Information Management

LAB Session: 16/10/2014

Some quick info about R

- a language and environment for statistical computing and graphics
- an integrated suite of software facilities for data manipulation, calculation and graphical display:
  - an effective data handling and storage facility,
  - a suite of operators for calculations on arrays, in particular matrices,
  - a large, coherent, integrated collection of intermediate tools for data analysis,
  - graphical facilities for data analysis and display either directly at the computer or on hardcopy,
  - a *well developed, simple and effective programming language (called `S')* which includes conditionals, loops, user defined recursive functions and input and output facilities. (Indeed most of the system supplied functions are themselves written in the S language.)

  - IDEs: http://www.rstudio.com/

Exercise Session 1

- Open R
- Encapsulation philosophy
- Load and display data (read.csv)
- Access to column and row slices
- Contingency tables

Exercise Session 1

- Basic statistics:
  - Mean, median, min, max, midrange
  - Quartiles, IQR, variance, standard deviation
  - Mode
    - Table() → max → which.max
- Visualization (see wdb.data)
  - Histogram (hist(q))
  - Boxplot (boxplot(q))
  - Scatter plot (plot(q))
  - Quantile - Quantile plot (qqplot(q))
  - Quantile plot (to be implemented with loops, quantile() and plot())
  -

Exercise Session 1

- S programming
  - Assignments, comparisons
  - Conditional statements
  - Loops
  - Definition of functions
- Write functions for
  - Mean, variance, stddev, median, mode

Exercise Session 1

- Proximity measures:
  - Build a contingency table and a dissimilarity matrix according to symmetric binary vars distance and Jaccard coefficient
  - Build a function for Minkowsky distance computation
  - Build a function for Cosine similarity

Exercise Session 1

- Data cleaning:
  - Pick a database (bcw attribute 7) and replace '?' with 'null' (replace + mapply)
- Correlation analysis:
  - Chi-square calculation (simple example ON DIAGNOSIS)
  - Covariance and Pearson

List of data

- Bcw: integer, good for tables (e.g. 3 and 4 are somehow correlated)
- Bcw 7 has missing data
- Diagnosis → reals (good for quantile, plots) ...