Instructions (Readme first)

Mike A. Marin

August 15, 2016

Introduction

In the spirit of literate programming and reproducible research, this document **will eventually** contain the description and all the steps required to process the files and reproduce the calculations used to analyse the data collected by the CMMN complexity metrics survey.

To reproduce the analysis for this research, you need,

- Software,
  - LaTeX distribution
  - R
  - RStudio (optional but nice to have)
  - R Markdown
  - knitr – from RStudio just run install.packages(knitr)
- The data files,
  - Survey-raw-data-from-LimeSurvey\results-survey338792.csv, data file exported from LimeSurvey
  - work\in-survey-var-names.csv, file exported from LimeSurvey
  - work\in-independent-variables.csv, file describing the independent variables (created manually)
  - work\in-independent-variables-map.csv, mapping of independent variables to each of the 30 groups (created manually)
  - work\in-weights.csv, file containing the weights to calculate CMMN Complexity (CC) (created manually)
- Sources and script files,
  - Instructions(read-me-first).Rmd – The file that generates this pdf file
  - work\Daily.BAT – main build script that will execute all other scripts (except the instructions)
  - work\daily.r – main R script that calls all *.Rmd scripts
  - work\copy-and-fix-file.r – Script used to copy and fix the LimeSurvey exported file
  - work\CMMN-Convert-File.Rmd – Script that generates the *dataset-all.csv*, and *dataset-clean.csv* files
  - work\CMMN-Sample.Rmd – compares the data set against the expected sample size for each experiment
  - work\CMMN-basic-stats.Rmd – Generate basic demographic statistics
  - work\CMMN-Weights.Rmd – Recalculate CC (iv.A.CC, iv.B.CC, and iv.C.CC) and generates the *dataset-clean-post.csv*

Files in this directory

The main files in this directory are:

- **Instructions(read-me-first).pdf** – this file
- **Empirical-Validation.pdf** – describe the experiments (basic methodology description)
- Survey-raw-data-from-LimeSurvey\results-survey338792.csv – original raw data file from LimeSurvey
- Survey-raw-data-from-LimeSurvey\results-survey338792 (description).pdf – Description of the raw data file
- work\Daily.BAT – Main script that do execute all other scripts to generate all the output files and reports
- work\daily.r – main R script
- work\CMMN-basic-stats.Rmd – R script that calculate basic demographics statistics and produces a report
- work\CMMN-Convert-File.Rmd – R script that convert the raw data file into dataset-all.cvs and produces a report
- work\CMMN-Sample.Rmd – R script that anlyses the data set to evaluate sample sizes and produces a report
- work\CMMN-Weights.Rmd – R script to generate a new dataset (dataset-clean-post.cvs), where the CC metric has been recalculated based on observed weights. Columns iv.A.CC, iv.B.CC, and iv.C.CC are the recalculated variables
- work\copy-and-fix-file.r – R script that converts the raw data set into a usable dataset (fixes a LimeSurvey issue)
- work\dataset-all(description).pdf – Describes the dataset variables
- work\in-independent-variables-map.csv – input file with some independent variables
- work\in-independent-variables.csv – input file with most of the independent variables
- work\in-survey-var-names.csv – file from LimeSurvey listing the raw variable names
- work\in-weights.csv – input file with weight independent variables
- work\pics\by-sa.png – creative commons icon

The generated files are:

- work\dataset-all.csv – converted data set containing all the collected data (includes incompleted surveys)
- work\dataset-clean.csv – converted data set useful for statistical analysis. This file is the same as *dataset-all.csv*, but removing all the rows with variable valid.row different than 1.
- work\dataset-clean-post.csv – converted data set useful for post-hoc statistical analysis with new CC variables
- work\out-comments.txt – File containing the comments that sujects provided in the survey
- work\CMMN-basic-stats.pdf – report generated by CMMN-basic-stats.Rmd
- work\CMMN-Convert-File.pdf – report generated by CMMN-Convert-File.Rmd
- work\CMMN-Sample.pdf – report generated by CMMN-Sample.Rmd
- work\CMMN-Weights.pdf – report generated by CMMN-Weights.Rmd

How to run the scripts

First, be sure you have all the required software installed. Second, you must modify the *Daily.bat* to adjust the configuration section to your environment. Now, you are ready to run the analysis, which you can do by just executing the *Daily.bat*.

You must run the *Daily.bat* at least one to create the *out-clean-data.csv* file that is used by all the scripts that do statistical calculations. After that, you can run individual scripts in RStudio or R console. In either case, you will be running the script in the following way,

```
library(knitr)
library(markdown)
render("script.Rmd", "pdf_document")
```

The process executed by *daily.bat* can be summarized as follows:

1. Executes *copy-and-fix-file.r* (input: *results-survey338792.csv*, output: *in-survey-data-file.csv*)
2. Executes *daily.r*, which in turns executes the following scripts,
    a. CMMN-Convert-File.Rmd (input: *in-survey-data-file.csv*, output: *dataset-all.csv*, and *dataset-clean.csv*)
    b. CMMN-Sample.Rmd (input: *dataset-all.csv*)
    c. CMMN-basic-stats.Rmd (input: *dataset-all.csv*)
    d. CMMN-Weights.Rmd (input: *dataset-all.csv*, output: *dataset-clean-post.csv*)

Note: *Instructions(read-me-first).Rmd* is the only script not executed by *daily.bat*.

Data-set for statistical analysis

The file you want to use for statistical analysis is the *dataset-clean.csv* file. This file contains the clean data that can be used for analysis. This file is the same as *dataset-all.csv*, but removing all the rows with variable valid.row different than 1.

In addition, the *dataset-clean-post.csv* file, contains the same data, but the variables iv.A.CC, iv.B.CC, and iv.C.CC have been recalculated based on the observed weights.