DDI Data Description Statistics Protection Software

Sebastian Kočar
Social Science Data Archives
University of Ljubljana

Privacy in Statistical Databases 2014,
Eivissa, September 18th 2014

The content of the presentation

- What DDI Data Description Statistics Protection Software **is**?
- What DDI Data Description Statistics Protection Software **is not**?
- The **background** of developing the software
- The **purpose** of developing the software
- DDI Data Description Statistics Protection Software **features**
- Quick **demonstration** of using the tool
- Discussion about the official statistics univariate statistics protection

What is DDI Data Description Statistics Protection Software?

- It is a relatively simple tool for protecting **aggregated data**
- It is a tool which works with **XML files**
- It protects aggregated data by automatically changing XML code
- It protects **DDI** Data Description displayed **univariate statistics**
- It enables various kinds of data protection – using techniques similar to microdata and tabular data protection ones

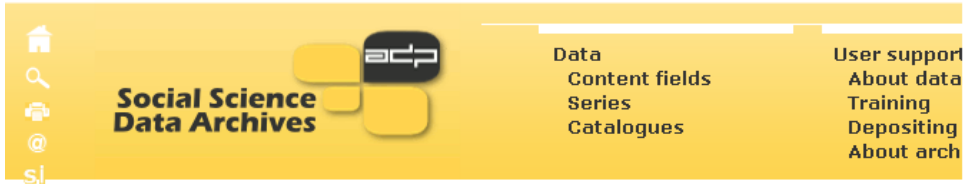
What DDI Data Description Statistics Protection Software is not?

- It is **not a microdata protection** tool (such as sdcMicro, μ-ARGUS)
- Datasets could be imported, but only to calculate certain univariate statistics
- It is **not a tabular data protection** tool (such as τ-ARGUS)
- It does not work with bivariate categorical data
- It is not meant to be used for protection of regularly published official statistics research results

The background/purpose of developing the software

- **Data without Boundaries** project's initiatives
- **Collaboration** between Social Science Data Archives and National Statistics Institutes
- Promotion of using official statistics data for **scientific research** purposes
- **Distribution** of aggregated official statistics data (with no restrictions)
- **Protection** of univariate statistics to minimize the dislosure risk

XML DDI Data Description Display



Labour force survey, 2011: Annual data - no confidentiality restrictions

Study description | Data description | Materials and publications | Download data

Data file description

ADS11 - Labour Force Survey, 2011 [data file]

File ID: ADS11_P1_SL_V1_R1

*.txt - TEXT

- Number of variables: 214
- Number of units: 61888

Version: August 2012

Variable list

val Round

Value		Frequency
1	1.	17797
2	2.	13491
3	3.	11580
4	4.	9892
5	5.	9128
-2	Missing value	0

Valid cases	Invalid cases	Minimum	Maximum	Arithmetic mean	Standard deviation
61888	0				

Valid range from 1 to 5

starost Age

Value		Frequency
-2	Missing value	

Valid cases	Invalid cases	Minimum	Maximum	Arithmetic mean	Standard deviation
61888	0	0		42.034	21.877

Valid range from 0 to Protected value

```
<dataDscr>
  <var ID="V5" name="val" files="F1" dcml="0" intrvl="discrete">
    <location width="1"/><labl>Round</labl>
    <valrng><range min="1" max="5"/></valrng>
    <invalrng><item VALUE="-2"/></invalrng>
    <sumStat type="vald">61888</sumStat>
    <sumStat type="invd">0</sumStat>
    <catgry><catValu>1</catValu>
      <labl>1.</labl>
      <catStat type="freq">17797</catStat>
    </catgry>
    <catgry><catValu>2</catValu>
      <labl>2.</labl>
      <catStat type="freq">13491</catStat>
    </catgry>
    <catgry><catValu>3</catValu>
      <labl>3.</labl>
      <catStat type="freq">11580</catStat>
    </catgry>
    <catgry><catValu>4</catValu>
      <labl>4.</labl>
      <catStat type="freq">9892</catStat>
    </catgry>
    <catgry><catValu>5</catValu>
      <labl>5.</labl>
      <catStat type="freq">9128</catStat>
    </catgry><catgry missing="Y">
      <catValu>-2</catValu>
      <labl>Missing value</labl>
      <catStat type="freq">0</catStat>
    </catgry>
    <varFormat type="numeric" schema="other"/>
  </var>
  <var ID="V8" name="starost" files="F1" dcml="0" intrvl="contin">
     <location width="2"/>
     <labl>Age</labl>
     <valrng>
       <range min="0" max="Protected value"/>
     </valrng>
     <invalrng><item VALUE="-2"/></invalrng>
        <sumStat type="vald">61888</sumStat>
        <sumStat type="invd">0</sumStat>
        <sumStat type="min">0</sumStat>
        <sumStat type="max">Protected value</sumStat>
        <sumStat type="mean">42.034</sumStat>
        <sumStat type="stdev">21.877</sumStat>
        <catgry missing="Y"><catValu>-2</catValu>
          <labl>Missing value</labl>
        </catgry>
        <varFormat type="numeric" schema="other"/>
      </var>
</dataDscr>
```

Protecting DDI DD Data

Software features and quick demonstration of using them

- Deleting variables/variable information
- Protecting descriptive statistics (min, max, mean, SD)
- Calculating replacement univariate statistics
- Top- and bottom- coding
- Bracketing/recoding
- Minimum frequency protection
- Displaying DDI Data Description statistics (importing XML)
- Guidelines and data protection report

Official statistics univariate statistics protection

- How much should univariate statistics (aggregated data) be protected to be publically distributed?
- Using the minimum frequency rule, what should be the threshold for one dimension table data - frequencies for values of one variable only?
- Which data protection functions should be added to the presented tool and which should be considered unnecessary?

Social Science Data Archives

Thank you for your attention!

Sebastian Kočar
sebastian.kocar@fdv.uni-lj.si
http://www.adp.fdv.uni-lj.si/

Privacy in Statistical Databases 2014,
Eivissa, September 18th 2014