

Test Series 1

Develop properly working R code to solve the problems described in the sequel. Work in the Tinn-R editor and save the R code and your answers to the questions into a file.

To submit your test: **Send the file – before you leave! – by e-mail to schwierz@stat.math.ethz.ch. Add “Using-R Test 1” in the subject-line.**

If you have any technical difficulties or other questions, **please ask us** (not your neighbours).

Good Luck!

You need the following dataset:

```
d.pcb <- read.table("http://stat.ethz.ch/~stahel/courses/R/pcb.txt",
  header=TRUE)
```

The `pcb` dataset reports 122 measurements of Polychlorinated biphenyls (PCB) concentrations in the sediments of the North Sea outward of the Dutch coast. The dataset contains the following variables:

- the year of the measurements (`year`)
- the UTM coordinates `x`, `y` (unit m),
- the shortest distance of the location where the measurement was taken to the coast (`coast`, unit m),
- the water depth at the measurement location (`depth`, unit m), and
- the PCB content (`pcb`, normalized concentration).

1. Vectors and Matrices.

a) Generate the following vectors:

- 10 9 8 7 6 5 4 3 2 1 0
- 1 1 1 4 4 4 9 9 9 16 16 16 ... 100 100 100
- "a1" "b2" "c3" "d4" ... "p16"

Hint: use the vector `letters`.

b) Generate the matrix

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]  5.5   6  6.5   7  7.5   8  8.5   9  9.5  10
[2,] 10.5  11 11.5  12 12.5  13 13.5  14 14.5  15
```

and (matrix-) multiply it by its transpose.

2. Selecting elements. Now take a look at the PCB dataset.

- a) Type `summary(d.pcb)` to gain an overview of the data. How many variables and observations (lines) does the data set contain?
- b) Select the coordinates and the PCB values of the 8th to the 20th observation.
- c) How many PCB values in the data set are larger than 5? For those observations, find
 - the observation number (i.e. line),
 - the coordinates, and
 - the PCB values.
- d) Calculate (and report) the median of all PCB values.
- e) Calculate (and report) the medians of the PCB values for each year.

Boxplots and Statistical Tests.

3. a) First, we study the frequency distribution of the PCB measurements. Create a boxplot of the PCB content. Add a title and label the axis with the variable name.
- b) Create separate boxplots of the PCB content measured in the years 1991, 1996 and 2000. Use logarithmic scale and add notches. By visual comparison of the notches, can you determine whether there are significant differences between the years? If yes, between which years?
- c) Test the hypothesis that the PCB values of the years 1996 and 2000 are essentially the same by a Wilcoxon rank sum test (`wilcox.test`) and by a `t.test`.
Hint: First create two new vectors by selecting the PCB values for each year. Then you can use the two vectors as the first and the second arguments in the statistical tests.
(*) Do you need the paired or unpaired version of the test?
You need to examine the dataset more closely to answer this question. Do not spend too much time to find out.

4. Scatterplots. Next, we explore the spatial distribution of `pcb`.

- a) Create a plot of the positions of the measurement locations such that both axes display a given distance by the same length. Use different colors or different symbols to discriminate between years.
Add a legend to explain the meaning of the different color or types of symbol.
- b) Now create a “bubble plot” for the PCB content which shows the position of the sampling locations and the PCB content by the *area* of circular symbols.
Add a legend with the symbols corresponding to a normalized PCB content of 0.2, 0.5, 1, 2, 5, 10. Describe in 1–2 sentences the spatial distribution pattern of `pcb`. Note that the mainland lies to the right of the points in SW part of the displayed area.
- c) Create a scatterplot showing the dependence of the PCB concentration on the distance to the coast, for the years 1991 and 2000 only. Use a logarithmic scale on both axes, and use different colors or different symbols for the two years.
Add two horizontal lines to mark the `pcb` concentrations 2 and 5, respectively.
How many measurements exceeded concentration 2 in the two years? Include an R statement that yields this result (rather than counting points on the plot).