Martin Losert and Katharina Naumann
Research Methods and Mathematical Psychology
Department of Psychology
University of Tuebingen

Problems in R

Data Structures and Data Manipulation

a) In preparation of tomorrow's lecture on copulas, install the `R` package `copBasic`. Ensure that the package is working properly by loading it into the workspace and executing an example of the function `COP()` via `example(COP)`.

b) Create a vector $\mathbf{x} = (1\ 5\ 8\ 3\ 7\ 2\ 6)'$. Use `seq()` to create another vector $\mathbf{y}$ of the same length containing the odd numbers $1, 3, 5, \ldots$ Compute a vector $\mathbf{z}$ as linear combination of $\mathbf{x}$ and $\mathbf{y}$: $\mathbf{z} = 4\mathbf{x} + 2\mathbf{y}$.

c) Combine the column vectors $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ into a matrix $\mathbf{A}$. Then create a matrix $\mathbf{B}$ with $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ as row vectors. Compute the matrix product $\mathbf{BA}$.

d) Create a data frame for 60 subjects and three factor variables:

id subject ID with levels from 1 to 60

type with levels "old" and "new" (mind the ordering!)

condition with levels "A", "B", and "C".

Type and condition are crossed factors in a balanced design, so there are 10 subjects in each cell. Use the following functions on the resulting data frame to check its structure:

- `str()`
- `summary()`
- `table()` or `xtabs()`

e) In an experiment, each subject has to respond to one stimulus and the subject's reaction time is recorded. Assume that reaction time is normally distributed as $RT \sim N(\mu = 400, \sigma^2 = 625)$. Using `rnorm()`, simulate reaction times for all subjects and store them in an additional variable in the previously created data frame.

Then use `aggregate()` to find the average reaction time in each of the six experimental conditions. Are there any suspiciously high or low cell means? Find out the minimum and maximum reaction time in each cell.

f) Go to
`http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-3E/`
and download the Vocabulary.txt data set (Fox, 2008). Store it in a local folder on your computer. Set that folder as your working directory using `setwd()`. Use `read.table()` to load these data into R's working memory. Hint: With `?read.table` you can get documentation for this function.

Use appropriate R functions to find out the number of observations, number of variables, variable names, descriptive statistics for the dependent variables, and the levels of the factor variables.

g) Extract the `vocabulary` variable from this dataframe and calculate its mean and standard deviation.

Look at the 217th row of the data frame and find out the following information: When was this person tested? For how long did this person go to school? What score did she have in the vocabulary test?

Extract only those cases who are male, were tested in 1974, and scored below 2 in the vocabulary test.

h) Sort the data frame by "years of education" and within that by "vocabulary test score." Hint: Use `order()`. On the sorted data frame, apply the functions `head()` and `tail()` to display the first and last 20 observations, respectively. What do you hypothesize about the relationship between "years of education" and "vocabulary test score?"

i) Are there sex differences in "years of education" and "vocabulary test score?" To find out, calculate the mean values of both variables for men and women separately. Hint: Use `aggregate()`.

Check if "years of education" have increased over the last decades. Calculate their means for every "year of testing."

Reference

Fox, J. (2008). *Applied regression analysis and generalized linear models.* Thousand Oaks: Sage Publications.