

Obtaining (ϵ, δ) -differential privacy guarantees when using a Poisson mechanism to synthesize contingency tables

James Jackson¹[0000-0002-4832-6638], Robin Mitra²[0000-0001-9584-8044],
Brian Francis¹[0000-0001-7926-9085], and Iain Dove³[0000-0002-1145-2999]

¹ Lancaster University, Lancaster, UK

² Department of Statistical Science, UCL, London, UK

³ Office for National Statistics, Titchfield, UK

Abstract. We show that differential privacy type guarantees can be obtained when using a Poisson synthesis mechanism to protect counts in contingency tables. Specifically, we show how to obtain (ϵ, δ) -probabilistic differential privacy guarantees via the Poisson distribution’s cumulative distribution function. We demonstrate this empirically with the synthesis of an administrative-type confidential database.

Keywords: contingency tables · differential privacy · Poisson mechanism · synthetic data ·

1 Introduction

Differential privacy (DP) [7] is a property of a perturbation mechanism that formally quantifies how accurately any individual’s true values can be established, given all other individuals’ true values are known. Originally developed as a way to protect the privacy of summary statistics (queries), it soon expanded as a way to protect entire data sets. Differentially private data synthesis (DIPS) has since become a popular area of research; see, for example, [1], [12], [5], [13], [4], [14], [6].

In [11], [10], we proposed a synthesis approach for contingency tables that uses saturated count models. This approach effectively uses a count distribution to apply noise to the counts in the original data’s contingency table, and therefore shares traits with DP mechanisms which apply noise in a similar way. Note that as microdata composed entirely of categorical variables can be expressed in contingency table format, this approach is suitable in the case of categorical data more generally.

In this paper, we consider the ability to obtain DP-guarantees when using the Poisson distribution to synthesize counts in contingency tables. We show that although ϵ -DP cannot be satisfied, (ϵ, δ) -DP guarantees can be obtained through the use of the Poisson’s cumulative distribution function (CDF).

The motivation behind this work is that, with the exception of [14], the use of count distributions has largely been overlooked as a way to satisfy DP.

An obvious benefit of using count distributions is that negative counts cannot be obtained. As the Poisson has only one parameter and hence is likely to be sub-optimal, the intention is that in the future the Poisson could be replaced with more complex count distributions, such as the (discretised) gamma family distribution, where additional parameters provide scope for fine-tuning.

The paper is structured as follows. Section 2 introduces some terminology and definitions. Section 3 looks at existing DP mechanisms for contingency tables, such as the (discretised) Laplace and Gaussian mechanisms. Section 4 gives our novel contribution, the ability to obtain (ϵ, δ) -DP guarantees when using a Poisson synthesis mechanism. Section 5 gives an empirical example using an administrative database. Section 6 gives some concluding remarks.

2 Terminology and definitions

Rinott et al. [15] set out how DP extends into a contingency table setting. Following their notation, let $\mathbf{a} = (a_k, \dots, a_K) \in \mathcal{A}$ and $\mathbf{b} = (b_k, \dots, b_K) \in \mathcal{B}$ denote vectors of counts in the original and synthetic data's contingency tables, respectively, where K denotes the number of cells and \mathcal{A} and \mathcal{B} denote the range of obtainable original and synthetic counts (respectively). For contingency tables, we suppose that $\mathcal{A} = \mathcal{B} = \mathbb{Z}_{\geq 0}^K$, where $\mathbb{Z}_{\geq 0}$ is the set of non-negative integers.

Moreover, we describe \mathbf{a} and \mathbf{a}' as neighbours, denoted by $\mathbf{a} \sim \mathbf{a}'$, whenever all but one of the counts in \mathbf{a} and \mathbf{a}' are identical and the differing count differs by exactly one. Henceforth, without loss of generality, we suppose \mathbf{a} and \mathbf{a}' differ in their k th element only, i.e. $a'_k = a_k - 1$ and $a_i = a'_i$ for $i = 1, \dots, K, i \neq k$. Thus \mathbf{a}' represents the data held by the intruder (who knows all but one of the individuals' true values) and \mathbf{a} represents the completed data where the "unknown individual" has been added to the cell in which they truly belong.

The ϵ -DP definition revolves around the likelihood ratio, or, more accurately, around a series of likelihood ratios.

Definition 1 (ϵ -DP). *A perturbation mechanism \mathcal{M} satisfies ϵ -DP ($\epsilon > 0$) if:*

$$\exp(-\epsilon) \leq \frac{\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b})} \leq \exp(\epsilon), \quad (1)$$

$$\forall \mathbf{a} \sim \mathbf{a}' \in \mathcal{A}, \forall \mathbf{b} \in \mathcal{B}.$$

Definition 1 is the special case of the standard DP definition, given in [7], for when the range of \mathcal{A} and \mathcal{B} are discrete. Although we appreciate that in some instances the denominator in (1) could be equal to zero, for the mechanisms we consider here this probability is always non-zero.

For any \mathbf{a} , \mathbf{a}' and \mathbf{b} , whenever the ratio $\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b})/\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b})$ is either small or large, relatively too much is gleaned about the unknown individual's true values. It is worth noting, too, that the above definition considers all possible synthetic data sets in \mathcal{B} , illustrating that DP is not a risk metric for a particular synthetic data set but rather a property of a synthesis mechanism.

Somewhat confusingly, there are two similar but different relaxations of ϵ -DP. The first is (ϵ, δ) -differential privacy [8]. The second is known as (ϵ, δ) -probabilistic differential privacy [12]. These are given below in Definitions 2 and 3. In the remainder of this paper, we focus on (ϵ, δ) -probabilistic DP. Yet whenever (ϵ, δ) -probabilistic DP is satisfied, (ϵ, δ) -DP is also satisfied [9].

Definition 2 ((ϵ, δ) -DP). *A perturbation mechanism \mathcal{M} satisfies (ϵ, δ) -DP ($\epsilon > 0; 0 \leq \delta \leq 1$) if:*

$$\frac{\mathbb{P}(\mathcal{M}(\mathbf{a}) = b) - \delta}{\mathbb{P}(\mathcal{M}(\mathbf{a}') = b)} \leq \exp(\epsilon) \quad \text{and} \quad \frac{\mathbb{P}(\mathcal{M}(\mathbf{a}') = b) - \delta}{\mathbb{P}(\mathcal{M}(\mathbf{a}) = b)} \leq \exp(\epsilon) \quad (2)$$

$$\forall \mathbf{a} \sim \mathbf{a}' \in \mathcal{A}, \mathbf{b} \in \mathcal{B}.$$

Definition 3 ((ϵ, δ) -probabilistic DP). *A perturbation mechanism \mathcal{M} satisfies (ϵ, δ) -probabilistic DP ($\epsilon > 0; 0 \leq \delta \leq 1$) if:*

$$\mathbb{P} \left[\frac{1}{\exp(\epsilon)} \leq \frac{\mathbb{P}(\mathcal{M}(\mathbf{a}) = b)}{\mathbb{P}(\mathcal{M}(\mathbf{a}') = b)} \leq \exp(\epsilon) \right] > 1 - \delta \quad \forall \mathbf{a} \sim \mathbf{a}' \in \mathcal{A}, \mathbf{b} \in \mathcal{B}. \quad (3)$$

Theorem 1 ((ϵ, δ) -probabilistic DP implies (ϵ, δ) -DP). *If a perturbation mechanism \mathcal{M} satisfies (ϵ, δ) -probabilistic DP, then it also satisfies (ϵ, δ) -DP. (Proof: see [9])*

3 Examples of existing DP mechanisms

We now give examples of existing DP mechanisms suitable for synthesizing counts in contingency tables. Note that for the Laplace and Gaussian mechanisms, discretised noise needs to be added (unless one is willing to accept non-integer “counts”). This can simply involve adding continuous noise before rounding the adjusted values to the nearest integer. Similarly, negative values can be rounded to zero.

Example 1 (The Laplace mechanism). A random variable $X \sim \text{Laplace}(\mu, d)$ has probability density function f_L :

$$f_L(x; \mu, d) = \frac{1}{2d} \exp \left(-\frac{|x - \mu|}{d} \right).$$

The Laplace mechanism \mathcal{M}_L satisfies ϵ -DP by using the Laplace distribution to add random noise to the original counts \mathbf{a} . Specifically, for every original count a_i , the Laplace mechanism generates a $\text{Laplace}(a_i, 1/\epsilon)$ random variate. To show that this mechanism does indeed satisfy DP, we suppose that $a_i = a'_i$ for $i = 1, \dots, k-1, k+1, \dots, K$ and that $a'_k = a_k - 1$ (i.e. the assumptions made in Section 2). Firstly, when $b_k > a_k$:

$$\begin{aligned} \frac{\mathbb{P}(\mathcal{M}_L(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}_L(\mathbf{a}') = \mathbf{b})} &= \frac{\exp(-\epsilon|b_k - a_k|)}{\exp(-\epsilon|b_k - a'_k|)} \\ &= \frac{\exp(-\epsilon|b_k - a_k|)}{\exp(-\epsilon|b_k - (a_k - 1)|)} \\ &= \exp(\epsilon). \end{aligned} \quad (4)$$

Similarly, when $a_k > b_k$, (4) is equal to $\exp(-\epsilon)$, and when $a_k = b_k$ it is equal to $\exp(0)$. Hence the DP definition in (1) holds.

Example 2 (The Gaussian mechanism). A random variable $X \sim \text{Normal}(\mu, \sigma^2)$ has probability density function f_G :

$$f_G(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

In a similar way to the Laplace mechanism, the Gaussian mechanism, say \mathcal{M}_G , applies $\text{Normal}(0, \sigma^2)$ random noise to the original counts, resulting in a mechanism that satisfies (ϵ, δ) -differential privacy. Using the same assumptions and notation as previous, it follows that:

$$\begin{aligned} \frac{\mathbb{P}(\mathcal{M}_G(\mathbf{a}) = b)}{\mathbb{P}(\mathcal{M}_G(\mathbf{a}') = b)} &= \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{b_k - a_k}{\sigma}\right)^2\right]}{\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{b_k - a_k + 1}{\sigma}\right)^2\right]} \\ &= \exp\left[-\frac{1}{2\sigma^2}(2a_k - 2b_k - 1)\right]. \end{aligned}$$

Recall that (ϵ, δ) -probabilistic DP is satisfied whenever

$$\frac{1}{\exp(\epsilon)} \leq \frac{\mathbb{P}(\mathcal{M}(\mathbf{a}) = b)}{\mathbb{P}(\mathcal{M}(\mathbf{a}') = b)} \leq \exp(\epsilon) \quad \text{with probability } 1 - \delta,$$

which, in this instance, occurs whenever

$$-\epsilon \leq -\frac{1}{2\sigma^2}(2a_k - 2b_k - 1) \leq \epsilon \quad \text{with probability } 1 - \delta.$$

The probability $1 - \delta$ can be obtained from Φ , the normal distribution's CDF [2], as $b_k \sim \text{Normal}(a_k, \sigma^2)$.

$$\begin{aligned} 1 - \delta &= \mathbb{P}\left(-\epsilon \leq -\frac{1}{2\sigma^2}(2a_k - 2b_k - 1) \leq \epsilon\right) \\ &= \mathbb{P}(a_k - \sigma^2\epsilon - 1/2 \leq b_k \leq a_k + \sigma^2\epsilon - 1/2) \\ &= \Phi\left(\frac{a_k + \sigma^2\epsilon - 1/2 - a_k}{\sigma}\right) - \Phi\left(\frac{a_k - \sigma^2\epsilon - 1/2 - a_k}{\sigma}\right) \\ &= \Phi(\sigma\epsilon - 1/(2\sigma)) - \Phi(-\sigma\epsilon - 1/(2\sigma)) \end{aligned}$$

Example 3 (Multinomial-Dirichlet synthesizer). A multinomial-Dirichlet synthesis mechanism [1], say \mathcal{M}_{MD} , can also yield DP guarantees. The original counts \mathbf{a} can be converted to cell probabilities $\boldsymbol{\pi}$ simply by dividing by n (the number of individuals in the data). A Dirichlet prior with concentration parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_K)$ is placed on $\boldsymbol{\pi}$ (see [1] for more on this approach). Using the same “without loss of generality” assumptions as previous, it follows that

$$\begin{aligned}
\frac{\mathbb{P}(\mathcal{M}_{MD}(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}_{MD}(\mathbf{a}') = \mathbf{b})} &= \frac{\Gamma(b_k + a_k + \alpha_k)}{\Gamma(a_k + \alpha_k)} \cdot \frac{\Gamma(a'_k + \alpha_k)}{\Gamma(b_k + a'_k + \alpha_k)} \\
&= \frac{\Gamma(b_k + a_k + \alpha_k)}{\Gamma(a_k + \alpha_k)} \cdot \frac{\Gamma(a_k - 1 + \alpha_k)}{\Gamma(b_k + a_k - 1 + \alpha_k)} \\
&= \frac{b_k + a_k - 1 + \alpha_k}{a_k - 1 + \alpha_k}. \tag{5}
\end{aligned}$$

Recall again that DP is satisfied whenever

$$\frac{1}{\exp(\epsilon)} \leq \frac{\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b})} \leq \exp(\epsilon).$$

As the expression in (5) is always greater than or equal to one, and hence always greater than $1/\exp(\epsilon)$, DP is satisfied whenever

$$\frac{b_k + a_k - 1 + \alpha_k}{a_k - 1 + \alpha_k} \leq \exp(\epsilon).$$

As $a_k \geq 1$ and $b_k \leq n$, this simplifies to

$$\frac{n + \alpha_k}{\alpha_k} \leq \exp(\epsilon) \quad \Rightarrow \quad \alpha_k \geq \frac{n}{\exp(\epsilon) - 1}.$$

Considering all counts a_1, \dots, a_K gives that DP is satisfied whenever

$$\max_i \alpha_i \geq \frac{n}{\exp(\epsilon) - 1}, \quad \text{a result from [12].}$$

4 Satisfying (ϵ, δ) -probabilistic DP with a Poisson synthesis mechanism

When using saturated count models to synthesize contingency tables, as set out in [11], a count distribution, e.g. the Poisson, applies noise to original counts. We assume that a constant pseudocount $\alpha > 0$ is added to every element of \mathbf{a} (i.e. to *all* original counts, not just to zero counts as in [11]), which opens up the possibility that original counts of zero can be synthesized to non-zeros. When using the Poisson we apply the following mechanism, which we denote by \mathcal{M}_P , to obtain a set of synthetic counts:

$$\begin{aligned}
&b_i \mid a_i, \alpha \sim \text{Poisson}(a_i + \alpha), \quad i = 1, \dots, K, \\
\text{i.e. } \mathbb{P}(\mathcal{M}_P(a_i) = b_i) &= \frac{\exp(-a_i - \alpha)(a_i + \alpha)^{b_i}}{b_i!}, \quad i = 1, \dots, K.
\end{aligned}$$

Supposing once again that \mathbf{a} and \mathbf{a}' differ in their k th element only, we have:

$$\frac{\mathbb{P}(\mathcal{M}_P(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}_P(\mathbf{a}') = \mathbf{b})} = \exp(-1) \left(\frac{a_k + \alpha}{a_k - 1 + \alpha} \right)^{b_k}. \tag{6}$$

This quantity is bounded below by $\exp(-1)$, with this minimum occurring when $b_k = 0$. It is unbounded above, however, as b_k can take any integer up to infinity; i.e. the expression in (6) tends to infinity as b_k tends to infinity. Thus ϵ -DP cannot be satisfied.

Instead, we now consider the (ϵ, δ) -probabilistic DP relaxation, first considering the left-hand inequality of the DP definition (Def. 1):

$$\frac{1}{\exp(\epsilon)} \leq \frac{\mathbb{P}(\mathcal{M}_P(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}_P(\mathbf{a}') = \mathbf{b})} \Rightarrow b_k \geq \frac{1 - \epsilon}{\log\left(\frac{a_k + \alpha}{a_k - 1 + \alpha}\right)}.$$

When $\epsilon \geq 1$, this inequality holds with probability 1. When $0 < \epsilon < 1$, the probability that this inequality holds can be determined through the Poisson's CDF, since b_k is a realization from a Poisson random variable. This probability is given as:

$$1 - F_{a_k + \alpha}^P \left[\frac{1 - \epsilon}{\log\left(\frac{a_k + \alpha}{a_k - 1 + \alpha}\right)} \right], \quad (7)$$

where $F_{a_k + \alpha}^P$ is the CDF of the Poisson distribution with mean $a_k + \alpha$.

We next consider the right-hand inequality of Def. 1:

$$\frac{\mathbb{P}(\mathcal{M}_P(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}_P(\mathbf{a}') = \mathbf{b})} \leq \exp(\epsilon) \Rightarrow b_k \leq \frac{1 + \epsilon}{\log\left(\frac{a_k + \alpha}{a_k - 1 + \alpha}\right)}.$$

For all ϵ , this inequality holds with probability

$$F_{a_k + \alpha}^P \left[\frac{1 + \epsilon}{\log\left(\frac{a_k + \alpha}{a_k - 1 + \alpha}\right)} \right]. \quad (8)$$

Recall that in (ϵ, δ) -probabilistic DP, $1 - \delta$ is the probability that DP is satisfied, i.e. the probability that both inequalities hold. A non-trivial question when $0 < \epsilon < 1$ is how to combine the probabilities given in (7) and (8) and hence compute δ ? This is an area of future research.

When $\epsilon > 1$, however, the left-hand inequality of Def. 1 always holds, thus we need only focus on (8). Although non-trivial for any $\epsilon \geq 1$ and $\alpha > 0$, (8) is minimised when $a_k = 1$ (when $a'_k = 0$). Note, a formal proof has been omitted here but extensive empirical simulation results have been undertaken. Thus,

$$1 - \delta = F_{1 + \alpha}^P \left[\frac{1 + \epsilon}{\log\left(\frac{1 + \alpha}{\alpha}\right)} \right]. \quad (9)$$

This also demonstrates the role of α as a tuning parameter for risk. In general, a larger α value corresponds to a lower δ value. Yet δ is not a decreasing function of α . For a very brief explanation, this is because increasing α increases the value

of the expression inside the squared bracket in (9), but it also increases the mean of the Poisson random variable from which a synthetic count is drawn. Figure 1 illustrates the nature of the relationship between α and δ for different values of ϵ . For example, setting $\alpha = 0.1$ satisfies approximately (3,0.3)-probabilistic DP and (1.5,0.6)-probabilistic DP.

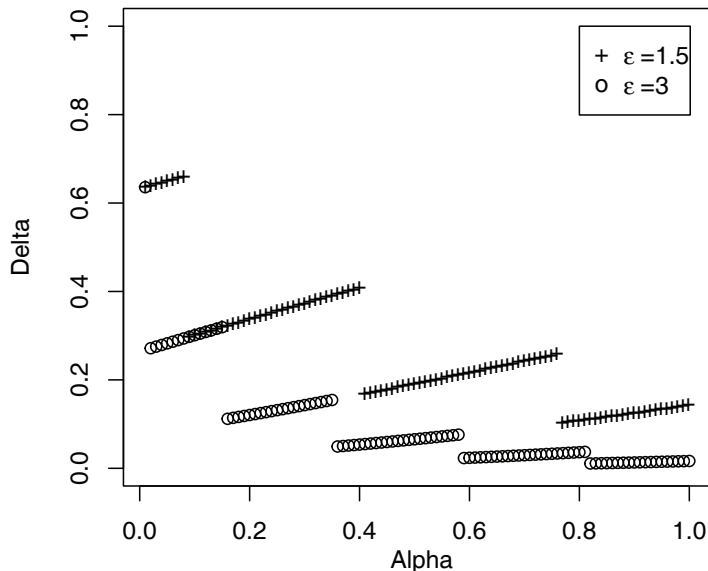


Fig. 1. The relationship between α and δ in the Poisson synthesis mechanism for $\epsilon = 1.5$ and $\epsilon = 3$.

In contingency tables where there are no zero counts, a (ϵ, δ) -DP guarantee can be obtained when $\alpha = 0$. In this instance, δ is determined by the smallest original count, i.e.:

$$1 - \delta = F_{a_i + \alpha}^P \left[\frac{1 + \epsilon}{\log \left(\frac{\min_i a_i + 1}{\min_i a_i} \right)} \right]. \quad (10)$$

In a sense, in this example we have violated the traditional (ϵ, δ) -probabilistic DP definition given in (3) because δ is dependent on a particular set of original counts \mathbf{a} – not all original counts.

We can easily replace the Poisson with any other count distribution (e.g. the negative binomial, Poisson inverse-Gaussian, Delaporte, Sichel, etc.), which of course would lead to a different expression for the ratio in (6).

5 An empirical example

5.1 The English School Census administrative database

The English School Census (ESC) is a large administrative database belonging to the UK’s Department for Education (DfE), which holds information about pupils attending state-funded schools in the UK. Owing to the presence of sensitive data, strict privacy guarantees would be required for data from the ESC to be made available to researchers. There is therefore great appeal to DP-type approaches, where more formal guarantees of privacy can be obtained.

Access to the real ESC data is currently restricted, even for the sake of demonstrating the effectiveness of privacy methods. For this reason, staff at the Office for National Statistics (ONS) created a substitute data set using publicly-available data sources, such as published ESC data and 2011 UK census data. A key feature of this data set, ESC_{rep} , is that it replicates some of the statistical properties present in the actual ESC. We take a subset of this data which has approximately 8×10^6 individuals (rows) and 5 categorical variables (columns). As all variables are categorical, the data set can be expressed as a contingency table with around 3.5×10^6 cells. More information about the data set – as well as the data set itself – is available at [3].

5.2 Applying the Poisson synthesis mechanism

We now apply the Poisson synthesis mechanism to the ESC_{rep} data, considering different values of α , and considering $\epsilon > 1$ values.

Figure 2 gives combinations of (ϵ, δ) values that can be achieved for the ESC_{rep} data when using α values of 0.1, 0.2, 0.5 and 1. For example, when $\epsilon = 2$, an α value of 1 is required to obtain a δ value of 0.05; when $\alpha = 0.1$, a δ value of 0.05, is obtained only for ϵ values greater than 6.

DP methods, in general are known to have a detrimental effect on utility. To gain a simple insight into general utility [16], the boxplots in Figure 3 compare the percentage differences between original and synthetic counts for various values of α , and for original counts between 1 and 10. Unsurprisingly, increasing α increasing the percentage differences, i.e. has an adverse effect on utility. This loss of utility is more magnified in specific analyses, especially when the analyst wishes to quantify uncertainty.

6 Discussion

To summarise, in this paper we have shown how to obtain (ϵ, δ) -DP guarantees when using a Poisson synthesis mechanism to protect the privacy of counts in contingency tables. For a given $\epsilon > 1$, the corresponding value of δ that is achievable with the Poisson is relatively high; much higher than that which is achievable with other DP mechanisms. Going forward, we believe other count distributions, such as the negative binomial, are likely to be more favourable

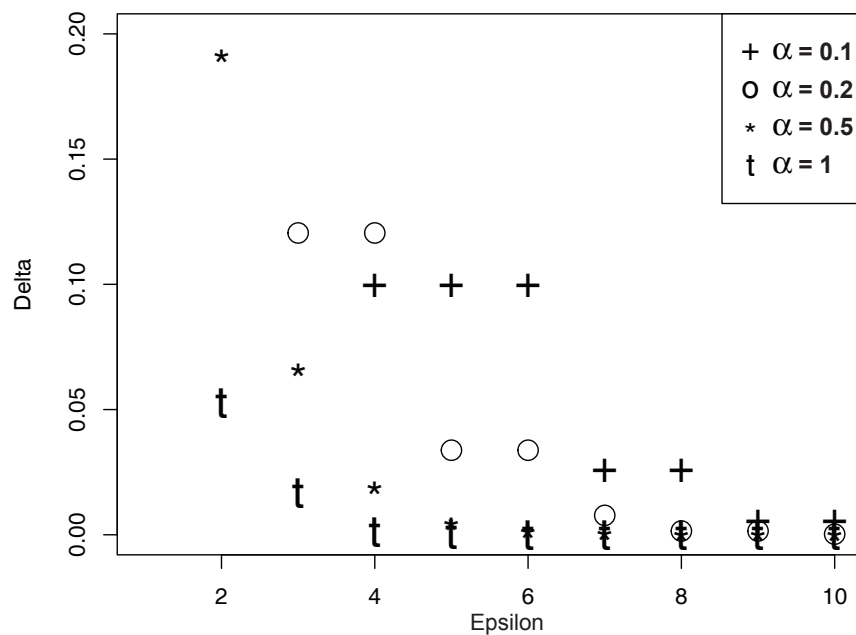


Fig. 2. Combinations of δ such that (ϵ, δ) -probabilistic DP is achieved when the Poisson is used, for various $\max_i a_i$ and ϵ equal to 1.5, 2, 2.5 and 3.

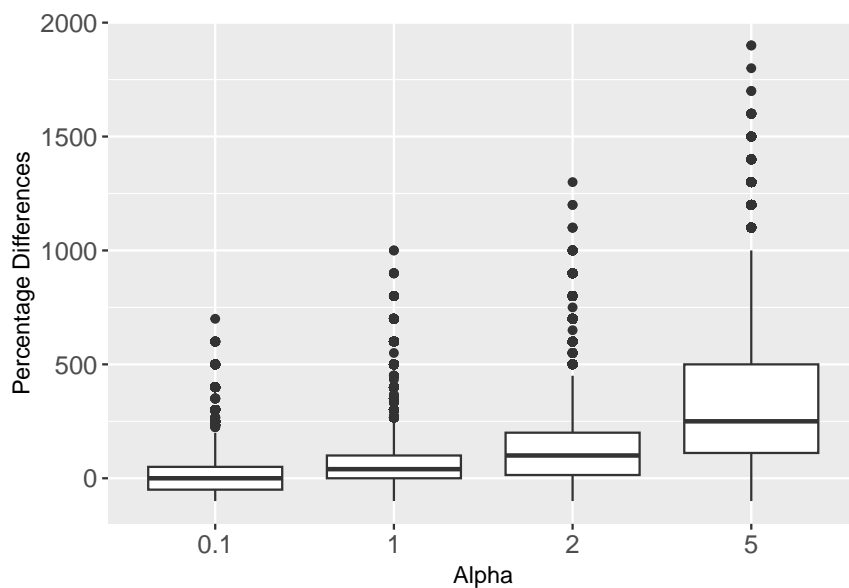


Fig. 3. For different values of α , boxplots showing percentage differences between original and synthetic counts (utility) for original counts in the range 1–10.

(i.e. will give better utility results), while also providing the same DP-type risk guarantees, because such distributions would introduce further tuning parameters in addition to α . Previous work suggests that such tuning parameter apply noise in a more efficient fashion [10]. These tuning parameters could be set to obtain certain ϵ or δ values.

We end with an interesting note in relation to DP. Somewhat counterintuitively, the reason why multinomial-based synthesis mechanisms (e.g. the multinomial Dirichlet synthesizer) can satisfy ϵ -DP – but the Poisson cannot – is because with multinomial mechanisms have a maximum synthetic count that any original count can take, namely n . With count distributions, any original count can be synthesized to any non-negative integer. To help explain why this causes the DP definition to fail, recall that, with contingency tables, DP definitions effectively assume that the intruder is trying to locate the cell to which just one individual belongs; i.e. in the intruder’s data set one, and only one, cell count is one less than it actually is. Suppose that a particular count in the intruder’s data set is equal to 1, but that the corresponding synthetic count – generated by simulating from the Poisson with $\alpha = 0$ – has a count of 5. It is 11.7 times more likely that this synthetic count originated from a cell with a count of 2 than from a count of 1, therefore the intruder can infer that that particular cell is a likely origin of the target. It is interesting therefore that, with DP, disclosure risk is deemed to be at its greatest when the scope for potential movement between original and synthetic counts is at its greatest. This largely goes against the objectives of traditional SDC methods, which typically reduce risk by increasing the divergence from the original counts.

7 Declaration

This version of the contribution has been accepted for publication, after peer review (when applicable) but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/978-3-031-69651-0_7. Use of this Accepted Version is subject to the publisher’s Accepted Manuscript terms of use <https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms>.

References

1. Abowd, J.M., Vilhuber, L.: How Protective Are Synthetic Data? In: Domingo-Ferrer, J., Saygin, Y. (eds.) *Privacy in Statistical Databases 2008*. pp. 239–246. Springer, Berlin, Heidelberg (2008)
2. Balle, B., Wang, Y.X.: Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In: Dy, J., Krause, A. (eds.) *Proceedings of the 35th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 80, pp. 394–403. PMLR (10–15 Jul 2018), <https://proceedings.mlr.press/v80/balle18a.html>
3. Blanchard, S., Jackson, J.E., Mitra, R., Francis, B.J., Dove, I.: A constructed English School Census substitute (2022), [10.17635/lancastr/researchdata/533](https://doi.org/10.17635/lancastr/researchdata/533)

4. Bowen, C.M., Liu, F.: Comparative Study of Differentially Private Data Synthesis Methods. *Statistical Science* **35**(2), 280 – 307 (2020). <https://doi.org/10.1214/19-ST5742>, <https://doi.org/10.1214/19-ST5742>
5. Charest, A.S.: How can we analyze differentially-private synthetic datasets? *Journal of Privacy and Conf.* **2**(2) (2011). <https://doi.org/10.29012/jpc.v2i2.589>, <https://journalprivacyconfidentiality.org/index.php/jpc/article/view/589>
6. Drechsler, J.: Differential privacy for government agencies—are we there yet? *Journal of the American Statistical Association* **118**(541), 761–773 (2023). <https://doi.org/10.1080/01621459.2022.2161385>, <https://doi.org/10.1080/01621459.2022.2161385>
7. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) *Theory of Cryptography*. pp. 265–284. Springer, Berlin, Heidelberg (2006)
8. Dwork, C., Roth, A.: The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science* **9**(3–4), 211–407 (2014). <https://doi.org/10.1561/04000000042>, <http://dx.doi.org/10.1561/04000000042>
9. Goetz, M., Machanavajjhala, A., Wang, G., Xiao, X., Gehrke, J.: Publishing Search Logs - A Comparative Study of Privacy Guarantees. *IEEE Trans. Knowl. Data Eng.* **24**, 520–532 (03 2012). <https://doi.org/10.1109/TKDE.2011.26>
10. Jackson, J., Mitra, R., Francis, B., Dove, I.: On integrating the number of synthetic data sets m into the a priori synthesis approach. In: Domingo-Ferrer, J., Laurent, M. (eds.) *Privacy in Statistical Databases 2022*. pp. 205–219. Springer International Publishing, Cham (2022)
11. Jackson, J., Mitra, R., Francis, B., Dove, I.: Using Saturated Count Models for User-Friendly Synthesis of Large Confidential Administrative Databases. *Journal of the Royal Statistical Society Series A: Statistics in Society* **185**(4), 1613–1643 (08 2022). <https://doi.org/10.1111/rssa.12876>, <https://doi.org/10.1111/rssa.12876>
12. Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L.: Privacy: Theory meets practice on the map. In: 2008 IEEE 24th international conference on data engineering. pp. 277–286. IEEE (2008)
13. McClure, D., Reiter, J.P.: Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data. *Transactions on Data Privacy* **5**(3), 535—552 (2012)
14. Quick, H.: Generating Poisson-distributed differentially private synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **184**(3), 1093–1108 (2021). <https://doi.org/https://doi.org/10.1111/rssa.12711>, <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12711>
15. Rinott, Y., O’Keefe, C.M., Shlomo, N., Skinner, C., et al.: Confidentiality and Differential Privacy in the Dissemination of Frequency Tables. *Statistical Science* **33**(3), 358–385 (2018)
16. Snoke, J., Raab, G.M., Nowok, B., Dibben, C., Slavkovic, A.: General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **181**(3), 663–688 (2018)