

**Biostat II: Lab 1, Some Exploratory Data Analysis in R**  
**Date: 21 April 2008**

**1. Getting started in R**

(a) One of the most convenient things about R is that we can use it as a calculator. Try entering the following commands into R and see what happens:

- `2+3`
- `34*28`
- `sin(2*pi)`
- `exp(2)`
- `sqrt(100)`

(b) R is also very good at handling vectors, which is very important to handle data in statistics. Let's create some vectors. Type the following commands into R and see what happens:

- `1:10`
- `c(1, 3, 5, 7, 9)`
- `seq(1, 10, by=2)`
- `seq(2, 5, by=0.4)`
- `seq(2, 5, length=15)`

Can you tell what the “c” command does? How about the “seq” command? To learn more about these commands, try using R's help capabilities. Type ? followed by the name of the command you want to learn about (i.e. “?seq”). The `help` command is one of the most useful ways to learn more about any function in R. Also, check out `help.search` to learn about new functions, (e.g. “`help.search("linear regression")`”).

**2. Reading in data, summary statistics, plots of a single variable**

(a) Of course, we really want to use R to deal with real data. To do so, we'll need to enter some data. We can use R's assignment operator “<-” to store anything we want into a variable. We'll enter some age data, and look at it in various ways:

- `age <- c(23, 16, 14, 44, 25, 62, 44, 58, 26, 30, 32, 29, 21)`
- `age`
- `age + 10`
- `2 * age`
- `age^2`

(b) We can also look at summary statistics of the ages that we input:

- `mean(age)`
- `median(age)`
- `sd(age)`
- `summary(age)`

(c) We can look at a stem and leaf plot of the ages:

- `stem(age)`

What does R mean when it writes `The decimal point is 1 digit(s) to the right of the |` in the output of the stem and leaf plot?

(d) We can make a histogram of the ages:

- `hist(age)`

Based on the histogram, how many people do we have who are between 20 and 30 years of age?

- (e) Of course, we won't always want to enter our data by hand. When working with larger data sets, it is much more convenient to keep the data in a file, and read that file from R. First, visit the following website to take a look at some data there:

<http://lib.stat.cmu.edu/DASL/Datafiles/Calcium.html>

These data contain results of an experiment to study the effects of calcium on blood pressure in African-American men. We can read the data into R using the command:

```
calcium.table <-
read.table("http://lib.stat.cmu.edu/DASL/Datafiles/Calcium.html",
header=T, skip=28, nrow=21)
```

Use the command `?read.table` to understand all the arguments in this command better. What is the “header” argument for? How about “skip”? “nrow”?

Now that we've got the data into R, we can look at it all by typing “`calcium.table`”, or we can just see a summary by typing “`summary(calcium.table)`”.

3. **Plots and Graphics** Now that we have our data loaded, we would like to visualize trends in the data. Let's look at the relationship between beginning and end blood pressure for the entire sample:

- `plot(calcium.table$Begin, calcium.table$End)`

We can easily change the look of our plot:

- `plot(calcium.table$Begin, calcium.table$End, xlab="Begin", ylab="End", pch=20, col="red", main="Blood Pressure")`

and we can visualize the trend separately for calcium and placebo groups:

- `plot(calcium.table$Begin, calcium.table$End, xlab="Begin", ylab="End", pch=20, col=c("red", "blue")[calcium.table$Treatment], main="Blood Pressure")`
- `legend(98, 132, pch=20, col=c("red", "blue"), legend=c("Calcium", "Placebo"))`

Finally, we can also compare decrease in blood pressure for individuals in the calcium group versus those in the placebo group by using a box-and-whisker plot (a boxplot).

- `boxplot(calcium.table$Decrease ~ calcium.table$Treatment)`

There is a lot more to learn about R. Take your time, and you'll become more and more familiar throughout this course. Remember, you can find tons of resources by searching on Google, as well as by using R's `help.search` command.