TALL SOM FORTELLER

Releasable inner cell frequencies by post-processing protected tabular data

Øyvind Langsrud

Statistics Norway

Oyvind.Langsrud@ssb.no, github.com/olangsrud

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

Statistisk sentralbyrå
Statistics Norway

Frequency table

	col1	col2	col3	Total
row1	3	6	2	**11**
row2	1	4	7	**12**
row3	5	8	27	**40**
Total	**9**	**18**	**36**	**63**

Inner cells

	col1	col2	col3	Total
row1	3	6	2	**11**
row2	1	4	7	**12**
row3	5	8	27	**40**
Total	**9**	**18**	**36**	**63**

Totals

9
18
36
11
12
40
63

Statistisk sentralbyrå
Statistics Norway

	col1	col2	col3	Total
row1	3	6	2	**11**
row2	1	4	7	**12**
row3	5	8	27	**40**
Total	**9**	**18**	**36**	**63**

Totals

9
18
36
11
12
40
63

Inner cells

3
1
5
6
4
8
2
7
27

In general

$$z = X^T y$$

Totals

9
18
36
11
12
40
63

$=$

1	0	0	**1**	0	0	**1**
1	0	0	0	**1**	0	**1**
1	0	0	0	0	**1**	**1**
0	**1**	0	**1**	0	0	**1**
0	**1**	0	0	**1**	0	**1**
0	**1**	0	0	0	**1**	**1**
0	0	**1**	**1**	0	0	**1**
0	0	**1**	0	**1**	0	**1**
0	0	**1**	0	0	**1**	**1**

^{T}

$\times$

Inner cells

3
1
5
6
4
8
2
7
27

In general

$$z = \boldsymbol{X}^T y + \text{noise}$$

Totals

9
18
36
11
12
40
63

$=$

	0	1	0	0	1
	0	0	0	1	0	1
1	0	0	0	0	1	1
0	1	0	1	0	0	1
0	1	0	0	1	0	1
0	1	0	0	0	1	1
0	0	1	1	0	0	1
0	0	1	0	1	0	1
0	0	1	0	0	1	1

^{T}

$\times$

Inner cells

3
1
5
6
4
8
2
7
27

In general

$$z = \boldsymbol{X}^T y + \text{noise}$$

Totals

11
18
38
11
13
45
61

$=$

	0	**1**	0	0	**1**
	0	0	0	**1**	0	**1**
1	0	0	0	0	**1**	**1**
0	**1**	0	**1**	0	0	**1**
0	**1**	0	0	**1**	0	**1**
0	**1**	0	0	0	**1**	**1**
0	0	**1**	**1**	0	0	**1**
0	0	**1**	0	**1**	0	**1**
0	0	**1**	0	0	**1**	**1**

^{T}

$\times$

Inner cells

3
1
5
6
4
8
2
7
27

$+$

noise

Estimate y from z by ordinary least squares regression?

$$z = \boldsymbol{X}^T y + \text{noise}$$

	col1	col2	col3	Total
row1	?	?	?	11
row2	?	?	?	13
row3	?	?	?	45
Total	11	18	38	61

- Too many variables/parameters
  - Many possible solutions
  - Collinearity must be handled.

Estimate y from z by Moore-Penrose generalized inverse

$$z = \boldsymbol{X}^T y + \text{noise}$$

	col1	col2	col3	Total
row1	-0.689	1.644	8.311	**11**
row2	-0.022	2.311	8.978	**13**
row3	10.644	12.978	19.644	**45**
Total	**11**	**18**	**38**	**61**

$$\hat{y} = (\boldsymbol{X}^T)^{\dagger} z$$

- Moore-Penrose inverse (denoted by †)
  - Least squares estimate
    - Sum of squared residuals minimized
  - When several solutions
    - Sum of squared parameters ($\hat{y}$) minimized

Estimate y from z when no noise added

	col1	col2	col3	Total
row1	-0.333	2.667	8.667	11
row2	0.000	3.000	9.000	12
row3	9.333	12.333	18.333	40
Total	9	18	36	63

$$z = X^T y$$

$$\hat{y} = (X^T)^\dagger z$$
$$= XX^\dagger y$$
$$= X\,(X^T X)^\dagger\, X^T y$$

- Moore-Penrose inverse (denoted by †)
  - Least squares estimate
    - Sum of squared residuals minimized
  - When several solutions
    - Sum of squared parameters ($\hat{y}$) minimized

$$y = X\beta + \text{noise}$$

Estimate y from z when no noise added

$$z = X^T y$$

$\hat{y}$ = Parameter estimates under one model

= Fitted values under another

$$\hat{y} = (X^T)^\dagger z$$
$$= XX^\dagger y$$
$$= X(X^T X)^\dagger X^T y$$

row1	-0.333	2.667	8.667	**11**
row2	0.000	3.000	9.000	**12**
row3	9.333	12.333	18.333	**40**
Total	**9**	**18**	**36**	**63**

- Moore-Penrose inverse (denoted by †)
  - Least squares estimate
    - Sum of squared residuals minimized
  - When several solutions
    - Sum of squared parameters ($\hat{y}$) minimized

$$y = X\beta + \text{noise}$$

Estimate suppressed cells from all safe cells

	col1	col2	col3	Total
row1	?	6	?	11
row2	?	4	?	12
row3	5	8	27	40
Total	9	18	36	63

$$z = \boldsymbol{X}^T y$$

$$\hat{y} = \left(\boldsymbol{X}^T\right)^{\dagger} z$$

z = all safe cells

y = all inner cells

but perfect fit for the 5 safe

Estimate suppressed cells from all safe cells

	col1	col2	col3	Total
row1	1.25	**6**	3.75	**11**
row2	2.75	**4**	5.25	**12**
row3	**5**	**8**	**27**	**40**
Total	**9**	**18**	**36**	**63**

$$z = \boldsymbol{X}^T y$$

$$\hat{y} = (\boldsymbol{X}^T)^{\dagger} z$$

z = all safe cells

y = all inner cells

but perfect fit for the 5 safe

Estimate suppressed cells and add information preserving residuals

	col1	col2	col3	Total
row1	0.633	**6**	4.367	**11**
row2	3.367	**4**	4.633	**12**
row3	**5**	**8**	**27**	**40**
Total	**9**	**18**	**36**	**63**

- Special application of regression-based synthetic data
  - As described in Langsrud (2019)
  - Special application of IPSO (Burridge, 2003) when residuals scaled as in original data
  - Implemented in R package RegSDC

- Langsrud, Ø. (2019), Information Preserving Regression-based Tools for Statistical Disclosure Control, *Statistics and Computing*, 29(5):965-976.
- Burridge, J. (2003), Information preserving statistical obfuscation, *Statistics and Computing*, 13(4):321-327.

Suppressed cells as decimal numbers

	col1	col2	col3	Total
row1	0.633	**6**	4.367	**11**
row2	3.367	**4**	4.633	**12**
row3	**5**	**8**	**27**	**40**
Total	**9**	**18**	**36**	**63**

- More information to users than just suppressed cells
  - Alterative to intervals.
- Only the inner cells need to be stored
  - Various totals can be calculated when needed.
  - When the result is a whole number, this means that this cell is not suppressed.
- Control the result of the suppression method
  - Perfect fit (whole number) of suppressed cells means that suppression method has failed.
- Permit calculation of arbitrary (user defined) sums of cells afterwards
  - When the result is a whole number, this means that this sum is not suppressed.

Estimate y from z non-negatively

$$z = \boldsymbol{X}^T y + \text{noise}$$

	col1	col2	col3	Total
row1	0.000	1.300	7.967	**11**
row2	0.000	2.300	8.967	**13**
row3	9.933	13.333	20.000	**45**
Total	**11**	**18**	**38**	**61**

- As Moore-Penrose inverse, but with non-negativity constraint
  - Least squares estimate
    - Sum of squared residuals minimized
  - When several solutions
    - Sum of squared parameters ($\hat{y}$) minimized

Estimate y from z non-negatively and recompute totals

$$z = \boldsymbol{X}^T y + \text{noise}$$

	col1	col2	col3	Total
row1	0.000	1.300	7.967	**9.267**
row2	0.000	2.300	8.967	**11.267**
row3	9.933	13.333	20.000	**43.267**
Total	**9.933**	**16.933**	**36.933**	**63.800**

- Additivity restored
  - But decimal numbers

- As Moore-Penrose inverse, but with non-negativity constraint
  - Least squares estimate
    - Sum of squared residuals minimized
  - When several solutions
    - Sum of squared parameters ($\hat{y}$) minimized

Estimate y from z non-negatively when no noise added

$$z = X^T y$$

	col1	col2	col3	Total
row1	0.000	2.500	8.500	**11**
row2	0.000	3.000	9.000	**12**
row3	9.000	12.500	18.500	**40**
Total	**9**	**18**	**36**	**63**

- Recapitulate
  - Without non-negativity constraint

    Fitted values under regression model

$$y = X\beta + \text{noise}$$

- Better model that ordinary regression model?

Estimate y from z by log-linear model when no noise added

$$z = \boldsymbol{X}^T y$$

	col1	col2	col3	Total
row1	1.571	3.143	6.286	**11**
row2	1.714	3.429	6.857	**12**
row3	5.714	11.429	22.857	**40**
Total	**9**	**18**	**36**	**63**

- $\hat{y}$ is fitted values under log-linear model
  - Estimated cell probabilities times the overall total

- Iterative proportional fitting is one estimation method
- Another is implemented in glm in R

- Estimate y from z non-negatively
- Recompute totals
- Refit by log-linear modelling

$$z = \boldsymbol{X}^T y + \text{noise}$$

	col1	col2	col3	Total
row1	?	?	?	11
row2	?	?	?	13
row3	?	?	?	45
Total	11	18	38	61

- **Estimate y from z non-negatively**
- **Recompute totals**
- **Refit by log-linear modelling**

$$z = \boldsymbol{X}^T y + \text{noise}$$

	col1	col2	col3	Total
row1	0.000	1.300	7.967	**11**
row2	0.000	2.300	8.967	**13**
row3	9.933	13.333	20.000	**45**
Total	**11**	**18**	**38**	**61**

- Estimate y from z non-negatively
- **Recompute totals**
- Refit by log-linear modelling

$$z = \boldsymbol{X}^T y + \text{noise}$$

	col1	col2	col3	Total
row1	0.000	1.300	7.967	**9.267**
row2	0.000	2.300	8.967	**11.267**
row3	9.933	13.333	20.000	**43.267**
Total	**9.933**	**16.933**	**36.933**	**63.800**

- Estimate y from z non-negatively
- Recompute totals
- **Refit by log-linear modelling**

$$z = \boldsymbol{X}^T y + \text{noise}$$

	col1	col2	col3	Total
row1	?	?	?	9.267
row2	?	?	?	11.267
row3	?	?	?	43.267
Total	9.933	16.933	36.933	63.800

- Estimate y from z non-negatively
- Recompute totals
- **Refit by log-linear modelling**

$$z = \boldsymbol{X}^T y + \text{noise}$$

	col1	col2	col3	Total
row1	1.443	2.459	5.364	**9.267**
row2	1.754	2.990	6.522	**11.267**
row3	6.736	11.484	25.047	**43.267**
Total	**9.933**	**16.933**	**36.933**	**63.800**

Differential privacy

- Estimate y from z non-negatively
- Recompute totals
- Possible refit by log-linear modelling

	col1	col2	col3	Total
row1	?	?	?	**19.757**
row2	?	?	?	**14.542**
row3	?	?	?	**40.470**
Total	**-2.429**	**25.266**	**34.867**	**?**

$$z = \boldsymbol{X}^T y + Lap(k/\varepsilon)$$

- Noise from Laplace distribution
- ε-differential privacy
  - ε=0.5 in this example
- *k* margins involved in z
  - Here *k*=2, since overall total not included (then *k*=3)

Differential privacy

- Estimate y from z non-negatively
- Recompute totals
- Possible refit by log-linear modelling

	col1	col2	col3	Total
row1	?	?	?	19.757
row2	?	?	?	14.542
row3	?	?	?	40.470
Total	-2.429	25.266	34.867	

$$z = \boldsymbol{X}^T y + Lap(k/\varepsilon)$$

- Noise from Laplace distribution
- ε-differential privacy
  - ε=0.5 in this example
- *k* margins involved in z
  - Here *k*=2, since overall total not included (then *k*=3)

Differential privacy

- Estimate y from z non-negatively
- Recompute totals
- Possible refit by log-linear modelling

	col1	col2	col3	Total
row1	0.000	6.856	10.056	**19.757**
row2	0.000	4.249	7.449	**14.542**
row3	0.415	17.005	20.205	**40.470**
Total	**-2.429**	**25.266**	**34.867**	

$$z = \mathbf{X}^T y + Lap(k/\varepsilon)$$

- Noise from Laplace distribution
- ε-differential privacy
  - ε=0.5 in this example
- *k* margins involved in z
  - Here *k*=2, since overall total not included (then *k*=3)

Differential privacy

- Estimate y from z non-negatively
- Recompute totals
- Possible refit by log-linear modelling

	col1	col2	col3	Total
row1	0.000	6.856	10.056	?
row2	0.000	4.249	7.449	?
row3	0.415	17.005	20.205	?
Total	?	?	?	?

$$z = \mathbf{X}^T y + Lap(k/\varepsilon)$$

- Noise from Laplace distribution
- ε-differential privacy
  - ε=0.5 in this example
- *k* margins involved in z
  - Here *k*=2, since overall total not included (then *k*=3)

Statistisk sentralbyrå
Statistics Norway

Differential privacy

- Estimate y from z non-negatively
- Recompute totals
- Possible refit by log-linear modelling

	col1	col2	col3	Total
row1	0.000	6.856	10.056	**16.913**
row2	0.000	4.249	7.449	**11.698**
row3	0.415	17.005	20.205	**37.625**
Total	**0.415**	**28.110**	**37.711**	**66.236**

$$z = \mathbf{X}^T y + Lap(k/\varepsilon)$$

- Noise from Laplace distribution
- ε-differential privacy
  - ε=0.5 in this example
- *k* margins involved in z
  - Here *k*=2, since overall total not included (then *k*=3)

Differential privacy

- Estimate y from z non-negatively
- Recompute totals
- Possible refit by log-linear modelling

	col1	col2	col3	Total
row1	?	?	?	**16.913**
row2	?	?	?	**11.698**
row3	?	?	?	**37.625**
Total	**0.415**	**28.110**	**37.711**	**66.236**

$$z = \boldsymbol{X}^T y + Lap(k/\varepsilon)$$

- Noise from Laplace distribution
- ε-differential privacy
  - ε=0.5 in this example
- *k* margins involved in z
  - Here *k*=2, since overall total not included (then *k*=3)

Differential privacy

- Estimate y from z non-negatively
- Recompute totals
- Possible refit by log-linear modelling

	col1	col2	col3	Total
row1	0.106	7.178	9.629	**16.913**
row2	0.073	4.965	6.660	**11.698**
row3	0.236	15.968	21.422	**37.625**
Total	**0.415**	**28.110**	**37.711**	**66.236**

$$z = \boldsymbol{X}^T y + Lap(k/\varepsilon)$$

- Noise from Laplace distribution
- ε-differential privacy
  - ε=0.5 in this example
- *k* margins involved in z
  - Here *k*=2, since overall total not included (then *k*=3)

Differential privacy

- Estimate y from z non-negatively
- Recompute totals
- Possible refit by log-linear modelling

	col1	col2	col3	Total
row1	0.106	7.178	9.629	**16.913**
row2	0.073	4.965	6.660	**11.698**
row3	0.236	15.968	21.422	**37.625**
Total	**0.415**	**28.110**	**37.711**	**66.236**

$$z = \boldsymbol{X}^T y + Lap(k/\varepsilon)$$

- Noise from Laplace distribution
- ε-differential privacy
  - ε=0.5 in this example
- *k* margins involved in z
  - Here *k*=2, since overall total not included (then *k*=3)

Differential privacy

+ assuming some true margins can be released

	col1	col2	col3	Total
row1	?	?	?	11
row2	?	?	?	12
row3	?	?	?	40
Total	3.286	21.633	35.433	63

- The non-perturbed totals are given high weight (1000)
  - Remain almost unchanged.

$Lap(k/\varepsilon)$ noise

added to k margins

- Noise from Laplace distribution
- ε-differential privacy
  - ε=0.5 in this example
- Noise added to k margins
  - Here k=1

Statistisk sentralbyrå
Statistics Norway

Differential privacy

+ assuming some true margins can be released

	col1	col2	col3	Total
row1	0.000	3.200	7.800	**11**
row2	0.000	3.700	8.300	**12**
row3	4.168	15.616	20.216	**40**
Total	**3.286**	**21.633**	**35.433**	**63**

- The non-perturbed totals are given high weight (1000)
  - Remain almost unchanged.

$Lap(k/\varepsilon)$ noise

added to k margins

- Noise from Laplace distribution
- ε-differential privacy
  - ε=0.5 in this example
- Noise added to *k* margins
  - Here *k*=1

Differential privacy

+ assuming some true margins can be released

	col1	col2	col3	Total
row1	0.000	3.200	7.800	?
row2	0.000	3.700	8.300	?
row3	4.168	15.616	20.216	?
Total	?	?	?	?

- The non-perturbed totals are given high weight (1000)
  - Remain almost unchanged.

$Lap(k/\varepsilon)$ noise

added to k margins

- Noise from Laplace distribution
- ε-differential privacy
  - ε=0.5 in this example
- Noise added to k margins
  - Here k=1

Differential privacy

+ assuming some true margins can be released

	col1	col2	col3	Total
row1	0.000	3.200	7.800	**11**
row2	0.000	3.700	8.300	**12**
row3	4.168	15.616	20.216	**40**
Total	**4.168**	**22.516**	**36.316**	**63**

- The non-perturbed totals are given high weight (1000)
  - Remain almost unchanged.

$Lap(k/\varepsilon)$ noise

added to k margins

- Noise from Laplace distribution
- ε-differential privacy
  - ε=0.5 in this example
- Noise added to *k* margins
  - Here *k*=1

Differential privacy

+ assuming some true margins can be released

	col1	col2	col3	Total
row1	?	?	?	11
row2	?	?	?	12
row3	?	?	?	40
Total	4.168	22.516	36.316	63

- The non-perturbed totals are given high weight (1000)
  - Remain almost unchanged.

$Lap(k/\varepsilon)$ noise

added to k margins

- Noise from Laplace distribution
- ε-differential privacy
  - ε=0.5 in this example
- Noise added to k margins
  - Here k=1

Differential privacy

+ assuming some true margins can be released

	col1	col2	col3	Total
row1	0.728	3.931	6.341	**11**
row2	0.794	4.289	6.917	**12**
row3	2.646	14.296	23.058	**40**
Total	**4.168**	**22.516**	**36.316**	**63**

- The non-perturbed totals are given high weight (1000)
  - Remain almost unchanged.

$Lap(k/\varepsilon)$ noise

added to k margins

- Noise from Laplace distribution
- ε-differential privacy
  - ε=0.5 in this example
- Noise added to *k* margins
  - Here *k*=1

Application of estimated inner frequencies

	col1	col2	col3
row1	0.728	3.931	6.341
row2	0.794	4.289	6.917
row3	2.646	14.296	23.058

- Improve (?) and additivity restore perturbed aggregates
- Calculate other aggregates than those perturbed (matrix mechanism)
- Publish directly as perturbed counts
- Use internally to calculate output on the fly
- Generate public microdata or synthetic data

Illustrate methodology - public microdata

• Five-way frequency table obtained from the German labour force survey

- **Large datasets handled by the R package glmnet. A sparse input matrix has dimension about 500000 × 40000**
  - **Inner cells:** *Lap(2)* noise is added to the inner cells
  - **3-way (no fitting):** *Lap(20)* is added to the three-way margins and those margins are released.
  - **3-way (zero replace):** As above, but negative values are replaced by zeros.
  - **3-way:** *Lap(20)* is added to the three-way margins. The inner cells are non-negatively estimated. Margins are recomputed.
  - **3-way (2-way exact):** *Lap(20)* is added to the three-way margins. The inner cells are non-negatively estimated from the perturbed three-way margins and upweighted (weight is 1000) non-perturbed two-way margins.

Mean absolute deviation

- **calculated from three-way margins**
  - **full dataset including cells with zeros**

    - **Inner cells:** *Lap(2)* noise is added to the inner cells

      23.13
    - **3-way (no fitting):** *Lap(20)* is added to the three-way margins and those margins are released.

      19.92
    - **3-way (zero replace):** As above, but negative values are replaced by zeros.

      13.36
    - **3-way:** *Lap(20)* is added to the three-way margins. The inner cells are non-negatively estimated. Margins are recomputed.

      9.95
    - **3-way (2-way exact):** *Lap(20)* is added to the three-way margins. The inner cells are non-negatively estimated from the perturbed three-way margins and upweighted (weight is 1000) non-perturbed two-way margins.

      6.38

Remarks

- In practise, model selection is important
  - Choosing margins to be perturbed and possible margins to be preserved
- Challenging numerical computations when weighted margins
  - Convergence threshold parameter in glmnet is important
- Exact margins fix bias caused by non-negative estimation
- Refit by log-linear modelling may improve the results
  - Since log-linear model assumptions are reasonable
  - But margins constrained by the model will not improve (three-way margins in above example)
  - Log-linear refit wasn't done for the large dataset in the paper
    - But the problem is manageable by iterative proportional fitting

Some R code can be found at https://github.com/olangsrud/experimentalRpackage

Extra slides

Database reconstruction

z = published statistics

y = database

Relation to matrix mechanism

- Want to protect statistics $X_A^T y$
- Add Laplace noise: $z_B = X_B^T y + Lap(k/\varepsilon)$
- Find $\hat{y} = (X_B^T)^\dagger z_B$
- Calculate protected statistics $z_A = X_A^T \hat{y}$
- Challenge: Optimally select the matrix X_B

Synthetic microdata

- Multinomial sample of size 63

  cell probabilities = (fitted inner frequencies)/63

	col1	col2	col3
row1	0.728	3.931	6.341
row2	0.794	4.289	6.917
row3	2.646	14.296	23.058

	col1	col2	col3
row1	p=0.012	p=0.062	p=0.101
row2	p=0.013	p=0.068	p=0.110
row3	p=0.042	p=0.227	p=0.366

	col1	col2	col3
row1	0	3	7
row2	1	4	9
row3	1	12	26

Public microdata another way

When inner frequencies are whole numbers, this is a representation of a microdataset.

- Keep whole numbers (size 58)
- Add multinomial sample of size 5

cell probabilities = decimals/5

	Col1	col2	col3
row1	0+0.728	3+0.931	6+0.341
row2	0+0.794	4+0.289	6+0.917
row3	2+0.646	14+0.296	23+0.058

	col1	col2	col3
row1	p=0.146	p=0.186	p=0.068
row2	p=0.159	p=0.058	p=0.183
row3	p=0.129	p=0.059	p=0.012

	col1	col2	col3
row1	0+0	3+1	6+0
row2	0+1	4+2	6+1
row3	2+0	14+0	23+0

Large datasets: R package glmnet

- The generalized inverse solution can be approximated by incorporating ridge shrinkage.

- Efficient computation by the R package glmnet
  - sparse x-matrix
    - Matrix in the example below has dimension about 500000 × 40000
  - ridge shrinkage
  - non-negative estimation