Generating synthetic data with the synthpop package for R

Belfast, Wednesday 20th June, Queen's University Belfast, Peter Froggatt Centre, (No 2 on map) Room 3.017 (map of campus)

Preparation for the course

1) Make a directory to hold the data from this course, perhaps call it *syncourse*
2) Install R (https://cloud.r-project.org/) and RStudio (https://www.rstudio.com/products/rstudio/download/) if not already installed on the laptop you will bring to the course
2) Start RStudio.
3) Install the synthpop package (tools menu install packages)
4) Open a new R script (file menu)

You should type the commands you are using in the script window and then save the script in your R course directory. To run commands you just do <Ctrl><Return> when your cursor is in the line you want to run, or highlight a whole set of lines to run.

Run the following commands to get started (text after # is comments)

```
rm(list=ls())  ### to clean out (remove) any items in the workspace

library(synthpop)  ### to make the synthpop package available
setwd("D://sls//synthetic//syncourse") ##### to set the course directory
                                       ##### change this to your own
#
# The synthpop package includes a data set SD2011 which is a data.frame
# a list of the variables in it are on page 3 of these notes;
# you can also see R help page for additional information (command ?SD2011)
#
```

A test data frame `SD2011` is available as part of the synthpop package. Go to its help page using command `?SD2011`, where you can find details about it. A list of the variables in it is also on page 2 of these notes. The examples in the help files for the synthpop functions use this data set. The first thing you will be asked to do in the course practical is to select four variables to use as your example data set for synthesising. Chose variables where you might be interested in examining relationships. Look at the data and try to decide on your four variables now (you can change later) :-

YOUR CHOICE____

If you are an R beginner see page 3 for some hints.

Codebook for data frame SD2011 that is part of the synthpop package

column	Variable name	class	N not missing	N missing	% missing	Number of distinct values
1	sex	factor	5000	0	0.00	2
2	age	numeric	5000	0	0.00	79
3	agegr	factor	4996	4	0.08	6
4	placesize	factor	5000	0	0.00	6
5	region	factor	5000	0	0.00	16
6	edu	factor	4993	7	0.14	4
7	eduspec	factor	4980	20	0.40	27
8	socprof	factor	4967	33	0.66	9
9	unempdur	numeric	5000	0	0.00	30
10	income	numeric	4317	683	13.66	406
11	marital	factor	4991	9	0.18	6
12	mmarr	numeric	3650	1350	27.00	12
13	ymarr	numeric	3680	1320	26.40	74
14	msepdiv	numeric	700	4300	86.00	12
15	ysepdiv	numeric	725	4275	85.50	50
16	ls	factor	4992	8	0.16	7
17	depress	numeric	4911	89	1.78	22
18	trust	factor	4963	37	0.74	3
19	trustfam	factor	4989	11	0.22	3
20	trustneigh	factor	4989	11	0.22	3
21	sport	factor	4959	41	0.82	2
22	nofriend	numeric	5000	0	0.00	44
23	smoke	factor	4990	10	0.20	2
24	nociga	numeric	5000	0	0.00	30
25	alcabuse	factor	4993	7	0.14	2
26	alcsol	factor	4918	82	1.64	2
27	workab	factor	4562	438	8.76	2
28	wkabdur	character	5000	0	0.00	33
29	wkabint	factor	4964	36	0.72	3
30	wkabintdur	factor	303	4697	93.94	5
31	emcc	factor	286	4714	94.28	17
32	englang	factor	4985	15	0.30	3
33	height	numeric	4965	35	0.70	64
34	weight	numeric	4947	53	1.06	90
35	bmi	numeric	4941	59	1.18	1387

For R beginners

A basic level of knowledge of R is required and if you need to get up to speed you need to access an online resource such as the free textbook *Introductory Statistics with R* by Dalgaard (Chapter 1-2) or other online resources. To learn the basics you can also use a free DataCamp course *Introduction to R*.

The main R objects you will be using are data frames (like data sets or tables in other packages). Each variable in a table has a data type, e.g. numeric, character, factor. An R workspace can contain many data frames and other R objects.

Here are some useful commands to use with data frames

```
summary(mydataframe)
dim(mydataframe)
names(mydataframe)
newdf <- mydataframe[1:1000, c(1,3,4)]
#
# The last command makes a new data frame with the first 1000 rows and
# columns 1, 3 and 4
```

The other R object you need to know about is a list. An R list allows you to group a whole set of R objects together, where the objects can be quite different things. The output of the syn() function for carrying out synthesis is a list with many components, you can see the names of the components in the help file for syn() (command ?syn): see the Value entry at the bottom of the help page. Individual elements of a list are accessed from their names using the syntax mylist$myelement. For example:

```
mysyn <- syn(SD2011[1:1000, c(1,3,4)])
mysyn$call              ## gives you the command you used to make mysyn
mysyndata <- mysyn$syn  ## this is the data frame of synthetic data
head(mysyndata)         ## shows the first 6 rows of the new data frame
```

Here is a list of a few useful R commands you should know

Command	What it does	Command	What it does
ls	lists R objects	plot	makes 2 way plots according to data type
dim	dimensions of a data frame or matrix	with	uses data frame
class	returns the class of an R object	lm	fits linear models
head	first few lines of a data frame or entries in a vector	glm	fits generalised linear models
tail	last few lines of a data frame or entries in a vector	names	names of an R object
table	makes tables (note it omits missings unless you specify -see help)	summary	summarises an R object; results depend on object class
hist	histograms	NA	R missing data code