eurostat

EUROPEAN MASTER IN OFFICIAL STATISTICS
EMOS

Organiser:
Department of Economics and Management
University of Pisa

10 March 2020

Statistical Disclosure Control:
Where do we go from here?

Natalie Shlomo, University of Manchester

Facilitator: Caterina Giusti, University of Pisa

EMOS Webinar

March 10th 2020

Statistical Disclosure Control: Where do we go from here?

Natalie Shlomo

Professor of Social Statistics

The University of Manchester

Topics

- Overview of types of disclosure risk in traditional forms of statistical data
- Common statistical disclosure limitation methods
- Disclosure risk-data utility paradigm
- Inferential disclosure and differential privacy
- New dissemination strategies:
    - Online flexible table builder
    - Other open data options
- Discussion

Traditional Statistical Outputs

- Survey Microdata
  - Social surveys (census/register and business survey microdata generally not released)
  - Available from data archives for registered users
- Tabular Data

Frequency Tables	Magnitude Tables
Census/registers (whole population) counts	Business Statistics, eg., total turnover
Weighted sample counts	

Types of Disclosure Risks

Identity Disclosure

Identification is widely referred to in confidentiality pledges and code of practice

Individual Attribute Disclosure

Confidential information about a data subject is revealed and can be attributed to the subject (Identity disclosure a necessary pre- condition)

Group Attribute Disclosure

Confidential information is learnt about a group and may cause harm

Common SDC Methods

Social Survey Microdata

Identity Disclosure (assume no response knowledge)-
rare categories of identifying variables (population unique)

Recoding/grouping identifying variables, eg. k-anonymity

Suppressing variables such as high level geographies

Sub-sampling, eg. census samples

Attribute disclosure - individual(s) identified and survey target variables learnt, eg. health, income

Top-coding sensitive variables

Recoding / Microaggregation, eg. l-diversity

Common SDC Methods

Frequency Tables (whole population counts)

Identity Disclosure –small cells

Table design, eg. spanning variables and grouped categories

Minimum population thresholds

Attribute disclosure - zeros in row/column and one populated cell

Pre-tabular and/or post-tabular perturbation to introduce ambiguity in zero cells

Nested tables to avoid disclosure by differencing

Common SDC Methods

Magnitude Tables (Business statistics)

Assumptions:

- Intruders are competitors in the cell and can form coalitions
- Businesses in a cell are known
- The ranking of the businesses with respect to their size is known

Table design

Minimum population thresholds

Cell suppression: primary and secondary

(mathematical programming and optimization)

Attribute disclosure - What can a competitor learn with sufficient precision

Disclosure Risk and Data Utility

Disclosure risk

Frequency tables:

Whole population counts and disclosure risk is visible: small cells, placement of zero cells

Let $F = \{F_1, F_2,\ldots, F_K\}$

$H\left(\frac{F}{N}\right) = -\sum_k \frac{F_k}{N}\log(\frac{F_k}{N})$ and

$1 - [\frac{H(\frac{F}{N})}{logK}]$

Microdata:

Set of cross-classified quasi-identifiers defined by k=1,…,K

$$\sum_k I(f_k = 1, F_k = 1)$$

where

f_k sample count

F_k population count

Probabilistic modelling for estimation: Poisson-log linear modelling

Magnitude tables (Business statistics):

Let $T_k = \sum_{i\epsilon k} x_i$ in cell k

(n,p) Dominance Rule classifies cell as disclosive if

$$x_{(1)} + \ldots + x_{(n)} \geq (p/100) \times T_k$$

Disclosure Risk and Data Utility

Utility

- Impact on variance
- Impact on bias

Distortions to distributions:
distance metrics, eg.
Hellinger's Distance*,
variation in propensity scores

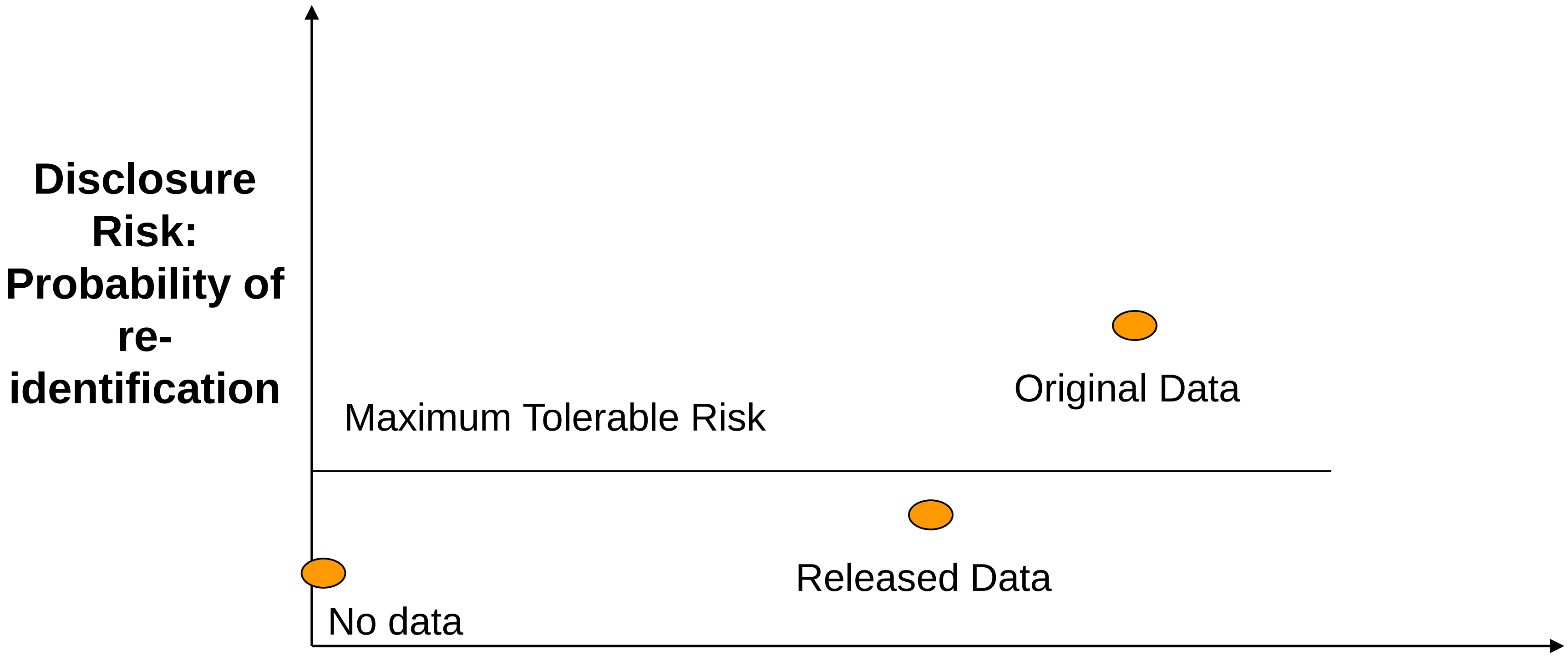
Changes in inference:
confidence interval overlap,
change in χ^2 or R^2

Changes in associations:
change in correlations and
rankings, Cramer's V

$$* \quad HD(F, F') = \sqrt{\frac{1}{2}\sum_{k=1}^{K}\left(\sqrt{F_k} - \sqrt{F'_k}\right)^2}$$

Disclosure Risk and Data Utility

R-U Confidentiality Map (Duncan, et.al. 2001)



Data Utility: Quantitative measure on the statistical quality

Questions

Inferential Disclosure

Confidential information may be revealed exactly or to a close approximation with high confidence from statistical properties of released and combined data

Examples:

Survey microdata – a good prediction model with very high R^2

Census tables – disclosure by differencing and linking tables

This type of disclosure has largely been ignored and dealt with through strict control of data that is released

- Microdata deposited in archives for registered users
- Strict control of tabular data, eg. review boards for special request tabulations

Where do we go from here?

- Traditional forms of statistical data and their confidentiality protection rely heavily on assumptions that may no longer be relevant

Digitalization of all aspects of our society leading to new and linked data sources offering opportunities for research and evidence-based policies

With detailed personal information easily accessible from the internet, traditional SDL may no longer be sufficient and agencies relying more on restricting and licensing data

- Growing demand for more open and accessible data via web-based applications
- Need for more rigorous data protection mechanisms with stricter privacy guarantees
- Collaborations with computer scientists through scientific programs

Differential Privacy

- Computer Science differential privacy (Dwork and Roth 2014): the intruder has knowledge of entire database except for one target unit ("worst case" scenario)

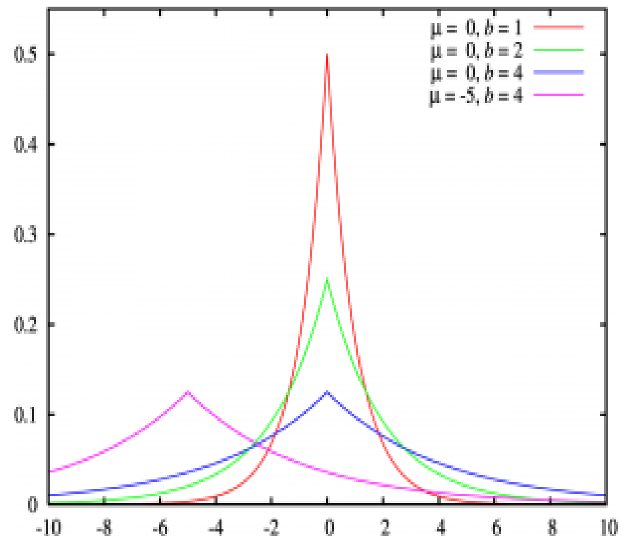
Definition: Mechanism M satisfies (ε, δ)-differential privacy if for all neighbouring databases D,D' differing by one individual, all possible queries q and $S \subseteq Range(M)$ all possible outputs:

$$P(M(q(D))\epsilon S) \leq e^{\varepsilon} P(M(q(D'))\epsilon S) + \delta$$

and the probability is taken over the randomness of the mechanism

If $\delta = 0$ then we have ε-differential privacy

Example of Differential Privacy Mechanism



Laplace Mechanism

Calibrating noise: what scale of noise b is large enough to ensure privacy on a query q?

q(D)+Z and Z sampled from Lap(0,b)

Amount of noise depends on ε and sensitivity of q
denoted Δq

$$\Delta q = max_{D,D'} |q(D) - q(D')|$$

where D D' **any** neighbouring databases

Example:

query	Δq
count	1
max(age)	120
avg(age)	120/n

Theorem: setting scale (b) of Laplace noise to $\Delta q/\varepsilon$ ensures ε-differential privacy

Mechanisms in Differential Privacy

Non-interactive Mechanism

Data custodian produces a 'safe' object, such as a synthetic database or collection of summary statistics

After this *release* all post-perturbative analyses are safe (no privacy budget spent after the original object)

Interactive Mechanisms

Data analyst sends queries (functions applied to a database) adaptively, deciding which query to pose next based on observed responses to previous queries

Accuracy will deteriorate with the number of questions asked, and providing accurate answers to all possible questions will be infeasible

Differential Privacy in the SDC Tool-kit at Statistical Agencies

Non-interactive mechanisms as agencies unable to monitor queries

DP useful when perturbative methods are needed with stricter privacy guarantees such as outputs disseminated via the internet where agencies relinquish control of the releases

Examples: flexible table builder, synthetic data, and multiple data products released from survey microdata

Agencies should still maintain 'safe data' and 'safe access' SDC approaches, eg. Data Labs for 'trusted' users

Differential Privacy vs. SDC

No distinction between key variables and sensitive variables, types of disclosure risks, sample or population or prior intruder knowledge

Designed for output perturbation and in this case a sum/average is disclosive and needs to be protected (same as disclosure by differencing)

Zeroes need to be perturbed

Perturbation mechanism not hidden and can be used to correct statistical analysis

Questions

Online Flexible Table Builder

Online Flexible Table Builder

- Increasing demands for online dissemination and open access of census tables (ABS, USA, EU)

Australian Bureau of Statistics

  - Web-based platform (drop down lists) with restrictions: number of dimensions, population thresholds, no sparse tables
  - SDL on-the-fly: pre-tabular (hypercubes, swapping) and/or post-tabular methods (noise addition, rounding)
- Perturbation matrix $p_{ij} = P(perturb\ cell\ to\ j | original\ cell\ is\ i)$
- Change (or do not change) value according to p_{ij} and random draw

Online Flexible Table Builder

- Other principles in SDC:

Perturbations unbiased, bounded, maximal entropy, non-negative and zeros not perturbed

Microdata keys for same perturbations on same cells across tables (Fraser and Wooton 2005)

Additivity - probability perturbation matrix with property of 'invariance' (ensures margins in expectations) and IPF (Shlomo and Young 2008)

- Differential Privacy (DP) for flexible table builders (Rinott, O'Keefe, Shlomo and Skinner 2018)

Exponential Mechanism

Exponential mechanism defined by: given a, choose $b \in B$ (B: range of b) with probability proportional to: $e^{(\varepsilon/2)u/\Delta u}$ where

$$\Delta u = \max_{b \in B} \max_{a \sim a' \in A} |u(\mathrm{a},\mathrm{b}) - u(\mathrm{a}',\mathrm{b})|$$

Assuming additive loss functions and independent perturbations

Bound the perturbations $P(M(a) = b) < \delta$, then for all $a \sim a' \in A$, if $P(M(a') = b) = 0$ implies $|a_k - b_k| \le m, \forall k$ then $M(.)$ satisfies $DP(\varepsilon, \delta)$

Examples of Laplace perturbation vectors:

$\varepsilon = 1.5, \delta = 0.00002$

-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
0.00002	0.00008	0.00035	0.00157	0.00706	0.03162	0.14172	0.63516	0.14172	0.03162	0.00706	0.00157	0.00035	0.00008	0.00002

$\varepsilon = 0.5, \delta = 0.008$

-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
0.0076	0.0125	0.0206	0.0339	0.0559	0.0922	0.1520	0.2506	0.1520	0.0922	0.0559	0.0339	0.0206	0.0125	0.0076

Exponential Mechanism

Original Value	Range for ε=1.5 and δ=0.00002					Range for ε =0.5 and δ =0.008				
	$\pm$ 0	$\pm$ 1	$\pm$ 2	$\pm$ 3	$\pm$ 4	$\pm$0	$\pm$1	$\pm$2	$\pm$3	$\pm$4
	Laplace m=7					Laplace m=7				
0	0.82	0.96	0.99	1	1	0.63	0.78	0.87	0.93	0.96
1	0.64	0.96	0.99	1	1	0.25	0.77	0.86	0.92	0.95
2	0.64	0.92	0.99	1	1	0.25	0.55	0.85	0.91	0.94
3	0.64	0.92	0.98	1	1	0.25	0.55	0.74	0.88	0.92
4	0.64	0.92	0.98	1	1	0.25	0.55	0.74	0.85	0.88
$\geq$ 5	0.64	0.92	0.98	1	1	0.25	0.55	0.74	0.85	0.92
	Normal m=12					Normal m=10				
0	0.57	0.7	0.81	0.89	0.94	0.54	0.63	0.71	0.78	0.84
1	0.14	0.7	0.81	0.89	0.94	0.09	0.62	0.7	0.78	0.84
2	0.14	0.4	0.81	0.89	0.94	0.09	0.26	0.69	0.76	0.82
3	0.14	0.4	0.62	0.89	0.94	0.09	0.26	0.42	0.74	0.8
4	0.14	0.4	0.62	0.78	0.94	0.09	0.26	0.42	0.57	0.78
$\geq$ 5	0.14	0.4	0.62	0.78	0.88	0.09	0.26	0.42	0.57	0.69

*Negative values to 0

Exponential Mechanism

Implications:

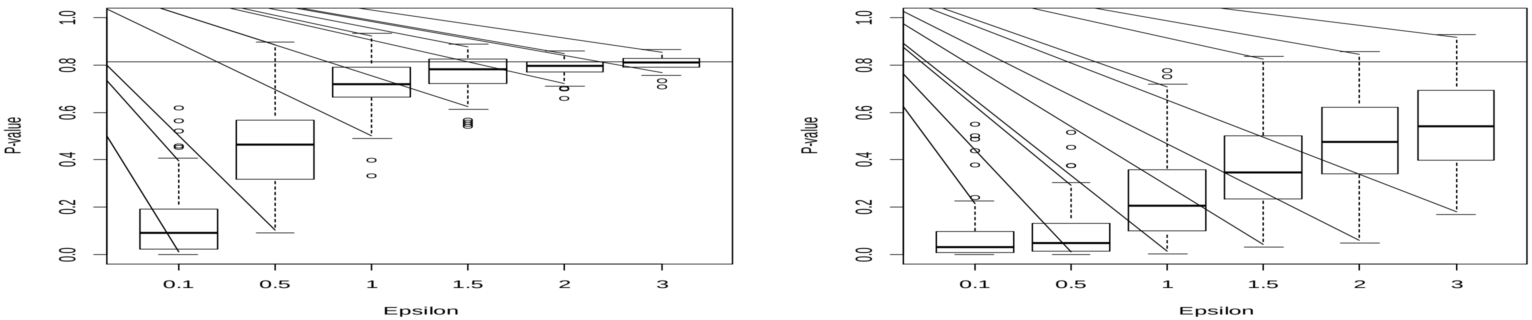
- DP leads to negative values, setting to zero still ensures DP but biased perturbations
- All (non-structural) zeroes must be perturbed
- If list-space has internal cells only $\Delta u = 1$, margins summed from internal cells DP but low utility
- In a *t*-way table all margins, $\Delta u = 2^t - 1$ (not including total) much larger perturbations implying smaller utility
- Margins can be perturbed (with appropriate sensitivity) and prorated to ensure additivity (post-processing does not violate DP)

Parameters of Differential Privacy not secret and can be used to adjust statistical analysis

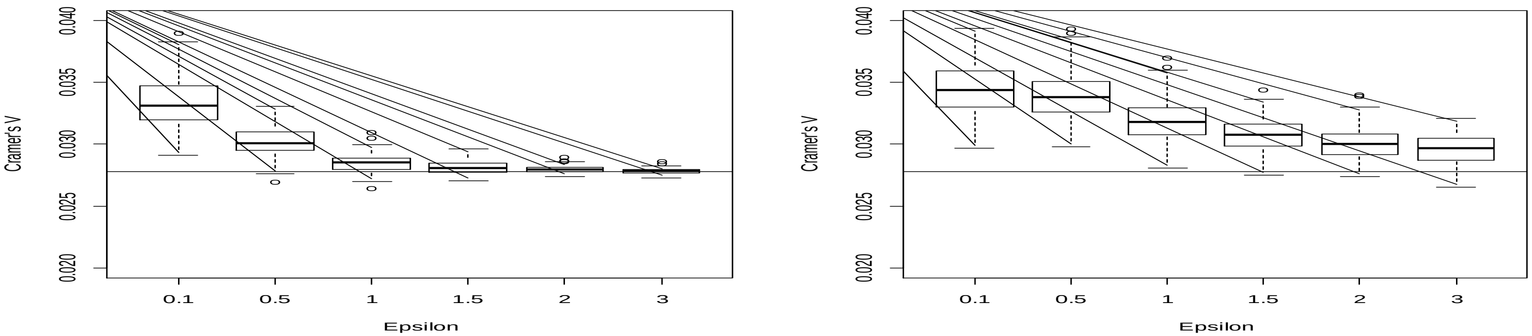
Generated independent table, N=10000, K=100 (average cell size=100)

Laplace Perturbations | **Normal Perturbations**

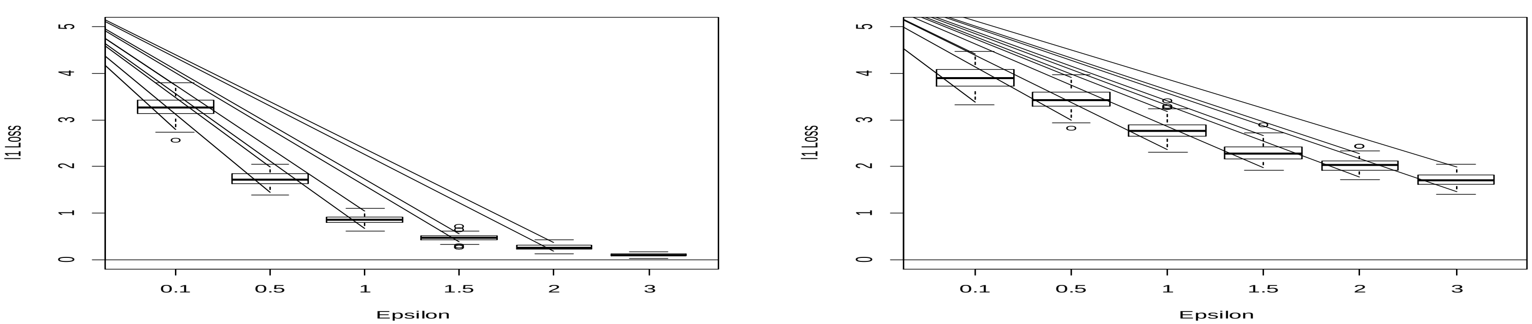
P-Value



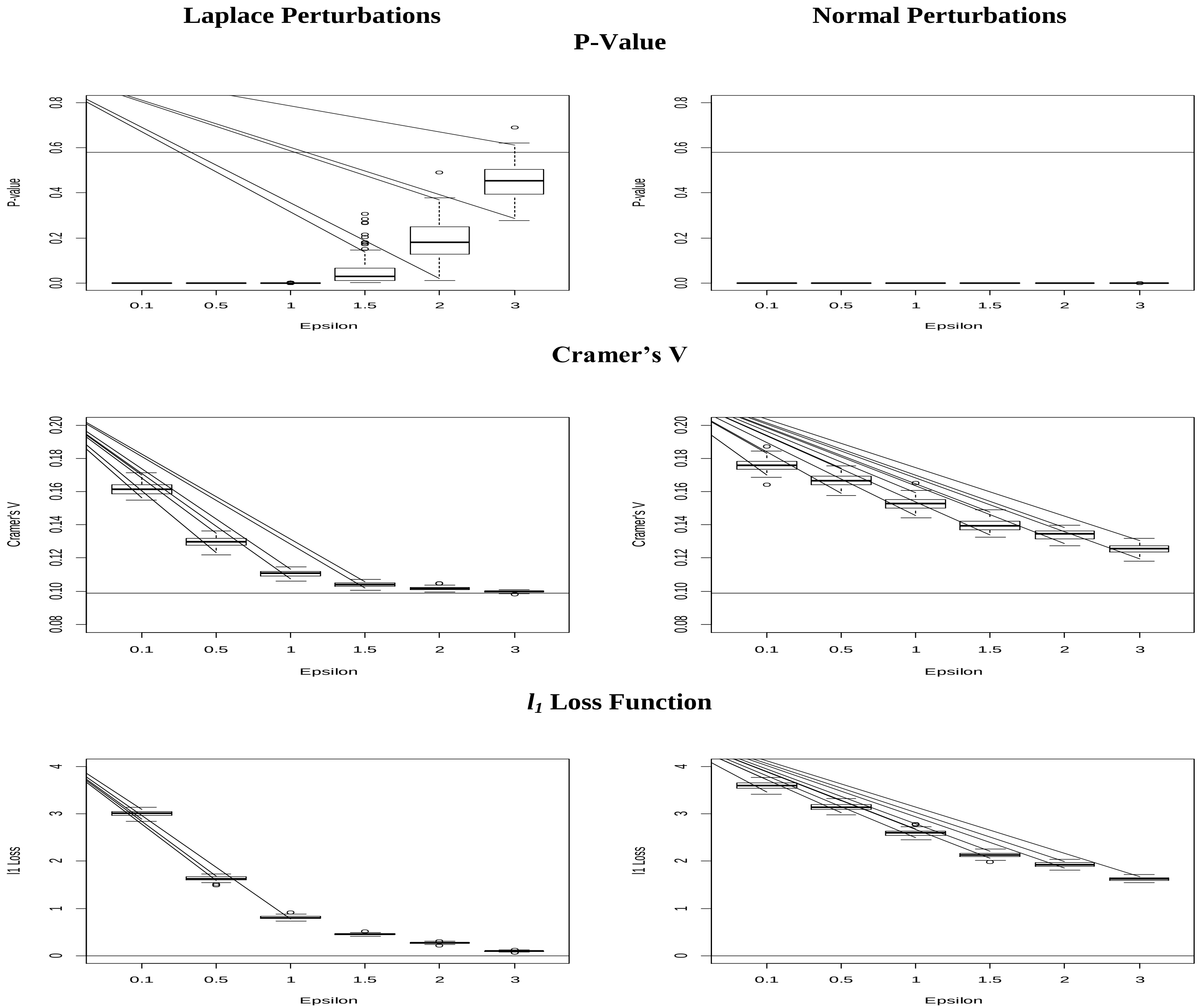
Cramer’s V



l_1 Loss Function



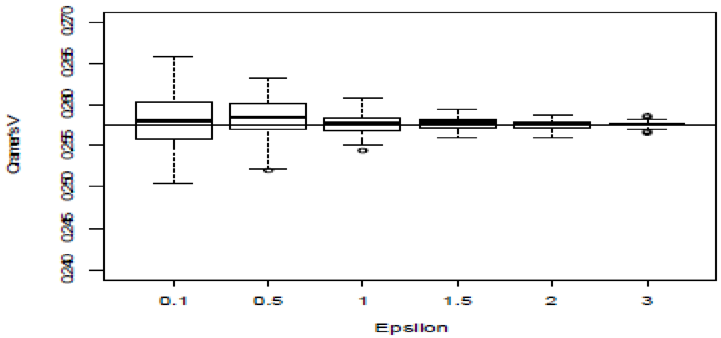
Generated independent table, N=10000, K=1000 (average cell size=10)

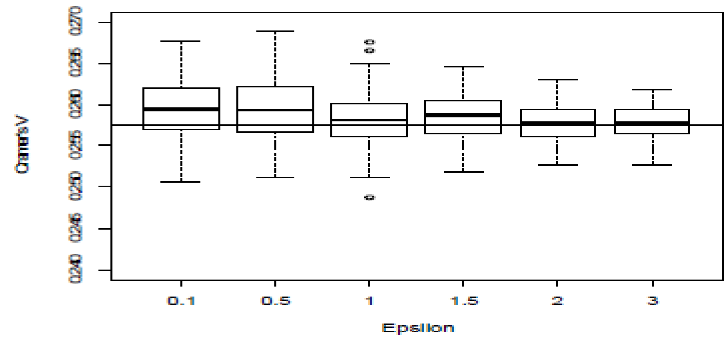


Real (dependent) Table from UK Census Data

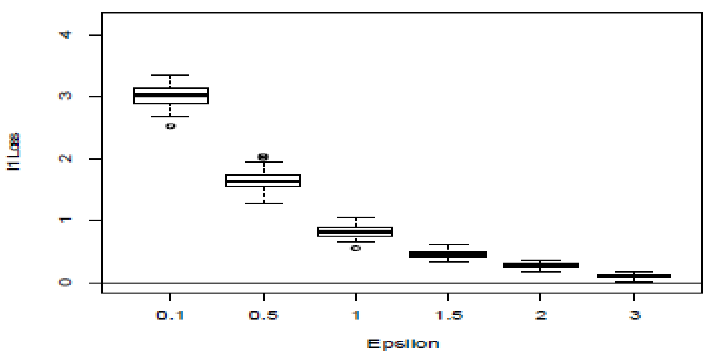
Laplace Perturbations | **Normal Perturbations**

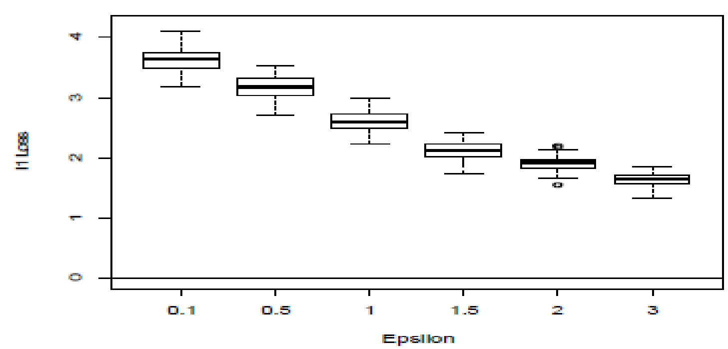
Cramer's V



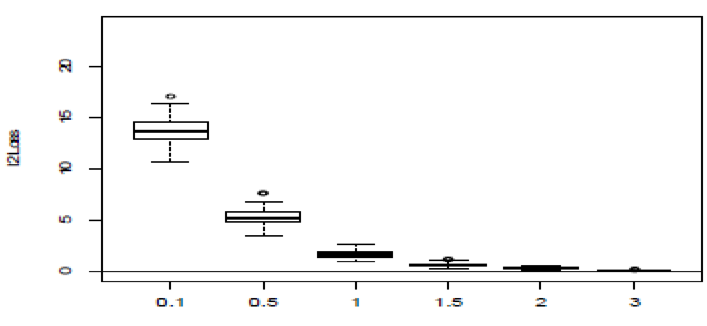


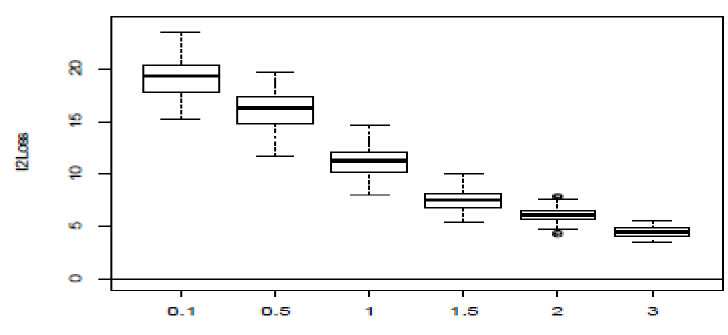
l_1 Loss Function





l_2 Loss Function





Other Open Data Options

Other Open Data

Synthetic Data

- Fit models from original data, eg. posterior predictive distributions Can be implemented on parts of data where a mixture is obtained of real and synthetic data
- Draw and release several samples to account for the uncertainty and obtain 'proper' variance estimates (Reiter 2005)

- In practice, difficult to capture all conditional relationships between variables and within sub-populations
  - If models of interest are sub-models of the synthesis model, then the analysis of (multiple) synthetic samples should give valid inferences

Differential Privacy for Synthetic Data

- Synthetic data

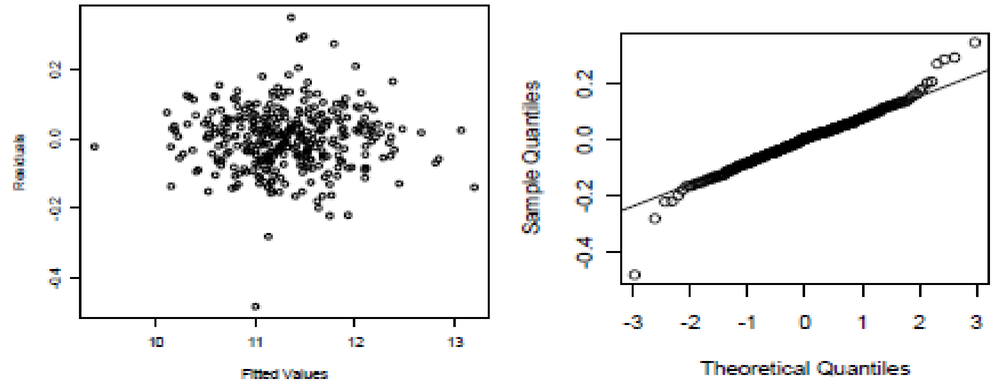
Ongoing Research:

- Bayesian Modeling with differentially private priors
- Current work on adding noise to estimating equations and also looking at ridge regression to regularize linear regression by adding a constraint to likelihood function: use in Sequential Regression modeling (Ragunathan et al. 2001)
- Reproducing microdata from differentially private counts

Other Open Data

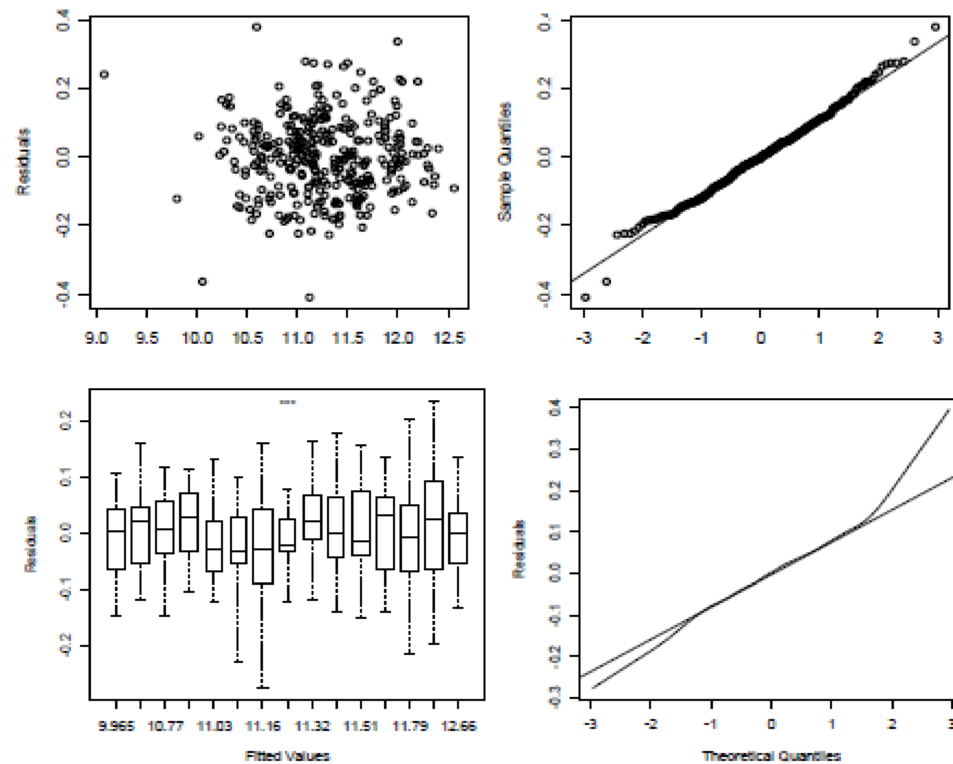
Remote Analysis

- Initial research in developing platforms for remote analysis or allowing researchers to submit code
- Aim to protect outputs without the need for human intervention



(a) Residuals vs Fitted Values

(b) Normal Q-Q Plot of Residuals



(a) Residuals by fitted values

(b) Normal QQ Plot of Residuals

Challenges and Discussion

Differential Privacy with formal privacy guarantees may provide solutions for SDC

Allows statistical agencies to consider new ways of disseminating open data via the internet

It provides a formal 'by-design' privacy guarantee against inferential disclosure

Combined with other SDC approaches of coarsening, subsampling, variable suppression etc. impacts on the privacy budget

Further research is needed to set these privacy budgets

Additive noise perturbation of DP can provide more utility than other additive SDC noise perturbations

Agencies should release parameters of the perturbation and DP parameters are not secret and can be used to adjust analyses

References

Abowd, J.M. and Vilhuber, L., (2008). How Protective Are Synthetic Data? In *PSD'2008 Privacy in Statistical Databases*, (Eds. J.Domingo-Ferrer and Y. Saygin), Springer LNCS 5262, 239-246.

Antal, L., Shlomo, N. and Elliot, M. (2014). Measuring Disclosure Risk with Entropy in Population Based Frequency Tables. In Privacy in Statistical Databases 2014, (Ed. J. Domingo-Ferrer), Springer LNCS 8744, pp. 62-78.

Dandekar, R.A. and Cox L. H. (2002). Synthetic Tabular Data: An Alternative to Complementary Cell Suppression. *Manuscript, Energy Information Administration*, U. S. Department of Energy.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science 9, 211-407.

Fraser, B. and Wooton, J. (2005). A Proposed Method for Confidentialising Tabular Output to Protect Against Differencing. *Joint UNECE/Eurostat work session on statistical data confidentiality*, Geneva, 9-11 November.

McSherry, F. and Talwar, K. (2007). Mechanism Design via Differential Privacy. In *Foundations of Computer Science*, 2007, *FOCS'07*, *48th Annual IEEE Symposium on 94-103*. IEEE, New York.

O'Keefe, C.M. and Shlomo, N. (2012). Comparison of Remote Analysis with Statistical Disclosure Control for Protecting the Confidentiality of Business Data. *Transactions on Data Privacy*, Vol. 5, Issue 2, 403-432.

Raghunathan T.E., Lepkowksi J.M., van Hoewyk J., Solenbeger P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology, Vol. 27, 85-95.

Reiter, J.P. (2005), Releasing Multiply Imputed, Synthetic Public-Use Microdata: An Illustration and Empirical Study. *Journal of the Royal Statistical Society*, A, Vol.168, No.1, 185-205.

Rinott, Y., O'Keefe, C., Shlomo, N., and Skinner, C. (2018). Confidentiality and Differential Privacy in the Dissemination of Frequency Tables. *Statistical Sciences*, Vol. 33, No. 3, 358-385.

Shlomo, N. and Skinner. C.J. (2012). Privacy Protection from Sampling and Perturbation in Survey Microdata. *Journal of Privacy and Confidentiality*, Vol. 4, Issue 1.

Shlomo, N. and Young, C. (2008). Invariant Post-tabular Protection of Census Frequency Counts. In *PSD'2008 Privacy in Statistical Databases*, (Eds. J.Domingo-Ferrer and Y. Saygin), Springer LNCS 5262, 77-89.

Questions