Safe Access to Register Microdata

UNECE/Eurostat Work Session on Statistical Confidentiality
The Hague, Netherlands
29-31 October 2019
Authors

Dr. Johan Heldal
Senior Adviser | Division for Methods
Department for Digitalisation and shared business services | Statistics Norway
johan.heldal@ssb.no

Svein Johansen
Senior Adviser | Division for Access to Microdata
Department of Social Statistics | Statistics Norway
svein.johansen@ssb.no

Ørnulf Risnes
Head of Digital Development
Norwegian Centre for Research Data (NSD)
ornulf.risnes@nsd.no
For researchers, PhD and master students in approved institutions
- Fast access
- No application for data
- Web interface with self service
- Microdata invisible
microdata.no

• Launched in March 2018.
• 30 institutions with some 300 users are connected to the system.
• Present content: 142 register variables on population, education, labour market, income, and welfare benefits.
• Event history data on 10,2 million individuals.
  • All national IDs ever issued in Norway.
• Remote web-access. Invisible microdata.
• Developed 2012–2018 jointly by Statistics Norway and NSD – Norwegian Centre for Research data.
  • Funded by The Norwegian Research Council (NRC).
  • New funding from NRC from 2020 to go further.
• Own interface with Stata-like commands.
Research is about results, not records

- *Metadata* are *open* for everybody.
- *Microdata* records are *invisible*.
- Interface with metadata included
- Data base stored in Statistics Norway
- Safe login, three factor authentication
  - Norwegian national ID required.
- GSIM (Generic Statistical Information Model)
- Results: confidentially safe on the fly
- [https://microdata.no/en/](https://microdata.no/en/)
microdata.no makes it easier to analyse register data

- Researchers and students can process and analyse all available register variables.
- Data is available on the population, education, income, the labour market and welfare benefits.
- Institutions that have signed an agreement can manage their own users.
- The solution has an anonymisation interface with embedded privacy protection.

Data protection award for microdata.no

The prize “Built-in data protection in practice 2018” granted by The Norwegian Data Protection Authority was awarded to microdata.no.

Read more
<table>
<thead>
<tr>
<th>Wildcard</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utdanningens art NUS2000</td>
<td>NUS-kode for höyeste fullførte utdanning etter definisjon av utdanningsnivå utarbeidet i 2006. Se h...</td>
</tr>
<tr>
<td>Grunnstanad fra folketrygden, mottak</td>
<td>Datering av mottaksperiode</td>
</tr>
<tr>
<td>Enslig forsørger stønad til barneblys, mottak</td>
<td>Datering av mottaksperiode</td>
</tr>
<tr>
<td>Inntekt etter skatt</td>
<td>Yrkesinntekter, kapitalinntekter, skattepliktige og skattefrie overføringer, i løpet av kalenderåre...</td>
</tr>
<tr>
<td>Personens hovedarbeidsforhold - virksomhet</td>
<td>Variabelen Identifiserer virksomheten der personen har sitt hovedarbeidsforhold (arbeidsforholdet...</td>
</tr>
<tr>
<td>Alderspensjon fra folketrygden, uttaksgrad</td>
<td>Angir uttaksgrad av alderspensjon</td>
</tr>
<tr>
<td>Pensjon, tillegspensjonsfaktor</td>
<td>Viser et fast tall for den tillegspensjonen som stønadsmottakeren har krav på (inkludert garantit...</td>
</tr>
<tr>
<td>Fødeand</td>
<td>Variabelen viser landkode for personens fødeand. Det vil si morens bostedsland ved fødselen.</td>
</tr>
<tr>
<td>Økonomisk sosialhjelp, antall måneder i året</td>
<td>Antall måneder personen mottok sosialhjelp i rapporteringsåret</td>
</tr>
<tr>
<td>Tettsted</td>
<td>En hussamling skal registreres som et tettsted dersom det 1. Bor minst 200 personer der og 2. Avst...</td>
</tr>
<tr>
<td>Bostedskommune ved fylte 16 år</td>
<td>Bostedskommune ved fylte 16 år. Uoppgitt (ELLER BLANK) for personer født før 1958.</td>
</tr>
<tr>
<td>Hjelpesnadar fra folketrygden, mottak</td>
<td>Datering av mottaksperiode</td>
</tr>
</tbody>
</table>
Infrastructure layout – metadata driven, remote access

Metadata

DataStore

User workspace

import var1 yyyy-mm-dd as v1
import var2 yyyy-mm-dd as v2
import var1 yyyy-mm-dd as v3
Confidentially safe output

Subsetting user workspace to study population

drop if ...
keep if ...

---
Temporality structure, an example.

import `SIVSTANDFDT_SIVSTAND 2002-01-01 as marstat02`

import-event `SIVSTANDFDT_SIVSTAND 2005-01-01 to 2015-12-31 as marstat05_15`
Metadata plays important roles

• It informs users about definitions, data types, temporal nature and codes.
• It drives most technical components. Data access only through metadata.
• Supportive function. Interactive assistance in User Interface.
Statistical Disclosure Control - Foundation

• An interpretation of the Five Safes model (e.g. Desai, Ritchie & Welpton, 2016) into our context.

• Safe people, achieved through access policy
  • Only researchers in approved research institutions can have access.

• Safe settings
  • Safe login (three factor authentication), Invisible microdata, logging of all activity, communication with users.
  • Norwegian national ID required for access.

• Safe Data
  • Invisible microdata, automatic winsorization of numerical variables.

• Safe outputs
  • Statistical Output Disclosure Control, on the fly.

• Safe projects
  • Less relevant. Approval of projects not needed.
Statistical Disclosure Control (SDC)

• Invisibility of microdata is not enough!
• Challenges:
  • Users can manipulate population definitions and variables quite freely.
  • This is essential to do interesting research.
  • But which smart attacks can a smart adverse user hit on?
  • We have to play Devil’s Advocate to imagine.
• But we have to weigh risks against utility.
• SN and NSD have near 40 years experience giving access to microdata on contract.
• We have never seen intentional breaches.
• Detected breaches will hit researchers institution.
Present SDC measures – Output SDC

• Minimum size of populations to be analysed: At least 1 000

• *Constant* unbiased restricted maximum entropy noise ($X$) on counts. ($|X \leq 5|$, $X$ whole number)
  • Applies Cell Key method. (Same cell, same perturbation within same population).
  • Unbiased noise, $EX = 0$. Zeros not perturbed.
  • No negative perturbed counts and no perturbed counts 1-4.

• Automatic 2 % winsorization of all numerical variables when subsetting and importing (1 % at each end).
  • A dynamic form of top and bottom coding. Hides extreme values
  • Winsorized magnitude totals and adjusted proportionally to perturbed counts.

• Scatter plots are smoothed with hexbin plots.
Noise on frequencies. Requirements.

1. A perturbation $X$ on a count $n$ must be an integer random variable on $\{-5, -4, ..., 4, 5\}$ with a distribution $p_n(x) = P(X = x|n)$.
2. Perturbed value $Y = X + n$.
3. Perturbations must be unbiased: $EX = 0$.
4. Perturbed counts $Y$, must be 0 or $Y \geq 5$.
5. Constant noise: Repeated counts of the same set of individuals must always result in the same perturbed count $Y$.
6. The noise distribution $p_n(x)$ must maximize entropy among all distributions satisfying conditions 1-3, that is

$$\mathcal{E}(p_n) = -E(\log(p_n(X))) = -\sum_{x=-5}^{5} p_n(x)\log(p_n(x)).$$
Present SDC-measures continued

• So far: Priority to simple descriptive statistics.
• All activity on microdata.no is logged
  • On suspicion we can close access and reproduce all behaviour.

• Along with new funding, the SDC measures will be evaluated and reconsidered based on experiences and new knowledge.
• One aspect is to evaluate how elements of Differential Privacy will fit in.
Some understandings

• Some differencing attacks are possible because noise is constant.
  • We have already done some adjustments.

• Winsorization.
  • Has large impact on some studies with age, income and wealth variables.

• In tables: Zeroes are disclosive as much as small counts.
  • Our noise creates more zeroes and uncertainty, but is it enough?
  • Should zeroes be perturbed? Differential privacy noise does.

• Presently: Tables are not additive
  • We have to introduce an additivity module for tables.

• Some users take results face value, disregarding noise uncertainty.
  • and disregarding present nonadditivity of tables.
Some topics

• Should zeros be perturbed?
  • Yes, but what then about structural zeros?

• Can we accept perturbations to somewhere out of range of the data/parameter?
  • Can we accept negative counts?
  • No. Users will not accept them. Difficult if input to new analyses.

• Can we accept biased perturbations?
  • Yes. If zeros are perturbed and non-negative counts are not accepted we have to.

• Do perturbed counts have to be integers?
  • No why? Perturbed counts should look perturbed not to be taken face value.

• Is Differential Privacy a solution?
  • DP-noise (Laplace noise) can do these things.
  • Rinott et al. 2018 suggests that Australian Table Builder should consider DP.
  • Microdata.no and TB have much in common.
Differential Privacy and microdata.no

• Microdata.no is not a closed data universe
  • Users select some part of the data in time and scope to analyse
  • microdata.no will expand in time, units and variables.
  • microdata.no is not only a TB, but also an analysis tool.
• DP noise alone does not make microdata.no DP system.
  • DP requires a budget ($\epsilon$) for consumption of privacy.
  • We are not in a position to establish or handle a privacy budget.
  • What should the budget $\epsilon$ be?
• DP is at odds with the Five Safes and the ideas for data access that microdata.no are built on.
• DP is to some extent also at odds with constant noise.
• But DP (Laplace) noise can be an option/inspiration when looking for better noise.
Some further proposals

• Keep constant noise (same cell same perturbation)
  • Constant noise limits consumption of privacy.
  • Required for user trust in results.
  • When sharing a script with a colleague or a peer reviewer with access, the two should see the same result.

• Make non-structural zeros invisible by perturbing them to positive numbers.

• Let noise on counts be continuous, not integer
  • Then no users will take the perturbed counts face value.
  • Additivity model will generate non-integer counts anyway.

• Winsorization: Serves in a way as a replacement for the DP sensitivity parameter $\Delta = \max_{x \sim x' \in D} |f(x) - f(x')|$, but must be modified to take effect only on relevant variables.
Some ideas for noise – comments are welcome

• \( n \) true count
• \( X \) Laplace noise \( \sim f(x) = \frac{\epsilon}{2} e^{-\epsilon|x|} \) (eventually with bounded \( x \))
• \( Y = |n + X| \geq 0 \)

Then the noise will have some bias that quickly approaches zero as \( n \) increases

\[
E(Y|n) = n + \frac{1}{\epsilon} e^{-\epsilon n}
\]

\[
\lim_{n \to \infty} MSE(Y|n, \epsilon) = MSE(Y|0, \epsilon) = \frac{2}{\epsilon^2} \geq MSE(Y|n, \epsilon)
\]

\[
= \frac{2}{\epsilon^2} - \frac{2}{\epsilon} e^{-\epsilon n} \geq \frac{2}{\epsilon^2} (1 - e^{-1})
\]
Uncertainty

• Uncertainty can be expressed for integer $n$s by the likelihood function

$$L(n|Y, \epsilon) = \frac{\epsilon}{2} \left( e^{-\epsilon|Y-n|} + e^{-\epsilon(Y-n)} \right)$$

which does not depend on the true $n$.

• It should be possible to see $L(n|Y, \epsilon)$ by clicking on the perturbed counts on the screen.
Winsorization

- Represents a kind of dynamic top and bottom coding
- Serves in a way as a replacement for the DP sensitivity parameter
  \[ \Delta = \max_{x \sim x' \in D} |f(x) - f(x')|, \ x \sim x' \text{ datasets that differ by one unit only.} \]
- But must be modified to take effect only on relevant variables.
Analyses

• Microdata.no is not just a table builder. One can manipulate data and do analyses.
  • Until now primarily regressions and log linear regressions.

• Winsorization is so far the only SDC measure that affects analyses.
  • Takes effect on both numerical regressors and dependent variables
  • It has considerable effects on regressions

• O’Keefe & Chipperfield (2013) summarizes some attack scenarios with regressions and proposes measures.
  • Winsorization and lower bound for $n$ (1000) prevents some of them.
  • But winsorization can be replaced by leverage restrictions or trimming.
  • OLS can be replaced with robust regression.
Some challenges

- Constant perturbations of zeros require cell keys for empty cells
  - Can be done by means of cell keys for marginal cells
- Structural zeros.
  - Impossible to keep track of all possible structural zeros in crosstabulations in the system
  - Possible solution: Give users the option to set some cells to zero apriori.
- Additivity module
  - Iterative Proportional Fitting to perturbed marginals.
  - IPF stepwise top down since marginals are inconsistent at all levels.
  - \( L(n|Y, \epsilon) \) will be more complex.
- Winsorization: Now used on all numerical variables. Too much.
With new funding from 2020 we want to:

- Improve the SDC methods.
- Include variables from new sources
  - E.g. health registers (pilot)
  - Requires distributed Data Store and encrypted import to Work Space.
- Make more statistical methods safely available.
- Give safe access to new groups, e.g. public administration.
- International access to researchers
  - Requires log-on solution without Norwegian ID
- Cooperate with other stakeholders home or abroad on use with other kinds of data.
Thank you
Appendix: 1. Record keys

- Microdata.no operates with encrypted person IDs.
- They are 64 bits random binary numbers.
- As a record key, $U_i$ for individual no. $i$, we take 8 of the 64 digits
- With decimal representation, $0 \leq U_i \leq 255$. 
Appendix: 2. Cell keys

- $c$ is a cell, or equivalently: a set of individuals in microdata.no.
- $U_i$ is a record key, an encrypted national ID. 64 bits.
- Cell key: $U_c = \sum_{i \in c}^{\text{XOR}} U_i$, is used in a **look-up table**.
- Bitwise **XOR** addition: Addition of binary integers without using carry over digits.
- An example with 8 bits.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_1$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>203</td>
</tr>
<tr>
<td>$U_2$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>102</td>
</tr>
<tr>
<td>$U_3$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>61</td>
</tr>
<tr>
<td>$U_4$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>239</td>
</tr>
<tr>
<td>$U_c$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>115</td>
</tr>
</tbody>
</table>

1. A perturbation $X$ on a count $n$ must be an integer random variable on $\{-5, -4, \ldots, 4, 5\}$ with a distribution $p_n(x) = P(X = x|n)$. Perturbed value $Y = X + n$.

2. Perturbations must be unbiased: $EX = 0$.

3. Perturbed counts $Y$, must be 0 or $Y \geq 5$.

4. Constant noise: Repeated counts of the same set of individuals must always result in the same perturbed count $Y$.

5. The noise distribution $p_n(x)$ must maximize entropy among all distributions satisfying conditions 1-3, that is

$$\mathcal{E}(p_n) = -E(\log(p_n(X))) = -\sum_{x=-5}^{5} p_n(x)\log(p_n(x)).$$
## Appendix: 4. Probability distributions

<table>
<thead>
<tr>
<th>$p(x)$</th>
<th>$n = 0$</th>
<th>$n = 1$</th>
<th>$n = 2$</th>
<th>$n = 3$</th>
<th>$n = 4$</th>
<th>$n = 5$</th>
<th>$n = 6$</th>
<th>$n = 7$</th>
<th>$n = 8$</th>
<th>$n = 9$</th>
<th>$n &gt;= 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x = -5$</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,2987</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0909</td>
</tr>
<tr>
<td>$x = -4$</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,3988</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,1296</td>
</tr>
<tr>
<td>$x = -3$</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,5175</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,1908</td>
<td>0,1219</td>
</tr>
<tr>
<td>$x = -2$</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,6555</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,2940</td>
<td>0,1634</td>
<td>0,1147</td>
<td>0,0909</td>
</tr>
<tr>
<td>$x = -1$</td>
<td>0,0000</td>
<td>0,8149</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,4880</td>
<td>0,2145</td>
<td>0,1400</td>
<td>0,1079</td>
</tr>
<tr>
<td>$x = 0$</td>
<td>1,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
</tr>
<tr>
<td>$x = 1$</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
</tr>
<tr>
<td>$x = 2$</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
</tr>
<tr>
<td>$x = 3$</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
</tr>
<tr>
<td>$x = 4$</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
</tr>
<tr>
<td>$x = 5$</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
<td>0,0000</td>
</tr>
</tbody>
</table>

| | Sum($p(x)$) | $E(X|n)$ | $Var(X|n)$ | $E(p|n)$ |
|---|---|---|---|---|
| | 1,000 | 0,000 | 0,000 | 0,000 |
| | 1,000 | 0,000 | 0,000 | 0,000 |
| | 1,000 | 0,000 | 0,000 | 0,000 |
| | 1,000 | 0,000 | 0,000 | 0,000 |
| | 1,000 | 0,000 | 0,000 | 0,000 |
| | 1,000 | 0,000 | 0,000 | 0,000 |
| | 1,000 | 0,000 | 0,000 | 0,000 |
| | 1,000 | 0,000 | 0,000 | 0,000 |
| | 1,000 | 0,000 | 0,000 | 0,000 |
| | 1,000 | 0,000 | 0,000 | 0,000 |
| | 1,000 | 0,000 | 0,000 | 0,000 |
| | 1,000 | 0,000 | 0,000 | 0,000 |
| | 1,000 | 0,000 | 0,000 | 0,000 |
| | 1,000 | 0,000 | 0,000 | 0,000 |

Table C1: Maximum entropy probability distributions for noise given the value of the true count $n$.

Combinations where $Y = n + X < 0$.

Combinations where $1 \leq Y = n + X \leq 5$

Based on table C1, a look-up table is generated.