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

Anne-Sophie Charest, Université Laval

October 28, 2013

Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Statistics Canada, Ottawa

Context

- Synthetic datasets based on multiple imputations are becoming increasingly popular.
  (See DPMPM for a recent example.)
- It is not clear why simply sampling from posterior predictive distributions should guarantee privacy.
- We consider here a way to add a strict guarantee to such synthetic datasets.

Mesure of privacy

We use Differential Privacy :

A randomized function κ gives ε-differential privacy if and only if for all neighboring datasets D_1 and D_2, and for all $S \subseteq \text{range}(\kappa)$,

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

Two datasets are said to be neighboring datasets if they differ only on their values for one of the respondents.

Previous work

Machanavajjhala *et al.* (2008) modifies the prior distribution for a Dirichlet-Multinomial model of counts in such a way as to obtain differential privacy :

1. Sample parameters for the posterior distribution

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

2. Sample a synthetic dataset

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

This mechanism gives ε DP if and only if $\alpha_i \geq \frac{\tilde{n}}{\exp(\varepsilon)-1}$ $i \in 1, \ldots, p$

This approach however does not extend well to more complicated settings.

What we propose

Sample datasets proportionnaly to their posterior probability, but with probabilities modified in such a way as to guarantee differential privacy.

What we propose

Sample datasets proportionnaly to their posterior probability, but with probabilities modified in such a way as to guarantee differential privacy.

To modify the probabilities, we use the Exponential Mechanism (McSherry and Talwar (2007)).

Exponential Mechanism

Consider a dataset consisting of n inputs each from domain $\mathscr{X}$, with the goal to release an output in the range $\mathscr{R}$.

Define a score function $q : \mathscr{D}^n \times \mathscr{R} \to \mathbb{R}$, which assigns a real-valued score to any pair (x, r) in $\mathscr{X}^n \times \mathscr{R}$. 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

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

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

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

Example

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

$$P(\tilde{x}=a|x=b) = \int_\theta \underbrace{P(\tilde{x}=a|\theta=p)}_{\text{Bin}(n\,,\,p)} \underbrace{P(\theta=p|x=b)}_{\text{Beta}(\gamma_1+b\,,\,\gamma_2+n-b)} \, d\theta$$

$$= \binom{n}{a} \frac{\Gamma(n+\gamma_1+\gamma_2)}{\Gamma(b+\gamma_1)\Gamma(n-b+\gamma_2)} \frac{\Gamma(\gamma_1+a+b)\Gamma(\gamma_2+2n-a-b)}{\Gamma(\gamma_1+\gamma_2+2n)}$$

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

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

Other Possible Score Functions

TABLE: **Summary statistics of synthetic datasets obtained with the exponential mechanism for 3 different choices of score functions.** In this example, $n = 50$, $d = x = 30$, $\varepsilon = 0.5$, 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 dist. to true dataset	Expected post. 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

Implications

In general, we believe that the posterior predictive probability is the most appropriate choice of score function.

This however raises difficulties regarding the calculation of the sensitivity. We are working on solving these problems.

All comments/suggestions to improve the methodology are welcome.

Comments ?
Questions ?