Johannes Gussenbauer,
Alexander Kowarik,
Bernhard Meindl
Statistik Austria
November, 2018

Implementation of the Cell-Key Method & Targeted Record Swapping

```r
knitr::opts_chunk$set(size="footnotesize")
```

- Cell-Key Method and Targeted Record Swapping implemented in R-Packages
- Available on https://github.com/sdcTools
    - recordSwapping
    - cellKey
- Implementations are prototype-ready

- Two different approaches implemented
    - *Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics* (Thompson, Broadfoot, Elazar)
    - Approach developed by the Federal Statistical Office of Germany (Destatis)
- `cellKey` depends on R-package `ptable`

Main Features

- Methods `abs` and `destatis`
- `ck_generate_rkeys()` for generating record keys
    - perturbation for magnitude tables only for `abs`
- main function `pertubTable()`
- allows sampling weights
- define arbitrarily complex hierarchies like in `sdcTable`
    - improved functionality in `cellKey`
- various auxiliary methods implemented

Example

```
# load package
library(cellKey,verbose=FALSE)

## Loading required package: data.table

# load dummy data
dat <- ck_create_testdata()[,c("sex","age","savings",
                               "income","sampling_weight")]
dat[,cnt_males:=ifelse(sex=="male", 1, 0)]
dat[,cnt_highincome:=ifelse(income>=9000, 1, 0)]
```

$\rightarrow$ create a perturbed table of counts of variables sex by age

```
pert_params <- ck_create_pert_params(
  bigN=17312941,
  smallN=12,
  pTable=ck_create_pTable(D=5, V=3, pTableSize=70, type="abs"),
  sTable=ck_generate_sTable(smallC=12),
  mTable=c(0.6,0.4,0.2))
```

```
inp <- ck_create_input(
  dat=dat,
  def_rkey=15*nrow(dat),
  pert_params=pert_params)
print(class(inp))

## [1] "pert_inputdat"
## attr(,"package")
## [1] "cellKey"
```

Specify Dimensions

```r
# example for variable sex
dim.sex <- data.table(levels=c("@","@@","@@"),
                      codes=c("Total", "male", "female"))
print(dim.sex)


##    levels  codes
## 1:      @  Total
## 2:     @@   male
## 3:     @@ female
```

```r
# or alternatively
dim.sex2 <- ck_create_node(total_lab="Total")
dim.sex2 <- ck_add_nodes(dim.sex2, reference_node="Total",
                         node_labs=c("male","female"))
print(dim.sex2)


##    levelName
## 1 Total
## 2  Â¦--male
## 3  Â°--female
```

```r
# example for variable age
dim.age <- data.table(levels=c("@",rep("@@", 6)),
                      codes=c("Total", paste0("age_group",1:6)))

# or alternatively
dim.age2 <- ck_create_node(total_lab="Total")
dim.age2 <- ck_add_nodes(dim.age2, reference_node="Total",
                         node_labs=paste0("age_group",1:6))
print(dim.age2)

##         levelName
## 1 Total
## 2   Â¦--age_group1
## 3   Â¦--age_group2
## 4   Â¦--age_group3
## 5   Â¦--age_group4
## 6   Â¦--age_group5
## 7   Â°--age_group6
```

Compine Hierachies

```
dimList <- list(sex=dim.sex, age=dim.age2)
print(dimList)

## $sex
##    levels   codes
## 1:      @   Total
## 2:     @@    male
## 3:     @@  female
##
## $age
##          levelName
## 1 Total
## 2  Â¦--age_group1
## 3  Â¦--age_group2
## 4  Â¦--age_group3
## 5  Â¦--age_group4
## 6  Â¦--age_group5
## 7  Â°--age_group6
```

Perturb Table

```
tab1 <- perturbTable(inp=inp, dimList=dimList,
                     countVars=c("cnt_males", "cnt_highincome"),
   weightVar="sampling_weight", numVars=NULL)
print(tab1)

## The weighted 2-dimensional table consists of 21 cells. The results are
## The dimensions are given by the following variables
## o sex
## o age
##
## Type of pTable-used: 'abs'
## The following count-variables have been tabulated/perturbed:
## o Total
## o cnt_males
## o cnt_highincome
## No numeric variables have been tabulated/perturbed in this table
```

Perturbed Table

▶ returns tables with `ck_freq_table()`

```
# count table containing
# original, perturbed and (un)weighted values
ck_freq_table(tab1, vname="cnt_males")
```

▶ compute information loss measures with `ck_cnt_measures()`

```
ck_cnt_measures(tab1, vname="Total")
```

- ▶ perturbed table of continous data

```
tab2 <- perturbTable(inp=inp,dimList=dimList,weightVar="sampling_weight",
  countVars=c("cnt_males", "cnt_highincome"),
  numVars=c("savings","income"))

p_income <- ck_cont_table(tab2, vname="income", meanBeforeSum=TRUE)
head(p_income)

##      sex           age UW_income  pUW_income   WS_income  pWS_income  pWM_inco
## 1: Total       Total 22952978 22930849.1 1363517148 1362202587      5006.7
## 2: Total age_group1  9810547  9794206.1  583047435  582076283      4966.6
## 3: Total age_group2  5692119  5679273.0  336517278  335757827      4968.7
## 4: Total age_group3  4406946  4398044.8  261341079  260813219      5090.3
## 5: Total age_group4  2133543  2133021.2  128441306  128409893      5054.5
## 6: Total age_group5   848151   857505.6   49990424   50541786      5073.9
```

▶ perturbed table for a specific group → by="cnt_highincome"

```
tab3 <- perturbTable(inp=inp, dimList=dimList,
  weightVar="sampling_weight",
  numVars=c("savings"), by="cnt_highincome")
```