Resit: Statistical Computing with R

June 26, 2013

Instructions. Download the file `http://www.math.leidenuniv.nl/~gill/priming.csv` (alternatively, you can find it in the folder "Data sets" of the course drop-box folder) and then disconnect your computer from internet. You can connect again, at the end of the exam, in order to email your solutions *both* to Maarten Kampert and to Richard Gill (`gill@math.leidenuniv.nl`). Your solutions should be in the form of a well commented R script. Please include your name in the name of the file, e.g., `john.smith.R`.

You will get credit from writing code which is transparent and well-commented so that the examiner can understand your approach and easily find your answers.

1 Step 1 (getting the data).

The file `http://www.math.leidenuniv.nl/~gill/priming.csv` is a spreadsheet in csv format containing data from a controversial experiment in social psychology. The variables `creative` and `analytic` contain scores on two tests performed by 60 psychology students each subjected to one of three different experimental conditions, given by the variable `priming`. (In the appendix to the exam paper you will find some background information for your general interest – no need to read this now).

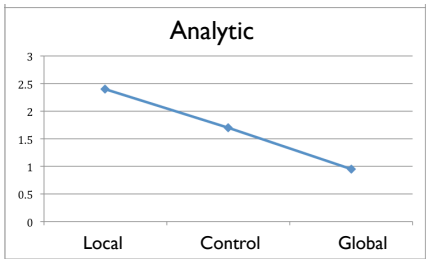
Create an R data-frame containing the 60 scores on each of these three variables. Make sure that `priming` is encoded as a factor with levels `local`, `control`, `global` *in that order*. Compute means and standard deviations of the scores on `analytic` and `creative` for the three groups. You should find

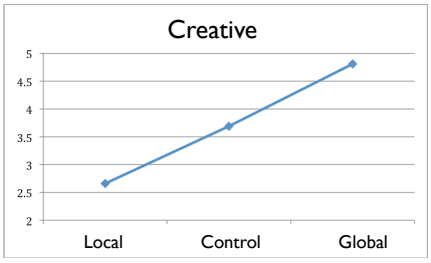
	local	control	global
means	2.40	1.70	0.95
sds	0.99	1.30	1.00

	local	control	global
means	2.66	3.69	4.81
sds	1.21	1.83	1.54

2 Step 2 (graphics).

Make plots of the average analytic and average creative score against priming level. You should be able to re-create the following two graphics:





Do your best to reproduce as many as possible of the graphical features of these two plots.

3 Step 3 (testing a hypothesis).

Suppose we assume that the score of the jth subject under the the ith experimental condition is independently and normally distributed with a mean depending on the group to which the subject belongs and constant variance: $Y_{ij} \sim N(\mu_i, \sigma^2)$, where $i = 1, 2, 3$, corresponding to local, control and global; and $j = 1, \ldots, 20$. As you know, this model can be fit in `R` using the command `lm(analytic~priming)`, provided that "priming" has been coded as a variable of type factor.

As a special case, we might assume that the three mean scores μ_i lie on a straight line $\mu_i = \alpha + \beta i$, as the graphics above suggest: fit with `lm(analytic~as.numeric(priming))`. The R command

`anova(lm(analytic~as.numeric(priming)),lm(analytic~priming))`

compares the fit of two different linear models. The command

`results <- anova(lm(analytic~as.numeric(priming)),lm(analytic~priming))`

saves the results of this analysis in an R list `results`. Where is the p-value of the F test in this object?

Write an R function which takes as arguments a numerical vector and a factor vector of the same length, and returns the p-value of the F-test of linearity as computed by the function "anova" in the way just shown. Applied to the data in question, you should find the following results: 0.9346 for analytic, 0.9124 for creative. Not statistically significant ... in fact, on the contrary, both tests return a strikingly *not* significant p-value.

4 Step 4 (resampling experiments)

The data is very clearly not normally distributed, so we may distrust the results just obtained. In this section we will use the bootstrap to get hopefully more reliable p-values. Use the function `set.seed()` to set the random seed of this sampling experiment, so that your results will be exactly reproduced when the examiner reruns your script.

Repeat the following 1000 times: (1) Take a random sample of size 20 *with replacement* from each of the three groups of 20 subjects. (2) Combine the samples from the three groups to create a new artificial data-set with 60 observations of "analytic", 60 of "creative", and a corresponding factor of length 60 with the three levels local, control, global. (3) Compute the p-values of the F-test of linearity for `analytic` $\sim$ `priming` and for `creative` $\sim$ `priming`, using the function you wrote in step 3.

How often do you see larger p-values than the two values found in the previous step? Draw a scatterplot of the pairs of p-values to investigate the dependence between the two.

Appendix: background

Some information just for background interest – not part of the examination.

According to psychological theory, the way people approach cognitive tasks can be influenced by getting their minds into different processing modes. The modes are called "local" and "global". Local processing involves attention to detail and carefully following rules; global processing involves creativity and association. The data in this examination comes from an experiment on 60 psychology students who were randomised into three groups of 20, called "local", "global" and "control". In the priming stage of the experiment, they were asked to taste several breakfast cereals and to report on the flavours. The differences between the breakfast cereals presented to the "local" group were designed to stimulate local processing mode, those presented to the "global" group were designed to stimulate global processing. The "control" group was given a mixture of both kinds of task.

After that, the students were given a completely unrelated analytic task involving logical reasoning, and a creative task involving inventing a funny caption to a cartoon. The experimenters showed that the "global" group did better at the caption competition than the "local" group, while with the analytic task, it was exactly the other way round. The control group, in both cases, was somewhere in between. The experiment resoundingly supported the theory of the psychologist.

Independent researchers later noticed a rather strange pattern in the data. The average score of the control group was not just "somewhere in between" the averages of the local and the global groups: it seemed to be exactly in between, in other words, one sees an almost perfect straight line relationship, though there is no reason at all to expect it in advance. This "too good to be true" pattern persisted over 20 or more similar experiments with priming carried out through all different senses (hearing, visual, touch, scent) and different analytic and creative tasks. Has the data been faked in some way?