University of California, Los Angeles
Department of Statistics

Statistics 13 **Instructor: Nicolas Christou**

Data analysis with R - Some simple commands

When you are in R, the command line begins with

```
>
```

To read data from a website:

```
> site="http://www.stat.ucla.edu/~nchristo/body_fat.txt"
> data <- read.table(file=site, header=T)
```

Another way to read data from a website is the following:

```
data <- read.table("http://www.stat.ucla.edu/~nchristo/body_fat.txt", header=TRUE)
```

This file contains data on percentage of body fat determined by underwater weighing and various body circumference measurements for 251 men. Here is the variable description:

Variable	Description
x_1	Density determined from underwater weighing
x_2	Percent body fat from Siri's (1956) equation
x_3	Age (years)
x_4	Weight (lbs)
x_5	Height (inches)
x_6	Neck circumference (cm)
x_7	Chest circumference (cm)
x_8	Abdomen 2 circumference (cm)
x_9	Hip circumference (cm)
x_{10}	Thigh circumference (cm)
x_{11}	Knee circumference (cm)
x_{12}	Ankle circumference (cm)
x_{13}	Biceps (extended) circumference (cm)
x_{14}	Forearm circumference (cm)
x_{15}	Wrist circumference (cm)

If the data file is on your computer (e.g. on your desktop), first you need to change the working directory by clicking on Misc at the top of your screen and then read the data as follows:

```
> data <- read.table("filename.txt", header=T)
```

Note: the expression <- is an assignment operator. The result of a read.table is a data frame (it looks like a matrix).

Useful commands:

- Extracting one variable from data (e.g. the second variable): `> data[,2]`
- Another way to extract one variable : `> data$x2`
- Similarly if we want to access a particular row in our data (e.g. first row): `> data[1,]`
- To list all the data simply type: `> data`
- To compute the mean of all the variables in the data set: `> mean(data)`
- To compute the mean of just one variable: `> mean(data$x2)`
- To compute the mean of variables 2 and 3: `> mean(data[,c(2,3)])`
- To compute the variance of one variable: `> var(data$x2)`
- To compute summary statistics for all the variables: `> summary(data)`.
- To construct stem-and-leaf plot, histogram, boxplot:

```
> stem(data$x2)
> boxplot(data$x2)
> hist(data$x2)
```

- To plot variable $x2$ against variable $x10$:

```
> plot(data$x2,data$x10)
```

- And you can give names to the axes and to your plot:

```
> plot(data$x2,data$x10, main="Scatterplot of percent body fat against
thigh circumference", xlab="Percent body fat",
ylab="Thigh circumference")
```

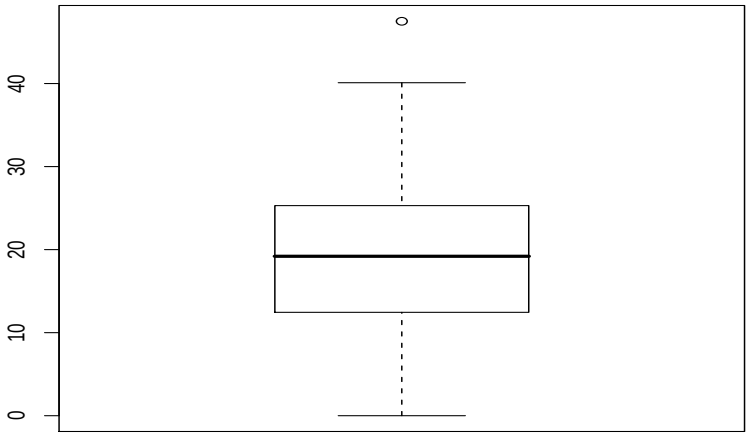
- To save a plot as a pdf file under the working directory (e.g. your desktop):

```
> pdf("box_x2.pdf")
> boxplot(x2)
> dev.off()
```

If you want to read more about a specific command (for example the histogram) at the command line you type the following:

```
> ?hist
> ?boxplot
```

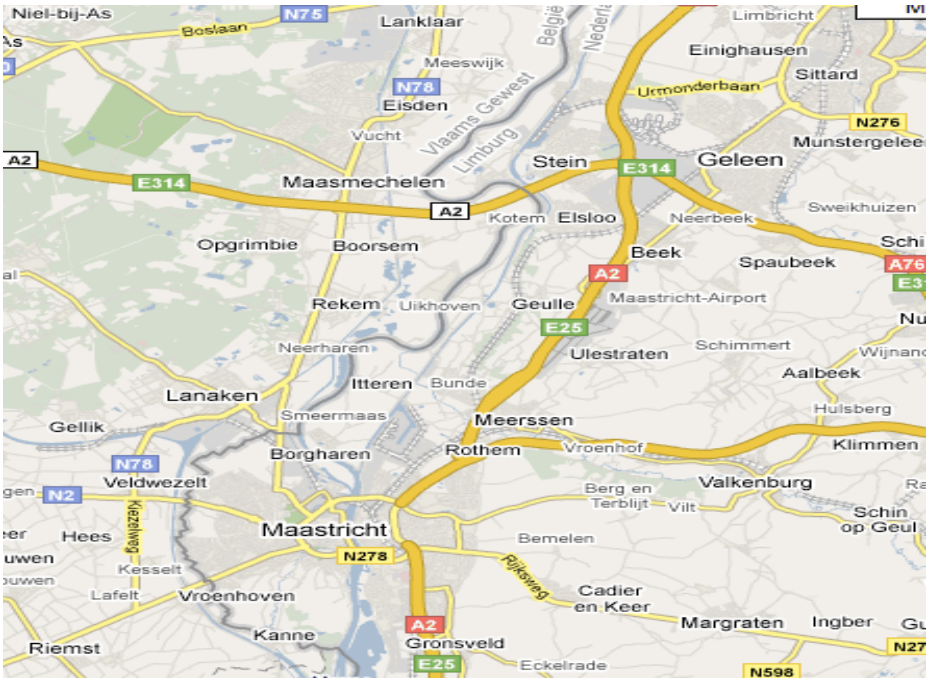
On your computer Desktop this is what you get (under the name "box_x2.pdf"):



- **Exercise:**
  Construct the same plots with different variables and save them on your desktop.

Another data set:

The following data were collected in the area west of the town Stein in the Netherlands near the river *Meuse* (Dutch *Maas*) river (see map below). The actual data set contains many variables but here we will use the x, y coordinates and the concentration of lead and zinc in *ppm* at each data point. The motivation for this study was to predict the concentration of heavy metals around the banks of the Maas river in this area. These heavy metals were accumulated over the years because of the river pollution. Here is the area of study:



Exercise:

a. You can access these data at

```
> soil <- read.table("http://www.stat.ucla.edu/~nchristo/statistics13/
          soil.txt", header=TRUE)
```

b. Construct the stem-and-leaf plot, histrogram, and boxplot for each one of the two variables (lead and zinc), and compute the summary statistics. What do you observe?

c. Transform the data in order to produce a symmetrical histrogram. Here is what you can do:

```
> log_lead <- log(soil$lead)
> log_zinc <- log(soil$zinc)
```

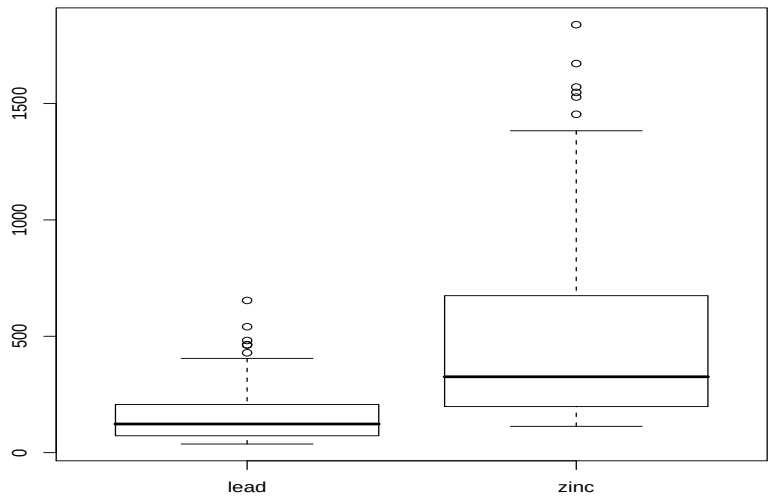
Construct the stem-and-leaf plot, histrogram, and boxplot for each one of the new variables (`log_lead` and `log_zinc`), and compute the summary statistics. What do you observe now.

Here is a side by side boxplot of the variables `lead` and `zinc`. First create a new data frame with only the variables lead and zinc:

```
soil_1 <- soil[,3:4]
```

Then you can construct a side by side boxplots of lead and zinc using:

```
> boxplot(soil_1)
```



Other useful commands in R:

- To create variables in R use <- or the equal sign =. Here are some examples:

```
> x <- c(1,2,3,4,5)
> y < c(10,20,30,40,50)

> q <- cbind(y,z)
```

And here is what you get:

```
> x
[1] 1 2 3 4 5

> q
     y  z
[1,] 1 10
[2,] 2 20
[3,] 3 30
[4,] 4 40
[5,] 5 50
```

- To rename variables:

```
> names(q) <- c("a", "b")
> q
  a  b
1 1 10
2 2 20
3 3 30
4 4 40
5 5 50
```