Advanced Survey Statistics: Disclosure Control

Part 4: Disclosure Risk

Matthias Templ

Institut für Datenanalyse und Prozessdesign
School of Engineering
Zürcher Hochschule für Angewandte Wissenschaften

FU-Berlin, 2019

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering
IDP Institut für Datenanalyse und Prozessdesign

Disclosure Risk

- most difficult part in SDC
- level: depends also who is the addressor (researcher, public, ...)
- to calculate the risk is easy for registers: all key variables are fully observed.
- It can be difficult for complex surveys
  - first step: disclosure scenario, i.e. defining the key variables. Which information is available in pubilc data bases or data bases which can be accessed through registering/payment (GfK data, ...) **to match with data to be delivered** to third parties. The decision on key variables is rather subjective (more on this in part 08)
  - second step: applying risk measures

First we discuss the methods, then we continue to have a closer look at disclosure scenarios.

Overview

- disclosure risk $r_i, i = 1, ..., n$ (or $i = 1, ..., N$), $r_i \geq 0$ for individuals
- global risk R. E.g. $R = \sum_{i=1}^{n} r_i$
- risk based on categorical or continuous variables

Overview

- disclosure risk $r_i, i = 1, ..., n$ (or $i = 1, ..., N$), $r_i \geq 0$ for individuals
- global risk R. E.g. $R = \sum_{i=1}^{n} r_i$
- risk based on categorical or continuous variables

Ignoring the/a sample design

- concept of uniqueness, k-anonymity
- l-diversity
- uniqueness on subsets (SUDA)

Overview

- disclosure risk $r_i, i = 1, ..., n$ (or $i = 1, ..., N$), $r_i \geq 0$ for individuals
- global risk R. E.g. $R = \sum_{i=1}^{n} r_i$
- risk based on categorical or continuous variables

Ignoring the/a sample design

- concept of uniqueness, k-anonymity
- l-diversity
- uniqueness on subsets (SUDA)

Risk from data with a complex sample designs

- individual risk approach
- through log-linear models

Notation

- sample $\mathbf{S}$ with n observations $\rightarrow$ sample frequency counts
- a finite population $\mathbf{U}$ with N observations $\rightarrow$ population frequency counts

$\mathbf{Z}^{N \times q} \in \mathbf{U}^{N \times p}$ ... categorical key variables

- $C_j = |\mathbf{Z_j}|$ ... number of categories of categorical key variable $\mathbf{Z_j}$.
- All combinations of categories in the key variables can be calculated by cross tabulation, this defines the **keys**.
- Frequency counts $\approx$ contingency table
- Each combination of values (=keys) defines a cell in the table.

Keys, Example

Example:

- two categorical key variables, one holding the information of gender, $\mathbf{Z}_1$, and the eye-color (blue, brown, green), $\mathbf{Z}_2$.
- $C_1 = |\mathbf{Z}_1| = 2$
- $C_2 = |\mathbf{Z}_2| = 3$
- $C = \prod_{i=1}^{q} C_i = 6$, with q the number of categorical key variables.
- This 6 keys are obviously: {„man“, „blue“}, {„man“, „brown“}, {„man“, „green“}, {„woman“, „blue“}, {„woman“, „brown“} and , {„woman“, „green“}
- given a sample, the frequency counts of these keys can be easily calculated, just by cross tabulation of these two variables.

Frequency counts

- unweighted (e.g. function `table()`)
- weighted (e.g. function `simPop::tableWt()`)
- with missing values in key variables (e.g. function `sdcMicro::freqCalc()`)

Frequencies, formal notation

- The subpopulation $\mathbf{U}_j \subseteq \mathbf{U}$ or subsample $\mathbf{S}_j \subseteq \mathbf{S}$ contains all observations belonging to the j-th key, with $j \in \{1, 2, ..., C\}$.
  - To give an example: if there are exactly five observations in a subpopulation $\mathbf{U}_j$ with the key: *woman*, *student*, *blue*. This key yields subpopulation $\mathbf{U}_{woman,student,blue} \subseteq \mathbf{U}$ with $|\mathbf{U}_{woman,student,blue}| = 5$.
- The population frequency counts F_j with $j \in \{1, ..., C\}$ are the numbers of observations belonging to subpopulation $\mathbf{U}_j$, i.e. $F_j = |\mathbf{U}_j|$.

Frequencies, formal notation

- Consider a random sample $\mathbf{S} \subseteq \mathbf{U}$ of size $n \leq N$ drawn from a finite population $\mathbf{U}$ of size N. Let π_j with $j \in \{1, 2, ..., N\}$ be the inclusion probabilities, which is the probability that a record $\mathbf{x}_j \in \mathbf{U}$ is chosen in the sample.
- The sample frequency counts are analogously defined as the population frequency counts F_j and denoted by f_j. Thus the sample frequency counts f_j with $j \in \{1, ..., C\}$ are the numbers of records belonging to subsample $\mathbf{S}_j$, i.e. $f_j = |\mathbf{S}_j|$.

Let's apply all this on two key variables in the eusilc data set expemlarely ...

```
library(laeken)
data(eusilc)
# ?eusilc
# str(eusilc)
# View(eusilc)
```

Frequency counts, unweighted

- usually applied to check k-anonymity on sample/population data with no missing values

```
table(eusilc$db040, eusilc$pb220a)
```

```
##
##                    AT    EU Other
##    Burgenland     453    16     7
##    Carinthia      842    19    26
##    Lower Austria 2207    26   107
##    Salzburg       655    25    83
##    Styria        1783    24    73
##    Tyrol          928    43    50
##    Upper Austria 2056    45   143
##    Vienna        1641    76   221
##    Vorarlberg     508     9    41
```

Frequency counts, unweighted

The same caluculations with other packages. . .

- As you probably know, package data.table is up to 100 times faster than base R!

```
library(data.table)
dt <- data.table(eusilc)
```

```
dt[, .N, by = list(db040, pb220a)]
```

- Alternative: package dplyr

```
library(dplyr)
eusilc %>%
  group_by(db040, pb220a) %>%
  summarize(count = n())
```

Frequency counts, weighted

- tableWt() can be applied to estimate population frequency counts without missing values

```
library(simPop)
tableWt(eusilc[, c("db040","pb220a")],
                  weights = eusilc$rb050)
```

```
##                   pb220a
## db040                 AT      EU   Other
##   Burgenland      215460    7824    3490
##   Carinthia       446052   10114   13276
##   Lower Austria  1229191   14326   57734
##   Salzburg        383897   15182   47265
##   Styria          920797   12623   36516
##   Tyrol           496769   23720   33034
##   Upper Austria  1051590   23206   72216
##   Vienna         1153368   52737  146336
##   Vorarlberg      265003    4694   20845
```

Frequency counts with missing information

Consider three categorical key variables and their (unweighted) frequencies

```
dt[, .N, by = list(db040, hsize, pb220a)]
```

```
##             db040 hsize pb220a   N
##   1:        Tyrol     3     AT 179
##   2:        Tyrol     3  Other   4
##   3:        Tyrol     3   <NA>  43
##   4:        Tyrol     4     AT 247
##   5:        Tyrol     4   <NA> 140
##  ---
## 221: Vorarlberg     2  Other   1
## 222:        Tyrol     8     AT   1
## 223:        Tyrol     7     AT   4
## 224:        Tyrol     7   <NA>   3
## 225: Burgenland     1     EU   1
```

Frequency counts with missing information

- There exists 225 keys given this example
- Missing values (NA) considered as own category
  - This would imply to underestimate the frequency counts, because, e.g. for pb220a at least with 1/3 probability the intruder can estimate the true category correctly (best case scenario)
  - In a worst case scenario an intruder can predict the true category correctly.

Frequency counts with missing information

However, more scenarios are thinkable. In the following five scenarios are reported how sample frequency counts can be calculated.

1. (current method) Missing values increase frequencies in other categories.
2. (conservative method) Misssing values do not increase frequencies in other categories but in those observation where a missing occurs.
3. (category size) Missing values do increase frequencies in other categories by a factor c. This method can be used as a general method to account for missing values in frequency calculations.
4. (conservative method 2) Misssing values do not increase frequencies in other categories
5. (own category) Same as method 4, but missings are treated like an own category

Frequency counts with missing information

key1	key2	key3
1	1	3
1	1	NA
2	1	3
NA	1	NA

$\longrightarrow$

key1	key2	key3	f_k
1	1	3	3
1	1	NA	3
2	1	3	2
NA	1	NA	4

. . . a rather extreme approach.

Frequency counts with missing information

key1	key2	key3
1	1	3
1	1	NA
2	1	3
NA	1	NA

$\longrightarrow$

key1	key2	key3	f_k
1	1	3	1
1	1	NA	3
2	1	3	1
NA	1	NA	4

Which approach is it?

Sample frequencies in sdcMicro

We switch to the Book, chapter 3.2.3. and the corresponding code theirein.

k-anonymity

One goal:

- to ensure that each distinct pattern of key variables is possessed by at least k records in the sample (k-anonymity)
- We now assign to each observation its frequency of its corresponding key.
- from now on, f_i is noted for any observation $i = 1, ..., n$.
- observation i is fulfilling k-anonymity if its key has frequency larger equal k
- a data set fulfils k-anonymity if $f_i \geq k$ and $\forall i \in \{1, ..., n\}$.

k-anonymity and l-diversity

Example inpatient records illustrating k-anonymity and l-diversity.

	key variables		f_k	sensitive variable	distinct l-diversity
	gender	age group		medical condition	
1	male	30s	3	cancer	2
2	male	30s	3	heart disease	2
3	male	30s	3	heart disease	2
4	female	20s	3	cancer	1
5	female	20s	3	cancer	1
6	female	20s	3	cancer	1

l-diversity

A group of observations with the same pattern of key variables is l-diverse if it contains at least l "well-represented" values for the sensitive variable.

Three different l-diversity measures, we only account for the first one.

distinct l-diversity: as the simplest definition that ensures that at least l distinct values for the sensitive field in each key;

entropy l-diversity: as the most complex definition, which defines entropy of a key

recursive l-diversity: the most common sensitive value does not appear too often in a key while less common sensitive values are ensured not to appear too infrequently in the same key.

See code page 59 (book)

Special uniques detection algorithm (SUDA)

- An observation is defined as a special unique with respect to a variable set Q, if it is sample unique both on Q and on a subset of Q.
- SUDA considers *Minimal Sample Uniques* (MSUs), which are unique variable sets without any unique subsets within a sample.
- SUDA scores:

1. the smaller the number of variables spanning the MSU within an observation, the higher the risk of the observation, and
2. the larger the number of MSUs in an observation, the higher the risk of the observation.

SUDA scores, cont.

For each MSU of size k contained in a given observation, a score is computed by

$$s_i = \begin{cases} \frac{1}{q!} \prod_{i=k}^{M} (q-i), & \text{if} \quad i \leq M. \\ 0, & \text{otherwise} \quad , \end{cases} \tag{1}$$

where M ($M > 0$) is the user-specified maximum size of MSUs, and q is the total number of categorical key variables in the data set.

- Note that if i is larger than M, the SUDA score should be set to zero.
- By definition, the smaller the size k of the MSU, the larger the score for the MSU (look at (q-i) in Equation~1).
- The final SUDA score for the observation is computed by multiplying the scores for each MSU (have a look at the product in Equation~1). In this way, observations with more MSUs are assigned a higher SUDA score.

SUDA scores, cont.

	age group	gender	income	education	f_k	score	DIS-SUD
1	20s	male	$\geq 50k$	high school	2	0	0
2	20s	male	$\geq 50k$	high school	2	0	0
3	20s	male	$\leq 50k$	high school	2	0	0
4	20s	male	$\leq 50k$	high school	2	0	0
5	20s	female	$\leq 50k$	university	1	1	0.0105
6	20s	female	$\leq 50k$	high school	1	0.5	0.0046
7	20s	female	$\leq 50k$	middle school	1	1.75	0.0203
8	60s	male	$\leq 50k$	university	1	2.25	0.0272

Observation 8 has two MSUs: *60s* of size 1, and (*male*, *university*) of size 2. Suppose the maximum size of MSUs is set at 3, the non-normalized score assigned to *60s* is computed by $\prod_{i=1}^{3}(4-i) = 6$, and the non-normalized score assigned to (*male*, *university*) is $\prod_{i=2}^{3}(4-i) = 2$. The normalized SUDA score: normalizing the scores and summation over these two norm. scores.

Simplified estimation of pop. freq. counts

Given a sample we want to estimate the population frequency counts.

- we already saw the tableWt() function, we continue here
- k-anonymity, l-diversity and SUDA do not consider sampling weights, thus only evaluate sample frequencies.
- However, when dealing with samples obtained with complex designs, we have to account for this.
- The sample weights are the best information how often a observation is represented in a population (according to its variables used for the complex design, calibration, etc.)

Simplified estimation of pop. freq. counts

Toy data set to illustrate the estimation of pop. frequency counts.

	Name	Year of birth	Gender	Citizenship	Occupation	Income	Weight
1	Max Mustermann	1978	m	AUT	Worker	35000	110
2	Josef Meier	1945	m	AUT	Pensioner	23500	70
3	Sabine Schnuller	1991	w	AUT	Student	7000	80
4	John Doe	1966	m	US	Employee	41200	120
5	Susan Rose	1989	w	AUT	Student	0	130
6	Markus Roller	1972	m	AUT	Employee	31100	90
7	Christoph Valon	1944	m	AUT	Pensioner	21400	150
8	Ulrike Mayer	1932	w	D	Pensioner	17600	150
9	Stefan Fuchs	1992	m	AUT	Worker	27500	130
10	Rainer Thomas	1950	m	AUT	Pensioner	25700	150
11	Julia Gross	1976	w	AUT	Employee	37000	140
12	Nadine Glatz	1987	w	AUT	Student	0	120
13	Makro Dilic	1990	m	AUT	Worker	21050	90
14	Sandra Stadler	1941	w	AUT	Pensioner	28500	80

- The simplest approach to estimate F_j under the assumption of simple random sampling without replacement is given by $\hat{F}_j = \frac{f_j}{f}$, where $f = \frac{n}{N}$ is the sampling fraction.
- general, samples from complex designs: sample freq. counts are multiplied by their sample weights and added together.

Simplified estimation of pop. freq. counts

- If the weights are known for every observation in the sample data set, then the population frequencies $\hat{F}_j$ are the sum of the weights of each record that has the same key combination.
- Thus $\hat{F}_j = \sum_{i \in |\mathbf{S}_j|} w_i$, where $|\mathbf{S}_j|$ is a subsample of $\mathbf{S}$ regarding key j and w_i are the weights of record $\mathbf{i}$ in subsample $\mathbf{S}_j$.

Why this is not good practice?

- Your turn: Copy and Paste the R code from page 63 (book)
- Answer the question why the simplified estimation of population frequencies is not good practice

The individual risk approach

- The fewer the individuals with whom an individual shares his or her combination of quasi-identifiers, the more likely the individual is to be correctly matched in another dataset that contains these quasi-identifiers.
- The individual risk values $r_i, i = 1, ..., n$ can also be interpreted as the **probability of disclosure for the individuals** or as the **probability for a successful match with individuals chosen at random from an external data** file with the same values of the key variables.
- This risk is often a worst-case scenario risk and does not imply that the person will be re-identified with certainty with this probability.
  - For instance, if an individual included in the microdata is not included in the external data file, the probability for a correct match is zero. Nevertheless, the risk measure computed based on the frequencies will be positive.

The individual risk approach

- To estimate the frequencies F_k in a population it is assumed that the population is drawn from a super-population $\rightarrow$ we assume that F_k is drawn from a certain distribution.
- Risk is estimated using quantiles of this distribution
- The estimation is just as good as the frequency counts of the population are modeled and how well the model assumptions are fulfilled.

The individual risk approach

The Benedetti-Franconi Model to estimate individual risks

- aim is to estimate $F_k|f_k$
- common assumption: $F_k \sim \text{Poisson}(N\pi_k)$ with π_k the inclusion probabilities of an individual selected from a sampling frame. N is assumed to be known.
- Benedetti-Franconi now assumes that $F_k|f_k$ is drawn from a negative binomial distribution.
- With a lot of math $\rightarrow$ Formula 3.4. in the book. We do not want to go into details here.
- Other authors made different assumptions, instead of the neg. bin. distr. using Poisson-Gamma, Dirichlet-multinom., Poisson-inverse Gaussian, ...
- Simulations shown that the Benedetti-Franconi model provides reasonalbe estimates, while in many situations others do not.

Individual risk with clusters

- Many micro-data sets have hierarchical, or multilevel, structures; for example, individuals who are situated in households.
- It is commonly assumed that the disclosure risk for a household is higher than or equal to the risk that at least one member of the household is re-identified.
- If one member of the household is re-identified, it is easier to identify the other members of the household.
- A household-level disclosure risk can be estimated by subtracting the probability that no person from the household is re-identified.
  - For example, if we consider a single household with three members, with individual disclosure risks of 0.1, 0.05 and 0.01, respectively, the disclosure risk for the entire household is calculated as $1 - (1 - 0.1) \cdot (1 - 0.05) \cdot (1 - 0.01) = 0.15355$.

The individual risk approach. Example

```
library(sdcMicro)
data(eusilc, package = "laeken") # ?eusilc
# create an sdc object
sdc1 <- createSdcObj(dat = eusilc,
                     keyVars = c("db040", "rb090", "pl030",
                     numVars = "eqIncome", weightVar = "rb0
                     hhId = "db030")
# extract risk
ir <- get.sdcMicroObj(sdc1, "risk")
names(ir)
```

```
## [1] "global"     "individual" "numeric"
```

```
head(ir$individual, 2)
```

```
##              risk fk       Fk  hier_risk
## [1,] 0.0009574837  3 1565.106 0.00154129
## [2,] 0.0004382825  5 2850.792 0.00154129
```

Global risk

1. Expected number of re-identifications. The easiest measure of global risk is to sum up the record-level individual disclosure risks, which gives the expected number of re-identifications.

In Software:

```
sum(ir$individual[, "risk"])
```

```
## [1] 25.01482
```

```
# or
ir$global$risk_ER
```

```
## [1] 25.01482
```

```
# or print(sdc1, "risk")
```

Global risk

2. Global risk measure based on log-linear models. This measure, defined as the probability that the number of sample uniques that are also population uniques, $\hat{\tau}_1 = \sum_{\{j:f_j=1\}} \mathbb{P}(F_j = 1|f_j = 1)$, is estimated using standard log-linear models. The population frequency counts, or the number of units in the population that possess a specific pattern of key variables, are assumed to follow a Poisson distribution.

- The global risk can then be estimated by a standard log-linear model, using the main effects (and interactions) of key variables.
- Log-linear models are used for modelling cell counts in contingency tables.
- These models declare how the expected cell count depends on levels of the categorical (key) variables.

Global risk, standard log-linear model

Let $\boldsymbol{\mu} = (\mu_1, ..., \mu_C)'$ denote the expected counts for the number of C cells of a contingency table. Multidimensional log-linear models for positive Poisson means have the following form:

$$log(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\lambda} \quad ,$$

where

- $log(\boldsymbol{\mu})$ is a $C \times 1$ vector containing the logarithms of the expected frequencies,
- $\mathbf{X}$ is a $C \times p$ model matrix and
- $\boldsymbol{\lambda}$ is a $p \times 1$ vector of model parameters.

Global risk

In Software (Standardmodell, weitere im Buch)

```
form <- as.formula(paste(" ~ ", "db040 + hsize + rb090 +
              age + pb220a + age:rb090 + age:hsize +
              hsize:rb090"))
sdc1 <- modRisk(sdc1, form = form)
get.sdcMicroObj(sdc1, "risk")$model
```

```
## The estimated model (using method 'default') was:
##  ~ db040 + hsize + rb090 + age + pb220a + age:rb090 + ag
## global risk-measures:
##  Risk-Measure 1: 0.261 (26.123 %)
##  Risk-Measure 2: 0.318 (31.765 %)
```

Global risk

3. Benchmark approach. This measure counts the number of observations with record-level risks higher than a certain threshold and higher than the main part of the data.

While the previous two measures indicate an overall re-identification risk for a microdata file, the benchmark approach is a relative measure that examines whether the distribution of record-level risks contains extreme values. For example, we can identify the number of records with individual risk satisfying the following conditions

$$r_i \geq 0.1 \wedge r_i \geq \tilde{\mathbf{r}} + 3 \cdot MAD(\mathbf{r}) \quad , \tag{2}$$

where $\mathbf{r}$ represents all record-level risks, and $MAD(\mathbf{r})$ is the median absolute deviation (the median of absolute distances of observations to it's median) of all record-level risks.

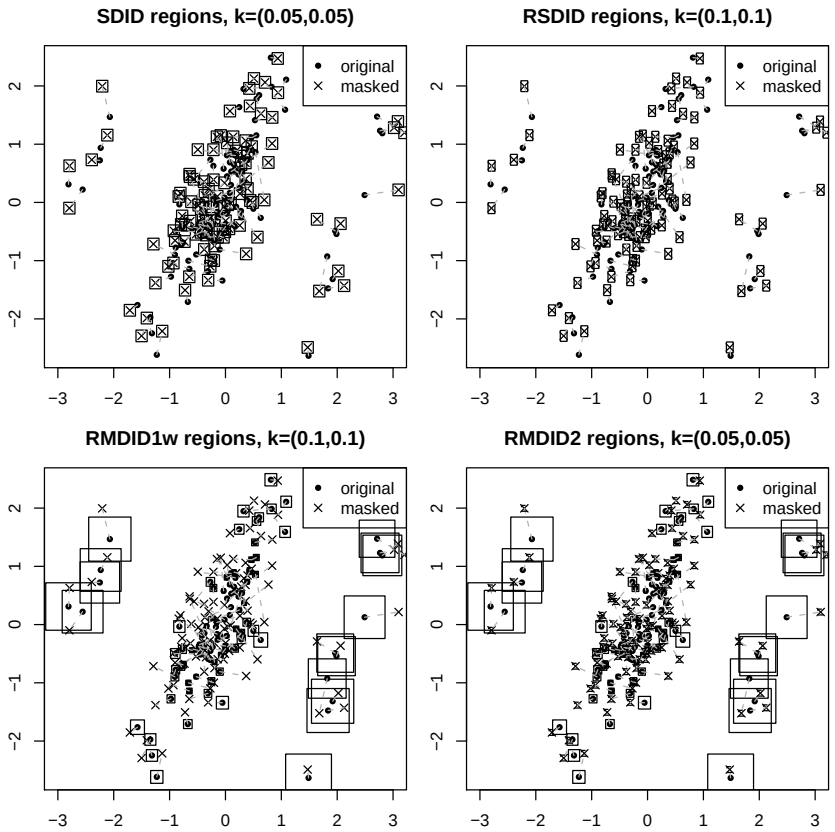
```
# see print(sdc1, "risk")
```

Disclosure Risk for continuous data

The intruder attempts to match records in the released data set with those in the external data set using common variables. $\rightarrow$ this is a record linkage problem.

- distance-based record linkage
  - Match anonymized data with (worst case) the non-anonymized data. Correct matches are highlighted. Different matching strategies are possible.
- probability-based record linkage
  - estimate the likelihood that two respondents belong to the same individual.
- interval-based methods
  - Simplification: intervals around the values of the non-anonymized continuous data. Is the anonymized observation within the interval? Interval size depends on outlyingness and standard deviation of values.

Disclosure Risk for continuous data, interval approach



Original and corresponding masked observations (perturbed by adding additive noise). In the bottom right graphic, small additional regions are plotted around the masked values for *RMDID2* procedures. The measure is mainly used for the (simplified) comparison of methods.

Disclosure Risk for continuous data

The interval disclosure approach:

```
print(sdc1, "numrisk")
```

```
## Numerical key variables: eqIncome
##
## Disclosure risk is currently between [0.00%; 100.00%]
##
## Current Information Loss:
##   - IL1: 0.00
##   - Difference of Eigenvalues: 0.000%
## ------------------------------------------------------------
```

High risk, because we haven't applied any anonymisation yet.

Disclosure Risk for continuous data

Considering outlyingness and neighborhood of perturbed values:

```
sdc1 <- dRiskRMD(sdc1)
get.sdcMicroObj(sdc1, "risk")$numericRMD$risk1
```

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

Summary

- k-anonymity for population data, but also as a first check on survey data
- suda2 is „just“ a detailed view on k-anonymity on subsets
- individual risk considers sampling weights
- global risk as a „summary“ of individual risks
- for continuous key variables, other approaches are used (record linkage)