, `row.names()` and `names()`. Try each of these functions with the data frames `allbacks`, `ant111b` and `tinting` (all in *DAAG*).
For getting information about the class of each column use e.g.

```
> library(DAAG)
> sapply(ant111b, class)
```

or

```
> unlist(sapply(ant111b, class))
```

This applies the function `class()` to each column of the data frame.
For each of these data frames, use `table()` to tabulate the number of values for each level.

2 Data Input

Exercise 2
The function `read.csv()` is a variant of `read.table()` that is designed to read in comma delimited files such as may be obtained from Excel. Use this function to read in the file **crx.data** that is available from the web page `http://mlearn.ics.uci.edu/databases/credit-screening/`.
Check the file **crx.names** to see which columns should be numeric, which categorical and which logical. Make sure that the numbers of missing values in each column are the number given in the file **crx.names**

For a first pass at reading in the data, try:

```
> crxpage <- "http://mlearn.ics.uci.edu/databases/credit-screening/crx.data"
> crx <- read.csv(url(crxpage), header=TRUE)
```

Exercise 3
For a challenging data input task, input the data from the file **bostonc.txt**. You can create this by attaching the *DAAG* package and entering `datafile("bostonc")` thus:

```
> datafile("bostonc")
```

Examine the contents of the initial lines of the file carefully before trying to read it in. You will need to change the parameters `comment.char` and `skip` from their defaults.

3 A Tabulation Exercise

Exercise 4
Tabulate the number of observations in each of the different districts in the data frame `rockArt` (*DAAGxtras*). Create a factor `groupDis` in which all `Districts` with less than 5 observations are grouped together into the category `other`.

```
> library(DAAGxtras)
> groupDis <- as.character(rockArt$District)
> tab <- table(rockArt$District)
> le4 <- rockArt$District %in% names(tab)[tab <= 4]
> groupDis[le4] <- "other"
> groupDis <- factor(groupDis)
```

4 A For Loop

Exercise 5
The following code uses a `for` loop to plot graphs that compare the relative population growth (here, by the use of a logarithmic scale) for the Australian states and territories.

```
> library(DAAG)
> oldpar <- par(mfrow=c(2,4))
> for (i in 2:9){
+ plot(austpop[,1], log(austpop[, i]), xlab="Year", ylab=names(austpop)[i],
+      pch=16, ylim=c(0,10))}
> par(oldpar)
```

Which Australian adminstration(s) showed the most rapid increase in the early years? Which showed the most rapid increase in later years?

5 Data Exploration – Distributions of Data Values

Exercise 6
The data frame `rainforest` (*DAAG* package) has data on four different rainforest species. Use `table(rainforest$species)` to check the names and numbers of the species present. In the following, attention will be limited to the species Acmena *smithii*.
The following plots a histogram showing the distribution of the diameter at base height:

```
> library(DAAG)      # The data frame rainforest is from DAAG
> Acmena <- subset(rainforest, species=="Acmena smithii")
> hist(Acmena$dbh)
```

By default, frequencies are used to label the the vertical axis.

An alternative is to use a density scale (`prob=TRUE`). The histogram is interpreted as a crude density plot. The density, which estimates the number of values per unit interval, changes in discrete jumps at the breakpoints (= class boundaries). The histogram can then be directly overlaid with a density plot, thus:

```
> hist(Acmena$dbh, prob=TRUE, xlim=c(0,50))   # Use a density scale
> lines(density(Acmena$dbh, from=0))
```

Why use the argument `from=0`? What is the effect of omitting it?

[Density estimates, as given by R's function `density()`, change smoothly and do not depend on an arbitrary choice of breakpoints, making them generally preferable to histograms. They do sometimes require tuning to give a sensible result. Note especially the parameter `bw`, which determines how the bandwidth is chosen, and hence affects the smoothness of the density estimate.]

6 Random Samples

Exercise 7
Histograms and density plots are, for “small” samples, notoriously variable under repeated sampling. This is true even for sample sizes as large as 50 or 100.

By taking repeated random samples from the normal distribution, and plotting the distribution for each such sample, one can get an idea of the effect of sampling variation on the sample distribution.

A random sample of 100 values from a normal distribution (with mean 0 and standard deviation 1) can be obtained, and a histogram and overlaid density plot shown, thus:

```
> y <- rnorm(100)
> hist(y, probability=TRUE)  # probability=TRUE gives a y density scale
> lines(density(y))
```

(a) Take 5 samples of size 25, then showing the plots.

(b) Take 5 samples of size 100, then showing the plots.

(c) Take 5 samples of size 500, then showing the plots.

(d) Take 5 samples of size 2000, then showing the plots.

Note: By preceding the plots with `par(mfrow=c(4,5))`, all 20 plots can be displayed on the one graphics page. (To bunch the graphs up more closely, make the further settings `par(mar=c(3.1,3.1,0.6,0.6), mgp=c(2.25,0.5,0))`)

Comment on the usefulness of a sample histogram and/or density plot for judging whether the population distribution is likely to be close to normal.

Exercise 8
This explores the function `sample()`, used to take a sample of values that are stored or enumerated in a vector. Samples may be with or without replacement; specify `replace = FALSE` (the default) or `replace = TRUE`. The parameter `size` determines the size of the sample. By default the sample has the same size (length) as the vector from which samples are taken. Take several samples of size 5 from the vector `1:5`, with `replace=FALSE`. Then repeat the exercise, this time with `replace=TRUE`. Note how the two sets of samples differ.

Exercise 9
If in Exercise 6 above a new random sample of trees could be taken, the histogram and density plot would change. How much might we expect them to change?

The boostrap approach treats the one available sample as a microcosm of the population. Repeated with replacement samples are taken from the one available sample. This is equivalent to repeating each sample value and infinite number of times, then taking random samples from the population that is thus created. The expectation is that variation between those samples will be comparable to variation between samples from the original population.

(a) Take repeated (5 or more) bootstrap samples from Acmena dataset of Exercise 6, and show the density plots. [Use `sample(Acmena$dbh, replace=TRUE)`].

(b) Repeat, now with the `cerealsugar` data from *DAAG*.

7 Smooth Curves

Exercise 10
The following compares three different smoothing functions. Comment on the different syntax and, in the case of `lowess()`, the different default output that is returned. Why, for the smooth obtained using `lowess()`, is it necessary to sort data in order of values of `dbh`? (Try omitting the ordering, and observe the result.)

```
> Acmena <- subset(rainforest, species=="Acmena smithii")
> ## Use lowess()
> plot(wood ~ dbh, data=Acmena)
> ord <- order(Acmena$dbh)
> with(Acmena[ord, ], lines(predict(loess(wood ~ dbh)) ~ dbh))
> ## Now use panel.smooth()
> plot(wood ~ dbh, data=Acmena)
> with(Acmena, panel.smooth(dbh, wood))
```

For each of the functions just noted, what are the parameters that control the smoothness of the curve? What, in each case, is the default?

8 Information on Workspace Objects

Exercise 11
An R workspace includes objects `possum1`, `possum2`, ...`possum5`. The folowing shows how to get the size of one of these objects one at a time.

```
> possum1 <- rnorm(10)
> object.size(possum1)
```

The names of the objects can be obtained with

```
> nam <- ls(pattern="^possum")
```

To get the sizes from the names that are held in `nam`, do

```
> sapply(nam, function(x)object.size(get(x)))
```

Create objects `possum2`, ...`possum5`, and enter this command. Explain the successive steps in the computation.
[Hint: Compare `class(possum1)` with `class("possum1")`, and `object.size(possum1)` with `object.size("possum1")`]

Exercise 12*
The function `ls()` lists, by default, the names of objects in the current environment. If used from the command line, it lists the objects in the workspace. If used in a function, it lists the names of the function's local variables. To get a listing of the contents of the workspace, do the following

```
> workls <- function()ls(name=".GlobalEnv")
> workls()
```

Exercise 12, continued*

(a) If `ls(name=".GlobalEnv")` is replaced by `ls()`, the function lists the names of its local variables. Modify `workls()` so that you can use it to demonstrate this.
[Hint: Consider adapting `if(is.null(name))ls()` for the purpose.]

(b) Write a function that calculates the sizes of all objects in the workspace, then listing the names and sizes of the largest ten objects.

9 Different Ways to Do a Calculation – Timings

Exercise 13
This exercise will investigate the relative times for alternative ways to do a calculation. First, we will create both matrix and data frame versions of a largish data set.

```
> xxMAT <- matrix(runif(480000), ncol=50)
> xxDF <- as.data.frame(xxMAT)
```

The function `system.time()` will provide timings. The first three numbers that are returned will be of interest; these are the user cpu time, the system cpu time, and the elapsed time.

Repeat each calculation several times, and note whether there is variation between repeats. If there is, make the setting `options(gcFirst=TRUE)`, and see whether this leads to more consistent timings. NB: If your computer chokes on these calculations, reduce the dimensions of `xxMAT` and `xxDF`

(a) The following compares the times taken to increase each element by 1:

```
> system.time(invisible(xxMAT+1))[1:3]
> system.time(invisible(xxDF+1))[1:3]
```

(b) Now compare the following alternative ways to calculate the means of the 50 columns:

```
> ## Use apply() [matrix argument], or sapply() [data frame argument]
> system.time(av1 <- apply(xxMAT, 2, mean))[1:3]
> system.time(av1 <- sapply(xxDF, mean))[1:3]
> ## Use a loop that does the calculation for each column separately
> system.time({av2 <- numeric(50);
+             for(i in 1:50)av[i] <- mean(xxMAT[,i])
+             })[1:3]
> system.time({av2 <- numeric(50);
+             for(i in 1:50)av[i] <- mean(xxDF[,i])
+             })[1:3]
> ## Matrix multiplication
> system.time({colOFones <- rep(1, dim(xxMAT)[2])
+              av3 <- xxMAT %*% colOFones / dim(xxMAT)[2]
+             })[1:3]
```

(c) Pick one of the above calculations. Vary the number of rows in the matrix, keeping the number of columns constant, and plot each of user CPU time and system CPU time against number of rows of data.

Suggest why the calculation that uses matrix multiplication is so efficient, relative to the other options.

10 Functions – Making Sense of the Code

*Exercise 14**

Data in the data frame `fumig` (*DAAGxtras*) are from a series of fumigation trials, in which produce was exposed to the fumigant over a 2-hour time period. Concentrations in the chamber were measured at times 5, 10, 30, 60, 90 and 120 minutes. Code given following this exercise calculates a concentration-time (c-t) product that measures exposure to the fumigant, leading to the measure `ctsum`.

Examine the code in the three alternative functions given below, and the data frame `fumig` (in the `DAAGxtras` package) that is given as the default argument for the parameter `df`. Do the following:

(a) Run all three functions, and check that they give the same result.

(b) Annotate the code for `calcCT1()` to explain what each line does.

(c) Are fumigant concentration measurements noticeably more variable at some times than at others?

(d) Which function is fastest? [In order to see much difference, it will be necessary to put the functions in loops that run perhaps 1000 or more times.]

Code for 3 functions that do equivalent calculations

```
> ## Function "calcCT1"
> "calcCT1" <-
+   function(df=fumig, times=c(5,10,30,60,90,120), ctcols=3:8){
+     multiplier <- c(7.5,12.5,25,30,30,15)
+     m <- dim(df)[1]
+     ctsum <- numeric(m)
+     for(i in 1:m){
+       y <- unlist(df[i, ctcols])
+       ctsum[i] <- sum(multiplier*y)/60
+     }
+     df <- cbind(ctsum=ctsum, df[,-ctcols])
+     df
+   }
> ##
> ## Function "calcCT2"
> "calcCT2" <-
+   function(df=fumig, times=c(5,10,30,60,90,120), ctcols=3:8){
+     multiplier <- c(7.5,12.5,25,30,30,15)
+     mat <- as.matrix(df[, ctcols])
+     ctsum <- mat%*%multiplier/60
+     cbind(ctsum=ctsum, df[,-ctcols])
+   }
> ##
> ## Function "calcCT3"
> "calcCT3" <-
+   function(df=fumig, times=c(5,10,30,60,90,120), ctcols=3:8){
+     multiplier <- c(7.5,12.5,25,30,30,15)
+     mat <- as.matrix(df[, ctcols])
+     ctsum <- apply(mat, 1, function(x)sum(x*multiplier))/60
+     cbind(ctsum=ctsum, df[,-ctcols])
+   }
```

11 *Use of sapply() to Give Multiple Graphs

Exercise 15 (Optional extra I)*
Here is code for the calculations that compare the relative population growth rates for the Australian states and territories, but avoiding the use of a loop:

```
> oldpar <- par(mfrow=c(2,4))
> invisible(
+ sapply(2:9, function(i, df)
+        plot(df[,1], log(df[, i]),
+             xlab="Year", ylab=names(df)[i], pch=16, ylim=c(0,10)),
+             df=austpop)
+ )
> par(oldpar)
```

Run the code, and check that it does indeed give the same result as the use of an explicit loop.
[By wrapping the code in the function `invisible()`, printed output that gives no useful information can be suppressed.]
Note that `lapply()` could be used in place of `sapply()`.

There are several subtleties here:

(i) The first argument to `sapply()` can be either a list (which is, technically, a non-atomic vector) or a vector.[1] Here, we have supplied the vector 2:9

(ii) The second argument is a function. Here we have supplied an anonymous function that has two arguments. The argument `i` takes as its values, in turn, the sucessive elements in the first argument to `sapply`

(iii) Where as here the anonymous function has further arguments, they are supplied as additional arguments to `sapply()`. Hence the parameter `df=austpop`.

12 *The Internals of R – Functions are Pervasive

Exercise 16 (Optional extra II)*
This exercise peeks into the internals of the way in which R structures arithmetic and related computations. Those internals are close enough to the surface that users can experiment with their use.

The binary arithmetic operators `+`, `-`, `*`, `/` and `^` are implemented as functions. (R is a functional language; albeit with features that compromise its purity as a member of this genre!) Try the following:

```
> "+"(2,5)
> "-"(10,3)
> "/"(2,5)
> "*"("+"(5,2), "-"(3,7))
```

[1] By "vector" we usually mean an atomic vector, with "atoms" that are of one of the modes ”logical”, ”integer”, ”numeric”, ”complex”, ”character”' or ”raw”. (Vectors of mode ”raw” can for our purposes be ignored.)

Exercise 16, continued*

There are two other binary arithmetic operators – `%%` and `%/%`. Look up the relevant help page, and explain, with examples, what they do. Try

```
> (0:25) %/% 5
> (0:25) %% 5
```

Of course, these are also implemented as functions. Write code that demonstrates this.

Note also that `[` is implemented as a function. Try

```
> z <- c(2, 6, -3, NA, 14, 19)
> "["(z, 5)
> heights <- c(Andreas=178, John=185, Jeff=183)
> "["(heights, c("Jeff", "John"))
```

Rewrite these using the usual syntax.

Use this syntax to extract, from the data frame `possumsites` (*DAAG*), the altitudes for Byrangery and Conondale.

Note: Expressions in which arithmetic operators appear as explicit functions with binary arguments translate directly into postfix reverse Polish notation, introduced in 1920 by the Polish logician and mathematician Jan ÅĄukasiewicz. Postfix notation is widely used in the interpreters and compilers that translate computer language code into machine or assembly language instructions. See the Wikipedia article “Reverse Polish Notation”.

Part III
Informal and Formal Data Exploration

1 Informal Data Exploration

These exercises explore data that will later be used for exercises in error rate estimation.

Exercise 1
Look up the help page for the data frame `Pima.tr2` (*MASS* package), and note the columns in the data frame. The eventual interest is in using use variables in the first seven column to classify diabetes according to `type`. Here, we explore the individual columns of the data frame.

(a) Several columns have missing values. Analysis methods inevitably ignore or handle in some special way rows that have one or moe missing values. It is therefore desirable to check whether rows with missing values seem to differ systematically from other rows.

Determine the number of missing values in each column, broken down by `type`, thus:

```
> library(MASS)
> ## Create a function that counts NAs
> count.na <- function(x)sum(is.na(x))
> ## Check function
> count.na(c(1, NA, 5, 4, NA))
> ## For each level of type, count the number of NAs in each column
> for(lev in levels(Pima.tr2$type))
+   print(sapply(subset(Pima.tr2, type==lev), count.na))
```

The function `by()` can be used to break the calculation down by levels of a factor, avoiding the use of the `for` loop, thus:

```
> by(Pima.tr2, Pima.tr2$type, function(x)sapply(x, count.na))
```

(b) Create a version of the data frame `Pima.tr2` that has `anymiss` as an additional column:

```
> missIND <- complete.cases(Pima.tr2)
> Pima.tr2$anymiss <- c("miss","nomiss")[missIND+1]
```

For remaining columns, compare the means for the two levels of `anymiss`, separately for each level of `type`. Compare also, for each level of `type`, the number of missing values.

Exercise 2

(a) Use strip plots to compare values of the various measures for the levels of `anymiss`, for each of the levels of `type`. Are there any columns where the distribution of differences seems shifted for the rows that have one or more missing values, relative to rows where there are no missing values?
Hint: The following indicates how this might be done efficiently:

```
> library(lattice)
> stripplot(anymiss ~ npreg + glu | type, data=Pima.tr2, outer=TRUE,
+           scales=list(relation="free"), xlab="Measure")
```

Exercise 2, continued

(b) Density plots are in general better than strip plots for comparing the distributions. Try the following, first with the variable `npreg` as shown, and then with each of the other columns except `type`. Note that for `skin`, the comparison makes sense only for `type=="No"`. Why?

```
> library(lattice)
> ## npreg & glu side by side (add other variables, as convenient)
> densityplot( ~ npreg + glu | type, groups=anymiss, data=Pima.tr2,
+              auto.key=list(columns=2), scales=list(relation="free"))
```

Exercise 3

Better than either strip plots or density plots may be Q-Q plots. Using `qq()` from *lattice*, investigate their use. In this exercise, we use random samples from normal distributions to help develop an intuitive understanding of Q-Q plots, as they compare with density plots.

(a) First consider comparison using (i) a density plot and (ii) a Q-Q plot when samples are from populations in which one of the means is shifted relative to the other. Repeat the following several times,

```
> y1 <- rnorm(100, mean=0)
> y2 <- rnorm(150, mean=0.5)  # NB, the samples can be of different sizes
> df <- data.frame(gp=rep(c("first","second"), c(100,150)), y=c(y1, y2))
> densityplot(~y, groups=gp, data=df)
> qq(gp ~ y, data=df)
```

(b) Now make the comparison, from populations that have different standard deviations. For this, try, e.g.

```
> y1 <- rnorm(100, sd=1)
> y2 <- rnorm(150, sd=1.5)
```

Again, make the comparisons using both density plots and Q-Q plots.

Exercise 4

Now consider the data set `Pima.tr2`, with the column `anymiss` added as above.

(a) First make the comparison for `type="No"`.

```
> qq(anymiss ~ npreg, data=Pima.tr2, subset=type=="No")
```

Compare this with the equivalent density plot, and explain how one translates into the other. Comment on what these graphs seem to say.

(b) The following places the comparisons for the two levels of `type` side by side:

```
> qq(anymiss ~ npreg | type, data=Pima.tr2)
```

Comment on what this graph seems to say.

NB: With `qq()`, use of "+" to get plots for the different columns all at once will not, in the current version of *lattice*, work.

2 Bootstrap sampling

Exercise 5
The following takes a with replacement sample of the rows of `Pima.tr2`.

```
> rows <- sample(1:dim(Pima.tr2)[1], replace=TRUE)
> densityplot(~ bmi, groups=type, data=Pima.tr2[rows, ],
+           scales=list(relation="free"), xlab="Measure")
```

Repeat, but using `anymiss` as the grouping factor, and with different panels for the two levels of `type`. Repeat for several different bootstrap samples. Are there differences between levels of `anymiss` that seem consistent over repeated bootstrap samples?

Exercise 6
Exercise 2 compared density plots, for several of the variables, between rows that had one or more missing values and those that had no missing values. We can use the bootstrapping idea to check the copnsistency of apparent differences across repeated bootstrap samples.

The distribution for `bmi` gives the impression that it has a different shape, between rows where one or more values was missing and rows where no values were missing, at least for `type=="Yes"`. One way to check whether this difference is consistent under repeated sampling is to treat the sample as representative of the population, and take repeated with replacement ("bootstrap") samples from it.

The following takes a bootstrap sample, then showing the Q-Q plot

```
> rownum <- 1:dim(Pima.tr2)[1]  # generate row numbers
> chooserows <- sample(rownum, replace=TRUE)
> qq(anymiss ~ bmi | type, data=Pima.tr2[chooserows, ],
+    auto.key=list(columns=2))
```

Wrap these lines of code in a function. Repeat the formation of the bootstrap samples and the plots several times. Does the shift in the distribution seem consistent under repeating sampling?

Judgements based on examination of graphs are inevitably subjective. They do however make it possible to compare differences in the shapes of distributions. Here, the shape difference is of more note than any difference in mean or median.

Exercise 7
In the data frame `nswdemo` (*DAAGxtras* package), compare the distribution of `re78` for those who received work training (`trt==1`) with controls (`trt==0`) who did not.

```
> library(DAAGxtras)
> densityplot(~ re78, groups=trt, data=nswdemo, from=0,
+             auto.key=list(columns=2))
```

The distributions are highly skew. A few very large values may unduly affect the comparison.

A reasonable alternative is to compare values of `log(re78+23)`. The value 23 is chosen between it is half the minimum non-zero value of `re78`. Here is the density plot.

```
> unique(sort(nswdemo$re78))[1:3]  # Examine the 3 smallest values
> densityplot(~ log(re78+23), groups=trt, data=nswdemo,
+             auto.key=list(columns=2))
```

Do the distribution for control and treated have similar shapes?

Exercise 8
Now examine the displacement, under repeated bootstrap sampling, of one mean relative to the other. Here is code for the calculation:

```
> twoBoot <- function(n=999, df=nswdemo, ynam="re78", gp="trt"){
+   d2 <- numeric(n+1)
+   fac <- df[, gp]
+   if(!is.factor(fac))fac <- factor(fac)
+   if(length(levels(fac)) != 2) stop(paste(gp, "must have 2 levels"))
+   y <- df[, ynam]
+   d2[1] <- diff(tapply(y, fac, mean))
+   for(i in 1:n){
+     chooserows <- sample(1:length(y), replace=TRUE)
+     faci <- fac[chooserows]
+     yi <- y[chooserows]
+     d2[i+1] <- diff(tapply(yi, faci, mean))
+ }
+   d2
+ }
> ##
> d2 <- twoBoot()
> quantile(d2, c(.025,.975))   # 95% confidence interval
```

Note that a confidence interval should not be interpreted as a probability statement. It takes no account of prior probability. Rather, 95% of intervals that are calculated in this way can be expected to contain the true probability.

Exercise 9
Another possibility is to work with a permutation distribution. If the difference between treated and controls is entirely due to sampling variation, then permuting the treatment labels will give another sample from this same distribution. Does the observed difference between treated and controls seem "extreme", relative to this permutation distribution? Here is code that may be used.

```
> dnsw <- numeric(1000)
> y <- nswdemo$re78
> treat <- nswdemo$trt
> dnsw[1] <- mean(y[treat==1]) - mean(y[treat==0])
> for(i in 2:1000){
+   trti <- sample(treat)
+   dnsw[i] <- mean(y[trti==1]) - mean(y[trti==0])
+ }
> sum(dnsw > dnsw[1])/length(dnsw)

[1] 0.029

> 2*min(sum(d2<0)/length(d2), sum(d2>0)/length(d2))  # 2-sided comparison

[1] 0.062
```

Compare the result with that from the bootstrap approach.

Replace `re78` with `log(re78+23)` and repeat the calculations.

Note: In formalizing the result for a test of hypothesis, note that the difference between `treat==1` and `treat==1` might go in either direction.

Part IV
Models and Model Accuracy

1 Straight Line Regression

Exercise 1
A plot of heart weight (`heart`) versus body weight (`weight`), for Cape Fur Seal data in the data set `cfseal` (*DAAG*) shows a relationship that is approximately linear. Check this. However variability about the line increases with increasaing weight. It is better to work with `log(heart)` and `log(weight)`, where the relationship is again close to linear, but variability about the line is more homogeneous. Such a linear relationship is consistent with biological allometry, here across different individuals. Allometric relationships are pairwise linear on a logarithmic scale.

Plot `log(heart)` against `log(weight)`, and fit the least squares regression line for `log(heart)` on `log(weight)`.

```
> cflog <- log(cfseal[, c("heart", "weight")])
> names(cflog) <- c("logheart", "logweight")
> plot(logheart ~ logweight, data=cflog)
> cfseal.lm <- lm(logheart ~ logweight, data=cflog)
> abline(cfseal.lm)
```

Exercise 2
In the training/test sample appraoch, the data are split into two parts. The model is trained on one part (the training set), and tested on the other part. Run the following code:

```
> fold <- sample(1:2, dim(cflog)[1], replace=TRUE)
> train <- subset(cflog, fold==1)
> test <- subset(cflog, fold==2)
> cfseal.lm <- lm(logheart ~ logweight, data=train)
> hat <- predict(cfseal.lm, newdata=test)
> ## Compare predictions for test data with obsewrvations
> plot(hat, test$logheart)
> ## Mean square difference between test predictions  & test observations
> meansquare <- sum((hat - test$logheart)^2)/dim(test)[1]
> sd <- sqrt(meansquare)
> sd
```

Values for `sd` should be of the same order of magnitude as `summary(cflog.lm)$sigma)` Repeat the calculations 100 times, and store the 100 values of `sd`

Exercise 3
We could interchange the training and test set, and repeat the calculations, averaging the two mean squares. But why limit ourselves to a split into two parts. Why not split the data into an arbitrary number of parts?

(a) In cross-validation, the data are split into perhaps 10 folds. Each of the ten parts is used in turn as the test data, with the remaining nine parts used as training data. Here is skeleton code for the calculations.

Exercise 3, continued

```
> `cvskeleton` <-
+   function(data, nfolds=3){
+     n <- dim(data)[1]
+     folds <- sample(rep(1:nfolds, length.out=n))  # Balanced sample
+     ufold <- unique(folds)
+     for(i in ufold){
+       trainrows <- (1:n)[i!=folds]
+       ## Fit model to training rows
+       testrows <- (1:n)[i==folds]
+       ## Make predictions for test rows
+       print(testrows)
+     }
+     ## Return a list that holds the nfolds sets of row numbers
+     invisible(split(1:n, folds))
+   }
```

Run the function and check that each row does indeed appear in exactly one of the test sets.

(b) Here is a function that implements cross-validation calculations. For the regression calculation, it will be necessary to supply it with functions that do the training and test calculations, thus:

```
> ## Note that cftrainfun must have arguments formula & traindata
> cftrainfun <- function(formula=logheart ~ logweight, traindata)
+                         lm(formula, traindata)
> ## Note that cftestfun must have arguments obj and newdata
> cftestfun <- function(obj, newdata)predict(obj, newdata)
```

Here then is the function:

```
> `cvpred` <-
+   function(formula = type ~ ., data, trainfun, testfun, nfolds=2,
+            balanced=TRUE){
+     n <- dim(data)[1]
+     pred <- numeric(n)
+     if (balanced) folds <- sample(rep(1:nfolds, length.out=n)) else
+     folds <- sample(1:nfolds, n, replace=T)
+     ufold <- unique(folds)
+     for(i in ufold){
+       testrows <- i==folds
+       traindata <- subset(data, !testrows)
+       trained.obj <- trainfun(formula, traindata=traindata)
+       testdata <- data[testrows, ]
+       pred[testrows] <- testfun(obj=trained.obj, newdata=testdata)
+     }
+     invisible(pred)
+   }
```

Now run the function, and compare predictions with observed values

```
> predobs <- cvpred(logheart ~ logweight, data=cflog, trainfun=cftrainfun,
+                   testfun=cftestfun)
> sd <- sqrt(sum((predobs-cflog$logheart)^2)/dim(cflog)[1])
```

Run the calculation several times.

Exercise 4
For each of (1) `nfolds=2` and (2) `nfolds=10`, run the calculations of Exercise 3 100 times, calculating a standard deviation at each run. Compare the two sets of standard deviation estimates, both using a density plot and using a Q-Q plot. Which calculation would you expect to give the more consistent (less variable) results? Why? Do the graphs bear this out?

Exercise 5
The following fits a simple linear discriminant model to the data frame `Pima.tr`, then using `Pima.te` to test the predictions. We are using the training/test set approach.

```
> Pima.lda <- lda(type ~ ., data=Pima.tr)
> predclass <- predict(Pima.lda, newdata=Pima.te)$class
> tab12 <- table(Pima.te$type, predclass)        # Confusion matrix
> tab12
> sum(tab12[row(tab12)==col(tab12)])/sum(tab12)  # Overall accuracy estimate
```

Note that we are regarding `Pima.tr` as data set 1, and texttttPima.te as data set 2. The confusion matrix when we train on data set 1 and test on data set 2 is therefore called `tab12`.
Now use `Pima.te` as the training data, and `Pima.tr` as the test data, and repeat the exercise, leading to the confusion matrix `tab21`.

You should find that the confusion matrices differ substantially, and that the overall accuracies differ. What might explain this?

Exercise 6
Now work with `Pima.tr` alone, and `Pima.te` alone, calculating cross-validated predictions and confusion matrices. As before functions will be be defined to handle the training and testing. Use is then made of the function `cvpred()` that was defined above.

(a) Calculate cross-validated estimates, with `Pima.tr`.

```
> Pimatrainfun <- function(formula, traindata)lda(formula, traindata)
> Pimatestfun <- function(obj, newdata)predict(obj, newdata)$class
> predclass <- cvpred(type~., data=Pima.tr, nfolds=10,
+                     trainfun=Pimatrainfun, testfun=Pimatestfun)
> tab11 <- table(Pima.tr$type, predclass)        # Confusion matrix
> sum(tab11[row(tab11)==col(tab11)])/sum(tab11)  # Overall accuracy estimate
```

(b) Now calculate cross-validated estimates for `Pima.te`, leading to the confusion matrix `tab22`.

Repeat each of these calculations several times, to get an idea of the statistical variation in the confusion matrices.
Compare `tab11`, `tab21`, `tab22` and `tab12`, and comment.

Exercise 7
The resubstitution measure of accuracy compares predictions for the training data with observations from the same training data. For example:

```
> Pima.lda <- lda(type ~ ., data=Pima.tr)
> predclass <- predict(Pima.lda)$class
> tab1resub <- table(Pima.tr$type, predclass)
> tab1resub
> sum(tab1resub[row(tab1resub)==col(tab1resub)])/sum(tab1resub)
```

Such resubstitution measures of accuracy can be grossly optimistic. Why? Do they seem to be optimistic for the models that have been fitted to these data.

Exercise 8
A feature of the resubstitution measure is that it cannot decrease as more terms are added to the model. For example, try adding in all second order interaction terms, i.e., we have all factors and interactions involving three or fewer columns.

```
> Pima.lda <- lda(type ~ .^3, data=Pima.tr)
> predclass <- predict(Pima.lda)$class
> tab1resub3 <- table(Pima.tr$type, predclass)
> tab1resub3
> sum(tab1resub3[row(tab1resub3)==col(tab1resub3)])/sum(tab1resub3)
```

There is a large increase in the resubstitution measure of predictive accuracy, but does it mean anything. In order to check, we do a cross-validation calculation.

```
> predclass <- cvpred(type~.^3, data=Pima.tr, nfolds=10,
+                     trainfun=Pimatrainfun, testfun=Pimatestfun)
> tab11cv3 <- table(Pima.tr$type, predclass)        # Confusion matrix
> sum(tab11cv3[row(tab11cv3)==col(tab11cv3)])/sum(tab11cv3)
```

Compare the resubstitution measure with the cross-validation measure, and comment.

Exercise 9

Most of the rows with missing values for Pima.tr2 arise from missing values for the column skin. Compare

```
> sum(!complete.cases(Pima.tr2[,-4]))
> sum(!complete.cases(Pima.tr2))
```

The following omits entirely rows where columns other than skin are missing. It then creates an outcome variable with four categories: type=="No" & NA or not for skin, type=="Yes" & NA or not for skin.

```
> mostlyOK <- complete.cases(Pima.tr2[, -4])
> Pima.tr2A <- Pima.tr2[mostlyOK, ]
> type4 <- with(Pima.tr2A, paste(type, complete.cases(skin), sep=""))
> Pima.tr2A <- cbind(Pima.tr2A[, c(1:3,5:7)], type4=factor(type4))
```

Now try

```
> PimaA.lda <- lda(type4 ~ ., data=Pima.tr2A)
> PimaA.lda
```

We can get a plot of the results:

```
> plot(PimaA.lda, dimen=2)
> ## More helpful plot
> scores <- predict(PimaA.lda)$x
> xyplot(scores[, 2] ~ scores[, 1], groups=type4, data=Pima.tr2A,
+        auto.key=list(columns=2),
+        par.settings=list(superpose.symbol=list(pch=16)))
```

The function cvpred() can be used as before to estimate the confusion matrix. Or use the setting CV=TRUE when the function lda is called, which uses the simple leave-one-out form of cross-validation.

```
> table(Pima.tr2A$type4, lda(type4 ~ ., data=Pima.tr2A, CV=TRUE)$class)
```

2 Analysis Using *randomForest*

Exercise 10
Repeat exercise 9, but now using `randomForest()`. For now, we use this as a black box.

```
> library(randomForest)
> Pima.rf <- randomForest(type4 ~ ., data=Pima.tr2A)
> Pima.rf

Call:
 randomForest(formula = type4 ~ ., data = Pima.tr2A)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 53.17%
Confusion matrix:
         NoFALSE NoTRUE YesFALSE YesTRUE class.error
NoFALSE        3     45        2       5   0.9454545
NoTRUE        19     92        2      19   0.3030303
YesFALSE       5     10        0      14   1.0000000
YesTRUE        3     25        2      38   0.4411765
```

Note that OOB = Out of Bag Error rate, calculated using an approach that has much the same effect as cross-validation, applied to the data specified by the `data` parameter. Notice that `randomForest()` will optionally give, following the one function call, an assessment of predictive error for a completely separate set of test data that has had no role in training the model.

```
> type4te <- with(Pima.te, paste(type, TRUE, sep=""))
> type4te <- factor(type4te, levels=levels(Pima.tr2A$type4))
> randomForest(type4 ~ ., data=Pima.tr2A, xtest=Pima.te[c(1:3, 5:7)],
+              ytest=type4te)
```

Where such a test set is available, this provides a reassuring check that `randomForest()` is not over-fitting.

Note: The function `tuneRF()` can be used for such limited tuning as `randomForest()` allows. Look up `help(tuneRF()`, and run this function in order to find an optimal value for the parameter `mtry`. Then repeat the above with this optimum value of `mtry`, and again compare the OOB error with the error on the test set.

References

Andrews D F and Herzberg A M, 1985. *Data. A Collection of Problems from Many Fields for the Student and Research Worker.* Springer-Verlag. (pp. 339-353)

Charig, C. R., 1986. Comparison of treatment of renal calculi by operative surgery, percutaneous nephrolithotomy, and extracorporeal shock wave lithotripsy. *British Medical Journal*, 292:879–882.

Gordon, N. C. et al.(1995): 'Enhancement of Morphine Analgesia by the GABA$_B$ against Baclofen'. *Neuroscience 69: 345-349.*

Intersalt Cooperative Research Group. 1988. Intersalt: an international study of electrolyte excretion and blood pressure: results for 24 hour urinary sodium and potassium excretion. British Medical Journal 297: 319-328.

Maindonald, J. H.; Waddell, B. C.; and Petry, R. J., 2001. Apple cultivar effects on codling moth (Lepidoptera: Tortricidae) egg mortality following fumigation with methyl bromide. *Postharvest Biology and Technology*, 22:99–110.

Meyer, M.C. and Finney, T. (2005): 'Who wants airbags?'. *Chance* 18:3-16.

Wilkinson, G. N. & Rogers, C. E. 1973. *Symbolic description of models in analysis of variance.* Applied Statistics 22: 392-399.