

# Package ‘SmallCountRounding’

October 18, 2019

**Type** Package

**Title** Small Count Rounding of Tabular Data

**Version** 0.4.0

**Date** 2019-10-17

**Author** Øyvind Langsrud [aut, cre],  
Johan Heldal [aut]

**Maintainer** Øyvind Langsrud <oyl@ssb.no>

**Depends** Matrix, SSBtools

**Imports** methods

**Suggests** sdcHierarchies

**Description** A statistical disclosure control tool to protect frequency tables in cases where small values are sensitive. The function `PLSrounding()` performs small count rounding of necessary inner cells so that all small frequencies of cross-classifications to be published (publishable cells) are rounded. This is equivalent to changing micro data since frequencies of unique combinations are changed. Thus, additivity and consistency are guaranteed. The methodology is described in Langsrud and Heldal (2018) <<https://www.researchgate.net/publication/327768398>>.

**License** Apache License 2.0 | file LICENSE

**URL** <https://github.com/statisticsnorway/SmallCountRounding>

**BugReports** <https://github.com/statisticsnorway/SmallCountRounding/issues>

**RoxygenNote** 6.1.1

**Encoding** UTF-8

**NeedsCompilation** no

**Repository** CRAN

**Date/Publication** 2019-10-17 22:30:13 UTC

## R topics documented:

SmallCountRounding-package . . . . .	2
HD . . . . .	2

PLS2way . . . . .	3
PLSRounding . . . . .	4
print.PLSrounded . . . . .	6
RoundViaDummy . . . . .	7
SmallCountData . . . . .	9

<b>Index</b>	<b>11</b>
--------------	-----------

---

SmallCountRounding-package

*Small Count Rounding of Tabular Data*

---

### Description

A statistical disclosure control tool to protect frequency tables in cases where small values are sensitive. The main function, `PLSRounding`, performs small count rounding of necessary inner cells (Heldal, 2017) so that all small frequencies of cross-classifications to be published (publishable cells) are rounded. This is equivalent to changing micro data since frequencies of unique combinations are changed. Thus, additivity and consistency are guaranteed. This is performed by an algorithm inspired by partial least squares regression (Langsrud and Heldal, 2018).

### References

Heldal, J. (2017): “The European Census Hub 2011 Hypercubes - Norwegian SDC Experiences”. In: *Work Session on Statistical Data Confidentiality*, Skopje, The former Yugoslav Republic of Macedonia, September 20-22, 2017.

Langsrud, Ø. and Heldal, J. (2018): “An Algorithm for Small Count Rounding of Tabular Data”. Presented at: *Privacy in statistical databases*, Valencia, Spain. September 26-28, 2018. <https://www.researchgate.net/publication/327768398>

---

HD

*Hellinger Distance (Utility)*

---

### Description

Hellinger distance (HD) and a related utility measure (`HDutility`) described in the reference below. The utility measure is made to be bounded between 0 and 1.

### Usage

`HD(f, g)`

`HDutility(f, g)`

**Arguments**

f                    Vector of original counts  
g                    Vector of perturbed counts

**Details**

HD is defined as  $\sqrt{\text{sum}((\sqrt{f} - \sqrt{g})^2)/2}$  and HDutility is defined as  $1 - \text{HD}(f, g) / \sqrt{\text{sum}(f)}$ .

**Value**

Hellinger distance or related utility measure

**References**

Shlomo, N., Antal, L., & Elliot, M. (2015). Measuring Disclosure Risk and Data Utility for Flexible Table Generators, *Journal of Official Statistics*, 31(2), 305-324. doi: <https://doi.org/10.1515/jos-2015-0019>

**Examples**

```
f <- 1:6
g <- c(0, 3, 3, 3, 6, 6)
print(c(
  HD = HD(f, g),
  HDutility = HDutility(f, g),
  maxdiff = max(abs(g - f)),
  meanAbsDiff = mean(abs(g - f)),
  rootMeanSquare = sqrt(mean((g - f)^2))
))
```

---

 PLS2way

*Two-way table from PLSrounding output*


---

**Description**

Two-way table from PLSrounding output

**Usage**

```
PLS2way(obj, variable = c("rounded", "original", "difference", "code"))
```

**Arguments**

obj                    Output object from [PLSrounding](#)  
variable                One of "rounded" (default), "original", "difference" or "code".

**Details**

When parameter "variable" is "code", output is coded as "#" (publish), "." (inner) and "&" (both).

**Value**

A data frame

**Examples**

```
# Making tables from PLSrounding examples
z <- SmallCountData("e6")
a <- PLSrounding(z, "freq", formula = ~eu * year + geo)
PLS2way(a, "original")
PLS2way(a, "difference")
PLS2way(a, "code")
PLS2way(PLSrounding(z, "freq", formula = ~eu * year + geo * year), "code")
eHrc2 <- list(geo = c("EU", "@Portugal", "@Spain", "Iceland"), year = c("2018", "2019"))
PLS2way(PLSrounding(z, "freq", hierarchies = eHrc2))
```

---

PLSRounding

*PLS inspired rounding*

---

**Description**

Small count rounding of necessary inner cells are performed so that all small frequencies of cross-classifications to be published (publishable cells) are rounded. The publishable cells can be defined from a model formula, hierarchies or automatically from data.

**Usage**

```
PLSRounding(data, freqVar, roundBase = 3, hierarchies = NULL,
  formula = NULL, maxRound = roundBase - 1, ...)
```

**Arguments**

data	Input data as a data frame (inner cells)
freqVar	Variable holding counts (inner cells frequencies)
roundBase	Rounding base
hierarchies	List of hierarchies
formula	Model formula defining publishable cells
maxRound	Inner cells contributing to original publishable cells equal to or less than maxRound will be rounded
...	Further parameters sent to RoundViaDummy

## Details

This function is a user-friendly wrapper for `RoundViaDummy` with data frame output and with computed summary of the results. See [RoundViaDummy](#) for more details.

## Value

Output is a four-element list with class attribute "PLSRounded" (to ensure informative printing).

<code>inner</code>	Data frame corresponding to input data with the main dimensional variables and with cell frequencies (original, rounded, difference).
<code>publish</code>	Data frame of publishable data with the main dimensional variables and with cell frequencies (original, rounded, difference).
<code>metrics</code>	A named character vector of various statistics calculated from the two output data frames (" <code>inner_</code> " used to distinguish). See examples below and the function <a href="#">HDutility</a> .
<code>freqTable</code>	Matrix of frequencies of cell frequencies and absolute differences. For example, row "rounded" and column "pub.4+" is the number of rounded inner cell frequencies greater than or equal to 4.

## References

Langsrud, Ø. and Heldal, J. (2018): "An Algorithm for Small Count Rounding of Tabular Data". Presented at: *Privacy in statistical databases*, Valencia, Spain. September 26-28, 2018. <https://www.researchgate.net/publication/327768398>

## See Also

[RoundViaDummy](#), [PLS2way](#)

## Examples

```
# Small example data set
z <- SmallCountData("e6")
print(z)

# Publishable cells by formula interface
a <- PLSRounding(z, "freq", roundBase = 5, formula = ~geo + eu + year)
print(a)
print(a$inner)
print(a$publish)
print(a$metrics)
print(a$freqTable)

# Recalculation of maxdiff, HDutility, meanAbsDiff and rootMeanSquare
max(abs(a$publish[, "difference"]))
HDutility(a$publish[, "original"], a$publish[, "rounded"])
mean(abs(a$publish[, "difference"]))
sqrt(mean((a$publish[, "difference"])^2))

# Four lines below produce equivalent results
```

```

# Ordering of rows can be different
PLSrounding(z, "freq")
PLSrounding(z, "freq", formula = ~eu * year + geo * year)
PLSrounding(z[, -2], "freq", hierarchies = SmallCountData("eHrc"))
PLSrounding(z[, -2], "freq", hierarchies = SmallCountData("eDimList"))

# Define publishable cells differently by making use of formula interface
PLSrounding(z, "freq", formula = ~eu * year + geo)

# Define publishable cells differently by making use of hierarchy interface
eHrc2 <- list(geo = c("EU", "@Portugal", "@Spain", "Iceland"), year = c("2018", "2019"))
PLSrounding(z, "freq", hierarchies = eHrc2)

# Package sdchHierarchies can be used to create hierarchies.
# The small example code below works if this package is available.
if (require(sdchHierarchies)) {
  z2 <- cbind(geo = c("11", "21", "22"), z[, 3:4], stringsAsFactors = FALSE)
  h2 <- list(
    geo = hier_compute(inp = unique(z2$geo), dim_spec = c(1, 1), root = "Tot", as = "df"),
    year = hier_convert(hier_create(root = "Total", nodes = c("2018", "2019")), as = "df")
  )
  PLSrounding(z2, "freq", hierarchies = h2)
}

# Use PLS2way to produce tables as in Langsrud and Heldal (2018)
# and to demonstrate parameters maxRound,
# zeroCandidates and identifyNew (see RoundViaDummy)
exPSD <- SmallCountData("exPSD")
set.seed(12345) # To guarantee same output as in reference/comments
a <- PLSrounding(exPSD, "freq", 5, formula = ~rows + cols)
PLS2way(a, "original") # Table 1
PLS2way(a) # Table 2
set.seed(12345)
a <- PLSrounding(exPSD, "freq", 5, formula = ~rows + cols, identifyNew = FALSE)
PLS2way(a) # Table 3
set.seed(12345)
a <- PLSrounding(exPSD, "freq", 5, formula = ~rows + cols, maxRound = 7)
PLS2way(a) # Values in col1 rounded
set.seed(12345)
a <- PLSrounding(exPSD, "freq", 5, formula = ~rows + cols, zeroCandidates = TRUE)
PLS2way(a) # (row3, col4): original is 0 and rounded is 5

```

---

```
print.PLSrounded
```

```
Print method for PLSrounded
```

---

## Description

Print method for PLSrounded

**Usage**

```
## S3 method for class 'PLSrounded'
print(x, digits = max(getOption("digits") - 3, 3),
      ...)
```

**Arguments**

x	PLSrounded object
digits	positive integer. Minimum number of significant digits to be used for printing most numbers.
...	further arguments sent to the underlying

**Value**

Invisibly returns the original object.

---

 RoundViaDummy

*Small Count Rounding of Tabular Data*


---

**Description**

Small count rounding via a dummy matrix and by an algorithm inspired by PLS

**Usage**

```
RoundViaDummy(data, freqVar, formula = NULL, roundBase = 3,
              singleRandom = FALSE, crossTable = TRUE, total = "Total",
              maxIterRows = 1000, maxIter = 1e+07, x = NULL,
              hierarchies = NULL, xReturn = FALSE, maxRound = roundBase - 1,
              zeroCandidates = FALSE, forceInner = FALSE, identifyNew = TRUE,
              step = 0, ...)
```

**Arguments**

data	Input data as a data frame (inner cells)
freqVar	Variable holding counts (name or number)
formula	Model formula defining publishable cells. Will be used to calculate x (via <a href="#">ModelMatrix</a> ). When NULL, x must be supplied.
roundBase	Rounding base
singleRandom	Single random draw when TRUE (instead of algorithm)
crossTable	When TRUE, cross table in output and caculations via FormulaSums()
total	String used to name totals
maxIterRows	See details
maxIter	Maximum number of iterations

x	Dummy matrix defining publishable cells
hierarchies	List of hierarchies, which can be converted by <a href="#">AutoHierarchies</a> . Thus, a single string as hierarchy input is assumed to be a total code. Exceptions are "rowFactor" or "", which correspond to only using the categories in the data.
xReturn	Dummy matrix in output when TRUE (as input parameter x)
maxRound	Inner cells contributing to original publishable cells equal to or less than maxRound will be rounded.
zeroCandidates	When TRUE, inner cells in input with zero count (and multiple of roundBase when maxRound is in use) contributing to publishable cells will be included as candidates to obtain roundBase value. With vector input, the rule is specified individually for each cell.
forceInner	When TRUE, all inner cells will be rounded. Use vector input to force individual cells to be rounded. Can be combined with parameter zeroCandidates to allow zeros and roundBase multiples to be rounded up.
identifyNew	When TRUE, new cells may be identified after initial rounding to ensure that no nonzero rounded publishable cells are less than roundBase.
step	When step>1, the original forward part of the algorithm is replaced by a kind of stepwise. After step steps forward, backward steps may be performed. The step parameter is also used for backward-forward iteration at the end of the algorithm; step backward steps may be performed.
...	Further parameters sent to <a href="#">Hierarchies2ModelMatrix</a>

### Details

Small count rounding of necessary inner cells are performed so that all small frequencies of cross-classifications to be published (publishable cells) are rounded. This is equivalent to changing micro data since frequencies of unique combinations are changed. Thus, additivity and consistency are guaranteed. The matrix multiplication formula is:  $y_{Publish} = t(x) \%* \% y_{Inner}$ , where x is the dummy matrix.

### Value

A list where the two first elements are two column matrices. The first matrix consists of inner cells and the second of cells to be published. In each matrix the first and the second column contains, respectively, original and rounded values. By default the cross table is the third element of the output list.

### Note

Iterations are needed since after initial rounding of identified cells, new cells are identified. If cases of a high number of identified cells the algorithm can be too memory consuming (unless singleRandom=TRUE). To avoid problems, not more than maxIterRows cells are rounded in each iteration. The iteration limit (maxIter) is by default set to be high since a low number of maxIterRows may need a high number of iterations.

### See Also

See the user-friendly wrapper [PLSrounding](#) and see Round2 for rounding by other algorithm



**Examples**

```
# See similar and related examples in PLSrounding documentation
RoundViaDummy(SmallCountData("e6"), "freq")
RoundViaDummy(SmallCountData("e6"), "freq", formula = ~eu * year + geo)
RoundViaDummy(SmallCountData("e6"), "freq", hierarchies =
  list(geo = c("EU", "@Portugal", "@Spain", "Iceland"), year = c("2018", "2019")))

RoundViaDummy(SmallCountData('z2'),
  'ant', ~region + hovedint + fylke*hovedint + kostragr*hovedint, 10)
mf <- ~region*mnd + hovedint*mnd + fylke*hovedint*mnd + kostragr*hovedint*mnd
a <- RoundViaDummy(SmallCountData('z3'), 'ant', mf, 5)
b <- RoundViaDummy(SmallCountData('sosialFiktiv'), 'ant', mf, 4)
print(cor(b[[2]]), digits=12) # Correlation between original and rounded

## Not run:
# Demonstrate parameters maxRound, zeroCandidates and forceInner
# by tabulating the inner cells that have been changed.
z4 <- SmallCountData("sosialFiktiv")
for (forceInner in c("FALSE", "z4$ant < 10"))
  for (zeroCandidates in c(FALSE, TRUE))
    for (maxRound in c(2, 5)) {
      set.seed(123)
      a <- RoundViaDummy(z4, "ant", formula = mf, maxRound = maxRound,
        zeroCandidates = zeroCandidates,
        forceInner = eval(parse(text = forceInner)))
      change <- a$yInner[, "original"] != a$yInner[, "rounded"]
      cat("\n\n-----\n")
      cat("      maxRound:", maxRound, "\n")
      cat("zeroCandidates:", zeroCandidates, "\n")
      cat("      forceInner:", forceInner, "\n\n")
      print(table(original = a$yInner[change, "original"], rounded = a$yInner[change, "rounded"]))
      cat("-----\n")
    }

## End(Not run)
```

---

SmallCountData

*Function that returns a dataset*


---

**Description**

Function that returns a dataset

**Usage**

```
SmallCountData(dataset, path = NULL)
```

**Arguments**

dataset            Name of data set within the SmallCountRounding package  
path                When non-NULL the data set is read from "path/dataset.RData"

**Value**

The dataset

**Note**

Except for "europe6", "eHrc", "eDimList" and "exPSD", the function returns the same datasets as is included in the package easySdcTable.

**See Also**

[SSBtoolsData](#), [Hrc2DimList](#)

**Examples**

```
SmallCountData("z1")  
SmallCountData("e6")  
SmallCountData("eHrc")        # TauArgus coded hierarchies  
SmallCountData("eDimList")   # sdcTable coded hierarchies  
SmallCountData("exPSD")      # Example data in presentation at Privacy in statistical databases
```

# Index

## \*Topic **print**

print.PLSrounded, 6

AutoHierarchies, 8

HD, 2

HDutility, 5

HDutility (HD), 2

Hierarchies2ModelMatrix, 8

Hrc2DimList, 10

ModelMatrix, 7

PLS2way, 3, 5

PLSrounding, 2, 3, 4, 8

print.PLSrounded, 6

RoundViaDummy, 5, 7

SmallCountData, 9

SmallCountRounding

(SmallCountRounding-package), 2

SmallCountRounding-package, 2

SSBtoolsData, 10