Data Confidentiality in Practice

Privacy Seminar, LMU SoSe 2022

Anna-Carolina Haensch, Marcel Neunhoeffer, Leah von der Heyde, Frauke Kreuter

2022-06-21 and 2022-06-28

Outline

- A Very Brief History of Data Confidentiality
- Traditional Approaches to SDC and their Problems
  - Theory - Group work
  - sdcMicro and sdcTable R Packages
- Synthetic Datasets as an Alternative
  - Theory
  - synthpop R Package

Thank you!

Some of the slides are adapted from Jörg Drechsler's course on Statistical Disclosure Control at IPSDS. The extensive documentation for synthpop and sdcMicro and sdcTable are also helpful to dig further into some of the topics discussed here.

History of Data Confidentiality

- Data confidentiality is a hot topic, but only since the last 3-4 decades
- Personal information has been collected for thousands of years

Clay tablet with tabulation of persons and goods from 7thC BC, Babylonia

https://www.britishmuseum.org/collection/object/W_1881-0706-688

History of Data Confidentiality

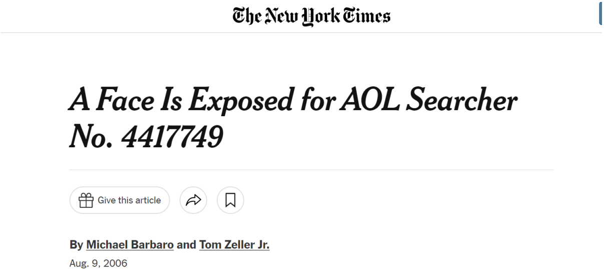
- In the early days most data collected by statistical agencies
  - Most information was published only in tables
  - Access to the microdata for external researchers was unthinkable and nobody else stored much data

History of data confidentiality

- Research on data confidentiality mainly focused on tabular data
  - Confidentiality for tabular data still a very important topic for statistical agencies
- Things changed with increase in computing power
  - Researchers requested access to the underlying microdata
  - Agencies started thinking about data dissemination strategies
- And later on also big industry players

Famous privacy breaches

- Identification of a city mayor in “anonymised” medical records in the USA
- A Face Is Exposed for AOL Searcher No. 4417749
- Netflix Spilled Your Brokeback Mountain Secret



The New York Times

A Face Is Exposed for AOL Searcher No. 4417749

By Michael Barbaro and Tom Zeller Jr.

Aug. 9, 2006

Figure 1: Snapchot from the New York Times article

Group work

Group work

Please divide into three groups

- Each group will work on different historical approaches to data confidentiality
- We will come together in ∼30 minutes and you will present your results to the other groups
- Results will be shared in written form after the session

Reading 1

Reiter, J. P. (2012). Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. Public Opinion Quarterly 76(1), 163–181. https://doi.org/10.1093/poq/nfr058

- Short overview of risk assessment and utility
- Discussion with lots of examples from Official Stats and discussion of impact on data utility
  - Recoding
  - Data Swapping
  - Adding Noise
- Short discussion of partially synthetic data

Group 1

Cell suppression: Outline

- Why are small cell counts in tables potentially problematic?
- Definition for primary and secondary cell suppression
- Problems for utility

Cell suppression

	Good Health	Fair Health	Bad Health	Very bad Health	Total
White	6	7	3	2	18
Mixed	2	2	3	1	8
Asian	1	0	5	0	6
Black	0	5	0	0	5
Other	0	0	0	1	1
Total	9	14	11	4	38

Cell suppression

- Definition for cell suppression:
- Primary suppression and second cell suppression
    - Primary suppression can be defined as **withholding the values of all risky cells from publication**, which means that their value is not shown in the table but replaced by a symbol such as xxx to indicate the suppression.
    - In frequency count tables cells containing **small counts** as understood as risky cells.
    - To reach the desired protection for risky cells, it is necessary to suppress additional non-risky cells, which is called complementary (secondary) suppression, for example total counts.

Cell suppression: Magnitude tables

When are cells in magnitude tables potentially problematic?

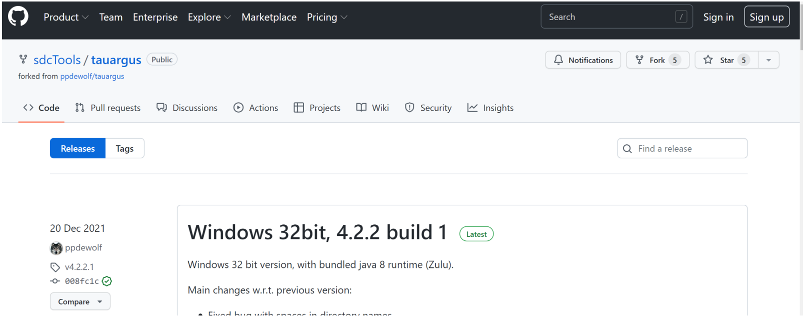
- Dominance rule
  - A cell is regarded as confidential, if the n largest units contribute more than k% to the cell total.
- p%-rule
  - The p% rule states that a cell, contained in a table of magnitudes, is sensitive if the value for any contributor can be calculated to within a given percentage.

Cell suppression: Magnitude tables

- pq-rule
  - It is assumed that out of publicly available information the contribution of one individual to the cell total can be estimated to within p per cent (p=error before publication); after the publication of the statistic the value can be estimated to within q percent (q=error after publication).
  - The pq-rule states that a cell, contained in a table of magnitudes, is sensitive if the value for any contributor can be calculated to within a given pq-threshold.

Cell suppression: Tau-Argus

Tau-ARGUS is a software program designed to protect statistical tables



Windows 32bit, 4.2.2 build 1

Windows 32 bit version, with bundled java 8 runtime (Zulu).

Main changes w.r.t. previous version:

- Fixed bug with spaces in directory names

Figure 2: Tau-ARGUS

Cell suppression: Problems with utility

- No guarantee that information is protected even if sensitivity rules are fulfilled
- Often tables are released sequentially
- Difficult to keep track of all released information
- Large number of cells need to be suppressed

k-anonymity

A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k - 1 individuals whose information also appear in the release.

Problems with k-anonymity

	Fair Health	Obesity	Hypertension	Diabetes	Total
White	6	7	5	6	24
Mixed	5	x	x	x	x
Asian	x	10	5	7	x
Black	0	0	0	5	5

Group 2

Local suppression

- Definition of local suppression
  - Sensitive values in the data are set to missing
- Potential inferential consequences of local suppression with a typical user:
  - Records that contain missings will be deleted
  - Bias in the obtained inferences
  - Suppressed values are often outliers that have strong impact on results
  - Risk that close approximation of the suppressed value is possible through imputation

Global and top coding

- Definition of global encoding
  - Continuous variables are discretized
- Definition of top encoding
  - Values above a predefined threshold are replaced
  - Sometimes values above the threshold are replaced with the average of the replaced values
    - Preserves total and mean
    - Variance is underestimated
    - Regression coefficients biased towards zero

PRAM method

- PRAM means that for each record in the data file the score of a number of variables is changed, according to a specified probability mechanism.

Group 3

Rounding

- Definition: All values are rounded to the closest rounding base.
- Inferential consequences:
  - Variance is underestimated
  - Correlations (and regression coefficients) are biased towards zero

Swapping

- Definition: Individual values are swapped between records for sensitive variables
- Very low protection since swapping rate usually small to minimize negative effects on analytical validity
- Inferential consequences:
  - Strongly negative on data quality, almost all multivariate relationships are strongly distorted

Other forms of swapping

- Group swapping
  - Values for several variables are swapped simultaneously
  - Correlations within swapped variables preserved
  - Correlations between swapped and unswapped records distorted

Other forms of swapping

- Rank swapping
  - Works for continuous variables
  - Variable is sorted
  - Values can only be swapped within a p% range according to the rank ordering
  - Ensures that replacement values are not too far apart from originally observed value
  - Increases analytical validity
  - Substantially increases risk

Micro-aggregation

- Definition
- Records are grouped in clusters of size k, clusters are chosen to minimize variance in each cluster, values in each group are replaced with centroid of the cluster
- Inferential consequences
  - Variances from the protected dataset will be too small
  - Relationships between the variables will be distorted

Micro-aggregation

- Risk
  - Intruder might know that all units in a cluster share (almost) the same value for the sensitive attributes
  - Risk of disclosure still high unless k is large

Multi-variate micro-aggregation vs. uni-variate aggregation

- Multivariate extension, aggregates several variables simultaneously
- Replaces observed values in each cell with multivariate centroid
- Finding optimal clustering solution is computationally hard but approximations exist
- Still considerable distortion of the correlation

R implementations

Some of these methods can be easily coded by hand, some require more specialized programs

There are two prominent packages in R:

- sdcTable for tabular data
- sdcMicro for microdata

We will not discuss the implementations but want to point to the extensive documentation available for both packages, e.g.

sdcpractice.readthedocs.io/en/latest/anon_methods.html

sdcMicro

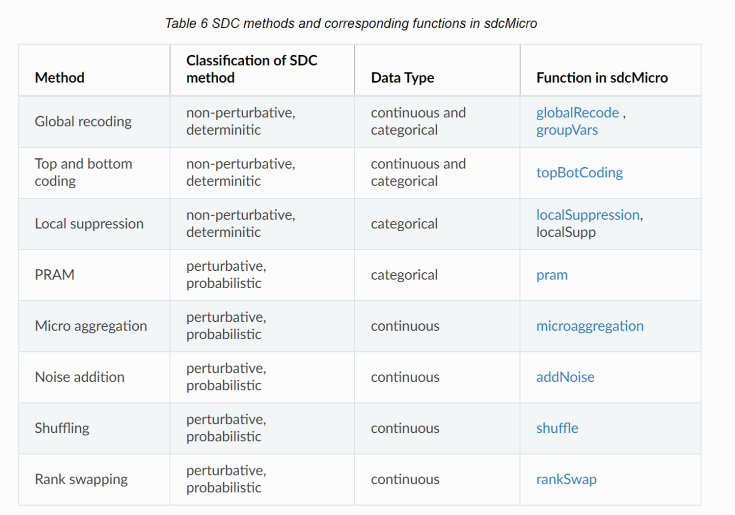


Table 6 SDC methods and corresponding functions in sdcMicro

Method	Classification of SDC method	Data Type	Function in sdcMicro
Global recoding	non-perturbative, determinitic	continuous and categorical	globalRecode , groupVars
Top and bottom coding	non-perturbative, determinitic	continuous and categorical	topBotCoding
Local suppression	non-perturbative, determinitic	categorical	localSuppression, localSupp
PRAM	perturbative, probabilistic	categorical	pram
Micro aggregation	perturbative, probabilistic	continuous	microaggregation
Noise addition	perturbative, probabilistic	continuous	addNoise
Shuffling	perturbative, probabilistic	continuous	shuffle
Rank swapping	perturbative, probabilistic	continuous	rankSwap

Data synthesis

First ideas

- The idea of using synthetic datasets for greater data confidentiality was first proposed by Don Rubin in 1993.
- General idea:
  - Treating the units that were not sampled for the survey as missing data.
  - Multiply impute this missing information for a sample
  - **Fully synthetic approach**
- Later works also extended the approach to only **partially synthetic data**
  - Only impute part of the data

Fully vs. partially synthetic data

Original dataset

👤	💵	🩺	📍	💼

Fully synthetic dataset

👤	💵	🩺	📍	💼
🟨	🟨	🟨	🟨	🟨
🟨	🟨	🟨	🟨	🟨
🟨	🟨	🟨	🟨	🟨
🟨	🟨	🟨	🟨	🟨
🟨	🟨	🟨	🟨	🟨

Partially synthetic dataset

👤	💵	🩺	📍	💼
	🟨	🟨		
	🟨	🟨		
	🟨	🟨		
	🟨	🟨		
	🟨	🟨		

How does this work?

y_{obs} is the part of the dataset to be synthesized, x_{obs} is the part of the dataset that can remain unchanged (can be empty).

- We define the joint distribution as a series of conditional distributions
- One column of y_{obs} is selected and the distribution of this variable, conditional on x_{obs} is estimated
- The distribution of the next column is estimated conditional on x_{obs} and the column of y_{obs} already synthesized and so on:
- **In parallel**: Each column of the synthetic data is generated from the assumed distribution, conditional on x_{obs}, the fitted parameters of the conditional distribution and the synthesized values of all the previous columns of y_{obs}.

Synthesizing data - In short

- Generating synthesis always consists of two steps
  - Generate a random draw from the posterior distribution of the model parameters for the synthetic data model given the data $f(\theta|D_{obs})$ (or take the fitted values)
  - Generate synthetic values by drawing from the specified data distribution given the drawn model parameters

Implications for the legitimate data user

- Allows analyses for population estimands using straightforward statistical tools
- Data is analyzable using the full range of standard complete-data statistical tools
- Valid inferences are easy to obtain
  - Since data synthesis was historically developed from multiple imputation, variance estimates first relied on creating multiple synthetic datasets
  - Solution by Raab (2016) now available for single synthetic datasets

Seems trivial but real data applications offer additional challenges

- Semi-continuous variables
- Skip patterns
- Bound variables
- Logical constraints

Intro to the R Package synthpop

synthpop R Package

```r
library("synthpop")
```

The synthpop package includes a data frame SD2011 with individual microdata that will be used for illustration.

The dataset is a subset of survey data collected in 2011 on objective and subjective quality of life in Poland.

Look at the SD2011 dataset

```
# Variables to be included
vars <- c("sex", "age", "marital",
          "income", "ls", "wkabint")

#Reduce dataset
origdat <- SD2011[, vars]
```

The dataset SD2011

```
## Rows: 3
## Columns: 6
## $ sex     <fct> FEMALE, MALE, FEMALE
## $ age     <dbl> 57, 20, 18
## $ marital <fct> MARRIED, SINGLE, SINGLE
## $ income  <dbl> 800, 350, NA
## $ ls      <fct> PLEASED, MOSTLY SATISFIED, PLEASED
## $ wkabint <fct> "NO", "NO", "NO"
```

Lets try synthesizing some data

```r
my.seed <- 12031992
syn.object <- syn(origdat, seed = my.seed)
```

```
## Synthesis
## -----------
##  sex age marital income ls wkabint
```

What type of information is contained in syn.object?

```
names(syn.object)
```

```
##  [1] "call"              "m"
##  [3] "syn"               "method"
##  [5] "visit.sequence"    "predictor.matrix"
##  [7] "smoothing"         "event"
##  [9] "denom"             "proper"
## [11] "n"                 "k"
## [13] "rules"             "rvalues"
## [15] "cont.na"           "semicont"
## [17] "drop.not.used"     "drop.pred.only"
## [19] "models"            "seed"
## [21] "var.lab"           "val.lab"
## [23] "obs.vars"          "numtocat"
## [25] "catgroups"
```

What type of information is contained in syn.object?

```
syn.object$method
```

```
##      sex      age  marital   income       ls  wkabint
## "sample"   "cart"   "cart"   "cart"   "cart"   "cart"
```

synthpop has sensible method defaults for different variable types. Just make sure that categorical unordered variables have the data type factor. We will see how to manually change the methods in a second.

What type of information is contained in syn.object?

```
syn.object$visit.sequence
```

```
##      sex     age marital  income      ls wkabint
##        1       2       3       4       5       6
```

The visit sequence can also be inspected (and changed).

What type of information is contained in syn.object?

```
syn.object$predictor.matrix
```

```
##         sex age marital income ls wkabint
## sex       0   0       0      0  0       0
## age       1   0       0      0  0       0
## marital   1   1       0      0  0       0
## income    1   1       1      0  0       0
## ls        1   1       1      1  0       0
## wkabint   1   1       1      1  1       0
```

The variables used to predict and generate synthetic values for a sensitive variable depend on the visit sequence. The resulting predictor matrix can also be inspected or manually changed.

Let’s tweak the synthesis procedure a little bit

Let's tweak the synthesis procedure a little bit

Goals:

- Let's make sure that there are no records with age<18 and marital=married
- Let's exclude sex (sex) from the predictors of life satisfaction (ls)
- Let's use marital status (marital) as predictor but do not synthesize the variable itself
- Let's use random forest to generate income (income) instead of the default cart method

Synthesis with logical constraints

In Poland, marriage is allowed only at age 18. People younger than 18 should therefore have the status **SINGLE** for variable **marital**. We can implement these and other logical constraints using **rules** and **rvalues**.

```r
rules.marital <- list(marital = "age < 18")
rvalues.marital <- list(marital = "SINGLE")
syn.object.married <- syn(origdat,
                        rules = rules.marital,
                        rvalues = rvalues.marital,
                        seed = my.seed)
```

```
## Synthesis
## -----------
##  sex age marital income ls wkabint
```

Results: synthesis with logical constraints

```
table(syn.object$syn$age < 18,
      syn.object$syn$marital == "SINGLE")
```

```
##
##          FALSE TRUE
##   FALSE  3790 1108
##   TRUE      4   92
```

```
table(syn.object.married$syn$age < 18,
      syn.object.married$syn$marital == "SINGLE")
```

```
##
##          FALSE TRUE
##   FALSE  3770 1124
##   TRUE      0   96
```

Excluding and including variables for generating variables

- Let's exclude sex (sex) from the predictors of life satisfaction (ls)

```
# With m=0 we do not generate any data,
# just create an syn.object which we can
# use to modify the generating procedure

syn.object.ini <- syn(origdat,
                      seed = my.seed,
                      m = 0)

syn.object.ini$predictor.matrix["ls",
                                "sex"] <- 0
```

Excluding and including variables for generating variables

```
syn.object.ini$predictor.matrix
```

```
##          sex age marital income ls wkabint
## sex        0   0       0      0  0       0
## age        1   0       0      0  0       0
## marital    1   1       0      0  0       0
## income     1   1       1      0  0       0
## ls         0   1       1      1  0       0
## wkabint    1   1       1      1  1       0
```

```
my.pred.matrix <- syn.object.ini$predictor.matrix
syn.object.predchanged <- syn(data = origdat,
                              predictor.matrix =
                                my.pred.matrix)
```

Excluding variables from data synthesis

- Let's use marital status (marital) as a predictor but do not synthesize marital status (marital) variable itself

```
syn.object.ini$method["marital"] <- ""
my.method1 <- syn.object.ini$method
syn.object.methodchanged1 <- syn(data = origdat,
                                 method = my.method1)
```

Change methods used for synthesis

- Let’s use random forests to synthesize income (income) instead of the default cart method

```r
syn.object.ini$method["income"] <- "rf"
my.method2 <- syn.object.ini$method
syn.object.methodchanged2 <- syn(data = origdat,
                                 method = my.method2)
```

Decision Tree

- Linear regression and logistic regression models fail in situations where the relationship between variables and outcome is nonlinear or where variables interact with each other (and this interaction is not included).
- Time to shine for the decision tree!

Decision Tree

- Tree based models split the data multiple times according to certain cutoff values in the variables.
- Through splitting, different subsets of the dataset are created, with each instance belonging to one subset.
- The final subsets are called terminal or leaf nodes and the intermediate subsets are called internal nodes or split nodes.
- To predict the outcome in each leaf node, the average outcome of the training data in this node is used.
- Trees can be used for classification and regression.

Decision Tree

- There are various algorithms that can grow a tree.
- They differ in the possible structure of the tree (e.g. number of splits per node), the criteria how to find the splits, when to stop splitting and how to estimate the simple models within the leaf nodes.
- The classification and regression trees (CART) algorithm is probably the most popular algorithm for tree induction.

Advantages

- The tree structure is ideal for **capturing interactions** between features in the data.
- The data ends up in **distinct groups** that are often easier to understand than points on a multi-dimensional hyperplane as in linear regression.
- The interpretation is arguably pretty simple.
- The tree structure also has a **natural visualization**, with its nodes and edges.

Disadvantages

- **Trees fail to deal with linear relationships**. Any linear relationship between an input feature and the outcome has to be approximated by splits, creating a step function.
- This goes hand in hand with **lack of smoothness**. Slight changes in the input feature can have a big impact on the predicted outcome, which is usually not desirable.
- Trees are also quite **unstable**. A few changes in the training dataset can create a completely different tree.

Passive synthesis

Variables that are exact functions of other variables that are synthesized should not be synthesized themselves.

- Ex: weight and height are synthesized and the bodymass-index is also included in the dataset
- Use the **syn.passive** method

```
origdat2 <- SD2011[, c("height", "weight", "bmi")]
origdat2 <- origdat2[!is.na(origdat2$weight),]
origdat2 <- origdat2[!is.na(origdat2$bmi),]
meth <- c("sample", "cart",
          "~I(weight / height^2 * 10000)")
syn.object.passive <- syn(origdat2, method = meth)
```

Beware: Function only works if there is no missing data :(

Evaluating the utility of synthetic datasets

Evaluating the utility of synthetic datasets

Possible approach to judge the utility of a synthesis method:

- Visualize all the one-way tables with **compare()**.
- Next visualize all two-way pMSE ratios with **utility.tables()**.
- If all the standardized pMSE ratios are below 10, or better still below 3, it is probably not necessary to do anything more as the utility seems acceptable.
- At each of the steps above you should try to improve the utility by changing the default parameters of syn(), checking whether methods used are appropriate, logical constraints are implemented and so on

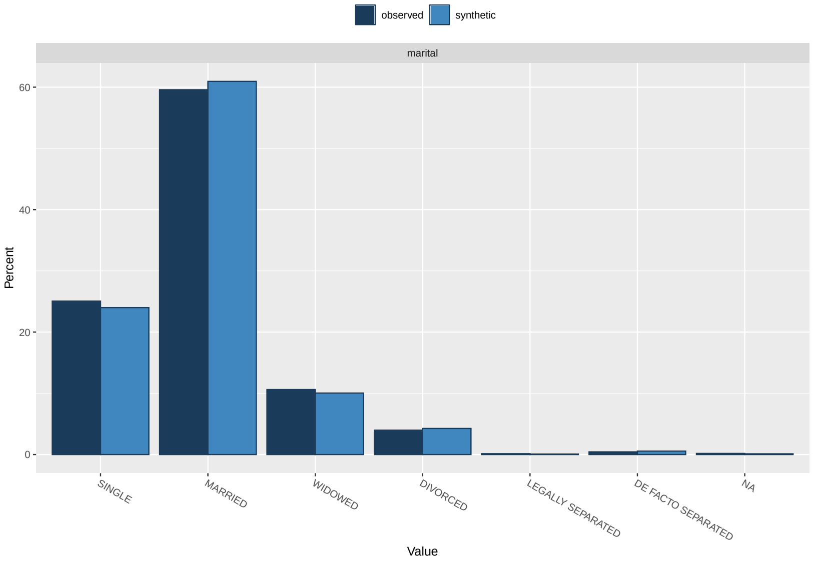
One-way marginals with compare()

```
compare_martial <- compare(data = origdat,
                           object = syn.object,
                           vars = "marital")
compare_martial$tab.utility[,"pMSE"]
```

```
## [1] 0.0001335128
```

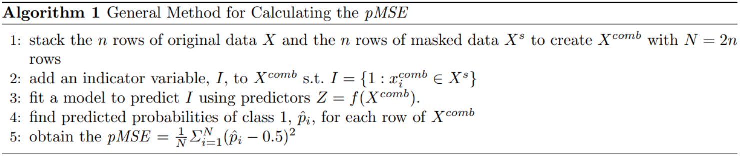
One-way marginals with compare()

```
compare_martial$plots
```



The pMSE measure

Let p_i, $i = 1, \ldots, N$ with $N = n_{org} + n_{syn}$ denote the predicted value obtained from the model for record i in the stacked dataset. The *pMSE* is calculated as $1/N \sum_N (p_i - c)^2$, with $c = n_{syn}/N$.



Algorithm 1 General Method for Calculating the *pMSE*

```
stack the $n$ rows of original data $X$ and the $n$ rows of masked data $X^s$ to create $X^{comb}$ with $N = 2n$ rows
add an indicator variable, $I$, to $X^{comb}$ s.t. $I = \{1 : x_i^{comb} \in X^s\}$
fit a model to predict $I$ using predictors $Z = f(X^{comb})$.
find predicted probabilities of class 1, $\hat{p}_i$, for each row of $X^{comb}$
obtain the $pMSE = \frac{1}{N}\Sigma_{i=1}^{N}(\hat{p}_i - 0.5)^2$
```

The pMSE measure

The smaller the *pMSE* the higher the analytical validity of the synthetic data (note that $p_i \rightarrow c$ if the model cannot discriminate between the original data and the synthetic data).

```
compare_martial$tab.utility[,"pMSE"]
```

```
## [1] 0.0001335128
```

Standardized pMSE

- Standardized *pMSE* ratio which is computed as the empirical *pMSE* divided by its expected value under the null hypothesis (no difference between syn and orig data).
  - In Synthpop called **S_pMSE**
  - Will have an expected value of 1 if the synthesizing model is correct.
  - From experience: Values below 10 are okay, below 3 are optimal.

```r
compare_martial$tab.utility[,"S_pMSE"]
```

```
## [1] 1.780171
```

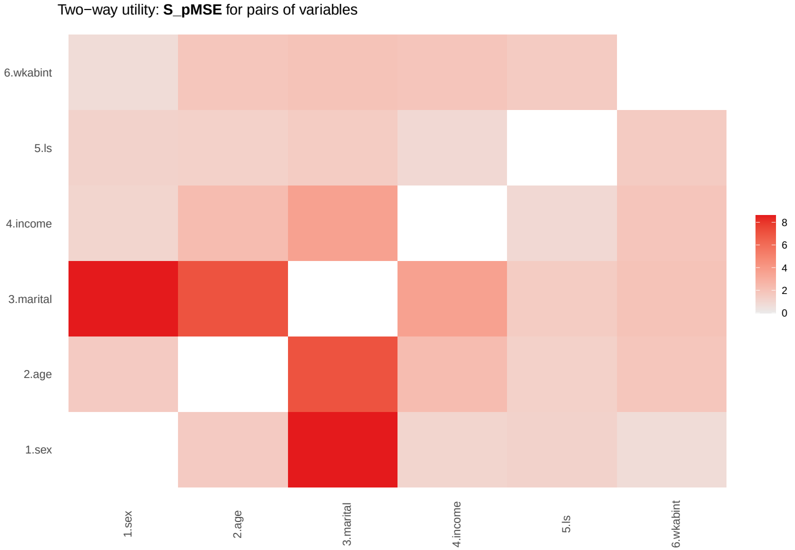
Two-way marginals

Let us take a look at the two-way utility

```
utility.twoway <- utility.tables(data = origdat,
                                 object = syn.object)
```

Results: Two-way utility

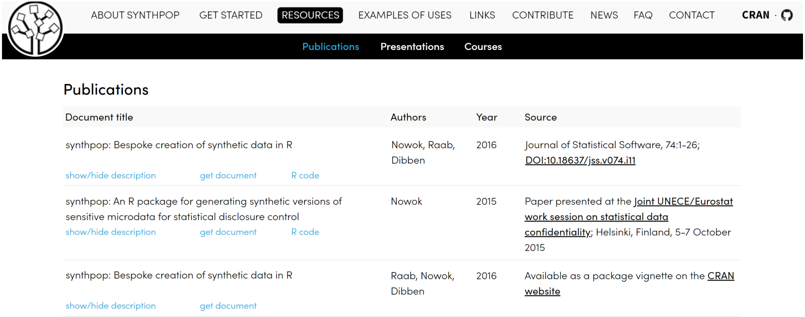
```
utility.twoway$utility.plot
```



More resources

The documentation of synthpop is excellent, check out

https://www.synthpop.org.uk/resources.html



Publications

Document title	Authors	Year	Source
synthpop: Bespoke creation of synthetic data in R show/hide description get document R code	Nowok, Raab, Dibben	2016	Journal of Statistical Software, 74:1-26; DOI:10.18637/jss.v074.i11
synthpop: An R package for generating synthetic versions of sensitive microdata for statistical disclosure control show/hide description get document R code	Nowok	2015	Paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality; Helsinki, Finland, 5-7 October 2015
synthpop: Bespoke creation of synthetic data in R show/hide description get document	Raab, Nowok, Dibben	2016	Available as a package vignette on the CRAN website

Reading 1

Reiter, J. P. (2012). Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. Public Opinion Quarterly 76(1), 163–181. https://doi.org/10.1093/poq/nfr058

- Short overview of risk assessment and utility
- Discussion with lots of examples from Official Stats and discussion of impact on data utility
    - Recoding
    - Data Swapping
    - Adding Noise
- Short discussion of partially synthetic data

Reading 2

Raghunathan, T. E. (2021). Synthetic Data Annual Review of Statistics and Its Application 8, 129-140. https://doi.org/10.1146/annurev-statistics-040720-031848

- General idea of synthetic data and background in Bayesian Statistics
- Fully and partially synthetic data
- Synthetic data and differential privacy (see also next week!)
- **Problems with synthetic datasets**

What questions do you have?