Test Series 2

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. Plots do not need to be saved! We will check your R-code to see if it generates the right plots.

There are 4 questions, each counting for 5-10 points. A minimum of 15 points will be necessary to pass the test.

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

IMPORTANT: Name your file according your name (LastnameFirstname_Test2.R), e.g. "MusterPeter_Test2.R".
To submit your test: Send the file by e-mail to `schwierz@stat.math.ethz.ch`. Add "Using-R Test 2" in the subject-line of the email. BEFORE YOU LEAVE, check with Conny Schwierz that your email has arrived.

Good Luck!

You need the following dataset:

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

The `dornach` data set reports measurements of three heavy metals (copper, cadimum, zinc) at 115 locations around a metal smelter. The data set contains the following variables:

- the coordinates `x`, `y` (unit m),
- a factor `'survey'` which codes the project in which the data were collected,
- `cu`, `zn`, `cd` metal concentrations in the topsoil (unit mg/kg).

1. **Generating a graphical summary of copper data [10 points].** Display the following plots on the same graphics device. All plots should have a title and the axes must be labelled. The plots should be arranged in the following way: plot **a**) → top left, **b**) → top right and **c**) → bottom left [**1 point**].

   a) Generate a scatter plot of `cu` vs. `x`. Use the color scale `rainbow` for displaying the `cu` content also by the color of the points (smallest concentration should be displayed in red and the largest in purple). [**2 points**]

   b) Generate of boxplot of `cu`. Use a logarithmic scale for the axis. Add horizontal lines to display the *guide* (40 mg/kg), *trigger* (150 mg/kg) and the *intervention thresholds* (1000 mg/kg). Use different colors for these lines. [**2 points**]

c) Lastly, generate a scatterplot of `y` vs. `cu`. Note that the Cu content should be displayed on the x-axis. Add again lines to display the guide, trigger and intervation thresholds as **b)**. Add a legend that explains the meaning of the lines. What can you now say about the spatial distribution of the Cu content? At what position is the metal smelter most probably located? **[3 points]**

d) Create the plots **a**) to **c**) again but this time choose `'black'` as the background color and `'white'` as foreground color and as the color for the axes and the tick marks of the axes. Increase the size of the lettering in the title by a factor of 2 and display all titles in `'yellow'` and the axis labels and tick mark annotations in `'cyan'`. **[2 points]**
Note: First save your current settings `oldpar <- par()` and reload the old settings when you have finished this questions by `par(oldpar)`.

2. Distributions [5 points].

a) What is the meaning/significance/interpretation of a *cumulative distribution*? **[1 point]**

b) Determine the values of the cumulative standard normal distribution (i.e. expected value $\mu = 0$ and standard deviation $\sigma = 1$) for the values $x = 1$ and $x = 2.5$. What is the interpretation of these numbers? **[2 points]**

c) Plot the cumulative probability distribution for an expected value $\mu = 5$ and standard deviation $\sigma = 3$. To achieve this: First generate a vector for the x-range from 0 to 20. Then create the results of the cumulative distribution for these values. Make a line plot, label the axes and create a suitable title. **[2 points]**

3. Functions [8 points]. Write a function that simulates n random numbers from a uniform distribution and that automatically calculates and outputs the mean, median and the standard deviation of the simulated numbers. The function should take the size of the random sample and the parameters of the uniform distribution as input.
Follow these steps:

a) Inform yourself about the arguments to the function `runif`. **[1 point]**

b) Try out the simulation of $n = 10$ random numbers from a uniform distribution with $min = 0$ and $max = 1$. List your results in the answers. **[1 point]**

c) Calculate the mean, median and the standard deviation of the 10 random numbers. Report the results. **[2 points]**

d) Now create a function that performs all these steps and takes the size of the random sample and the parameters of the uniform distribution as input variables. **[3 points]**
Hint: To output the 3 values for mean, median and standard deviation together, you need to link them together in a vector or list.

e) Finally, use your function with the following parameters:
$n = 10000, min = -10, max = 10$ and report the results for mean, median and the standard deviation. **[1 point]**

4. Missing Data [8 points].

a) Read in the file
`d.dornach2 <-`
`  read.table("http://stat.ethz.ch/~stahel/courses/R/dornach2.txt",...)`.
Note that this file contains missing values, which are coded as -999. **[1 points]**
Hint: Use the argument `na.strings=c(-999)`.

b) Find the number of missing values in your data frame. Find the lines and rows of the missing values in your data frame. Report those results in your answer. **[2 points]**

c) Calculate and report the average of the observed `cu`, `zn` and `cd` concentrations from `d.dornach2`. Compare them to the results of the averages from `d.dornach`. **[2 points]**

d) Test whether the `cu` contents recorded in the data frame `d.dornach2` differs significantly with respect to the two surveys `gs` and `wirz` using a Wilcoxon test. What is the result? Should you use a paired-sample test? **[3 points]**
Hints: First create two data vectors that contain the `cu` content for one survey only, by selecting the correct elements. Then use those two vectors in the test.