

sdCLog

**Utilities for statistical disclosure control in the
context of research data centers**

Matthias Gomolka

Deutsche Bundesbank, Research Data and Service Center

What do I do?

I work at the Bundesbank's Research Data Center. We enable researchers to work with confidential microdata.

At our premises we supply formally anonymized data to researchers.

Before researchers receive their research results, we need to check if the results contain any confidential information.

This is where **sdCLog** comes into play.

A tiny bit of theory

Two simple rules:

1. Each result must be based on at least 5 different entities.
2. The largest two entities must account for less than 85% of a result (dominance).

An example // `sdcdescriptives()`

```
head(DT)
#>      id sector year      val_1      val_2
#> 1:   A     S1 2019         NA 9.477642
#> 2:   A     S1 2020 94.174449 5.856641
#> 3:   B     S1 2019  4.349115 3.697140
#> 4:   B     S1 2020  2.589011 6.796527
#> 5:   C     S1 2019  6.155680 7.213390
#> 6:   C     S1 2020  7.183206 5.948330
```

```
# results wanted
DT[, .(mean = mean(val_1, na.rm = TRUE)), by = "sector"]
#>      sector      mean
#> 1:      S1 15.42511
#> 2:      S2 24.43726
```

```
# show that this result is fine
sdcdescriptives(DT, id_var = "id", val_var = "val_1", by = "sector")
#> [ OPTIONS:  sdc.n_ids: 5 | sdc.n_ids_dominance: 2 | sdc.share_dominance: 0.85 ]
#> [ SETTINGS: id_var: id | val_var: val_1 | by: sector | zero_as_NA: FALSE ]
#> Output complies to RDC rules.
```

An example // sdc_descriptives() (cont.)

```
sdc_descriptives(DT, id_var = "id", val_var = "val_1", by = c("sector", "year"))
#> Warning: Potential disclosure problem: Not enough distinct entities.
#> Warning: Potential disclosure problem: Dominant entities.
#> [ OPTIONS:  sdc.n_ids: 5 | sdc.n_ids_dominance: 2 | sdc.share_dominance: 0.85 ]
#> [ SETTINGS: id_var: id | val_var: val_1 | by: sector, year | zero_as_NA: FALSE ]
#> Not enough distinct entities:
#>   sector year distinct_ids
#> 1:     S1 2019           4
#> 2:     S1 2020           5
#> 3:     S2 2019           5
#> 4:     S2 2020           5
#> Dominant entities:
#>   sector year value_share
#> 1:     S2 2020   0.9056314
#> 2:     S1 2020   0.8776852
#> 3:     S1 2019   0.6815011
#> 4:     S2 2019   0.5506965
```

Another example // sdc_model()

```
# results wanted  
model_1 <- lm(y ~ x_1 + x_2, data = DT)
```

```
# show that this result is fine  
sdc_model(data = DT, model = model_1, id_var = "id")  
#> [ OPTIONS:  sdc.n_ids: 5 | sdc.n_ids_dominance: 2 | sdc.share_dominance: 0.85 ]  
#> [ SETTINGS: id_var: id ]  
#> Output complies to RDC rules.
```

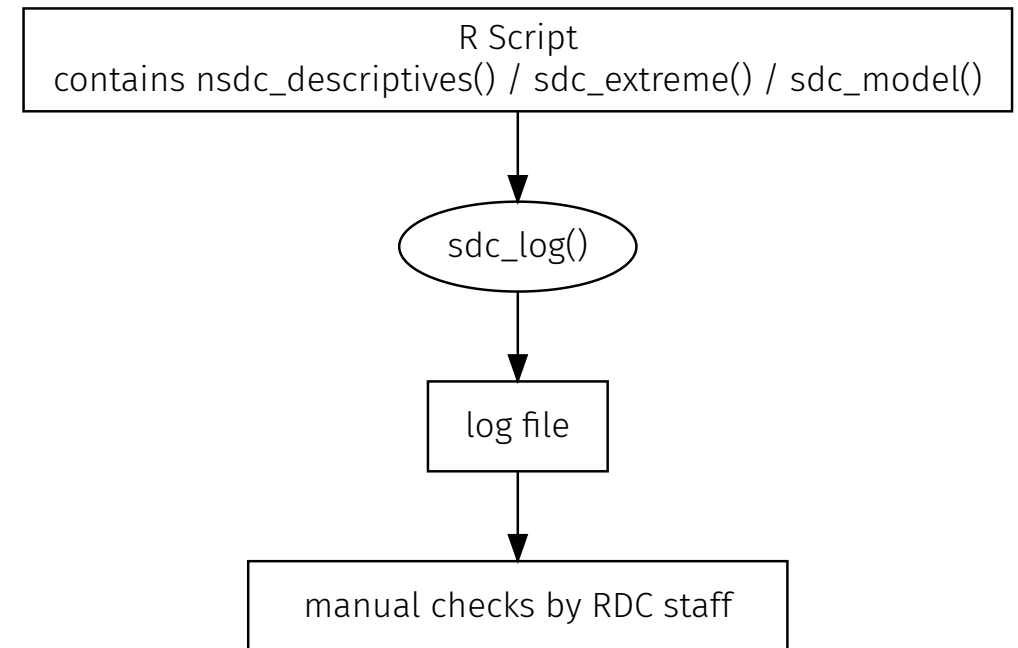
Another example // sdc_model() (cont.)

```
model_2 <- lm(y ~ x_1 + x_2 + dummy_3, data = DT)
sdc_model(data = DT, model = model_2, id_var = "id")
#> Warning: Potential disclosure problem: Not enough distinct entities.
#> [ OPTIONS:  sdc.n_ids: 5 | sdc.n_ids_dominance: 2 | sdc.share_dominance: 0.85 ]
#> [ SETTINGS: id_var: id ]
#> $dummy_3
#> Not enough distinct entities:
#>      dummy_3 distinct_ids
#> 1:      FR           4
#> 2:      BE          10
#> 3:      DE          10
#> 4:      ES          10
```

Why is it called sdcLog?

At our Research Data Center (and any other I know), researchers need to provide log files which document what they have done.

Based on these log files, Research Data Centers check if the output can be released.



How to install?

```
# CRAN  
install.packages("sdcLog")  
  
# GitHub  
remotes::install_github("https://github.com/matthiasgomolka/sdcLog")
```

About me



@matthiasgomolka



@matthiasgomolka



<https://github.com/matthiasgomolka/sdcLog>