Working Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Ottawa, Canada, 28-30 October 2013)

Topic (ii): New methods for protection of microdata

Synthetic Contingency Tables for Confidentiality Protection: Offering Guarantees when Sampling from Posterior Distributions

Prepared by Anne-Sophie Charest, Université Laval, Canada

Synthetic Contingency Tables for Confidentiality Protection: Offering Guarantees when Sampling from Posterior Distributions

Anne-Sophie Charest*

* Department of Mathematics and Statistics, Université Laval, Québec city, QC, Canada, anne-sophie.charest@mat.ulaval.ca

Abstract. We consider the problem of releasing microdata from categorical variables in the form of large confidentialised contingency tables. Several algorithms have been designed for this purpose, most of which are based on the idea of multiple imputation for nonresponse. While some offer some strong guarantee of privacy protection, such as differential privacy, others merely rely on the published observations not corresponding directly to real respondents. We consider here the difficulty of offering confidentiality guarantee when creating synthetic datasets by sampling from posterior predictive distributions, and propose a possible general mechanism for releasing differentially-private synthetic datasets using this method.

1 Introduction

Data collectors usually have a responsibility to protect the personal information of their respondents when they publish datasets or results from statistical analyses. This responsibility is sometimes imposed by law, as for the census data, or the consequence of a confidentiality promise made to the respondent during data collection.

Varied approaches are used to offer this confidentiality protection. The interested reader can see Duncan et al. (2011) and Hundepool et al. (2012) for a survey of the methods, and references therein for details. One may broadly divide these Statistical Disclosure Control (SDC) methods into two categories: those which publish protected versions of summary statistics or statistical output, and those which publish protected versions of the microdata. Our preference is for methods of the second kind, as they do not require that the data disseminators know which models are of interest for all of the data users. This paper is thus concerned with the creation of such protected datasets, which we refer to as synthetic datasets. We moreover focus on the case of contingency tables, though the main ideas may be applicable for other types of data.

A general approach to creating synthetic datasets is based on the idea of multiple imputation for non-response, and was proposed in Rubin (1993). It simply involves

modeling the dataset with some appropriate Bayesian model, and then sampling a synthetic dataset from the posterior predictive distribution. This offers a very flexible framework for the creation of synthetic datasets, and has been the source of an important amount of research in the SDC community (see Reiter (2011) and Dreschler (2012) for reviews of SDC using synthetic datasets).

The privacy guarantee of such mechanism has however not been well-established. It was originally assumed that the fact that the individuals in the released dataset did not correspond to any individual in the real population was sufficient to ensure confidentiality protection. While this helps reduce re-identification of the respondents, inferential disclosure is still a concern, that is one could infer with certainty or with high probability the characteristics of an individual of interest in the population from the published synthetic dataset and possibly auxiliary information.

In the reminder of this paper, we will discuss an approach to generating synthetic contingency tables with provable confidentiality guarantee protection, which is based on the rigorous criterion of differential privacy. We first present different models for generating synthetic contingency tables in section 2. Section 3 introduces the criterion of differential privacy and its application to the context of generating synthetic data from a posterior predictive distribution, and our proposal to create differentially-private synthetic datasets. We offer a few concluding remarks in section 4.

2 Synthetic Contingency Tables Generation

The only published model for generating synthetic contingency tables from a posterior predictive distribution is the simple Dirichlet-Multinomial model. In this case, we simply consider the contingency table as a vector of counts $X = (x_1, \ldots, x_p)$, where $\sum_{i=1}^{p} X_i = n$, which we model with a Multinomial(n, π) likelihood and the conjugate prior Dirichlet(α). Creating a synthetic dataset is as follows:

1. Sample parameters for the posterior distribution

$$\tilde{\pi} \sim \text{Dirichlet}(\alpha + X)$$

2. Sample a synthetic dataset

$$\tilde{X} \sim \text{Multinomial}(\tilde{n}, \tilde{\pi})$$

The confidentiality guarantee of such a mechanism will depend on the choice of the hyperparameter α: a larger α represents more *a priori* information and thus reduce the influence of the real dataset on the distribution of $\tilde{X}$, thereby limiting the influence that any one individual can have on the published output, and thus increasing the confidentiality guarantee. This will be made more rigorous in section 3.

The Dirichlet-Multinomial model clearly does not take the relationships between the different variables in the dataset into account, and may create synthetic datasets with bad inferential properties (see for e.g. Charest (2012)). A natural extension of this model is to use a bayesian log-linear model for the counts in the contingency table and create a synthetic dataset by sampling from its posterior predictive distribution. This method was briefly considered in Charest (2012), but we are not aware of other published work using log-linear models to generate synthetic contingency tables. The main difficulty in this case is the choice of the specific log-linear model to use, which is very difficult for high-dimensional datasets, particularly if the data is sparse as is often the case in real applications.

A much more flexible nonparametric bayesian model was proposed in Dunson and Xing (2009) and has recently been used for nonresponse imputation in Reiter and Si (2013), namely the Dirichlet Process Mixture of Product Multinomial (DPMPM). Consider a contingency table with d variables, and d_j categories per variables. Let X_{ij} be the value of variable j for individual i, for $i = 1, \ldots, n$ and $j = 1, \ldots, d$. The values of X_{ij} are assumed without loss of generality to be in the set $1, 2, \ldots, d_j$. The DPMPM is then as follows:

$$
\begin{aligned}
X_{ij}|z_i, \phi &\sim \text{Multinomial}(\phi_{z_i j 1}, \ldots, \phi_{z_j j d_j}), \forall i, j \\
z_i|\pi &\sim \text{Multinomial}(\pi_1, \ldots, \pi_\infty), \forall i \\
\pi_h &= V_h \prod_{g<h} (1 - V_g), \text{for } h = 1, \ldots, \infty \\
V_h &\sim \text{Beta}(1, \alpha) \\
\alpha &\sim \text{Gamma}(a_\alpha, b_\alpha) \\
\phi_{hj} &= (\phi_{hj1}, \ldots, \phi_{hjd_j})
\end{aligned}
$$

This model presents several advantages from the point of view of the creation of synthetic contingency tables. As shown in Dunson and Xing (2009), this model "denes a prior with full support on the space of distributions for multiple unordered categorical variables" and thus allow for any dependence structure between the variables. This also means that model selection need not be performed prior to the creation of the synthetic dataset, but is driven by the data as part of model fitting. In addition, posterior distributions for all of the parameters can be obtained using a simple Gibbs sampler. The sampler proposed in Reiter and Si (2013) is particularly simple to implement as it truncates the infinite mixture to a set (large) number of components to reduce the computation burden. We refer readers to their paper for implementation details.

We believe that it model is especially well designed for the task of releasing synthetic contingency tables. Indeed, simulation results by the author show promising results for the use of this model to create synthetic datasets with regards to the utility of such datasets, and should be the topic of future publication.

As a consequence, we think that the use of posterior predictive distributions to create synthetic contingency tables may gain more interest in the SDC literature. It will thus become more important to understand and quantify the confidentiality guarantee offered by such methods, and perhaps to modify them in order to offer strict confidentiality guarantees. This is the motivation for the ideas presented in the next section.

3 Privacy Guarantee

A significant portion of the literature on privacy guarantees for SDC methods focuses on estimating the probability of re-identification of individuals using various assumptions about the information available to the intruder, and these methods have been used successfully in several contexts. In the case of synthetic datasets however, the concern is not re-identification per say, but on the risk of inferential disclosure. A notable attempt at estimating this risk can be found in section 4 of McClure and Reiter (2012), and we believe that similar work will be necessary and useful for different data generation process. For this paper, we will however approach privacy guarantee from the point of view of the rigorous criterion of differential privacy.

Differential Privacy

Differential privacy was proposed in Dwork (2006) as a criterion to measure the confidentiality guarantee offered by any SDC mechanism. It is a very strict criterion which protects the information of any respondent in the database against an adversary with arbitrary auxiliary information, and in particular complete knowledge of the rest of the dataset. Formally, we say that a randomized function κ gives ϵ-differential privacy if and only if for all neighboring datasets D_1 and D_2, and for all $S \subseteq \text{range}(\kappa)$,

$$e^{-\epsilon} \leq \frac{Pr[\kappa(D_1) \in S]}{Pr[\kappa(D_2) \in S]} \leq e^{\epsilon}.$$

Two datasets are said to be neighboring datasets if they differ only on the values of one of the respondents. Also, ϵ characterizes the level of protection, smaller values of ϵ indicating greater protection.

Differentially-Private Synthetic Datasets

The definition of differential privacy allows any type of function κ. In the context of the generation of synthetic datasets, we can consider κ to take as input a true dataset from the set of possible datasets, and return a synthetic dataset, again from the space of possible datasets.

For the Dirichlet-Multinomial model described in section 2, we can achieve differential privacy simply by through the choice of the hyperparameter α. Indeed, one

can show this mechanism will give ϵ-differential if and only if

$$\alpha_i \geq \frac{\tilde{n}}{\exp(\epsilon) - 1} \; i \in 1, \ldots, p$$

where $\alpha = (\alpha_1, \ldots, \alpha_p)$. (Machanavajjhala et al. (2008))

No similar results have been obtained for synthetic datasets based on posterior distributions from log-linear models, or the DPMPM. In fact, the calculations required to show the result in the case of the Dirichlet-Multinomial model may prove too difficult in these cases. Instead, we present here a general method to obtain differentially-private synthetic contingency tables based on posterior predictive distributions. This method relies on the exponential mechanism, which we describe below.

Exponential Mechanism

The exponential mechanism was proposed in McSherry and Talwar (2007) as a general algorithm to sample outputs for publication in a differentially-private manner. Consider a dataset consisting of n inputs each from domain $\mathcal{X}$, with the goal to release an output in the range $\mathcal{R}$. Then, we must define a score function $q : \mathcal{D}^n \times \mathcal{R} \to \mathbb{R}$, which assigns a real-valued score to any pair (x, r) in $\mathcal{X}^n \times \mathcal{R}$. Any function $q(x, r)$ may be used, but in all cases a high score should mean that output r is an appealing output to release for the input x. An ϵ-differentially-private output r can be obtained by sampling r with probability proportional to

$$\exp\left(\frac{\epsilon}{2\Delta q} q(x, r)\right) \times \mu(r)$$

where μ is a base measure on $\mathcal{R}$ and

$$\Delta q = \max_{r \in \mathcal{R}, x_1, x_2 \text{neighbors}} |q(x_1, r) - q(x_2, r)|.$$

Proposed Methodology

Let $\mathcal{X} = (x_1, \ldots, x_n)$ be the database that we want to publish, and $l(x_i; \theta)$ represent the log-likelihood function evaluated at x_i, where θ is a vector of unknown parameters. Also, assume a prior distribution $p(\theta)$ for θ. Without the requirement of differential privacy, one could sample a synthetic dataset $\tilde{x}$ to publish directly from the posterior predictive distribution

$$\begin{aligned} p(\tilde{x}|x_1, \ldots, x_n) &= \int p(\tilde{x}, \theta \mid x_1, \ldots, x_n) d\theta \\ &= \int p(\tilde{x} \mid \theta, x_1, \ldots, x_n) p(\theta \mid x_1, \ldots, x_n) d\theta \end{aligned}$$

according to one of the models presented in section 2.

To satisfy differential privacy, we suggest to use the exponential mechanism and incorporate the posterior predictive distribution in the score. An obvious choice for the score function is simply to set

$$q(d, r) = q(x_1, \ldots, x_n, \tilde{x}) = p(\tilde{x}|x_1, \ldots, x_n).$$

The sampling step required in the exponential mechanism can be carried out using rejection sampling.

Simple Illustration

Consider the case where $x_i \in \{0, 1\}$ for $i = 1, \ldots, n$. We assume that $x_i \sim \text{Binomial}(1, \theta)$ and suppose a prior $p(\theta) = \text{Beta}(\gamma_1, \gamma_2)$. The posterior predictive distribution is then

$$\begin{aligned} P(\tilde{x} = a|x = b) &= \int_\theta \underbrace{P(\tilde{x} = a|\theta = p)}_{\text{Bin}_{(n\ ,\ p)}} \underbrace{P(\theta = p|x = b)}_{\text{Beta}_{(\gamma_1+b\ ,\ \gamma_2+n-b)}} \, d\theta \\ &= \binom{n}{a} \frac{\Gamma(n + \gamma_1 + \gamma_2)}{\Gamma(b + \gamma_1)\Gamma(n - b + \gamma_2)} \frac{\Gamma(\gamma_1 + a + b)\Gamma(\gamma_2 + 2n - a - b)}{\Gamma(\gamma_1 + \gamma_2 + 2n)} \end{aligned}$$

Hence, for a given input x, if $q(d, r) = p(\tilde{x}|x_1, \ldots, x_n)$, we must sample an output r with probability

$$\frac{\exp\left(\frac{\epsilon}{\Delta q}\binom{n}{r}\frac{\Gamma(n+\gamma_1+\gamma_2)}{\Gamma(x+\gamma_1)\Gamma(n-x+\gamma_2)}\frac{\Gamma(\gamma_1+r+x)\Gamma(\gamma_2+2n-r-x)}{\Gamma(\gamma_1+\gamma_2+2n)}\right)}{\sum_{r=0}^{n}\exp\left(\frac{\epsilon}{\Delta q}\binom{n}{r}\frac{\Gamma(n+\gamma_1+\gamma_2)}{\Gamma(x+\gamma_1)\Gamma(n-b+\gamma_2)}\frac{\Gamma(\gamma_1+r+x)\Gamma(\gamma_2+2n-r-x)}{\Gamma(\gamma_1+\gamma_2+2n)}\right)}$$

This will ensure that the most probably outcome to be sampled by the exponential mechanism is the one with the highest posterior probability, but not that it will be the expected value of the exponential mechanism. Indeed, numerical experiments indicate that it is better to use the log of the posterior predictive probabilities in this case.

Set $n = 50$, $d = x = 30$, $\epsilon = 0.5$, and $\gamma_1 = \gamma_2 = 1$. We calculate the probabilities for each of the possible synthetic datasets $\{0, 1, \ldots, 50\}$ directly using the exponential mechanism and 3 choices of score functions. Table 1 gives summary statistics of these distributions. Using the log posterior probabilities yields datasets with both higher posterior probabilities and smaller expected distance to the true dataset than the raw probabilities. We also see that $q = -|d - r|$ outperforms both score functions in all metrics. However, we believe that this is caused by the simplicity of the dataset under consideration. For a contingency table, datasets with high posterior probabilities may not be the closest in Euclidean distance to the true dataset and we expect $q(d, r) = \log(P(\tilde{x} = r|x = d))$ our method to outperform $q = -|d - r|$. More work is needed to extend our conclusions to other models.

Note that the choice of the log probabilities instead of the actual posterior predictive probabilities complicates the application of the exponential mechanism. Indeed, the score function can take any real value, and so we may no longer bound Δq by 1. We thus need to calculate

$$\Delta q = \max_{\tilde{x} \in \{0,1,\ldots,n\}} \max_{\substack{d_1, d_2 \in \{0,1,\ldots,n\} \\ s.t. |d_1 - d_2| = 1}} |P(\tilde{x}|d_1) - P(\tilde{x}|d_2)|$$

While this is easy to do in this simple illustration, it will not scale well to the size of the dataset and the complexity of the model.

Table 1: **Summary statistics of synthetic datasets obtained with the exponential mechanism for 3 different choices of score functions.** In this example, $n = 50$, $d = x = 30$, $\epsilon = 0.5$, and $\gamma_1 = \gamma_2 = 1$, and the output space is $\{0, 1, \ldots, 50\}$.

Score function $(q(d, r))$	Most probable outcome	Expected outcome	Expected distance to true dataset	Expected posterior probability
$-\lvert d - r \rvert$	30	29.99962	1.918678	0.07231692
$P(\tilde{x} = r \lvert x = d)$	30	25.04738	13.142660	0.01998924
$\log(P(\tilde{x} = r \lvert x = d))$	30	29.53010	5.308439	0.04820520

Extension to Other Models

The proposed methodology can in principle be extended to work with any bayesian model which we want to fit to the contingency table, in particular the DPMPM which we proposed to use for synthetic data generation in section 2. Of course, it will in general not be possible to write down the posterior probabilities in closed form as in the simple illustration, but since MCMCs are usually required to fit these more complicated model we will be able to obtain Monte Carlo estimate for them.

The main challenge of our propose methodology lies in calculating Δq, and this is area of current research. In particular, we are considering an approximation to the value for a subset of the sample space which we believe to have a higher prior probability, which would yield a privacy guarantee akin to that of the $\delta - \epsilon$ differential privacy, a relaxation to the stricter ϵ-differential privacy.

4 Discussion

In this paper, we first presented a new nonparametric bayesian model for large contingency tables, which we believe has a lot of potential for the generation of synthetic contingency tables. We then argued for the need to measure the confidentiality guarantee offered by synthetic datasets generated from posterior predictive distributions in such a way. Our approach is based on the criterion of differential privacy, and allows to sample an output for publication with a high posterior probability, while achieving the rigorous criteria of differential privacy. We presented an illustration in a very simple case, and offered suggestions for extension to more complicated models, but much work is needed to apply the methodology in real case scenarios.

Note that while this work is based on the criterion of differential privacy, we understand that it may prove too rigorous in many circumstances. Other approaches to quantifying the risk of inferential disclosure from synthetic datasets must also be studied. In particular, the method for risk assessment proposed in McClure and Reiter (2012), which has some of the flavor of differential privacy, is a promising avenue for further research.

References

Charest, A-S. (2012) "Empirical Evaluation of Statistical Inference from Differentially-Private Contingency Tables", *Privacy in Statistical Databases 2012*, Lecture Notes in Computer Science, Springer.

Duncan, G. T., Elliot, M. and Salazar-González, J.J. (2011) *Statistical Confidentiality: Principles and Practice*, Springer.

Dunson, D.B. and Xing, C. (2009) "Nonparametric Bayes Modeling of Multivariate Categorical Data", *Journal of the American Statistical Association*, **104/487**, 1042–1051.

Dwork, C. (2006), "Differential Privacy", *Automata, languages and programming*, 1–12.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Shulte Nordholt, E., Spicer, K. and Wolf, P.P. (2012) *Statistical Disclosure Control*, Wiley Series in Survey Methodology, Wiley.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008) "Privacy: Theory Meets Practice on the Map." in *Data Engineering, ICDE 2008, IEEE 24th International Con- ference on*, 277–286.

McClure, D. and Reiter, J.P., (2012) "Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data", *Transactions on Data Privacy*, **5/3**, 535–552.

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

Reiter, J.P. (2011) "Data Confidentiality", *Wiley Interdisciplinary Reviews: Computational Statistics*, **3/5**, 450–456.

Si, Y. and Reiter, J.P. (2013) "Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Assessment Surveys", *Journal of Educational and Behavioral Statistics*, **38/5**, 599–521.

Donald B. Rubin, (1993) "Statistical Disclosure Limitation", *Journal of Official Statistics*, **9/2**, 461–468.