Bayesian Inference on Ratios Subject to Differentially Private Noise

by

Linlin Li

Department of Statistical Science
Duke University

Date: _____

Approved:

_____

Jerome Reiter, Advisor

_____

Rebecca Steorts

_____

Yue Jiang

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the Department of Statistical Science
in the Graduate School of
Duke University

2021

ABSTRACT

Bayesian Inference on Ratios Subject to Differentially Private Noise

by

Linlin Li

Department of Statistical Science
Duke University

Date: _____

Approved:

_____

Jerome Reiter, Advisor

_____

Rebecca Steorts

_____

Yue Jiang

An abstract of a thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the Department of Statistical Science
in the Graduate School of
Duke University

2021

Copyright © 2021 by Linlin Li
All rights reserved

Abstract

Data privacy is a long-term issue for data sharing, especially in health-related data. Agencies need to find a balance between data utility and the privacy of the respondents. Differential privacy provides a solution by ensuring that the statistic of interest is basically the same, regardless of whether an individual is included or excluded. Due to this differentially private perturbation, analysts need to infer the true value from the released value. In this thesis, I propose Bayesian inference methods to infer the posterior distribution of ratios of two counts given the released values. I illustrate the Bayesian inference methods under several scenarios and with two commonly used differentially private mechanisms and prior distributions. The Bayesian inference method not only provides a point estimate, but it also provides posterior intervals. Simulation studies show that the Bayesian inference method can generate accurate inferences with close to nominal coverage rates, and have small values of bias and mean squared error.

Acknowledgements

I would like to express my sincere gratitude to my advisor, Professor Jerome Reiter, for his patience, motivation, enthusiasm, immense knowledge, and his continuous guidance and support throughout my time here at Duke University. I would also like to thank the other members of my committee, Dr. Rebecca Steorts and Dr. Yue Jiang, for their encouragement and insightful comments. Finally, I would like to thank my family for their long-term support.

Contents

Abstract **iv**

Acknowledgements **v**

List of Figures **viii**

List of Tables **ix**

1 Introduction **1**

2 Background **3**

2.1 Laplace Mechanism 4

2.2 Geometric Mechanism 4

3 Methodology **7**

3.1 Overview 7

3.2 True Count in Numerator and Noisy Count in Denominator 8

3.2.1 Laplace Mechanism 9

3.2.2 Truncated Geometric Mechanism 10

3.3 Estimate in Numerator and Noisy Count in Denominator 12

3.3.1 Normal Approximation for Numerator 12

3.4 Estimate in Numerator and Estimate with Noise Added in Denominator 14

3.4.1 Normal Approximation for b_{est} 14

3.5 Estimates With Noise Added in Both the Numerator and Denominator 15

4 Simulation Studies **16**

4.1 Simulation 1: Known Value of a and Noisy Value of b 16

4.2 Simulation 2: Estimated Value of a and Noisy Value of b 19

4.3 Simulation 3: Estimated Values of a and Estimated Values with Noise Added b 22

4.4 Simulation 4: Estimated Values with Noise Added to Both 24

5 Conclusion **28**

A Random Sampling **29**

A.1 Random Sampling From SGeom Distribution 29

A.2 Random Sampling From DGamma Distribution 31

A.3 Random Sampling From Piecewise Distributions 35

B Random Sampling From the Posterior Distribution **39**

B.1 Random Sampling From $p(a, b \mid a_{\text{rel}}, b_{\text{rel}})$ when $a_{\text{rel}} = a_{\text{est}} \neq a$ and $b_{\text{rel}} = b_{\text{est}} + \delta_b$ 39

B.2 Random Sampling From $p(a, b \mid a_{\text{rel}}, b_{\text{rel}})$ when $a_{\text{rel}} = a_{\text{est}} + \delta_a$ and $b_{\text{rel}} = b_{\text{est}} + \delta_b$ 40

List of Figures

2.1 The probability density functions of some zero-centered Laplace distributions. 5

2.2 The probability mass functions of some zero-centered symmetric geometric distributions. 6

A.1 The upper bound for $|C-1|$ under different values of (α, β). The red point represents the setting used in the simulation ($\alpha = 45,000$, $\beta = 1.5$). 33

A.2 Relative difference between $\Pr\left(k - \frac{1}{2} \leq X_{\alpha,\beta} < k + \frac{1}{2}\right)$ and $\Pr(x = k)$ for $k \in [25,000, 35,000]$ under the setting used in the simulation ($\alpha = 45,000$, $\beta = 1.5$). 35

List of Tables

3.1 Summary of notation. Note that $a_{\text{est}} = a$ and $b_{\text{est}} = b$ in a census. . . 7

4.1 The percent bias (PBIAS), root-mean-square error (RMSE), and coverage of 95% posterior intervals of the ratio for Laplace mechanism when the numerator is the true value and the denominator is a noisy value. $t = 10,000$ simulation replications were drawn, for each replication, $n = 1,000$ samples were drawn from the posterior distribution to obtain the 95% posterior interval. 18

4.2 The percent bias (PBIAS), root-mean-square error (RMSE), and coverage of 95% posterior intervals of the ratio for the truncated geometric mechanism when the numerator is the true value and the denominator is a noisy value. $t = 10,000$ simulation replications were drawn, for each replication, $n = 1,000$ samples were drawn from the posterior distribution to obtain the 95% posterior interval. 18

4.3 The percent bias (PBIAS), root-mean-square error (RMSE), and coverage of 95% posterior intervals of the ratio for the Laplace mechanism when the numerator is an estimate and the denominator is an noisy value. $t = 10,000$ simulation replications were drawn, for each replication, $n = 1,000$ samples were drawn from the posterior distribution to obtain the 95% posterior interval. 21

4.4 The percent bias (PBIAS), root-mean-square error (RMSE), and coverage of 95% posterior intervals of the ratio for the truncated geometric mechanism when the numerator is an estimate and the denominator is a noisy value. $t = 10,000$ simulation replications were drawn, for each replication, $n = 1,000$ samples were drawn from the posterior distribution to obtain the 95% posterior interval. 21

4.5 The percent bias (PBIAS), root-mean-square error (RMSE), and coverage of 95% posterior intervals of the ratio for Laplace mechanism when the numerator is the an estimate and the denominator is an estimate with noise added. $t = 10,000$ simulation replications were drawn, for each replication, $n = 1,000$ samples were drawn from the posterior distribution to obtain the 95% posterior interval. 23

4.6 The percent bias (PBIAS), root-mean-square error (RMSE), and coverage of 95% posterior intervals of the ratio for the geometric mechanism when the numerator is the an estimate and the denominator is an estimate with noise added. $t = 10,000$ simulation replications were drawn, for each replication, $n = 1,000$ samples were drawn from the posterior distribution to obtain the 95% posterior interval. 24

4.7 The percent bias (PBIAS), root-mean-square error (RMSE), and coverage of 95% posterior intervals of the ratio for Laplace mechanism when the numerator is the an estimate and the denominator is an estimate with noise added to both. $t = 10,000$ simulation replications were drawn, for each replication, $n = 1,000$ samples were drawn from the posterior distribution to obtain the 95% posterior interval. 26

4.8 The percent bias (PBIAS), root-mean-square error (RMSE), and coverage of 95% posterior intervals of the ratio for the geometric mechanism when the numerator is the an estimate and the denominator is an estimate with noise added to both. $t = 10,000$ simulation replications were drawn, for each replication, $n = 1,000$ samples were drawn from the posterior distribution to obtain the 95% posterior interval. . 27

Chapter 1

Introduction

Health disparity refers to the discrepancy in health status between one group of individuals and another group of individuals (Carter-Pokras and Baquet, 2002), as observed in American Indians and Alaska Natives who experience certain cancers at higher rates than non-Hispanic white individuals (Jones, 2006; White et al., 2014). A common measure of health disparity is the ratio of the rate of occurrences in a group of people over the rate of occurrences in another group (such as the entire population) (Pearcy and Keppel, 2002; Carter-Pokras and Baquet, 2002). The estimated occurrence rates for different groups of people come from different agencies, including the Indian Health Service (IHS), Agency for Healthcare Research and Quality (AHRQ), and the Centers for Disease Control and Prevention (CDC).

With the large-scale popularization of the internet and social networks, all kinds of data are stored, transmitted, and processed through the internet (Borders and Prakash, 2009). Data sharing through the internet often suffers from an increased risk of unauthorized data access and malicious data theft (Antón et al., 2010). These breaches may give rise to serious consequences as many datasets contain sensitive variables, including variables that are private and personal such as religious beliefs (Li et al., 2008). In order to preserve the privacy of research participants and avoid such consequences, de-identification, the process of anonymizing a dataset before publishing it, has been the main paradigm for data sharing (Polonetsky et al., 2016; Malin et al., 2011; Roden et al., 2008). However, numerous released datasets that are supposed anonymous have been subject to re-identification attacks (Rothstein, 2010; Ohm, 2009; Loukides et al., 2010; Rocher et al., 2019), indicating that de-identification is not enough to preserve privacy.

Instead, differential privacy offers a solution. On the one hand, differentially private algorithms provide strong privacy protections to prevent potential attacks (Dwork, 2008). These algorithms rely on incorporating random noise, making all the information received by an adversary noisy and inaccurate, so it is much more difficult to breach privacy. On the other hand, "differential privacy addresses the paradox of learning nothing about an individual while learning useful information about a population" (Dwork et al., 2014, p. 5). For example, a study that investigates the most popular routes people take walking through a park may teach us that a proportion of people prefer to take a short-cut on the grass. This result may affect the road planning of the lawn. However, the impact of the study is independent of whether an individual was present or absent in the study (Dwork et al., 2014). With such promising advantages, differentially private algorithms are used by many institutions to share data while preserving people's privacy (Abowd, 2018; Yang et al., 2012; Dankar and El Emam, 2012, 2013; Li et al., 2016; Reiter, 2019).

However, due to the introduction of random noise in the calculation, the analysis and inference subject to differentially private perturbation vary from traditional statistical analysis (Wood et al., 2018). The development and application of methods for differentially private inference is still in the early stage (Wood et al., 2018). With the added noise, there might be some issues of directly applying existing methods to estimate the ratio, as the naive method to ignore the added noise and use the ratio of the two estimates fails to quantify the uncertainty brought by the noise.

To alleviate this problem, I propose Bayesian inference for ratios with differential privacy in this thesis. One advantage of the Bayesian method is that one can incorporate knowledge of the noise distribution into the inference. Another advantage is that an interval can be obtained along with a point estimate from the posterior distribution. Therefore, the Bayesian methods is a powerful tool to infer the ratio upon observance of the released numerator and denominator.

Specifically, I present the Bayesian inference method for the ratio under several scenarios. First, I suppose that the numerator of the ratio is known exactly and not subject to noise, whereas the denominator is a census value subject to differentially private perturbation. Second, I suppose the numerator is an estimated value and not subject to noise, while the denominator is a census value subject to noise. Third, I suppose the numerator is an estimated value without differentially private noise, and the denominator is an estimated value subject to noise. Finally, I consider the scenario where both the numerator and denominator are estimated values subject to differentially private noise.

The remainder of the thesis is organized as follows. Chapter 2 introduces differential privacy, including the Laplace mechanism and the geometric mechanism. Chapter 3 introduces the Bayesian inference method on the ratio under these differentially private mechanisms for the various scenarios. Chapter 4 shows its effectiveness using simulation studies. Chapter 5 summarizes the thesis with a discussion.

Chapter 2

Background

Differential privacy is a *definition* or *criterion* used to quantify privacy threats, where the degree of privacy risks is quantified by the parameter ϵ. The smaller the value of ϵ, the higher the degree of privacy protection, and the less accurate the responses (Dwork et al., 2014).

To formally state the definition of differential privacy, we borrow the definitions from Cynthia Dwork et al.'s book "The Algorithmic Foundations of Differential Privacy" (Dwork et al., 2014). As stated in (Dwork et al., 2014, p. 17), we will consider databases D, which consist of observations from a universe $\mathcal{D}$. Each entry D_i represents the number of elements in the database D of *type* $i \in \mathcal{D}$(Dwork et al., 2014, p. 17). We can measure the distance between two databases D and D′ using the ℓ_1 distance.

Definition 1 (Def. 2.3 in Dwork et al. (2014))**.** The ℓ_1 norm of a database D is defined as

$$\|\mathrm{D}\|_1 = \sum_{i=1}^{|\mathcal{D}|} |\mathrm{D}_i|.$$

The ℓ_1 distance between two databases D and D′ is $\|\mathrm{D} - \mathrm{D}'\|_1$.

Note that $\|\mathrm{D} - \mathrm{D}'\|_1$ is a measure of the number of records that are different between D and D′. We called database D and D′ "neighboring databases" if $\|\mathrm{D} - \mathrm{D}'\|_1 \leq 1$ (i.e., there is only one row that is different and other rows are identical). With this, a formal definition of *differential privacy* is given.

Definition 2 (Modified from Def. 2.4 in Dwork et al. (2014))**.** A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{D}|}$ is ϵ-differentially private if for all $S \subseteq \mathrm{Range}(\mathcal{M})$ and for all $\mathrm{D}, \mathrm{D}' \in \mathbb{N}^{|\mathcal{D}|}$ such that $(\mathrm{D}, \mathrm{D}')$ is a pair of neighboring databases:

$$\Pr(\mathcal{M}(\mathrm{D}) \in S) \leq \exp(\epsilon)\Pr(\mathcal{M}(\mathrm{D}') \in S),$$

where the probability space is over of the sample space of the mechanism $\mathcal{M}$.

ϵ-differential privacy ensures that the distributions of the outputs of the mechanism $\mathcal{M}$ are (almost) the same for any pair of neighboring databases (Dwork et al., 2014), where $\exp(\epsilon)$ is a measure of similarity between the distributions (Reiter, 2019). ϵ is called the *privacy loss*, which controls the degree of privacy protection offered by the mechanism $\mathcal{M}$ (Dwork et al., 2014).

A number of algorithms have been proposed to achieve ϵ-differential privacy, including the Laplace mechanism, the exponential mechanism and the geometric mechanism (Dwork et al., 2014).

2.1 Laplace Mechanism

For real valued statistics, the Laplace mechanism is commonly used for ensuring ϵ-differential privacy (Holohan et al., 2019; Quan et al., 2015; Wang et al., 2016; Sarathy and Muralidhar, 2011). The Laplace mechanism simply calculates the statistic and then perturbs it with noise drawn independently from the Laplace distribution, where ϵ controls the spread of the distribution (Dwork et al., 2014).

Definition 3 (Def. 3.3 from Dwork et al. (2014))**.** Given any function $f : \mathbb{N}^{|\mathcal{D}|} \to \mathbb{R}^k$, the Laplace mechanism is defined as:

$$\mathcal{M}_L(\mathrm{D}, f(\cdot), \epsilon) = f(\mathrm{D}) + (\delta_1, ..., \delta_k)$$

where δ_i are i.i.d. random variables drawn from $\mathrm{Lap}(\Delta f/\epsilon)$, which is represented by its probability density function (p.d.f.)

$$p(\delta \mid \epsilon, \Delta f) = \frac{\epsilon}{2\Delta f} \exp\left(-\frac{\epsilon|\delta|}{\Delta f}\right),$$

and the sensitivity of f is defined as

$$\Delta f = \max_{\substack{\mathrm{D},\mathrm{D}' \in \mathbb{N}^{|\mathcal{D}|} \\ \|\mathrm{D}-\mathrm{D}'\|_1 = 1}} \|f(\mathrm{D}) - f(\mathrm{D}')\|_1.$$

For counting statistics, $\Delta f = 1$ since adding or removing an individual can only change the count value by at most 1 (Dwork et al., 2014). For such statistics, ϵ-differential privacy can be achieved by addition of noise drawn from $\mathrm{Lap}(1/\epsilon)$ (Dwork et al., 2014). Figure 2.1 shows the density functions of a few Laplace distributions. It can be seen that the larger the value of ϵ, the more concentrated the distribution, where the noise is closer to 0.

2.2 Geometric Mechanism

For integer statistics, the Laplace mechanism may provide unreasonable values because the noise drawn from a Laplace distribution is almost certainly not an integer. Instead, we can use the geometric mechanism, a discrete variant of the Laplace mechanism.

Definition 4 (Modified from Ghosh et al. (2012))**.** Given any function $f : \mathbb{N}^{|\mathcal{D}|} \to \mathbb{Z}^k$, the geometric mechanism is defined as:

$$\mathcal{M}_G(\mathrm{D}, f(\cdot), \epsilon) = f(\mathrm{D}) + (\delta_1, ..., \delta_k)$$

where δ_i are i.i.d. random variables drawn from a symmetric geometric distribution $\mathrm{SGeom}(\exp(\epsilon))$, which is represented by its probability mass function (p.m.f.)

$$\Pr(\delta = h \mid \epsilon) = \frac{1 - e^{-\epsilon}}{1 + e^{-\epsilon}} e^{-\epsilon|h|}, \ h \in \mathbb{Z}.$$

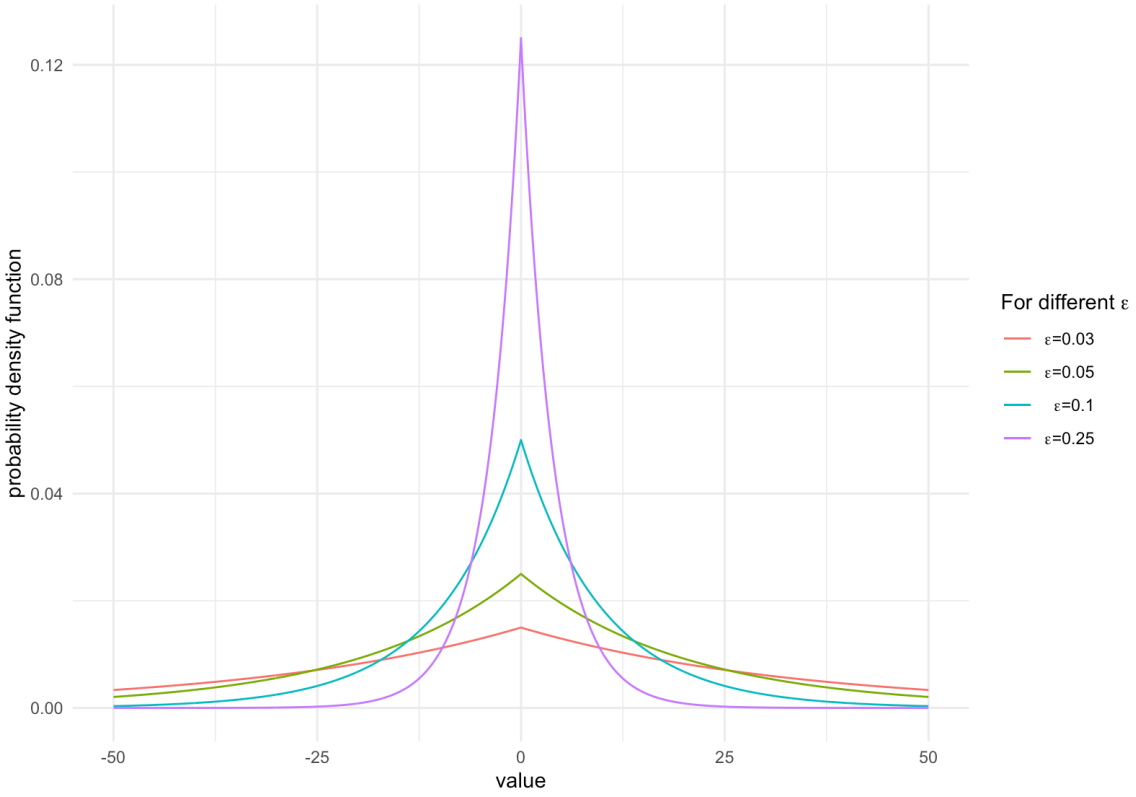

Figure 2.1: The probability density functions of some zero-centered Laplace distributions.

The ϵ-geometric mechanism is ϵ-differentially private for counting statistics (Ghosh et al., 2012). Figure 2.2 shows the p.m.f. of a few symmetric geometric distributions. Similar to the Laplace distribution, it can be seen that the symmetric geometric distribution is more concentrated with a larger value of ϵ. Thus, a larger value of ϵ results in a generally more accurate released value and less privacy protection.

Sometimes, the output value of the ϵ-geometric mechanism is unreasonable. For example, the output value of a count statistics should be at least 0 and at most N (i.e., the number in the population, when known) Ghosh et al. (2012). The truncated geometric mechanism with support $\{0, 1, ..., N\}$ can mitigate this problem by truncating all negative outputs to 0 and all outputs greater than N to N Ghosh et al. (2012). The p.m.f. of this truncated geometric distribution is still analytically tractable Ghosh et al. (2012).

Definition 5 (Modified from Ghosh et al. (2012))**.** Given any function $f : \mathbb{N}^{|\mathcal{D}|} \to \mathbb{Z}^k$, the truncated geometric mechanism with support $\{0, 1, ..., N\}$ is defined as:

$$\mathcal{M}_{TG}(\mathrm{D}, f(\cdot), \epsilon) = f(\mathrm{D}) + (\delta_1, ..., \delta_k)$$

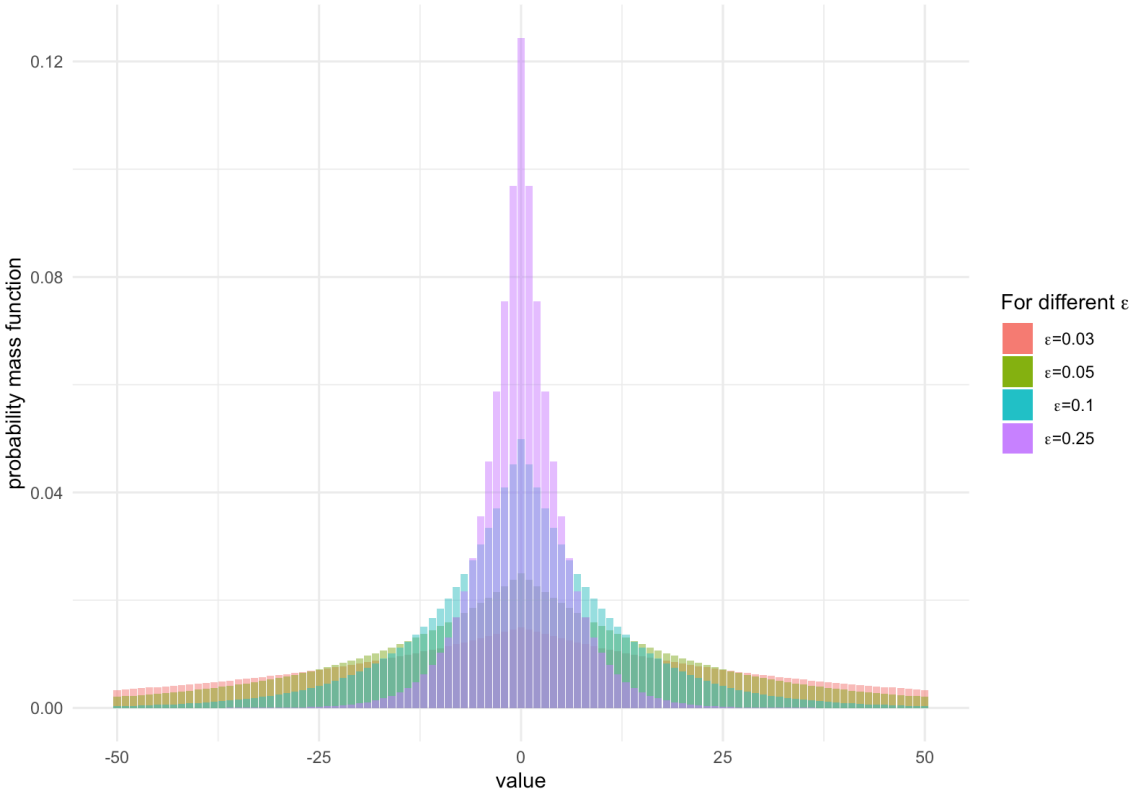

Figure 2.2: The probability mass functions of some zero-centered symmetric geometric distributions.

where δ_i are i.i.d. random variables drawn from a truncated symmetric geometric distribution $\text{TSGeom}(\exp(\epsilon); -f(\text{D})_k, N - f(\text{D})_k)$ with support $\{-f(\text{D})_k, -f(\text{D})_k + 1, ..., N - f(\text{D})_k - 1, N - f(\text{D})_k\}$, which is represented by its p.m.f.

$$\Pr(\delta = h \mid \epsilon) = \begin{cases} 0, & h \in \mathbb{Z} \cap ((-\infty, -f(\text{D})_k) \cup (N - f(\text{D})_k, +\infty)) \\ \frac{1}{1+\exp(-\epsilon)} e^{-\epsilon|h|} & h = -f(\text{D})_k \\ \frac{1-\exp(-\epsilon)}{1+\exp(-\epsilon)} e^{-\epsilon|h|}, & h \in \mathbb{Z} \cap (-f(\text{D})_k, N - f(\text{D})_k) \\ \frac{1}{1+\exp(-\epsilon)} e^{-\epsilon|h|} & h = N - f(\text{D})_k \end{cases}$$

and $f(\text{D})_k$ is the k-th element of $f(\text{D})$.

This truncation does not affect the differential privacy properties (Ghosh et al., 2012). Thus, the truncated ϵ-geometric mechanism is also ϵ-differentially private (Ghosh et al., 2012).

Chapter 3

Methodology

If the true values of the numerator and denominator are known, the true ratio can be trivially obtained. However, sometimes, the true values of the numerator or the denominator are not known, so that agencies need to estimate one or both using a survey. This introduces sampling error into the estimated quantities, which should be accounted for in inferences. On the other hand, releasing the original data to public may not be possible due to needs to protect privacy and confidentiality. Agencies may add differentially private noise onto the estimated values before releasing them. Therefore, the numerator and denominator obtained from the released data may have at least two parts of errors compared to the true values, the sampling error and the noise from privacy treatments. In this chapter, I describe methodologies to account for these sources of noise when estimating ratios.

3.1 Overview

In this section, I introduce some notations and definitions that are used in the following chapters. Let a and b be the true values of the numerator and denominator respectively. When a and b are unknown and need to be estimated using surveys, let a_{est} and b_{est} represent the estimated values obtained by the agencies based on the surveys. With a census, we assume that $a_{\text{est}} = a$ and $b_{\text{est}} = b$. For privacy protection, they may not release the estimated values to the public. Instead, they may add differentially private noise onto the estimated values, and the released values are represented by a_{rel} and b_{rel}, respectively. The variable of interest is the ratio a/b, which we also can obtain from the pair of (a, b). Table 3.1 summarizes the notations used for subsequent discussions.

Notation	Meaning
a	The true count of the numerator
b	The true count of the denominator
a_{est}	The estimate of true count of the numerator with no privacy protections
b_{est}	The estimate of true count of the denominator with no privacy protections
a_{rel}	The released count of the numerator
b_{rel}	The released count of the denominator

Table 3.1: Summary of notation. Note that $a_{\text{est}} = a$ and $b_{\text{est}} = b$ in a census.

We use a Bayesian approach to estimate the posterior distribution given the re-

leased values

$$p(a, b \mid a_{\text{rel}}, b_{\text{rel}}) = \frac{p(a, b)p(a_{\text{rel}}, b_{\text{rel}} \mid a, b)}{p(a_{\text{rel}}, b_{\text{rel}})}.$$

In the sections that follow, we make the following assumptions.

Assumption 1. $a_{\text{rel}} \perp\!\!\!\perp b_{\text{rel}} \mid a, b$.

Assumption 2. $a_{\text{rel}} \perp\!\!\!\perp b \mid a$.

Assumption 3. $b_{\text{rel}} \perp\!\!\!\perp a \mid b$.

When the released values of the numerator and the denominator come from different data sources, it is natural to assume Assumptions 1 because the agencies release their results independently. In addition, under the Laplace mechanism or truncated geometric mechanism, the differentially private noise is only related to the privacy loss parameter ϵ. In this case, Assumptions 2 and 3 are plausible.

Under Assumptions 1 to 3,

$$\begin{aligned} p(a, b \mid a_{\text{rel}}, b_{\text{rel}}) &= p(a \mid a_{\text{rel}}, b_{\text{rel}})p(b \mid a, b_{\text{rel}}, b_{\text{rel}}) \\ &\propto p(a)p(a_{\text{rel}} \mid a)p(b \mid a)p(b_{\text{rel}} \mid b). \end{aligned} \tag{3.1}$$

Therefore, we need to specify the prior distributions $p(a)$ and $p(b \mid a)$, and model the relationship between released values and true values through $p(a_{\text{rel}} \mid a)$ and $p(b_{\text{rel}} \mid b)$.

In some cases, the released numerator is the estimated value of the true value from a survey, which we write as $a_{\text{rel}} = a_{\text{est}}$. Similarly, the denominator may be an estimate of the true value, with a noise added to it for privacy concerns. In those cases, we write $b_{\text{rel}} = b_{\text{est}} + \delta_b$ where δ_b is a specific noise added.

The following sections are arranged as follows. In section 3.2, we assume for simplicity that both estimates are accurate: $a_{\text{est}} = a, b_{\text{est}} = b$. Then we remove these assumptions and discuss the situation where the estimates can be inaccurate in Section 3.3, 3.4 and 3.5.

3.2 True Count in Numerator and Noisy Count in Denominator

When both estimates are equal to the true values, we have $a_{\text{rel}} = a, b_{\text{rel}} = b + \delta_b$. Thus, 3.1 reduces to

$$p(a, b \mid a_{\text{rel}}, b_{\text{rel}}) = p(b \mid a)p(b_{\text{rel}} \mid b)\mathbb{1}_{a=a_{\text{rel}}}.$$

Therefore, we only need to consider the posterior distribution of b:

$$p(b \mid a_{\text{rel}}, b_{\text{rel}}) \propto p(b \mid a_{\text{rel}})p(b_{\text{rel}} \mid b). \tag{3.2}$$

We now discuss the choice of prior and the inference methods under two commonly used differentially private mechanisms.

3.2.1 Laplace Mechanism

As mentioned in Section 2.1, since $\Delta f = 1$ for count statistics, the addition of noise drawn from $\text{Lap}(1/\epsilon)$ preserves ϵ-differential privacy (Dwork et al., 2014). For $c \in \mathbb{R}$, we denote $X \sim \text{Lap}(c, 1/\epsilon)$ if $X - c \sim \text{Lap}(1/\epsilon)$. Then

$$b_{\text{rel}} = b + \delta_b \mid b \sim \text{Lap}(b, 1/\epsilon),$$

given the noise δ_b drawn from $\text{Lap}(1/\epsilon)$. We now show how to do inferences under two prior distributions, namely a gamma prior and a uniform prior.

Gamma Prior

Claim 1. Under certain conditions, when the prior is a Gamma distribution, under the Laplace mechanism, the posterior distribution of x is a piecewise Gamma distribution.

Proof. Suppose the prior distribution of a random variable x is $\text{Gamma}(\alpha, \beta)$ with support $\mathbb{R}^+$,

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \ \alpha, \beta > 0.$$

Given the released value $x_{\text{rel}} = x + \delta$, where δ is independently drawn from $\text{Lap}(1/\epsilon)$, the posterior distribution of x is given by

$$\begin{aligned} p(x \mid x_{\text{rel}}) &\propto p(x) p(x_{\text{rel}} \mid x) \\ &\propto x^{\alpha-1} e^{-\beta x} e^{-\epsilon |x_{\text{rel}} - x|}, \end{aligned}$$

which is a piecewise Gamma distribution as long as $\beta > \epsilon$. □

If we have a prior belief that the distribution of b is $\text{Gamma}(\alpha, \beta)$ with support $\mathbb{R}^+ \cup [a_{\text{rel}}, +\infty)$

$$p(b \mid a_{\text{rel}}) \propto \frac{\beta^\alpha}{\Gamma(\alpha)} b^{\alpha-1} e^{-\beta b} \mathbb{1}_{b \geq a_{\text{rel}}}, \ \alpha > 0, \beta > \epsilon,$$

the posterior distribution of b is

$$\begin{aligned} p(b \mid b_{\text{rel}}, a_{\text{rel}}) &\propto p(b \mid a_{\text{rel}}) p(b_{\text{rel}} \mid b) \\ &\propto b^{\alpha-1} e^{-\beta b} \mathbb{1}_{b \geq a_{\text{rel}}} e^{-\epsilon |b_{\text{rel}} - b|}, \end{aligned}$$

which is a piecewise Gamma distribution truncated on $[a_{\text{rel}}, +\infty)$ as long as $\beta > \epsilon$. Since ϵ is usually published and known, $\beta > \epsilon$ holds for a reasonable prior.

Analysts can use historical data of b to formulate a prior belief. When prior beliefs are available, analysts can convert them in terms of moments (e.g., mean and variance) to fit a Gamma distribution to these moments. With a Gamma prior on b, we can directly obtain posterior samples from the piecewise Gamma distribution (see Appendix A.3).

Uniform Prior

When little information about the denominator is known in advance, analysts can use a uniform prior. Suppose the prior distribution of b is a uniform distribution with support $\mathbb{R}^+ \cap [n_1, n_2]$ given by

$$p(b \mid a_{\text{rel}}) = \frac{1}{n_2 - n_1} \mathbb{1}_{b \geq a_{\text{rel}}} \mathbb{1}_{n_1 \leq b \leq n_2}, \ a_{\text{rel}} \leq n_1 \leq b \leq n_2.$$

The posterior distribution of b is

$$\begin{aligned} p(b \mid b_{\text{rel}}, a_{\text{rel}}) &\propto p(b \mid a_{\text{rel}}) p(b_{\text{rel}} \mid b) \\ &\propto \mathbb{1}_{b \geq a_{\text{rel}}} \mathbb{1}_{n_1 \leq b \leq n_2} e^{-\epsilon |b_{\text{rel}} - b|}, \end{aligned}$$

which is a special case of $\alpha = 1$ and $\beta = 0$ truncated to a narrower support.

Analysts may find historical data helpful for determining the range of the uniform prior. Although uniform is not a conjugate distribution for the Laplace mechanism, the posterior distribution is still easy to sample. Specifically, the posterior distribution is a truncated Laplace distribution, where we can directly obtain posterior samples.

3.2.2 Truncated Geometric Mechanism

As mentioned in Section 2.2, the ϵ-geometric mechanism preserves ϵ-differential privacy (Ghosh et al., 2012). For $c \in \mathbb{R}$, we denote $X \sim \text{SGeom}(c, \exp(\epsilon))$ if $X - c \sim \text{SGeom}(\exp(\epsilon))$. Then

$$b_{\text{rel}} = b + \delta_b \mid b \sim \text{SGeom}(b, \exp(\epsilon)),$$

given the noise δ_b drawn from $\text{SGeom}(\exp(\epsilon))$.

In practice, the magnitude of the noise should be limited to a reasonable range. Otherwise, extreme noise values should be discarded to preserve the reasonableness of the released value. To model this, we assume that the noise is actually sampled from a truncated symmetric geometric distribution, thus ensuring that $b + \delta_b$ is within a reasonable range. There might not be a universal rule to determine what a reasonable range is, and as we will show later, the posterior inference need not be sensitive to the choice of the range, which suggests that we can always choose a range that is large enough to include all reasonable values.

Discrete Gamma Prior

Define $\text{DGamma}(\alpha, \beta)$ distribution with support $\mathbb{N}$ by its p.m.f. at any $k \in \mathbb{N}$

$$\Pr(x = k) \propto k^{\alpha - 1} e^{-\beta k}, \ \alpha > 1, \ \beta > 0.$$

Some scholars have studied this kind of distribution as a discrete analogue of Gamma distribution (Chakraborty and Chakravarty, 2012; Kulasekera and Tonkyn, 1992; Doray and Luong, 1997; Chakraborty, 2015). This distribution is well-defined (proper), as the normalizing constant $Z = \sum_{k\in\mathbb{N}} k^{\alpha-1}e^{-\beta k}$ converges (Kulasekera and Tonkyn, 1992). The convergence can be shown easily from ratio test as (Kulasekera and Tonkyn, 1992)

$$\lim_{k\to\infty} \frac{(k+1)^{\alpha-1}e^{-\beta(k+1)}}{k^{\alpha-1}e^{-\beta k}} = \left(\frac{k+1}{k}\right)^{\alpha-1} e^{-\beta} = e^{-\beta} < 1.$$

The moments of DGamma distribution can be approximated by the moments of Gamma distribution (Kulasekera and Tonkyn, 1992).

Claim 2. Under certain conditions, when the prior is a DGamma distribution, under the truncated geometric mechanism, the posterior distribution of x is a piecewise DGamma distribution.

Proof. Suppose the prior distribution of x is DGamma(α, β) with support $\mathcal{X} \in \mathbb{N}$. Given the released value $x_{\text{rel}} = x + \delta$, where δ is independently drawn from SGeom$(\exp(\epsilon))$, the posterior distribution of x is given by

$$\begin{aligned}\Pr(x \mid x_{\text{rel}}) &\propto \Pr(x) \times \Pr(x_{\text{rel}} \mid x) \\ &\propto x^{\alpha-1}e^{-\beta x}e^{-\epsilon|x_{\text{rel}}-x|},\end{aligned}$$

which is a piecewise DGamma distribution as long as $\beta > \epsilon$. □

If we have a prior belief that the distribution of b is DGamma(α, β) with support $\mathbb{N} \cap [a_{\text{rel}}, +\infty)$, then we have

$$\Pr(b = k) \propto k^{\alpha-1}e^{-\beta k}\mathbb{1}_{k\geq a_{\text{rel}}},\ \alpha > 1,\ \beta > \epsilon,$$

and the posterior distribution of b is

$$\begin{aligned}\Pr(b \mid b_{\text{rel}}, a_{\text{rel}}) &\propto \Pr(b \mid a_{\text{rel}}) \times \Pr(b_{\text{rel}} \mid b) \\ &\propto b^{\alpha-1}e^{-\beta b}\mathbb{1}_{b\in\mathbb{N}\cap[a_{\text{rel}},+\infty)}e^{-\epsilon|b_{\text{rel}}-b|},\end{aligned}$$

which is a piecewise DGamma distribution truncated on $\mathbb{N} \cap [a_{\text{rel}}, +\infty)$.

Analysts can use historical data to formulate a prior belief. When prior beliefs are available, analysts can convert them in terms of moments (e.g., mean and variance) to fit a DGamma distribution to these moments. With a DGamma prior on b, we can directly obtain posterior samples from the piecewise DGamma distribution (see Appendix A.2, and A.3).

Discrete Uniform Prior

When little information about the denominator is known in advance, analysts can use a discrete uniform prior. Suppose the prior distribution of b is a discrete uniform distribution with support $\mathbb{N} \cap [n_1, n_2]$ given by

$$\Pr(b \mid a_{\text{rel}}) = \frac{1}{n_2 - n_1 + 1} \mathbb{1}_{b \geq a_{\text{rel}}} \mathbb{1}_{b \in \mathbb{N} \cap [n_1, n_2]}, \ a_{\text{rel}} \leq n_1 \leq b \leq n_2.$$

The posterior distribution of b is

$$\begin{aligned} \Pr(b \mid b_{\text{rel}}, a_{\text{rel}}) &\propto \Pr(b \mid a_{\text{rel}}) \Pr(b_{\text{rel}} \mid b) \\ &\propto \mathbb{1}_{b \geq a_{\text{rel}}} \mathbb{1}_{b \in \mathbb{N} \cap [n_1, n_2]} e^{-\epsilon |b_{\text{rel}} - b|}, \end{aligned}$$

which is a special case of $\alpha = 1$ and $\beta = 0$ truncated to a narrower support.

Analysts may find historical data helpful for determining the range of the discrete uniform prior. Although uniform is not a conjugate distribution for the (truncated) geometric mechanism, the posterior distribution is still easy to sample. We use an efficient approach to sample from the posterior distribution by sampling directly from the truncated piecewise symmetric geometric distribution (see Appendix A.1 and A.3).

3.3 Estimate in Numerator and Noisy Count in Denominator

When $a_{\text{rel}} = a_{\text{est}} \neq a$, and $b_{\text{est}} = b$, we have

$$\begin{aligned} p(a, b \mid a_{\text{rel}}, b_{\text{rel}}) &\propto p(a) p(a_{\text{rel}} \mid a) p(b \mid a) p(b_{\text{rel}} \mid b) \mathbb{1}_{a_{\text{rel}} = a_{\text{est}}} \\ &= p(a) p(a_{\text{est}} \mid a) p(b \mid a) p(b_{\text{rel}} \mid b). \end{aligned} \tag{3.3}$$

This can be determined by the priors and the conditional distribution of the released values. Compared with the situation when the true value of a is known, we need to additionally specify $p(a)$ and $p(a_{\text{est}} \mid a)$.

3.3.1 Normal Approximation for Numerator

A large family of estimates including MLE exhibit asymptotic normality as the sample size increases, so we assume that the conditional distribution of a_{est} is a normal distribution. A normal distribution is fully characterized by its expectation and variance. We assume that the agency publishes the estimated values of $\mathbb{E}[a_{\text{est}} \mid a]$ and $\text{Var}[a_{\text{est}} \mid a]$, for example, based on Horvitz-Thompson estimators.

Therefore, suppose the prior distribution of a is $N(\mu, \tau^2)$. Then, (3.3) can be computed as follows.

$$\begin{aligned} p(a, b \mid a_{\text{rel}}, b_{\text{rel}}) &\propto p(a)p(a_{\text{est}} \mid a)p(b \mid a)p(b_{\text{rel}} \mid b) \\ &\propto e^{-(a-\mu)^2/(2\tau^2)} e^{-(a_{\text{est}} - \mathbb{E}[a_{\text{rel}}|a])^2/(2\text{Var}[a_{\text{est}}|a])} p(b \mid a)p(b_{\text{rel}} \mid b). \end{aligned}$$

If we further assume that the estimator of a is unbiased so that $\mathbb{E}[a_{\text{rel}} \mid a] = a$, the posterior distribution of (a, b) can be reduced to

$$\begin{aligned} p(a, b \mid a_{\text{rel}}, b_{\text{rel}}) &\propto p(a)p(a_{\text{rel}} \mid a)p(b \mid a)p(b_{\text{rel}} \mid b)\mathbb{1}_{a_{\text{rel}}=a_{\text{est}}} \\ &\propto p(a)p(a_{\text{est}} \mid a)p(b \mid a)p(b_{\text{rel}} \mid b) \\ &\propto e^{-(a-\mu)^2/(2\tau^2)} e^{-(a_{\text{est}} - a)^2/(2\text{Var}[a_{\text{est}}|a])} p(b \mid a)p(b_{\text{rel}} \mid b). \end{aligned}$$

The posterior of a is a normal distribution, and the inference methods of the posterior of b mentioned in Section 3.2 can be used here to infer the last two terms. Plug in the estimated value of $\text{Var}[a_{\text{est}} \mid a]$, we then can infer the posterior distribution of (a, b) given the released values a_{rel} and b_{rel}, thereby inferring the posterior distribution of a/b.

When we do not have enough confidence in the prior beliefs, non-informative priors (such as uniform prior, Jeffreys' prior) may be useful. Take the Jeffreys' prior as an example. Assuming that $p(a) \propto 1$, (3.3) then reduces to

$$p(a, b \mid a_{\text{rel}}, b_{\text{rel}}) \propto e^{-(a_{\text{est}} - a)^2/(2\text{Var}[a_{\text{est}}|a])} p(b \mid a)p(b_{\text{rel}} \mid b).$$

The posterior of a is still a normal distribution, and the approach of sampling from the posterior of b mentioned in Section 3.2 can be used here to infer the last two terms, thereby inferring the posterior distribution of a/b. Algorithm 1 provides the details of sampling from the posterior of a/b given a_{rel} and b_{rel}.

Algorithm 1 *Algorithm for sampling from the posterior distribution of a/b when $a_{\text{rel}} = a_{\text{est}}$ and $b_{\text{rel}} = b + \delta_b$*

---

1. For $i = 1$ to n:
   - Generate a random sample from the posterior of a:
     $$a^{(i)} \sim p(a \mid a_{\text{rel}})$$
   - With the sample $a^{(i)}$, generate a sample from the posterior of b:
     $$b^{(i)} \sim p(b \mid a^{(i)}, a_{\text{rel}}, b_{\text{rel}})$$
   - Compute $r^{(i)} = a^{(i)}/b^{(i)}$
2. Output samples $r^{(1)}, \cdots, r^{(n)}$.

3.4 Estimate in Numerator and Estimate with Noise Added in Denominator

In this section, I discuss the choice of priors and inference methods when the numerator is an estimate and denominator is an estimate with noise added. In this case, $a_{\text{rel}} = a_{\text{est}}$ and $b_{\text{rel}} = b_{\text{est}} + \delta_b$. Thus, we have

$$p(a, b \mid a_{\text{rel}}, b_{\text{rel}}) \propto p(a)p(a_{\text{rel}} \mid a)p(b \mid a)p(b_{\text{rel}} \mid b)\mathbb{1}_{a_{\text{rel}}=a_{\text{est}}}.$$

3.4.1 Normal Approximation for b_{est}

For ease of computation and because of the typical nature of weighted estimates in surveys, we treat b_{est} as a continuous value. We assume that the conditional distribution of b_{est} given b is a normal distribution, relying on asymptotic normality. In this section, we assume that the Laplace mechanism is used for privacy protection. As mentioned in Section 3.3.1, the mean and the variance of the estimate ($\mathbb{E}[b_{\text{est}} \mid b]$ and $\text{Var}[b_{\text{est}} \mid b]$) are difficult to obtain unless the agency publishes them. If we assume that the estimate b_{est} is unbiased and its estimated variance is published or can be approximated from other sources like past surveys, then

$$\begin{aligned} p(b_{\text{est}} \mid b) &\sim N(b, \text{Var}[b_{\text{est}} \mid b]), \\ p(b_{\text{rel}} \mid b_{\text{est}}, b) &\sim \text{Lap}(b_{\text{est}}, 1/\epsilon). \end{aligned}$$

Thus, $p(b_{\text{rel}} \mid b) = \int_{b_{\text{est}}} p(b_{\text{rel}} \mid b_{\text{est}}, b)p(b_{\text{est}} \mid b)\, \mathrm{d}b_{\text{est}}$ can be approximated numerically. To do this, for some large M (I use $M = 4,000$ in the simulations), we first generate $b_{\text{est}}^{(1)}, ..., b_{\text{est}}^{(M)} \overset{i.i.d.}{\sim} N(b, \text{Var}[b_{\text{est}} \mid b])$. Then according to the weak law of large numbers, $\frac{1}{M}\sum_{m=1}^{M} p(b_{\text{rel}} \mid b_{\text{est}}^{(m)}, b) \xrightarrow{p} p(b_{\text{rel}} \mid b)$.

With the estimated $p(b_{\text{rel}} \mid b)$, we can use rejection sampling to generate samples from the posterior of b given $(a, a_{\text{rel}}, b_{\text{rel}})$. We can do this by generating each proposal of b from its prior $p(b \mid a)$, and then accept the proposal with probability $p(b_{\text{rel}} \mid b)$. However, this approach may be very inefficient. We use a more efficient approach to generate samples from the posterior of b (see Appendix B.1). Combined with the method of sampling from the posterior of a in Section 3.3, we can obtain samples from the posterior of (a, b) pair given a_{rel} and b_{rel}, thereby obtaining samples from the posterior of a/b.

If we assume that b_{est} is a discrete value, then we cannot directly use a normal distribution to approximate the conditional distribution of b_{est} given b. A more natural limiting distribution, discretized normal distribution $N^d(\mu, \sigma^2)$ with support on $\mathbb{Z}$, is defined by its probability mass function

$$\Pr(z = j) = \Pr\left(j - \frac{1}{2} \leq Z_{\mu,\sigma^2} < j + \frac{1}{2}\right), \ j \in \mathbb{Z},$$

where Z_{μ,σ^2} follows a normal distribution with mean μ and variance σ^2. This also is convenient when the truncated geometric mechanism can be used for privacy protection. If we assume that the estimator b_{est} is unbiased and its estimated variance is published, then

$$p(b_{\text{est}} \mid b) \sim N^d(b, \text{Var}[b_{\text{est}} \mid b]),$$
$$p(b_{\text{rel}} \mid b_{\text{est}}, b) \sim \text{SGeom}(b_{\text{est}}, \exp(\epsilon)).$$

Thus, $p(b_{\text{rel}} \mid b) = \sum_{b_{\text{est}}} p(b_{\text{rel}} \mid b_{\text{est}}, b)p(b_{\text{est}} \mid b)$ can be approximated numerically. To do this, we can first generate $b_{\text{est}}^{(1)}, ..., b_{\text{est}}^{(M)} \overset{i.i.d.}{\sim} N^d(b, \text{Var}[b_{\text{est}} \mid b])$, and then estimate $p(b_{\text{rel}} \mid b)$ by the mean of $p(b_{\text{rel}} \mid b_{\text{est}}^{(1)}, b), \ldots, p(b_{\text{rel}} \mid b_{\text{est}}^{(M)}, b)$. Similarly, we can use rejection sampling to generate samples from the posterior of b given $a, a_{\text{rel}}, b_{\text{rel}}$ (see Appendix B.1). We then can obtain samples from the posterior of (a, b) pair given a_{rel} and b_{rel}, thereby obtaining samples from the posterior of the ratio a/b.

3.5 Estimates With Noise Added in Both the Numerator and Denominator

In some cases, analysts may have to work with numerators that have been perturbed by differentially private noise. In this case, $a_{\text{rel}} = a_{\text{est}} + \delta_a$ and $b_{\text{rel}} = b_{\text{est}} + \delta_b$. We make an additional assumption.

Assumption 4. $\delta_a \perp\!\!\!\perp \delta_b$.

It is reasonable to assume Assumption 4 when the released values of the numerator and denominator come from different surveys. To evaluate (3.1), under Assumption 1 to 4, $p(a, b \mid a_{\text{rel}}, b_{\text{rel}})$ is proportional to

$$p(a) \int_{a_{\text{est}}} p(a_{\text{rel}} \mid a_{\text{est}})p(a_{\text{est}} \mid a)\, \mathrm{d}a_{\text{est}}\ p(b \mid a) \int_{b_{\text{est}}} p(b_{\text{rel}} \mid b_{\text{est}})p(b_{\text{est}} \mid b)\, \mathrm{d}b_{\text{est}}. \tag{3.4}$$

Unless agencies announce their sampling and estimation methods, it is difficult to infer the distribution of a_{est} conditioned on a and the distribution of b_{est} conditioned on b. Similar to Section 3.3.1 and 3.4.1, we assume that these two conditional distributions are normal distributions (or discretized normal distribution for discrete values). As the posterior of a is not related to b, we can first obtain samples from the posterior distribution of a given a_{rel}, and then apply the method mentioned in Section 3.4 to obtain samples from the posterior of b, thereby obtaining samples from the posterior of the ratio a/b (see Appendix B.2 for more details).

Chapter 4

Simulation Studies

To evaluate the performance of the Bayesian inference methods, we repeated sampling. The noise in the numerator and denominator are independently generated from Laplace or symmetric geometric distribution, and the true values are set artificially.

In each simulation, we set the true value of $r = a/b$ by setting the true values of a and b, and construct the noise-added observations a_{rel} and b_{rel}. Our proposed method provides a point estimate of r (which we denote as $\hat{r}$) and summarizes its uncertainty through the 95% posterior interval. To evaluate the performance of our method, we assess the proximity of the point estimate to the truth via percent bias and root-mean-square error, and the correctness in quantifying the uncertainty via coverage rates of 95% posterior intervals.

The percent bias (PBIAS) is defined as the expected relative discrepancy of the estimate over the true value, in percentage.

$$\text{PBIAS} = \mathbb{E}[(\hat{r} - r)/r].$$

The root-mean-sqare-error (RMSE) is defined as the square root of mean squared error of the estimate

$$\text{RMSE} = \sqrt{\mathbb{E}[(r - \hat{r})^2]}.$$

The coverage is defined as the probability that the 95% posterior interval would cover the true value

$$\text{Coverage} = \mathbb{E}[\mathbb{1}_{r \in CI}].$$

4.1 Simulation 1: Known Value of a and Noisy Value of b

Several settings are designed to evaluate the performance of the proposed inference method when the numerator is the true value and the denominator is true value with noise added. In this scenario, we have $a_{\text{rel}} = a$, and $b_{\text{rel}} = b + \delta_b$. For each configuration, $t = 10,000$ simulation replication were drawn. If not specifically mentioned, the prior distribution of b is as follows.

- For the Laplace mechanism, Gamma$(45000, 1.5)$ and Uniform$(25000, 35000)$ are used.
- For the (truncated) geometric mechanism, DGamma$(45000, 1.5)$ and discrete Uniform$(25000, 35000)$ are used.

For comparisons, we also show the properties of a naive method, in which we directly use the ratio of the released values as an estimate, that is, we use $\hat{r}_{\text{naive}} = a_{\text{rel}}/b_{\text{rel}}$.

Case 1 True values: $a = 300$ and $b = 30,000$

Privacy loss: $\epsilon = 0.05$.

Case 2 True values: $a = 30$ and $b = 3,000$

Privacy loss: $\epsilon = 0.05$

The prior distribution of b is either Gamma$(4500, 1.5)$ or Uniform$(2500, 3500)$ under the Laplace mechanism, and either DGamma$(4500, 1.5)$ or discrete Uniform$(2500, 3500)$ under the geometric mechanism.

Case 3 True values: $a = 300$ and $b = 30,000$

Privacy loss: $\epsilon = 0.25$.

Case 4 True values: $a = 300$ and $b = 30,000$

Privacy loss: $\epsilon = 0.05$

The prior distribution of b is Gamma$(48050, 1.55)$ under the Laplace mechanism and DGamma$(48050, 1.55)$ under the geometric mechanism.

Case 1 is the benchmark of simulation settings. Case 2 changes the true value of the (a, b) pair while retaining the true ratio and the privacy loss. Under the same true values of the numerator and denominator, the privacy loss of case 3 is greater than that of case 1. Case 4 demonstrates the performance of the inferences with poorly specified priors.

		PBIAS(%)	RMSE($\times 10^{-6}$)	Coverage(%)
	$\hat{r}_{\text{naive}}$	-.0010	9.38	-
Case 1.1	Gamma	-.0008	9.04	95.1
	Uniform	-.0009	9.39	94.7
	$\hat{r}_{\text{naive}}$	-.0009	95.1	-
Case 1.2	Gamma	.0003	70.8	98.7
	Uniform	-.0012	95.2	95.0
	$\hat{r}_{\text{naive}}$	.0001	1.89	-
Case 1.3	Gamma	.0001	1.89	94.8
	Uniform	.0001	1.89	94.7
	$\hat{r}_{\text{naive}}$	-.0015	9.26	-
Case 1.4	Biased Gamma	-.3870	39.1	61.8

Table 4.1: The percent bias (PBIAS), root-mean-square error (RMSE), and coverage of 95% posterior intervals of the ratio for Laplace mechanism when the numerator is the true value and the denominator is a noisy value. $t = 10,000$ simulation replications were drawn, for each replication, $n = 1,000$ samples were drawn from the posterior distribution to obtain the 95% posterior interval.

		PBIAS(%)	RMSE($\times 10^{-6}$)	Coverage(%)
	$\hat{r}_{\text{naive}}$	-.0015	9.61	-
Case 1.1	DGamma	-.0030	9.27	95.3
	Uniform	.0003	9.61	95.0
	$\hat{r}_{\text{naive}}$	.0035	95.2	-
Case 1.2	DGamma	.0091	70.5	98.9
	Uniform	.0208	95.3	94.7
	$\hat{r}_{\text{naive}}$	.0002	1.9	-
Case 1.3	DGamma	-.0015	1.91	95.2
	Uniform	.0021	1.92	95.8
	$\hat{r}_{\text{naive}}$	-.0006	9.54	-
Case 1.4	Biased DGamma	-.3884	39.2	62.7

Table 4.2: The percent bias (PBIAS), root-mean-square error (RMSE), and coverage of 95% posterior intervals of the ratio for the truncated geometric mechanism when the numerator is the true value and the denominator is a noisy value. $t = 10,000$ simulation replications were drawn, for each replication, $n = 1,000$ samples were drawn from the posterior distribution to obtain the 95% posterior interval.

The results for the Laplace mechanism and the (truncated) geometric mechanism are shown in Table 4.1 and Table 4.2, respectively. From the results, we observe that:

1. As long as the prior distributions are not poorly specified, the Bayesian inference method can provide posterior intervals with close to nominal coverage.

2. When the prior information is effective, where the prior mean is close to the true value, the Bayesian inference method has similarly small values of bias and mean squared error to the naive method.

3. The performance of the Bayesian inference method under the Laplace mechanism is similar to that under the truncated geometric mechanism.

4. Comparing Case 2 with Case 1, while the true ratio remains the same, reducing both a and b increases the impact of privacy noise, so the discrepancy between the true value and the estimated value is larger. This phenomenon is understandable because the influence of noise on the released values is larger, which makes it more difficult to infer the true values. The issue is that adding noise to the components of a ratio has a bigger effect when the components are small compared to when they are large.

5. In comparison with Case 1 and Case 3, the proposed inference method provides a more accurate estimate when increasing the privacy loss. This is reasonable because the variance of the noise added for privacy protection is smaller and thus the released value is closer to the true value.

6. Case 4 illustrates that with a largely biased prior, the Bayesian inference method with strong informative prior beliefs may not achieve satisfactory performance, while the proposed Bayesian inference method with non-informative priors provides a reasonable posterior interval. This reflects that our method has a certain discrepancy tolerance in prior beliefs, and the uniform prior is still a reliable choice when we are not confident enough in prior beliefs.

In short, Simulation 1 shows that the Bayesian method can provide posterior intervals with close to nominal coverage for the ratio, and has similarly small values of bias and mean squared error compared with the naive method.

4.2 Simulation 2: Estimated Value of a and Noisy Value of b

In this section, I present simulation results when the numerator is an estimate and the denominator is the true value with noise added. In this scenario, we have $a_{\text{rel}} = a_{\text{est}} \neq a$, and $b_{\text{rel}} = b + \delta_b$. As before, $t = 10,000$ simulation replication were drawn

for each configuration. Specifically, Jeffreys' prior is used on the numerator a. If not specifically mentioned, the prior distribution of b is as follows.

- For the Laplace mechanism, Gamma$(45000, 1.5)$ and Uniform$(25000, 35000)$ are used.
- For the (truncated) geometric mechanism, DGamma$(45000, 1.5)$ and discrete Uniform$(25000, 35000)$ are used.

Case 1 True values: $a = 300$ and $b = 30,000$

Privacy loss: $\epsilon = 0.05$,

Variance of the estimation for the numerator: $\text{Var}[a_{\text{est}} \mid a] = 200$.

Case 2 True values: $a = 30$ and $b = 3,000$

Privacy loss: $\epsilon = 0.05$

Variance of the estimation for the numerator: $\text{Var}[a_{\text{est}} \mid a] = 200$.

The prior distribution of b is either Gamma$(4500, 1.5)$ or Uniform$(2500, 3500)$ under the Laplace mechanism, and either DGamma$(4500, 1.5)$ or discrete Uniform$(2500, 3500)$ under the geometric mechanism.

Case 3 True values: $a = 300$ and $b = 30,000$

Privacy loss: $\epsilon = 0.25$,

Variance of the estimation for the numerator: $\text{Var}[a_{\text{est}} \mid a] = 200$.

Case 4 True values: $a = 300$ and $b = 30,000$

Privacy loss: $\epsilon = 0.05$,

Variance of the estimation for the numerator: $\text{Var}[a_{\text{est}} \mid a] = 200$.

The prior distribution of b is Gamma$(48050, 1.55)$ under the Laplace mechanism and DGamma$(48050, 1.55)$ under the geometric mechanism.

The design is similar to the simulation in Section 4.1. The results for the Laplace mechanism and the (truncated) geometric mechanism are shown in Table 4.3 and Table 4.4, respectively.

		PBIAS(%)	RMSE($\times 10^{-6}$)	Coverage(%)
	$\hat{r}_{\text{naive}}$	-.0843	477	-
Case 2.1	Gamma	-.0828	478	94.6
	Uniform	-.0829	478	94.4
	$\hat{r}_{\text{naive}}$	.1841	4665	-
Case 2.2	Gamma	.1987	4665	95.3
	Uniform	.1967	4665	95.2
	$\hat{r}_{\text{naive}}$	.0372	476	-
Case 2.3	Gamma	.0374	476	94.7
	Uniform	.0374	476	94.7
	$\hat{r}_{\text{naive}}$	.0194	471	-
Case 2.4	Biased Gamma	-.3656	470	94.6

Table 4.3: The percent bias (PBIAS), root-mean-square error (RMSE), and coverage of 95% posterior intervals of the ratio for the Laplace mechanism when the numerator is an estimate and the denominator is an noisy value. $t = 10,000$ simulation replications were drawn, for each replication, $n = 1,000$ samples were drawn from the posterior distribution to obtain the 95% posterior interval.

		PBIAS(%)	RMSE($\times 10^{-6}$)	Coverage(%)
	$\hat{r}_{\text{naive}}$	-.0907	464	-
Case 2.1	DGamma	-.0909	464	95.1
	Uniform	-.0877	464	95.2
	$\hat{r}_{\text{naive}}$	-.0961	4726	-
Case 2.2	DGamma	-.0747	4731	94.8
	Uniform	-.0613	4732	94.9
	$\hat{r}_{\text{naive}}$	.0052	477	-
Case 2.3	DGamma	.0038	477	94.4
	Uniform	.0073	477	94.4
	$\hat{r}_{\text{naive}}$	-.0193	469	-
Case 2.4	DGamma	-.4063	469	95.0

Table 4.4: The percent bias (PBIAS), root-mean-square error (RMSE), and coverage of 95% posterior intervals of the ratio for the truncated geometric mechanism when the numerator is an estimate and the denominator is a noisy value. $t = 10,000$ simulation replications were drawn, for each replication, $n = 1,000$ samples were drawn from the posterior distribution to obtain the 95% posterior interval.

The results are qualitatively similar to those in Table 4.1 and 4.2. Overall, the Bayesian inferences generate accurate inferences with reasonable coverage rates. The result for Case 4 is surprising, and this finding may not hold for other badly specified priors or other values of $\text{Var}[a_{\text{est}} \mid a]$. In fact, when I tried with a more biased prior (e.g. DGamma$(51200, 1.6)$), the coverage drops to 89%.

4.3 Simulation 3: Estimated Values of a and Estimated Values with Noise Added b

In this section, I present simulation results when the numerator is an estimate and the denominator is an estimate with noise added. In this scenario, we have $a_{\text{rel}} = a_{\text{est}} \neq a$, and $b_{\text{rel}} = b_{\text{est}} + \delta_b$. As before, $t = 10,000$ simulation replication were drawn for each configuration. Specifically, Jeffreys' prior is used on the numerator a. If not specifically mentioned, the prior distribution of b is as follows.

- For the Laplace mechanism, Gamma$(45000, 1.5)$ and Uniform$(25000, 35000)$ are used.
- For the (truncated) geometric mechanism, DGamma$(45000, 1.5)$ and discrete Uniform$(25000, 35000)$ are used.

Case 1 True values: $a = 300$ and $b = 30,000$

Privacy loss: $\epsilon = 0.05$

Variance of the estimation for the numerator: $\text{Var}[a_{\text{est}} \mid a] = 200$

Variance of the estimation for the denominator: $\text{Var}[b_{\text{est}} \mid b] = 40,000$

Case 2 True values: $a = 30$ and $b = 3,000$

Privacy loss: $\epsilon = 0.05$

Variance of the estimation for the numerator: $\text{Var}[a_{\text{est}} \mid a] = 200$

Variance of the estimation for the denominator: $\text{Var}[b_{\text{est}} \mid b] = 40,000$

The prior distribution of b is either Gamma$(4500, 1.5)$ or Uniform$(2500, 3500)$.

Case 3 True values: $a = 300$ and $b = 30,000$

Privacy loss: $\epsilon = 0.25$

Variance of the estimation for the numerator: $\text{Var}[a_{\text{rel}} \mid a] = 200$

Variance of the estimation for the denominator: $\text{Var}[b_{\text{est}} \mid b] = 40,000$

Case 4 True values: $a = 300$ and $b = 30,000$

Privacy loss: $\epsilon = 0.05$

Variance of the estimation for the numerator: $\text{Var}[a_{\text{rel}} \mid a] = 200$

Variance of the estimation for the denominator: $\text{Var}[b_{\text{est}} \mid b] = 40,000$

The prior distribution of b is Gamma$(48050, 1.55)$.

The design is similar to simulation in Section 4.1. The results for the Laplace mechanism and the (truncated) geometric mechanism are shown in Table 4.5 and Table 4.6, respectively.

		PBIAS(%)	RMSE($\times 10^{-6}$)	Coverage(%)
Case 3.1	$\hat{r}_{\text{naive}}$	.0399	478	-
	Gamma	.0462	475	94.6
	Uniform	.0452	478	94.6
Case 3.2	$\hat{r}_{\text{naive}}$	.6165	4768	-
	Gamma	.1351	4680	94.9
	Uniform	1.1189	4797	94.9
Case 3.3	$\hat{r}_{\text{naive}}$	-.0414	478	-
	Gamma	-.0508	473	94.7
	Uniform	-.0035	478	94.7
Case 3.4	$\hat{r}_{\text{naive}}$	-.0051	478	-
	Biased Gamma	-2.0706	509	92.4

Table 4.5: The percent bias (PBIAS), root-mean-square error (RMSE), and coverage of 95% posterior intervals of the ratio for Laplace mechanism when the numerator is the an estimate and the denominator is an estimate with noise added. $t = 10,000$ simulation replications were drawn, for each replication, $n = 1,000$ samples were drawn from the posterior distribution to obtain the 95% posterior interval.

		PBIAS(%)	RMSE($\times 10^{-6}$)	Coverage(%)
	$\hat{r}_{\text{naive}}$	-.0507	480	-
Case 3.1	Gamma	-.0584	475	94.9
	Uniform	-.0456	480	94.8
	$\hat{r}_{\text{naive}}$	.1465	4820	-
Case 3.2	Gamma	-.4148	4739	94.9
	Uniform	.6134	4848	94.7
	$\hat{r}_{\text{naive}}$	-.0053	478	-
Case 3.3	Gamma	-1.2265	475	94.6
	Uniform	.0001	479	94.7
	$\hat{r}_{\text{naive}}$	.0138	476	-
Case 3.4	Biased Gamma	-2.0920	507	92.6

Table 4.6: The percent bias (PBIAS), root-mean-square error (RMSE), and coverage of 95% posterior intervals of the ratio for the geometric mechanism when the numerator is the an estimate and the denominator is an estimate with noise added. $t = 10,000$ simulation replications were drawn, for each replication, $n = 1,000$ samples were drawn from the posterior distribution to obtain the 95% posterior interval.

The results are qualitatively similar to those in Table 4.3 and 4.4. Overall, the Bayesian inference method can generate accurate inferences with close to nominal coverage rates.

4.4 Simulation 4: Estimated Values with Noise Added to Both

In this section, I present simulation results when the numerator is an estimated value and the denominator is an estimated value with noise added to both. In this scenario, we have $a_{\text{rel}} = a_{\text{est}} + \delta_a$, and $b_{\text{rel}} = b_{\text{est}} + \delta_b$. As before, $t = 10,000$ simulation replication were drawn for each configuration. Specifically, Jeffreys' prior is used on the numerator a. If not specifically mentioned, the prior distribution of b is as follows.

- For the Laplace mechanism, Gamma$(45000, 1.5)$ and Uniform$(25000, 35000)$ are used.
- For the (truncated) geometric mechanism, DGamma$(45000, 1.5)$ and discrete Uniform$(25000, 35000)$ are used.

Case 1 True values: $a = 300$ and $b = 30,000$

Privacy loss: $\epsilon_a = \epsilon_b = 0.05$

Variance of the estimation for the numerator: $\text{Var}[a_{\text{est}} \mid a] = 200$

Variance of the estimation for the denominator: $\text{Var}[b_{\text{est}} \mid b] = 40,000$

Case 2 True values: $a = 30$ and $b = 3,000$

Privacy loss: $\epsilon_a = \epsilon_b = 0.05$

Variance of the estimation for the numerator: $\text{Var}[a_{\text{est}} \mid a] = 200$

Variance of the estimation for the denominator: $\text{Var}[b_{\text{est}} \mid b] = 40,000$

The prior distribution of b is either Gamma$(4500, 1.5)$ or Uniform$(2500, 3500)$.

Case 3 True values: $a = 300$ and $b = 30,000$

Privacy loss: $\epsilon_a = \epsilon_b = 0.25$

Variance of the estimation for the numerator: $\text{Var}[a_{\text{est}} \mid a] = 200$

Variance of the estimation for the denominator: $\text{Var}[b_{\text{est}} \mid b] = 40,000$

Case 4 True values: $a = 300$ and $b = 30,000$

Privacy loss: $\epsilon_a = \epsilon_b = 0.05$

Variance of the estimation for the numerator: $\text{Var}[a_{\text{est}} \mid a] = 200$

Variance of the estimation for the denominator: $\text{Var}[b_{\text{est}} \mid b] = 40,000$

The prior distribution of b is Gamma$(48050, 1.55)$.

Case 5 True values: $a = 300$ and $b = 30,000$

Privacy loss: $\epsilon_a = 0.05$, $\epsilon_b = 0.25$

Variance of the estimation for the numerator: $\text{Var}[a_{\text{est}} \mid a] = 200$

Variance of the estimation for the denominator: $\text{Var}[b_{\text{est}} \mid b] = 40,000$

The design of the first four case setting is similar to simulation in Section 4.1, while in Case 5, I evaluate the performance when the privacy loss for the numerator is different from the privacy loss for the denominator. The results for the Laplace mechanism and the (truncated) geometric mechanism are shown in Table 4.7 and Table 4.8, respectively.

		PBIAS(%)	RMSE($\times 10^{-6}$)	Coverage(%)
	$\hat{r}_{\text{naive}}$	.0341	1064	-
Case 4.1	Gamma	.0175	1062	94.9
	Uniform	.0398	1064	94.7
	$\hat{r}_{\text{naive}}$	.3714	10684	-
Case 4.2	Gamma	-.1025	10600	95.0
	Uniform	.7941	10742	95.1
	$\hat{r}_{\text{naive}}$	.0651	513	-
Case 4.3	Gamma	.0690	510	94.7
	Uniform	.0700	514	94.8
	$\hat{r}_{\text{naive}}$	.0051	1046	-
Case 4.4	Biased Gamma	-2.0569	1044	94.8
	$\hat{r}_{\text{naive}}$	.0954	1051	-
Case 4.5	Gamma	.0984	1050	94.9
	Uniform	.1068	1053	95.0

Table 4.7: The percent bias (PBIAS), root-mean-square error (RMSE), and coverage of 95% posterior intervals of the ratio for Laplace mechanism when the numerator is the an estimate and the denominator is an estimate with noise added to both. $t = 10,000$ simulation replications were drawn, for each replication, $n = 1,000$ samples were drawn from the posterior distribution to obtain the 95% posterior interval.

		PBIAS(%)	RMSE($\times 10^{-6}$)	Coverage(%)
Case 4.1	$\hat{r}_{\text{naive}}$	.0799	1078	-
	Gamma	.0721	1076	94.5
	Uniform	.0841	1078	94.6
Case 4.2	$\hat{r}_{\text{naive}}$	-.4126	10787	-
	Gamma	-.7402	10696	94.8
	Uniform	.0221	10843	94.7
Case 4.3	$\hat{r}_{\text{naive}}$	-.0022	514	-
	Gamma	-.0049	510	94.9
	Uniform	.0023	514	94.7
Case 4.4	$\hat{r}_{\text{naive}}$	.1880	1060	-
	Biased Gamma	-1.9280	1053	94.6
Case 4.5	$\hat{r}_{\text{naive}}$	-.1748	1064	-
	Gamma	-.1780	1064	94.3
	Uniform	-.1762	1065	94.5

Table 4.8: The percent bias (PBIAS), root-mean-square error (RMSE), and coverage of 95% posterior intervals of the ratio for the geometric mechanism when the numerator is the an estimate and the denominator is an estimate with noise added to both. $t = 10,000$ simulation replications were drawn, for each replication, $n = 1,000$ samples were drawn from the posterior distribution to obtain the 95% posterior interval.

The results are qualitatively similar to those in Table 4.5 and 4.6. Overall, the Bayesian inference method can generate accurate inferences with close to nominal coverage rates.

Chapter 5

Conclusion

For privacy considerations, the released values may have differentially private noise. When an estimate of the ratio between two indicators is desirable, the natural approach is to ignore the additional noise and directly use the released values to estimate the ratio. However, the Bayesian method can incorporate the knowledge of the differentially private mechanisms into inference and may provide a more accurate point estimate and posterior intervals for the ratio. When prior beliefs are available, the Bayesian inference can take them into account and provide a more accurate posterior intervals with reasonable coverage rates. When little information is known in advance or we do not have enough confidence in the prior belief, the Bayesian method with non-informative priors is a reliable choice. Future work may consider inferring the ratio under other differentially private mechanisms or other prior distributions and then make better use of the historical data. Additionally, future research may consider different statistic of interest, rather than ratios.

Appendix A

Random Sampling

This appendix describes how we sample from several distributions, including the DGamma distribution, the symmetric geometric distribution and the piecewise distribution. See the main text for a description of our overall inference strategy, formulating prior beliefs, generating samples from the posterior distribution, and providing point estimate and posterior intervals.

A.1 Random Sampling From SGeom Distribution

In this section, I present how to sample from a symmetric geometric distribution. The p.m.f. of $\text{SGeom}(c, \exp(\epsilon))$ is

$$\Pr(x = h \mid \epsilon) = \frac{1 - e^{-\epsilon}}{1 + e^{-\epsilon}} e^{-\epsilon|h-c|}, \ h \in \mathbb{Z}.$$

For notational simplicity, let $p = 1 - e^{-\epsilon}$. The p.m.f. of $\text{SGeom}(\exp(\epsilon))$ is

$$\Pr(x = h \mid \epsilon) = \frac{p}{2 - p}(1 - p)^{|h-c|}, \ h \in \mathbb{Z}.$$

Then we have

$$\sum_{h=c}^{+\infty} \Pr(x = h \mid \epsilon) = 1,$$

$$\sum_{h=c-1}^{-\infty} \Pr(x = h \mid \epsilon) = 1 - p.$$

Algorithm 2 provides the details of sampling from the $\text{SGeom}(c, \exp(\epsilon))$ distribution.

Algorithm 2 *Algorithm for sampling from the SGeom$(c, \exp(\epsilon))$ distribution*

1. Compute $p = 1 - e^{-\epsilon}$.
2. For $i = 1$ to n:
   - Generate a random indicator from the Bernoulli distribution:
   $$u^{(i)} \sim Bernoulli\left(\frac{1}{2-p}\right)$$
   - Generate a sample from the geometric distribution:
   $$y^{(i)} \sim Geo(p),$$
   where $\Pr(y = h \mid p) = p(1-p)^h, h \in \mathbb{N}$.
   - If $u^{(i)} = 1$, then $s^{(i)} = c + y^{(i)}$; otherwise, $s^{(i)} = c - 1 - y^{(i)}$.
3. Output samples $s^{(1)}, \cdots, s^{(n)}$.

I then will show that samples generated using Algorithm 2 follow SGeom$(c, \exp(\epsilon))$ distribution.

Proof. For $s - c \in \mathbb{N}$,

$$\begin{aligned}\Pr(x = s \mid \epsilon) &= \Pr(u = 1 \mid \epsilon)\Pr(y = s - c \mid \epsilon) \\ &= \frac{1}{2-p}p(1-p)^{(s-c)}.\end{aligned}$$

For $s - c \in \mathbb{Z}^-$,

$$\begin{aligned}\Pr(x = s \mid \epsilon) &= \Pr(u = 0 \mid \epsilon)\Pr(y = c - s - 1 \mid \epsilon) \\ &= \left(1 - \frac{1}{2-p}\right)p(1-p)^{(c-s-1)} \\ &= \frac{1}{2-p}p(1-p)^{(c-s)}.\end{aligned}$$

In this case,

$$\Pr(x = s \mid \epsilon) = \frac{1}{2-p}p(1-p)^{|s-c|},\ s \in \mathbb{Z}.$$

□

A.2 Random Sampling From DGamma Distribution

In this section, I present how to sample from a DGamma distribution. Some scholars have studied this kind of distribution as a discrete analogue of Gamma distribution (Chakraborty and Chakravarty, 2012; Kulasekera and Tonkyn, 1992; Doray and Luong, 1997; Chakraborty, 2015). The p.m.f. of DGamma(α, β) is

$$\Pr(x = k) \propto k^{\alpha-1} e^{-\beta k},\ \alpha > 1,\ \beta > 0,\ k \in \mathbb{N}$$

As $k \in \mathbb{N}$, it is difficult to sample from an infinite range. A natural way is to approximate $\Pr(x = k)$ by

$$\Pr(x = k) \approx \Pr\left(k - \frac{1}{2} \leq X_{\alpha,\beta} < k + \frac{1}{2}\right),\ k \in \mathbb{N},$$

where $X_{\alpha,\beta}$ follows a Gamma distribution with shape α and rate β.

Claim 3. The discrepancy between $\Pr(x = k)$ and $\Pr\left(k - \frac{1}{2} \leq X_{\alpha,\beta} < k + \frac{1}{2}\right)$ is relatively small.

Proof. Let $q(x)$ represent the p.d.f. of Gamma(α, β) evaluated at x, and $\Pr(x = k) = \frac{1}{C} q(k)$, where $C = \sum_{k=0}^{+\infty} q(k)$. To compute the difference, we first estimate the constant C. If C is close to 1, then $\Pr(x = k)$ is close to $q(k)$. Let $x_0 = \arg\max_{x \in \mathbb{R}^+} q(x) = (\alpha - 1)/\beta$. Since $\alpha > 1$, when $x \leq x_0$, $q(x)$ increases as x increases; otherwise, $q(x)$ decreases as x increases.

$$\begin{aligned}
|C - 1| &= \left| \sum_{k=0}^{+\infty} \Pr(x = k) - \int_0^{+\infty} q(x)dx \right| \\
&= \left| \sum_{k=0}^{+\infty} \int_k^{k+1} q(\lfloor x \rfloor)dx - \int_0^{+\infty} q(x)dx \right| \\
&= \left| \int_0^{+\infty} q(\lfloor x \rfloor)dx - \int_0^{+\infty} q(x)dx \right| \\
&= \left| \int_0^{+\infty} (q(\lfloor x \rfloor) - q(x))dx \right| \\
&\leq \left| \int_0^{\lfloor x_0 \rfloor} (q(\lfloor x \rfloor) - q(x))dx \right| + \left| \int_{\lfloor x_0 \rfloor}^{\lceil x_0 \rceil} (q(\lfloor x \rfloor) - q(x))dx \right| + \left| \int_{\lceil x_0 \rceil}^{+\infty} (q(\lfloor x \rfloor) - q(x))dx \right|,
\end{aligned}$$

where

$$\begin{aligned}\left|\int_0^{\lfloor x_0\rfloor}(q(\lfloor x\rfloor)-q(x))dx\right| &\le \int_0^{\lfloor x_0\rfloor}|q(\lfloor x\rfloor)-q(x)|dx\\ &\le \int_0^{\lfloor x_0\rfloor}|q(\lfloor x\rfloor)-q(\lceil x\rceil)|dx\\ &= \sum_{k=0}^{\lfloor x_0\rfloor-1}\int_k^{k+1}(q(\lceil x\rceil)-q(\lfloor x\rfloor))dx\\ &= \sum_{k=0}^{\lfloor x_0\rfloor-1}(q(k+1)-q(k))\\ &= q(\lfloor x_0\rfloor)-q(0)\\ &= q(\lfloor x_0\rfloor),\end{aligned}$$

$$\begin{aligned}\left|\int_{\lfloor x_0\rfloor}^{\lceil x_0\rceil}(q(\lfloor x\rfloor)-q(x))dx\right| &\le \left|\int_{\lfloor x_0\rfloor}^{x_0}(q(\lfloor x\rfloor)-q(x))dx\right|+\left|\int_{x_0}^{\lceil x_0\rceil}(q(\lfloor x\rfloor)-q(x))dx\right|\\ &\le |q(\lfloor x_0\rfloor)-q(x_0)|+\sup_{x\in[x_0,\lceil x_0\rceil]}|q(\lfloor x_0\rfloor)-q(x)|\\ &= q(x_0)-q(\lfloor x_0\rfloor)+\max\{q(x_0)-q(\lfloor x_0\rfloor),|q(\lfloor x_0\rfloor)-q(\lceil x_0\rceil)|\}\\ &\le q(x_0)-q(\lfloor x_0\rfloor)+|q(\lfloor x_0\rfloor)-q(x_0)|+|q(x_0)-q(\lceil x_0\rceil)|\\ &= 3q(x_0)-2q(\lfloor x_0\rfloor)-q(\lceil x_0\rceil),\end{aligned}$$

$$\begin{aligned}\left|\int_{\lceil x_0\rceil}^{+\infty}(q(\lfloor x\rfloor)-q(x))dx\right| &\le \int_{\lceil x_0\rceil}^{+\infty}|q(\lfloor x\rfloor)-q(x)|dx\\ &\le \int_{\lceil x_0\rceil}^{+\infty}|q(\lfloor x\rfloor)-q(\lceil x\rceil)|dx\\ &= \sum_{k=\lceil x_0\rceil}^{+\infty}\int_k^{k+1}(q(\lfloor x\rfloor)-q(\lceil x\rceil))dx\\ &= \sum_{k=\lceil x_0\rceil}^{+\infty}(q(k)-q(k+1))\\ &= q(\lceil x_0\rceil)-\lim_{n\to+\infty}q(n)\\ &= q(\lceil x_0\rceil).\end{aligned}$$

Thus, we have

$$\begin{aligned}|C-1| &\le \left|\int_0^{\lfloor x_0\rfloor}(q(\lfloor x\rfloor)-q(x))dx\right|+\left|\int_{\lfloor x_0\rfloor}^{\lceil x_0\rceil}(q(\lfloor x\rfloor)-q(x))dx\right|+\left|\int_{\lceil x_0\rceil}^{+\infty}(q(\lfloor x\rfloor)-q(x))dx\right|\\ &\le q(\lfloor x_0\rfloor)+3q(x_0)-2q(\lfloor x_0\rfloor)-q(\lceil x_0\rceil)+q(\lceil x_0\rceil)\\ &= 3q(x_0)-q(\lfloor x_0\rfloor).\end{aligned}$$

For a wide range of (α, β), $|C-1|$ is very small (Figure A.1). For example, when $\alpha = 45,000$ and $\beta = 1.5$, $|C-1|$ is less than 0.006.

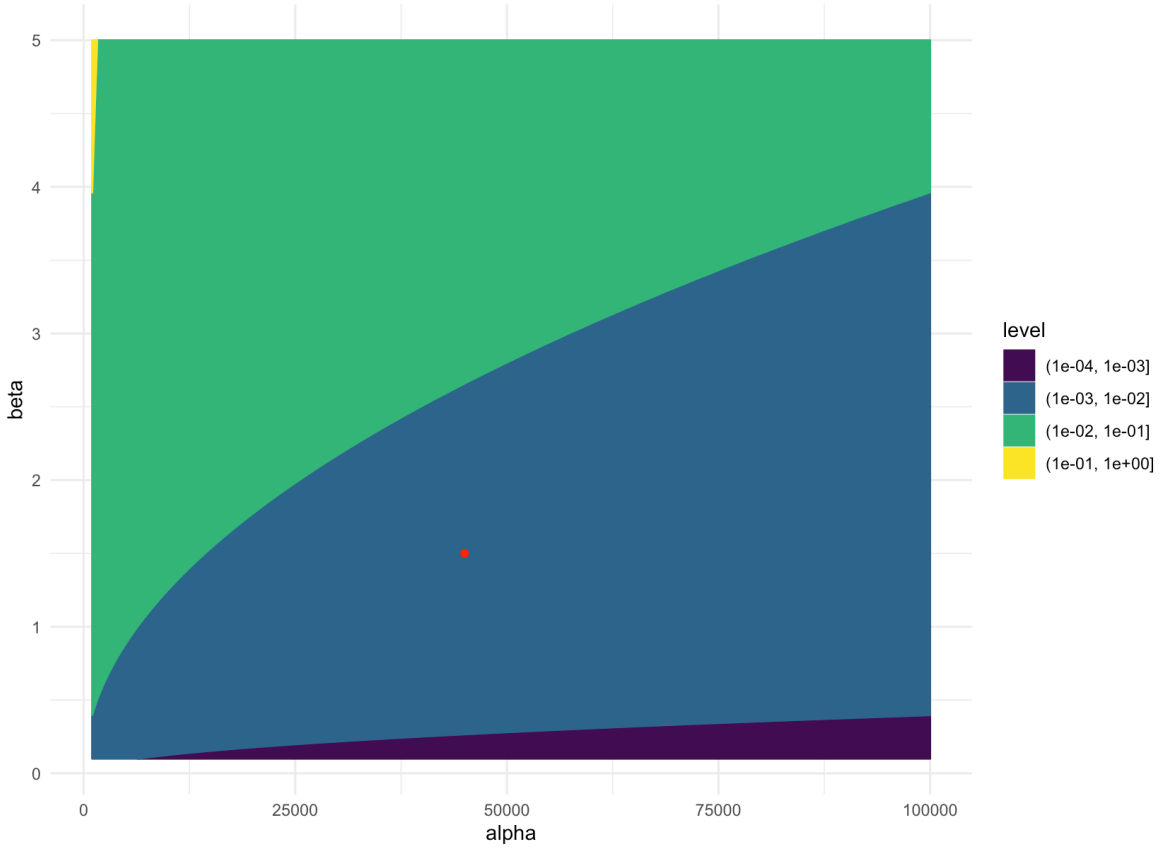

Figure A.1: The upper bound for $|C-1|$ under different values of (α, β). The red point represents the setting used in the simulation ($\alpha = 45,000$, $\beta = 1.5$).

Therefore,

$$\begin{aligned}\left|\Pr\left(k-\frac{1}{2} \le X_{\alpha,\beta} < k+\frac{1}{2}\right) - \Pr(x=k)\right| &= \left|\int_{k-0.5}^{k+0.5} q(x)dx - \frac{1}{C}q(k)\right| \\ &= \left|\int_{k-0.5}^{k+0.5} q(x)dx - q(k) + q(k) - \frac{1}{C}q(k)\right| \\ &\le \left|\int_{k-0.5}^{k+0.5} (q(x)-q(k))dx\right| + q(k)\frac{|C-1|}{C}.\end{aligned}$$

For the first term in the right, we use Taylor expansion around k to approximate $q(x)$,

$$q(x) = q(k) + \frac{q'(k)}{1!}(x-k) + \frac{q''(\theta_x)}{2!}(x-k)^2,$$

where

$$q'(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-2} e^{-\beta x}(\alpha - 1 - \beta x),$$
$$q''(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-3} e^{-\beta x}(\alpha^2 - 2\alpha\beta x - 3\alpha + \beta^2 x^2 + 2\beta x + 2),$$

and θ_x is related to x. Then, we have

$$\begin{aligned}
\left| \int_{k-0.5}^{k+0.5} (q(x) - q(k))dx \right| &= \left| \int_{k-0.5}^{k+0.5} \frac{q'(k)}{1!}(x-k) + \frac{q''(\theta_x)}{2!}(x-k)^2 dx \right| \\
&\leq \left| q'(k) \int_{k-0.5}^{k+0.5} (x-k)dx \right| + \left| \int_{k-0.5}^{k+0.5} q''(\theta_x) \frac{(x-k)^2}{2} dx \right| \\
&= \left| \int_{k-0.5}^{k+0.5} q''(\theta_x) \frac{(x-k)^2}{2} dx \right| \\
&\leq \int_{k-0.5}^{k+0.5} \left| q''(\theta_x) \right| \frac{(x-k)^2}{2} dx \\
&\leq \frac{1}{8} \sup_{x \in [k-0.5, k+0.5)} \left| q''(\theta_x) \right|.
\end{aligned}$$

The relative difference between $\Pr\left(k - \frac{1}{2} \leq X_{\alpha,\beta} < k + \frac{1}{2}\right)$ and $\Pr(x = k)$

$$\frac{\left| \Pr\left(k - \frac{1}{2} \leq X_{\alpha,\beta} < k + \frac{1}{2}\right) - \Pr(x = k) \right|}{\Pr(x = k)}$$

is bounded by

$$\sup_{x \in [k-0.5, k+0.5)} \left| \frac{q''(\theta_x)}{q(k)} \right| \frac{1 + B}{8} + B,$$

where B is the upper bound for $|C - 1|$. The upper bound of the relative difference is very small (Figure A.2). In the case when $\alpha = 45,000, \beta = 1.5$, the relative difference is below 2% for $x \in [25000, 35000]$.

□

In this case, we can first generate random samples from Gamma(α, β) and then round these values to get samples from DGamma(α, β).

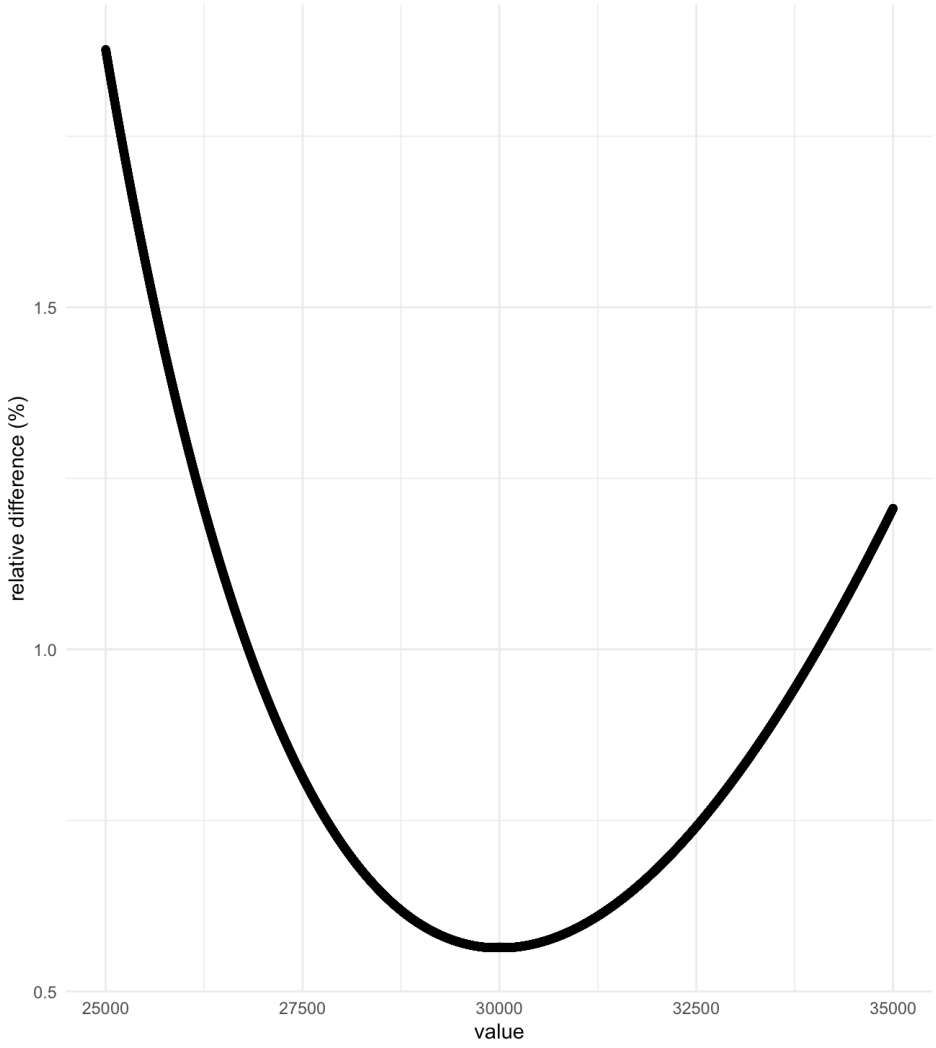

Figure A.2: Relative difference between $\Pr\left(k - \frac{1}{2} \leq X_{\alpha,\beta} < k + \frac{1}{2}\right)$ and $\Pr(x = k)$ for $k \in [25,000, 35,000]$ under the setting used in the simulation ($\alpha = 45,000$, $\beta = 1.5$).

A.3 Random Sampling From Piecewise Distributions

In this section, I present how to sample from a piecewise distribution. Some distributions are split into several parts, where the distribution in each part may be different. These types of distributions are called piecewise distributions. For simplicity, I consider only piecewise distributions with two parts. For piecewise distributions with more than two parts, analysts can apply my method multiple times and generate samples from them.

As an example, we consider a piecewise Gamma distribution, which can be split into two parts on $x = x_0$

$$p(x) = \begin{cases} C_1 p_1(x), & x \in \mathbb{R} \cap (0, x_0] \\ C_2 p_2(x), & x \in \mathbb{R} \cap [x_0, +\infty) \end{cases}$$

for $x \in \mathbb{R} \cup (0, +\infty)$, where $C_1 > 0$ and $C_2 > 0$ are two normalizing constants, $p_1(\cdot)$ and $p_2(\cdot)$ are the p.d.fs of the Gamma distribution for $x \in \mathbb{R} \cap (0, x_0)$ and $x \in \mathbb{R} \cap [x_0, +\infty)$, respectively. The first step is to ensure that $p(x)$ is a valid p.d.f. To do this, we need to ensure that $p(x)$ satisfies the following requirements:

1. $p(x) \geq 0$ for any $x \in \mathbb{R} \cup (0, +\infty)$.
2. $\int_{x \in \mathbb{R} \cup (0, +\infty)} p(x)\, dx = 1$.
3. $C_1 p_1(x_0) = C_2 p_2(x_0)$

The first requirement has been met since $C_1, C_2 > 0$ and $p_1(\cdot), p_2(\cdot) \geq 0$. For the other two requirements, we need to have

$$\begin{cases} C_1 \int_{x \in \mathbb{R} \cup (0, x_0)} p(x)\, dx + C_2 \int_{x \in \mathbb{R} \cup [x_0, +\infty)} p(x)\, dx = 1, \\ C_1 p_1(x_0) = C_2 p_2(x_0). \end{cases}$$

By solving these equations, we can obtain the values of C_1 and C_2, so that $p(x)$ can be calculated.

To generate a random sample from this distribution, we can first generate an indicator $u \sim Bernoulli\left(C_1 \int_{x \in \mathcal{X}_1} p_1(x)\, dx\right)$. If $u = 1$, then generate a random sample from $p_1(x)$; otherwise, generate a random sample from $p_2(x)$. Algorithm 3 provides the details of random sampling from a piecewise distribution with support on $\mathcal{X}_1 \cup \mathcal{X}_2$, where $\mathcal{X}_1 \cap \mathcal{X}_2 = x_0$.

Algorithm 3 *Algorithm for sampling from the piecewise distribution*

1. Compute C_1 and C_2 by solving equations

$$\begin{cases} C_1 \int_{x \in \mathcal{X}_1} p_1(x)\, dx + C_2 \int_{x \in \mathcal{X}_2} p_2(x)\, dx = 1, \\ C_1 p_1(x_0) = C_2 p_2(x_0). \end{cases}$$

2. For $i = 1$ to n:

   - Generate a random indicator from the Bernoulli distribution:

$$u^{(i)} \sim Bernoulli\left(C_1 \int_{x \in \mathcal{X}_1} p_1(x)\, dx\right)$$

   - If $u^{(i)} = 1$, then generate a sample from the distribution:

$$y^{(i)} \sim p_1(\cdot);$$

     otherwise, generate a sample from the distribution:

$$y^{(i)} \sim p_2(\cdot).$$

   - Set $s^{(i)} = y^{(i)}$.

3. Output samples $s^{(1)}, \cdots, s^{(n)}$.

I then will show that samples generated using Algorithm 3 follow the target distribution.

Proof. For $s \in \mathcal{X}_1$ and $s \neq x_0$,

$$\begin{aligned} p_x(s) &= \Pr(u = 1) p_{y \sim p_1}(s \mid s \in \mathcal{X}_1 \text{ and } s \neq x_0) \\ &= C_1 \int_{x \in \mathcal{X}_1} p_1(x)\, dx\; p_1(s \mid s \in \mathcal{X}_1 \text{ and } s \neq x_0) \\ &= C_1 p_1(s). \end{aligned}$$

For $s \in \mathcal{X}_2$ and $s \neq x_0$,

$$\begin{aligned} p_x(s) &= \Pr(u = 0) p_{y \sim p_2}(s \mid s \in \mathcal{X}_2 \text{ and } s \neq x_0) \\ &= \left(1 - C_1 \int_{x \in \mathcal{X}_1} p_1(x)\, dx\right) p_2(s \mid s \in \mathcal{X}_2 \text{ and } s \neq x_0) \\ &= C_2 p_2(s). \end{aligned}$$

In this case,

$$p(s) = \begin{cases} C_1 p_1(x), & x \in \mathcal{X}_1, \\ C_2 p_2(x), & x \in \mathcal{X}_2 \end{cases}$$

for $x \in \mathcal{X}_1 \cup \mathcal{X}_2$. □

Appendix B

Random Sampling From the Posterior Distribution

This appendix describes how we use importance sampling and rejection sampling to generate samples from the posterior distribution in a more efficient way. See the main text for a description of our overall inference strategy, formulating prior beliefs, generating samples from the posterior distribution, and providing point estimate and posterior intervals.

B.1 Random Sampling From $p(a, b \mid a_{\text{rel}}, b_{\text{rel}})$ when $a_{\text{rel}} = a_{\text{est}} \neq a$ and $b_{\text{rel}} = b_{\text{est}} + \delta_b$

In this section, I present how to sample from the posterior distribution of (a, b) pair when $a_{\text{rel}} = a_{\text{est}} \neq a$ and $b_{\text{rel}} = b_{\text{est}} + \delta_b$. Suppose we use a Jeffreys' prior on a (i.e., $p(a) \propto 1$). The posterior of (a, b) pair is

$$\begin{aligned} p(a, b \mid a_{\text{rel}}, b_{\text{rel}}) &\propto p(a)p(a_{\text{rel}} \mid a)p(b \mid a)p(b_{\text{rel}} \mid b)\mathbb{1}_{a_{\text{rel}}=a_{\text{est}}} \\ &= p(a)p(a_{\text{est}} \mid a)p(b \mid a)p(b_{\text{rel}} \mid b). \end{aligned}$$

As the posterior of a does not depend on b_{rel} or b, we first sample from $p(a \mid a_{\text{rel}})$ by directly sampling from $N(a_{\text{est}}, \text{Var}[a_{\text{est}} \mid a])$.

For b, if we use a uniform prior with support on $[n_1, n_2]$ (or a discrete uniform prior for discrete values), then we have $p(b \mid a) = \mathbb{1}_{b \geq a}\mathbb{1}_{b \in [n_1, n_2]}$. Thus, we can obtain a sample from the posterior distribution of b given (a, b_{rel}) by directly generating a sample from $N(b_{\text{est}}, \text{Var}[b_{\text{est}} \mid b])$, where b_{est} is generated from $\text{Lap}(b_{\text{rel}}, 1/\epsilon_b)$ (or $\text{SGeom}(b_{\text{rel}}, \exp(\epsilon_b))$). Since

$$\begin{aligned} p(b, b_{\text{est}} \mid a, b_{\text{rel}}) &\propto p(b_{\text{est}} \mid b_{\text{rel}})p(b \mid a, b_{\text{est}}, b_{\text{rel}}) \\ &= p(b_{\text{est}} \mid b_{\text{rel}})p(b \mid a, b_{\text{est}}) \\ &\propto p(b_{\text{est}} \mid b_{\text{rel}})p(b \mid a)p(b_{\text{est}} \mid b), \end{aligned}$$

we can obtain a sample of (b, b_{est}) given b_{rel}. In this case, we can simply treat the sample of b as the posterior sample of b, and ignore the sample of b_{est}. However, if we use a Gamma prior (or DGamma prior), this approach does not work since $p(b \mid a)$ is not a constant for $b \in [a, +\infty)$ and we need to account for $p(b \mid a)$. Instead, we can use rejection sampling to obtain samples from the posterior of b. A natural idea is to generate proposals of b from its prior $p(b \mid a)$, and then use the estimated $p(b_{\text{rel}} \mid b)$ as the acceptance probability. However, this approach may be very inefficient.

To improve efficiency, with a Gamma prior (or DGamma prior), I propose to re-parameterize $p(b \mid a)p(b_{\text{rel}} \mid b)$ before performing the rejection sampling. The idea is similar to importance sampling (Kahn, 1950a,b; Owen, 2013). Since

$$p(b \mid a)p(b_{\text{rel}} \mid b) = p(b \mid a)g(b) \times \frac{p(b_{\text{rel}} \mid b)}{g(b)},$$

where $g(\cdot)$ is a p.d.f. of another distribution and $g(b) > 0$ for all reasonable values of b, we can regard $p(b \mid a)g(b)$ as the proposal distribution and use $p(b_{\text{rel}} \mid b)/g(b)$ as the acceptance probability. The estimated $p(b_{\text{rel}} \mid b)$, as a function of b, has a symmetric and unimodal pattern, and it can be well approximated with a gamma distribution. In this case, we choose $g(\cdot)$ as the p.d.f. of a gamma distribution, where the mean is equal to b_{rel} and the variance is equal to $\text{Var}[b_{\text{est}} \mid b]$. Specifically, if we use $Gamma(\alpha, \beta)$ with support on $[a, +\infty)$ as a prior on b given a, we can generate proposals of b from the truncated Gamma distribution $Gamma(\alpha', \beta')$ with support on $[a, +\infty)$, where

$$\begin{aligned}\beta' &= \beta + b_{\text{obs}}/\text{Var}[b_{\text{est}} \mid b],\\ \alpha' &= \alpha + b_{\text{obs}}^2/\text{Var}[b_{\text{est}} \mid b] - 1.\end{aligned}$$

We then use $p(b_{\text{rel}} \mid b)/g(b)$ as the acceptance probability of proposals. Since $p(b_{\text{rel}} \mid b)/g(b)$ is much flatter than $p(b_{\text{rel}} \mid b)$, we can generate samples from the posterior of b more efficiently. Algorithm 4 provides the details of sampling from the posterior of a/b given a_{rel} and b_{rel}.

B.2 Random Sampling From $p(a, b \mid a_{\text{rel}}, b_{\text{rel}})$ when $a_{\text{rel}} = a_{\text{est}} + \delta_a$ and $b_{\text{rel}} = b_{\text{est}} + \delta_b$

In this section, I present how to sample from the posterior distribution of (a, b) pair when $a_{\text{rel}} = a_{\text{est}} + \delta_a$ and $b_{\text{rel}} = b_{\text{est}} + \delta_b$. Assume that the estimated values of $\text{Var}[a_{\text{est}} \mid a]$ and $\text{Var}[b_{\text{est}} \mid b]$ are provided. Suppose we use a Jeffreys' prior on a (i.e., $p(a) \propto 1$). The posterior of (a, b) pair is

$$p(a, b \mid a_{\text{rel}}, b_{\text{rel}}) \propto p(a)p(a_{\text{rel}} \mid a)p(b \mid a)p(b_{\text{rel}} \mid b).$$

Since the posterior of a does not depend on b_{rel} or b and $p(a) \propto 1$, we first sample from $p(a \mid a_{\text{rel}}$ by directly sampling from $N(a_{\text{est}}, \text{Var}[a_{\text{est}} \mid a])$, where a_{est} is generated from $\text{Lap}(a_{\text{rel}}, 1/\epsilon_a)$ (or $\text{SGeom}(a_{\text{rel}}, \exp(\epsilon_a))$ for discrete values). With a sample generated from the posterior of a, we can use the approach described in Appendix B.1 to generate a sample from the posterior of b. In this case, we can obtain a posterior sample of (a, b). Algorithm 5 provides the details of sampling from the posterior of a/b given a_{rel} and b_{rel}.

Algorithm 4 *Algorithm for sampling from the posterior distribution of a/b when $a_{\text{rel}} = a_{\text{est}}$ and $b_{\text{rel}} = b_{\text{est}} + \delta_b$*

1: Let $\sigma_a^2 = \text{Var}[a_{\text{est}} \mid a]$ and $\sigma_b^2 = \text{Var}[b_{\text{est}} \mid b]$.
2: For $i = 1$ to n:

- Generate a random sample of a given a_{est} where $a_{\text{est}} = a_{\text{rel}}$:

$$a^{(i)} \sim N(a_{\text{est}}, \sigma_a^2).$$

- With the sample $a^{(i)}$,
  - If we use a Gamma prior (or a DGamma prior),
    1. Generate a proposal of b from the prior distribution:

$$b^{prop} \sim p(b \mid a^{(i)})g(b).$$

    2. Estimate $p(b_{\text{rel}} \mid b^{prop})$ by $\frac{1}{M}\sum_{m=1}^{M} p(b_{\text{rel}} \mid b_{\text{est}}^m, b^{prop})$, where $b_{\text{est}}^1, \cdots, b_{\text{est}}^M$ are independently generated from $N(b^{prop}, \sigma_b^2)$.
    3. Generate a random indicator from the Bernoulli distribution:

$$u \sim Bernoulli\left(\min\left\{1, \frac{p(b_{\text{rel}} \mid b^{prop})}{g(b)}\right\}\right)$$

    4. If $u = 1$, then accept the proposal:

$$b^{(i)} = b^{prop};$$

    otherwise, repeat (a)-(c) until we find $b^{(i)}$.
  - If we use a uniform prior (or a discrete uniform prior),
    1. Generate a random sample of b_{est} given b_{rel}:

$$b_{\text{est}}^{(i)} \sim p(b_{\text{est}} \mid b_{\text{rel}}),$$

    where $p(b_{\text{est}} \mid b_{\text{rel}}) = \text{Lap}(b_{\text{rel}}, 1/\epsilon_b)$ (or the p.m.f. of $\text{SGeom}(b_{\text{rel}}, \exp(\epsilon_b))$).
    2. Generate a random sample of b given $b_{\text{est}}^{(i)}$:

$$b^{(i)} \sim N(b_{\text{est}}^{(i)}, \sigma_b^2).$$

- Compute $r^{(i)} = a^{(i)}/b^{(i)}$

3: Output samples $r^{(1)}, \cdots, r^{(n)}$.

Algorithm 5 *Algorithm for sampling from the posterior distribution of a/b when $a_{\text{rel}} = a_{\text{est}} + \delta_a$ and $b_{\text{rel}} = b_{\text{est}} + \delta_b$*

1: Let $\sigma_a^2 = \text{Var}[a_{\text{est}} \mid a]$ and $\sigma_b^2 = \text{Var}[b_{\text{est}} \mid b]$.
2: For $i = 1$ to n:

- Generate a random sample of a_{est} given a_{rel}:

$$a_{\text{est}}^{(i)} \sim p(a_{\text{est}} \mid a_{\text{rel}}),$$

  where $p(a_{\text{est}} \mid a_{\text{rel}}) = \text{Lap}(a_{\text{rel}}, 1/\epsilon_a)$ (or the p.m.f. of $\text{SGeom}(a_{\text{rel}}, \exp(\epsilon_a))$).

- Generate a random sample of a given $a_{\text{est}}^{(i)}$:

$$a^{(i)} \sim N(a_{\text{est}}^{(i)}, \sigma_a^2).$$

- With the sample $a^{(i)}$,
  - If we use a Gamma prior (or a DGamma prior),
    1. Generate a proposal of b from the prior distribution:

$$b^{prop} \sim p(b \mid a^{(i)})g(b).$$

    2. Estimate $p(b_{\text{rel}} \mid b^{prop})$ by $\frac{1}{M}\sum_{m=1}^{M} p(b_{\text{rel}} \mid b_{\text{est}}^m, b^{prop})$, where $b_{\text{est}}^1, \cdots, b_{\text{est}}^M$ are independently generated from $N(b^{prop}, \sigma_b^2)$.
    3. Generate a random indicator from the Bernoulli distribution:

$$u \sim Bernoulli\left(\min\left\{1, \frac{p(b_{\text{rel}} \mid b^{prop})}{g(b)}\right\}\right)$$

    4. If $u = 1$, then accept the proposal:

$$b^{(i)} = b^{prop};$$

    otherwise, repeat (a)-(c) until we find $b^{(i)}$.
  - If we use a uniform prior (or a discrete uniform prior),
    1. Generate a random sample of b_{est} given b_{rel}:

$$b_{\text{est}}^{(i)} \sim p(b_{\text{est}} \mid b_{\text{rel}}),$$

    where $p(b_{\text{est}} \mid b_{\text{rel}}) = \text{Lap}(b_{\text{rel}}, 1/\epsilon_b)$ (or the p.m.f. of $\text{SGeom}(b_{\text{rel}}, \exp(\epsilon_b))$).
    2. Generate a random sample of b given $b_{\text{est}}^{(i)}$:

$$b^{(i)} \sim N(b_{\text{est}}^{(i)}, \sigma_b^2).$$

- Compute $r^{(i)} = a^{(i)}/b^{(i)}$

3: Output samples $r^{(1)}, \cdots, r^{(n)}$.

Bibliography

Abowd, J. M. (2018). The US Census Bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2867–2867.

Andrés, M. E., Bordenabe, N. E., Chatzikokolakis, K., and Palamidessi, C. (2013). Geo-Indistinguishability: Differential Privacy for Location-Based Systems. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, pages 901–914.

Antón, A. I., Earp, J. B., and Young, J. D. (2010). How internet users' privacy concerns have evolved since 2002. *IEEE Security & Privacy*, 8(1):21–27.

Borders, K. and Prakash, A. (2009). Quantifying Information Leaks in Outbound Web Traffic. In *2009 30th IEEE Symposium on Security and Privacy*, pages 129–140. IEEE.

Carter-Pokras, O. and Baquet, C. (2002). What is a "health disparity" ? *Public Health Reports*, 117(5):426.

Chakraborty, S. (2015). Generating discrete analogues of continuous probability distributions-A survey of methods and constructions. *Journal of Statistical Distributions and Applications*, 2(1):6.

Chakraborty, S. and Chakravarty, D. (2012). Discrete Gamma Distributions: Properties and Parameter Estimations. *Communications in Statistics - Theory and Methods*, 41(18):3301–3324.

Chen, R., Mohammed, N., Fung, B. C., Desai, B. C., and Xiong, L. (2011). Publishing Set-valued Data via Differential Privacy. *Proceedings of the VLDB Endowment*, 4(11):1087–1098.

Dankar, F. K. and El Emam, K. (2012). The Application of Differential Privacy to Health Data. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pages 158–166.

Dankar, F. K. and El Emam, K. (2013). Practicing Differential Privacy in Health Care: A Review. *Transactions on Data Privacy*, 6(1):35–67.

Doray, L. G. and Luong, A. (1997). Efficient estimators for the good family. *Communications in Statistics - Simulation and Computation*, 26(3):1075–1088.

Dwork, C. (2008). Differential Privacy: A Survey of Results. In *International Conference on Theory and Applications of Models of Computation*, pages 1–19. Springer.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer.

Dwork, C., Roth, A., et al. (2014). *The Algorithmic Foundations of Differential Privacy*. Now Foundations and Trends.

Ghosh, A., Roughgarden, T., and Sundararajan, M. (2012). Universally Utility-Maximizing Privacy Mechanisms. *SIAM Journal on Computing*, 41(6):1673–1693.

Holohan, N., Antonatos, S., Braghin, S., and Mac Aonghusa, P. (2019). The Bounded Laplace Mechanism in Differential Privacy. *Journal of Privacy and Confidentiality*, 10(1).

Horvitz, D. G. and Thompson, D. J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, 47(260):663–685.

Jones, D. S. (2006). The Persistence of American Indian Health Disparities. *American Journal of Public Health*, 96(12):2122–2134.

Kahn, H. (1950a). Random Sampling (Monte Carlo) Techniques in Neutron Attenuation Problems. I. *Nucleonics*, 6(5):27–37.

Kahn, H. (1950b). Random Sampling (Monte Carlo) Techniques in Neutron Attenuation Problems. II. *Nucleonics*, 6(6):60–65.

Kulasekera, K. B. and Tonkyn, D. W. (1992). A new discrete distrlbution, with applications to survival, dispersal and dispersion. *Communications in Statistics - Simulation and Computation*, 21(2):499–518.

Li, J., Tao, Y., and Xiao, X. (2008). Preservation of Proximity Privacy in Publishing Numerical Sensitive Data. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 473–486.

Li, N., Lyu, M., Su, D., and Yang, W. (2016). *Differential Privacy: From Theory to Practice*. Morgan & Claypool Publishers.

Loukides, G., Denny, J. C., and Malin, B. (2010). The disclosure of diagnosis codes can breach research participants' privacy. *Journal of the American Medical Informatics Association*, 17(3):322–327.

Malin, B., Benitez, K., and Masys, D. (2011). Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. *Journal of the American Medical Informatics Association*, 18(1):3–10.

Ohm, P. (2009). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, 57:1701.

Owen, A. B. (2013). *Monte Carlo theory, methods and examples.*

Pearcy, J. N. and Keppel, K. G. (2002). A Summary Measure of Health Disparity. *Public Health Reports*, 117(3):273–280.

Polonetsky, J., Tene, O., and Finch, K. (2016). Shades of Gray: Seeing the Full Spectrum of Practical Data De-identification. *Santa Clara Law Review*, 56:593.

Quan, D., Yin, L., and Guo, Y. (2015). Enhancing the Trajectory Privacy with Laplace Mechanism. In *2015 IEEE Trustcom/BigDataSE/ISPA*, volume 1, pages 1218–1223. IEEE.

Reiter, J. P. (2019). Differential Privacy and Federal Data Releases. *Annual Review of Statistics and Its Application*, 6:85–101.

Rocher, L., Hendrickx, J. M., and De Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10(1):1–9.

Roden, D. M., Pulley, J. M., Basford, M. A., Bernard, G. R., Clayton, E. W., Balser, J. R., and Masys, D. R. (2008). Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clinical Pharmacology & Therapeutics*, 84(3):362–369.

Rothstein, M. A. (2010). Is Deidentification Sufficient to Protect Health Privacy in Research? *The American Journal of Bioethics*, 10(9):3–11.

Sarathy, R. and Muralidhar, K. (2011). Evaluating Laplace Noise Addition to Satisfy Differential Privacy for Numeric Data. *Transactions on Data Privacy*, 4(1):1–17.

Wang, Y., Wu, X., and Hu, D. (2016). Using Randomized Response for Differential Privacy Preserving Data Collection. In *EDBT/ICDT Workshops*, volume 1558.

White, M. C., Espey, D. K., Swan, J., Wiggins, C. L., Eheman, C., and Kaur, J. S. (2014). Disparities in cancer mortality and incidence among American Indians and Alaska Natives in the United States. *American Journal of Public Health*, 104(S3):S377–S387.

Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., Honaker, J., Nissim, K., OBrien, D. R., Steinke, T., and Vadhan, S. (2018). Differential privacy: A primer for a non-technical audience. *Vanderbilt Journal of Entertainment & Technology Law*, 21(1):209–275.

Yang, Y., Zhang, Z., Miklau, G., Winslett, M., and Xiao, X. (2012). Differential Privacy in Data Publication and Analysis. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 601–606.