

# Package ‘SDCNway’

September 22, 2020

**Version** 1.0.0

**Date** 2020-08-19

**Title** Tools to Evaluate Disclosure Risk

**Maintainer** John Riddles <JohnRiddles@westat.com>

**Depends** R (>= 3.6.0), Rdpack

**Imports** methods, plyr (>= 1.8.5), dplyr (>= 0.8.4), ggplot2 (>= 3.2.1), MASS (>= 3.6.0)

**RdMacros** Rdpack

**Description** Tools for calculating disclosure risk measures, including record-level measures primarily using exhaustive tabulation, as well as file-level measures using a loglinear model. Funded by the National Center for Education Statistics. See El Emam (2011) <doi:10.1186/gm239> for a description of risk measures.

**Note** This publication was prepared for NCES under Contract No. ED-IES-12-D-0009/0005 with Sanametrix and Westat. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

**License** GPL-2

**Encoding** UTF-8

**LazyData** true

**ByteCompile** true

**NeedsCompilation** no

**RoxygenNote** 7.1.0

**Author** John Riddles [aut, cre],  
Westat [cph]

**Repository** CRAN

**Date/Publication** 2020-09-22 08:40:07 UTC

R topics documented:

exampledata . . . . .	2
sdc_extabs . . . . .	2
sdc_loglinear . . . . .	5

<b>Index</b>	<b>8</b>
--------------	----------

---

exampledata	<i>A subset of the 1992 National Adult Literacy Study (NALS) prison study public-use microdata file.</i>
-------------	--

---

**Description**

A subset of the 1992 National Adult Literacy Study (NALS) prison study public-use microdata file. It has 20 variables and 182 records.

**Usage**

data(exampledata)

**Format**

An object of class "data.frame";

---

sdc_extabs	<i>Calculate risk measures through exhaustive tabulations, Mu-Argus, and other methods.</i>
------------	---

---

**Description**

This function primarily uses the exhaustive tabulation method to quantify disclosure risk. It tabulates cell counts for different combinations of variables provided by the user. Using these counts, this function identifies variable categories and records which are considered high risk for disclosure. File-level re-identification risk measures are also provided, e.g., Mu-Argus (Polettini 2003) and the risk metrics promosed in El Emam (2011).

**Usage**

```
sdc_extabs(  
  data,  
  ID = NULL,  
  weight = NULL,  
  varpool = names(data),  
  forcelist = character(0),  
  forcenum = 1,  
  missingdef = list(),
```

```

    mindim = 1,
    maxdim = 2,
    threshold = NULL,
    wgtthreshold = NULL,
    condition = NULL,
    output_filename = NULL,
    tau1 = 0.2,
    tau2 = 0.2,
    include_mu_argus = TRUE
)

## S3 method for class 'sdc_extabs'
print(x, cutoff = 50, summary_outfile = NULL, ...)

## S3 method for class 'sdc_extabs'
plot(x, plotpath = NULL, plotvar1 = character(0), plotvar2 = character(0), ...)

```

### Arguments

data	Data frame containing the data for which we are to measure disclosure risk. Unexpected behavior may result if any column name begins with a period.
ID	Name of column which identifies records. If NULL (default), an ID column named <code>.ROW_NUMBER</code> is created and used in reports.
weight	Column name for sampling weights. NULL or empty if none.
varpool	Vector of column names over which to form tables.
forcelist	Vector of variable names. Some are included in all tabulations. Optional.
forcenum	Number of variables in <code>forcelist</code> that are mandatory for all tabulations. That is, all tabulations will have a number of variables from <code>forcelist</code> exactly equal to <code>forcenum</code> .
missingdef	A named list specifying missing values. The names correspond to column names in data.
mindim	Integer specifying the minimum number of <code>varpool</code> variables (including <code>forcelist</code> variables) that can be used to form tables.
maxdim	Integer specifying the maximum number of <code>varpool</code> variables (including <code>code-forcelist</code> variables) that can be used to form tables.
threshold	Threshold to determine the number of violations in terms of cell counts. If the number of cases in a cell is less than <code>threshold</code> , the cell is flagged as a violation. If <code>threshold</code> is NULL and <code>wgthreshold</code> is not NULL, then only a weighted threshold will be used. If both are NULL, <code>threshold</code> will be set to 3 and the weighted threshold will not be used.
wgtthreshold	Threshold to determine violations in terms of weighted cell counts. If NULL, a weighted threshold will not be used.
condition	Character string describing how weighted and unweighted thresholds are combined when both are used. If used, it must be "and" or "or" (case insensitive). This parameter is ignored if <code>weight</code> is NULL.

output_filename	Name of the csv file to save the data set with violation counts and Mu-Argus scores attached. NULL if no output file is to be saved.
tau1	A threshold to compute the risk measure, pRa. See User Manual for more details.
tau2	A threshold to compute the risk measure, jRa. This parameter is ignored if weight is NULL. See User Manual for more details.
include_mu_argus	Flag indicating whether Mu-Argus and El-Emam metrics should be calculated.
x	An object of class sdc_extabs, as returned by the <a href="#">sdc_extabs</a> function.
cutoff	The number of variable categories with the highest percentage of cell violations for each table dimension. Default is 50.
summary_outfile	Name of summary output .txt file. If not NULL, console output is copied to the file. Default is NULL (no logging of output). Errors and warnings are not diverted (consider running in batch mode if logging of errors and warnings is needed).
...	Currently unused. For NextMethod compatibility.
plotpath	Directory to save plots. Plots are saved as <i>jpeg</i> files (quality = 100%). If the directory does not exist, it is first created. If plotpath is NULL (default), plots are not saved.
plotvar1	A vector of names of discrete variables for boxplots. If none, boxplots are not produced.
plotvar2	A vector of names of continuous variables for scatterplots. If none, scatterplots are not produced.

## Details

If a specified missing value contains only whitespace, it will match any element with only whitespace. NA values in data are treated as missing regardless of missingdef. If you do not want NA values to be treated as missing, please recode them before passing the data to this function.

Note that if a weight variable is not provided, the number of statistics and plots that are produced is significantly reduced.

## Value

An object of type sdc\_extabs. Internally, a named list of statistics.

**tabulation** Cell counts and violation flags. Represented as a list with each element corresponding to a varpool combination.

**data\_with\_statistics** The original data with new columns showing statistics such as violation counts and Mu-Argus score for each record.

**recoded\_data\_with\_statistics** Same as data\_with\_statistics but with missing value recodes.

**mu\_argus\_summary** Summary table of Mu-Argus by cell count. For this summary, all variables in varpool are used to define a cell. If weight is NULL, then this summary is omitted.

***el\_emam\_measures*** List of file-level re-identification risk measures.

***percent\_violations\_by\_var\_and\_level*** Table with percent of records that are in violation for each variable/category.

***percent\_violations\_by\_dim\_var\_and\_level*** Table with percent of cells that are in violation for each dimension/variable/category.

***options*** Options provided to sdc\_extabs by the user, such as missingdef, mindim, etc.

## Methods (by generic)

- **print:** S3 print method for sdc\_extabs objects  
Prints a nicely formatted version of the percent record violations by variable/category and percent cell violations by dimension/variable/category
- **plot:** S3 plot method for sdc\_extabs objects  
Produces boxplots and scatterplots of violation counts and mu-argus scores.

## References

El Emam K (2011). “Methods for the de-identification of electronic health records for genomic research.” *Genome medicine*, **3**(4), 25. doi: [10.1186/gm239](https://doi.org/10.1186/gm239), <https://doi.org/10.1186/gm239>.

Polettini S (2003). “Some remarks on the individual risk methodology.” *Joint ECE/EUROSTAT Work Session on Data Confidentiality, Luxembourg*. <https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2003/04/confidentiality/wp.18.s.e.pdf>.

## Examples

```
data(exampladata)
vars <- c("BIB1201", "BIC0501", "BID0101", "BIE0601", "BORNUSA", "CENREG",
          "DAGE3", "DRACE3", "EDUC3", "GENDER")
results <- sdc_extabs(exampladata,
                      ID="CASEID",
                      weight="WEIGHT",
                      varpool=vars,
                      mindim=2,
                      maxdim=3,
                      missingdef=list(BIE0601=5),
                      wgtthreshold=3000,
                      condition="or")
print(results, cutoff=15)
plot(results, plotvar1="BORNUSA", plotvar2="WEIGHT")
```

---

sdc\_loglinear

---

*Calculates file-level risk measures using a loglinear model.*


---

## Description

Calculates file-level risk measures using a loglinear model.

**Usage**

```

sdc_loglinear(
  data,
  weight,
  varpool,
  degree = 2,
  numiter = 40,
  epsilon = 0.001,
  blanks_as_missing = TRUE,
  output_filename = NULL
)

## S3 method for class 'sdc_loglinear'
print(x, summary_outfile = NULL, ...)

## S3 method for class 'sdc_loglinear'
plot(x, plotpath = NULL, plotvar1 = character(0), plotvar2 = character(0), ...)

```

**Arguments**

<code>data</code>	Data frame containing the data to be evaluated.
<code>weight</code>	Column name for sampling weights.
<code>varpool</code>	Vector of column names to be used in model.
<code>degree</code>	Highest degree of interaction terms to be used in the model.
<code>numiter</code>	Maximum number of iterations to run iterative proportional fitting for the log-linear model.
<code>epsilon</code>	Maximum deviation allowed between observed and fitted margins.
<code>blanks_as_missing</code>	If TRUE, character and factor variables that are blank or pure whitespace are treated as missing values.
<code>output_filename</code>	Name of the csv file to save the data set with record-level risk measures, <code>.tau1_rec</code> and <code>.tau2_rec</code> , attached. NULL if no output file is to be saved.
<code>x</code>	Object of class <code>sdc_loglinear</code> , as returned by <code>sdc_loglinear</code> .
<code>summary_outfile</code>	Name of summary output .txt file. If not NULL, console output is copied to the file. Default is NULL (no logging of output). Errors and warnings are not diverted (consider running in batch mode if logging is needed).
<code>...</code>	Currently unused. For NextMethod compatibility.
<code>plotpath</code>	Directory to save plots. Plots are saved as <i>jpeg</i> files (quality = 100%). If the directory does not exist, it is first created. If <code>plotpath</code> is NULL (default), plots are not saved.
<code>plotvar1</code>	A vector of names of discrete variables for boxplots. If none, boxplots are not produced.
<code>plotvar2</code>	A vector of names of continuous variables for scatterplots. If none, scatterplots are not produced.

**Details**

The data should not contain any missing values among varpool variables or the weight variable.

**Value**

An object of type `sdc_loglinear` containing calculated risk measures.

**Methods (by generic)**

- `print`: S3 print method for `sdc_loglinear` objects  
Prints tables of file-level reidentification risk measures.
- `plot`: S3 plot method for `sdc_loglinear` objects  
Produces boxplots and scatterplots of record-level risk measures, `tau1` and `tau2`, of the degree specified in the original call to `sdc_loglinear`.

**Examples**

```
data(exampladata)
vars <- c("BORNUSA", "CENREG", "DAGE3", "DRACE3", "EDUC3", "GENDER")
wgt <- "WEIGHT"

results <- sdc_loglinear(exampladata, wgt, vars, degree=3)
print(results)
plot(results, plotvar1="BORNUSA", plotvar2="WEIGHT")
```

# Index

## \* **datasets**

[exampledata](#), [2](#)

[exampledata](#), [2](#)

`plot.sdc_extabs (sdc_extabs)`, [2](#)

`plot.sdc_loglinear (sdc_loglinear)`, [5](#)

`print.sdc_extabs (sdc_extabs)`, [2](#)

`print.sdc_loglinear (sdc_loglinear)`, [5](#)

`sdc_extabs`, [2](#), [4](#)

`sdc_loglinear`, [5](#)