arXiv:2301.08324v2 [stat.ME] 11 Apr 2024

《电子统计学杂志》第18卷（2024年），
1455–1494页 ISSN：1935-7524 DOI：
https://doi.org/10.1214/24-EJS2234

分层随机抽样下比例的差分隐私置信区间*

Shurong Lin

波士顿大学数学与统计系，美国马萨诸塞州波士顿，电子邮箱： shrlin@bu.edu

Mark Bun 和 Marco Gaboardi

波士顿大学计算机科学系，美国马萨诸塞州波士顿，电子邮箱：
mbun@bu.edu; gaboardi@bu.edu

Eric D. Kolaczyk

麦吉尔大学数学与统计系，加拿大 电子邮箱： eric.kolaczyk@mcgill.ca

Adam Smith

波士顿大学计算机科学系，马萨诸塞州波士顿 电子邮箱： ads22@bu.edu

摘要： 置信区间是量化感兴趣参数不确定性的基本工具。随着数据隐私意识的增强，开发隐私版本的置信区间日益受到统计学家和计算机科学家的关注。差分隐私是分析发布由敏感数据计算得出的统计量时隐私损失的先进框架。近期已有关于差分隐私置信区间的研究，但据我们所知，在调查抽样背景下，尚未研究差分隐私置信区间的严谨方法。本文提出三种差分隐私算法，用于在分层随机抽样下构造比例置信区间。我们阐述了适用于分层抽样设计数据的两种差分隐私变体，并分别在其中一种变体下分析每种算法。我们建立了分析性的隐私保证和估计量的渐近性质。此外，我们开展模拟研究以评估所提出的隐私置信区间，并给出两个针对1940年人口普查数据的应用。

2020年数学学科分类： 主要分类 68P27、62G15；次要分类 62Dxx。

关键词与短语： 差分隐私、置信区间、分层抽样、总体比例。

*本文所述研究由美国人口普查局合作协议 CB20ADR0160001 资助。

目录

1 引言.1456
1.1 相关工作.1458
2 预备知识.1458
2.1 总体比例置信区间.1458
2.2 差分隐私.1459
3 方法.1461

3.1 使用公开样本量进行估计.1461
3.1.1 在层级添加噪声.1462
3.1.2 在总体层级添加噪声.1462
3.2 使用私有样本量进行估计.1464

4 理论结果.1467
4.1 隐私与覆盖保证.1467
4.2 方差比较.1470
4.2.1 外在方差.1470
4.2.2 与非隐私置信区间的比较：单层情形 . . . 1471

5 数值结果.1472
5.1 模拟.1472
5.1.1 正态性检验.1472
5.1.2 变化的关键参数.1473
5.2 应用.1476
5.2.1 失业率的置信区间 1476
5.2.2 收入水平差异的置信区间 . . 1477

6 讨论.1479
A 证明.1480
A.1 定理 3.1 的证明.1480
A.2 定理 4.1 的证明.1484
A.3 定理 4.2 的证明.1485
A.4 定理 4.3 的证明.1486
A.5 定理 4.4 的证明.1487
A.6 定理 4.5 的证明.1488

致谢.1492
参考文献.1492

1. 引言

随着现代信息时代隐私意识的增强，为统计学和机器学习建立隐私保护方法已成为一个活跃的研究领域。差分隐私是一种先进的隐私保护技术 [14]，，被视为严格隐私保证的黄金标准。它不仅在学术界引起了广泛关注 [15, 16]„ 还已被政府、企业及其他数据机构采用，例如美国人口普查局 [1]，、Google [20]，、Microsoft [8]，和 Apple [38]。

近日，美国人口普查局发布了其针对2020年人口普查的差分隐私披露规避系统（DAS）新版演示 [5, 24]。在差分隐私与统计学的交汇领域，统计学家和计算机科学家都在致力于开发统计推断程序的隐私版本。早期探讨统计学背景下差分隐私的研究包括 [17, 13, 42, 37]。近期研究则探索了在差分隐私约束下进行统计推断和估计 [6, 29, 31]。

作为统计推断中最基础的工具之一，置信区间在量化感兴趣参数的不确定性方面无处不在。本文提出了三种差分隐私算法，用于在分层随机抽样下构建总体比例的置信区间。据我们所知，这是首次在调查抽样背景下为差分隐私置信区间建立严谨方法论。调查抽样是统计学中的一个重要领域，涉及从目标总体中抽取个体样本进行调查。它能及时且经济高效地估计人们感兴趣的总体特征，并广泛应用于大规模数据收集，例如美国社区调查（ACS）、收入与项目参与调查（SIPP）以及当前人口调查（CPS）。

本文首次研究了基于分层抽样设计数据的差分隐私置信区间。具体而言：

- 我们阐述了两种适用于分层抽样设计数据的差分隐私具体变体。除了标准的差分隐私概念外，我们还考虑样本层样本量固定且公开的情形。后一种情形可使用更简单的算法，并获得更窄的置信区间。

- 我们提出了将应用差分隐私机制（添加随机噪声）所带来的不确定性传递至置信区间构造中的方法。为获得（渐近）无偏的方差估计，需要进行偏差校正。我们还给出了中心极限定理（CLT）类型的结论，以保证渐近置信水平。

- 我们通过理论分析和模拟评估所提出算法的性能。对非隐私方法与隐私方法进行比较的理论分析，可以帮助实践者在将算法应用于真实数据之前了解其表现。
- 为支持对其中一种算法的理论分析，我们深入研究了正态随机变量的倒数的行为。我们得到了条件矩的泰勒展开的一般形式，以解决其重尾性质导致矩不存在的问题。

本文结构安排如下。我们在第1.1节简要讨论现有的差分隐私置信区间相关研究。第2节介绍总体比例置信区间和差分隐私的预备知识。第3节讨论三种差分隐私算法的方法。第4节给出关于隐私和渐近覆盖保证的定理。

第5节开展数值实验，包括模拟研究以及针对1940年人口普查数据的两个应用。第6节讨论我们的方法所带来的启示，以及差分隐私置信区间领域的一般研究方向。

1.1. 相关工作

近期，研究者已针对其他情形研究了差分隐私置信区间。一些研究探讨了正态分布数据总体均值的差分隐私置信区间 [27, 11, 23] 。研究者还探索了置信区间方面的其他任务。Drechsler等人设计并评估了若干种用于获得中位数差分隐私置信区间的策略[10]。Wang等人针对使用目标扰动或输出扰动算法训练的差分隐私模型，提供了置信区间 [41]。

此外，自助法是构建更通用的差分隐私置信区间的一种常用技术。Ferrando等人提出了一种通用方法，用于构建总体参数的置信区间[21]。研究者还给出了均值差的数值置信区间 [9]。文献 [3]中探讨了非参数自助法。Covington等人描述了一种通过小批量自助法（BLB）生成均值和协方差估计分布的方法，进而可据此构建隐私置信区间[7]。

我们的工作首次研究了针对抽样的基于设计的方法。在基于设计的设定中，关注的数值被视为固定但未知的常数。随机性仅来自抽样设计。设计中引入的选择概率将用于估计。相反，在基于模型的设定中，则假定一个参数模型。在许多情况下，尤其是对于无法获得总体分布准确先验信息的自然总体，基于设计的抽样方法可能更令人放心。关于抽样中基于设计与基于模型的方法的更多讨论，见 [39]。

2. 预备知识

本节介绍总体比例估计和差分隐私方面的一些预备知识。我们首先回顾分层随机抽样下总体比例的经典 Wald置信区间。随后，我们针对分层数据定义差分隐私的概念。为第4节的理论分析做准备，我们还将重新考察差分隐私的一些性质。

2.1. 总体比例置信区间

在分层随机抽样中，由 N 个体组成的总体被划分为H 层，其中第 h 层有 N_h 个体，并在每层内进行简单随机抽样

从 n_h 个体中进行。在目标是估计总体中具有某种属性的个体比例时，可以用样本比例进行估计。设 y_{hi} 为相应的指示变量：当第 h 层中的个体 i 具有该属性时， $y_{hi}=1$ ；否则为 $y_{hi}=0$ 。可以估计总体比例

$$p=\frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h}y_{hi}$$

基于样本比例

$$\hat{p}=\frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{i=1}^{n_h}y_{hi}=\sum_{h=1}^{H}w_h\hat{p}_h$$

其中 $w_h \overset{def}{=} \frac{N_h}{N}$ 以及 $\hat{p}_h \overset{def}{=} \frac{1}{n_h}\sum_{i=1}^{n_h}y_{hi}$。其方差 $\mathrm{Var}(\hat{p})=\sum_{h=1}^{H}w_h^2\mathrm{Var}(\hat{p}_h)$，其中

$$\mathrm{Var}(\hat{p}_h)=\left(\frac{N_h-n_h}{N_h-1}\right)\frac{p_h(1-p_h)}{n_h}.$$

$\mathrm{Var}(\hat{p}_h)$ 的无偏估计量由该层中的样本方差给出

$$\widehat{\mathrm{Var}}(\hat{p}_h)=\left(\frac{N_h-n_h}{N_h}\right)\frac{\hat{p}_h(1-\hat{p}_h)}{n_h-1}. \tag{1}$$

因此，$\mathrm{Var}(\hat{p})$ 的无偏估计量由 $\widehat{\mathrm{Var}}(\hat{p})=\sum w_h^2\widehat{\mathrm{Var}}(\hat{p}_h)$。基于正态分布，可以构造 p 的近似 $100\%(1-\alpha)$ 置信区间：

$$\hat{p}\pm z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\hat{p})}, \tag{2}$$

其中 $z_{1-\frac{\alpha}{2}}$ 表示标准正态分布的 $1-\frac{\alpha}{2}$ 分位数。当所有样本量均为中等或较大时，正态近似是有用的。否则，通常使用具有适当自由度的 t 分布来替代标准正态分布。对于小样本量，已有多种专门的置信区间 [22]。

2.2. 差分隐私

差分隐私确保当数据集中的一条记录发生变化时，数据分析或查询的输出不会有太大差异，因此无法推断任何个人是否在数据集中。如果两个数据集 $\boldsymbol{x},\boldsymbol{x}'$ 相差一条记录，我们就称 $\boldsymbol{x},\boldsymbol{x}'$ 彼此相邻 或邻近，记作 $\boldsymbol{x}\sim\boldsymbol{x}'$。差分隐私的定义取决于邻接关系的定义。对于分层抽样下的分区数据，改变一条

记录有两种方式：（1）一种方式是在某一层内替换一条记录，同时保持各层的样本量不变。我们将这种邻接关系称为“层内替换一条记录关系”，记作 $\sim_{ss}$。这种关系对应于样本量公开且固定的情况；（2）另一种获得相邻数据集的方式是从某一层中移除或添加一条记录；我们将相应的关系称为“移除/添加一条记录关系”，记作 $\sim_{r}$。在这种情况下，其中一层的样本量将变化一条记录，整体样本量也会随之改变。这种关系对应于样本量为私有的情况。

在任一邻接关系下，我们都可以定义零集中差分隐私(ρ-零集中差分隐私)，如下 [4]:

定义 1 (ρ-零集中差分隐私)。 令 $\boldsymbol{X}^*$ 表示输入数据的空间，其维度可以是任意有限值。在邻接关系 $\sim$ 下，随机化算法 $M:\boldsymbol{X}^* \to Y$ 是 ρ-零集中差分隐私（ρ-零集中差分隐私）的，当且仅当对于每一对相邻数据集 $\boldsymbol{x} \sim \boldsymbol{x}' \in \boldsymbol{X}^*$ 以及所有 $\alpha \in (1,\infty)$，

$$\mathrm{D}_\alpha(M(\boldsymbol{x})\|M(\boldsymbol{x}')) \leq \rho\alpha,$$

其中 $\mathrm{D}_\alpha(M(\boldsymbol{x})\|M(\boldsymbol{x}'))$ 是 α-*Rényi*散度 *[40]*，表示 $M(\boldsymbol{x})$ 的分布与 $M(\boldsymbol{x}')$的分布之间的差异。

参数 ρ 表示隐私级别。 ρ 越小， $M(\boldsymbol{x})$ 与 $M(\boldsymbol{x}')$ 之间的距离控制就越严格。因此，两个相邻数据集上的输出更难区分，算法也能实现更高的隐私保护。当我们有意设计算法以满足 ρ-零集中差分隐私时，将 ρ 称为隐私预算。

根据相邻关系的定义，差分隐私分为两种类型：有界和无界差分隐私 [28]。定义1 在“移除/添加一条记录关系”下，对应于标准的无界差分隐私。添加或移除一条记录以获得相邻数据集时，数据集的样本量会发生变化。在“层内替换一条记录”关系下 $\sim_{ss}$，得到的概念对应于有界差分隐私，其中两个相邻数据集的大小相同。但它与标准的有界差分隐私概念略有不同，因为在后者中，替换可以跨层进行。也就是说，我们可以同时更改记录及其所属的层。

在差分隐私文献中，(ϵ，δ)-DP（[15] 定义 2.4）被视为经典概念。我们考虑 ρ-zCDP，原因有三：（1） ρ-zCDP 蕴含 (ϵ，δ)-DP（[4] 命题 1.3）；（2）采用高斯机制实现 zCDP 有助于理论分析；（3）ρ-zCDP 的组合简单明了。高斯机制是满足 zCDP 的典型机制示例，它通过添加高斯噪声来扰动真实值。我们在以下命题中给出高斯机制，以及 ρ-zCDP 的组合与后处理性质。所有命题均可在 [4] 中找到，并将在第4节中用于隐私保证的分析。

定义 2 (敏感度)。函数 $q: X^* \to \mathbb{R}$ 的敏感度为 Δ，当且仅当对于所有相邻数据集对 $x \sim x' \in X^*$，都有 $|q(x) - q(x')| \leq \Delta$。

命题1 (高斯机制 ρ-零集中差分隐私)。设 $q : X^* \to \mathbb{R}$ 是一个敏感度为Δ 的查询。考虑机制 $M : X^* \to \mathbb{R}$，它在输入 x时，从 $N(q(x)\ \Delta^2/(2\rho))$中释放一个样本。那么，$M$满足 ρ-零集中差分隐私。

较小的预算平均会给查询添加较大的噪声。因此，输出具有更强的隐私性。

命题2 (组合)。设 $M:X^* \to Y$和 $M':X^* \to Z$ 是两个随机化算法。假设 M 满足 ρ-零集中差分隐私，且 M' 满足 ρ'-零集中差分隐私，则算法 $M'' = (M, M\prime):X^* \to Y \times Z$ 满足$(\rho + \rho')$-零集中差分隐私。

命题3 (后处理)。设 $M : X^* \to Y$和 $f : Y \to Z$ 是随机化算法。若 M 满足 ρ-零集中差分隐私，则组合算法$M' = f \circ M:X^* \to Z$ 也满足该性质。

3. 方法

我们的目标是在分层随机抽样下，发布一个 ρ-零集中差分隐私总体比例置信区间p。为了像（2）那样构造置信区间，我们需要私下估计 p 以及估计量的方差。回顾一下，总体比例的非隐私估计量由样本比例给出

$$\hat{p} = \sum_{h=1}^{H} w_h \hat{p}_h.$$

我们假设层大小 N_h 均为公开且固定，因此 w_h也如此。要获得 p的隐私估计量，记为 $\tilde{p}$，我们可以在（非隐私）估计量 $\hat{}p$ 或估计量 $\hat{}p_h$的层级添加噪声。利用 $\tilde{p}$，我们进一步为 Var($\tilde{p}$)设计了一个隐私估计量。基于这一思路，我们在第3.1节针对公开样本量的情形，通过在层级或总体层级添加噪声，设计了两种算法。在第 3.2 节中，我们还提出了在样本量为隐私信息时于层级添加噪声的方法。全文中，符号$\hat{}\cdot$ 和 $\tilde{}\cdot$ 分别用于表示非隐私估计量和隐私估计量。

3.1. 使用公开样本量进行估计

当样本量 n_h 固定时，对 $\hat{}p$加噪有两种自然策略：向（1）层级统计量 $\hat{}p_h$ 加入高斯噪声，或向（2）总体统计量 $\hat{}p$加入高斯噪声。向 $\hat{}p_h$ 加入噪声的优点是能够同时为层级比例生成隐私估计量。

3.1.1. 在层级添加噪声

我们对每一层应用高斯机制，从而得到隐私估计量$\tilde{p}_h \overset{def}{=} \hat{p}_h + e_h$，其中 e_h 为高斯噪声。于是，总体比例的隐私估计量为

$$\tilde{p} \overset{def}{=} \sum_{h=1}^{H} w_h \tilde{p}_h.$$

因此，$\tilde{p}$ 的方差由两部分组成：通过$\hat{p}_h$估计p_h时固有的方差，以及添加噪声带来的额外变异：

$$\mathrm{Var}(\tilde{p}) = \sum_{h=1}^{H} w_h^2 \left(\mathrm{Var}(\hat{p}_h) + w_h^2 \,\mathrm{Var}(e_h)\right) \tag{3}$$

其中，$\mathrm{Var}(e_h)$ $h = 1, ..., H$是公开的，因为它们不依赖于数据。

要为$\hat{p}$获得隐私置信区间，我们需要对$\mathrm{Var}(\hat{p}_h)$进行隐私估计。注意，添加的噪声会使式 (1) 中$\mathrm{Var}(\hat{p}_h)$的非隐私估计量里的项$\hat{p}_h(1-\hat{p}_h)$产生偏倚。具体而言，$\mathbb{E}_e[\tilde{p}_h(1-\tilde{p}_h)] = \hat{p}_h(1-\hat{p}_h) - \mathrm{Var}(e_h)$，其中 $\mathbb{E}_e$ 表示对添加噪声的随机性取期望。于是，式 (3) 右侧$\mathrm{Var}(\hat{p}_h)$的隐私无偏估计量为

$$\widetilde{\mathrm{Var}}(\hat{p}_h) \overset{def}{=} \left(\frac{N_h - n_h}{N_h}\right) \frac{\tilde{p}_h(1-\tilde{p}_h) + \mathrm{Var}(e_h)}{n_h - 1}. \tag{4}$$

为估计 $\mathrm{Var}(\tilde{p})$，我们设定

$$\widetilde{\mathrm{Var}}(\tilde{p}) \overset{def}{=} \sum_{h=1}^{H} w_h^2 \left(\widetilde{\mathrm{Var}}(\hat{p}_h) + \mathrm{Var}(e_h)\right)$$

该方法在算法 1 中进行了表述，我们称之为 **StrNz-PubSz**（在层级添加噪声，并使用公开样本量）。关于隐私级别和渐近覆盖率的理论结果见定理 4.1 和4.2。

3.1.2. 在总体层级添加噪声

另一种策略是直接向 p 的非隐私估计量添加噪声，即 $\hat{p}$。$\hat{p}$ 的敏感度为

$$\Delta_p = \max_h \frac{w_h}{n_h}.$$

由于 w_h 和 n_h 是公开的，因此可以公开 Δ_p。我们设定 $\tilde{p} = \hat{p} + e$，其中 e 是标准差与 Δ_p成正比的高斯噪声。然后，$\tilde{p}$的方差变为

$$\mathrm{Var}(\tilde{p}) = \mathrm{Var}(\hat{p}) + \mathrm{Var}(e). \tag{5}$$

算法 1 在层级添加噪声，使用公开样本量，StrNz- PubSz

输入： $\hat{p}_h$, n_h, N_h, w_h, ρ, α. **输出：** ρ-零集中差分隐私（$1-\alpha$）总体比例置信区间。1：**对于** $h=1$ 到 H **执行**2：生成高斯噪声 $e_h \sim \mathcal{N}(0, \frac{1}{2\rho n_h^2})$，并令

$$\tilde{p}_h \leftarrow \hat{p}_h + e_h.$$

3：估计 $\mathrm{Var}(\tilde{p}_h)$，方法如下

$$\widetilde{V}_h \leftarrow \left(\frac{N_h - n_h}{N_h}\right) \frac{\tilde{p}_h(1-\tilde{p}_h) + \frac{1}{2\rho n_h^2}}{n_h - 1} + \frac{1}{2\rho n_h^2}.$$

4: **end for**

5: Estimate p by $\tilde{p} \leftarrow \sum_{h=1}^{H} w_h \tilde{p}_h$ and $\mathrm{Var}(\tilde{p})$ by $\widetilde{V} \leftarrow \sum_{h=1}^{H} w_h^2 \widetilde{V}_h$.

6：Return

$$\tilde{p} \pm z_{1-\alpha/2}\sqrt{\widetilde{V}},$$

其中 $z_{1-\alpha/2}$ 是标准正态分布的（$1-\alpha/2$）分位数。

回顾一下，

$$\widehat{\mathrm{Var}}(\hat{p}) = \sum_{h=1}^{H} w_h^2 \left(\frac{N_h - n_h}{N_h}\right) \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}$$

是 $\mathrm{Var}(\hat{p})$ 的无偏估计量。为了获得 $\mathrm{Var}(\tilde{p})$ 的隐私估计量，我们再次基于 $\mathrm{Var}(\hat{p})$ 的敏感度，对 $\mathrm{Var}(\hat{p})$ 应用高斯机制：

$$\Delta_V = \max_h \left(\frac{C_h}{n_h}\left(1 - \frac{1}{n_h}\right)\right),$$

其中 $C_h = w_h^2 \frac{N_h - n_h}{N_h} \frac{1}{n_h - 1}$。

由于我们两次应用高斯机制，因此应将隐私预算分为两部分：$\rho = \rho_1 + \rho_2$ 用于向 $\hat{p}$和 $\mathrm{Var}(\hat{p})$分别添加噪声。组合性质（命题2）确保总隐私级别为 ρ。所得算法 **PopNz-PubSz**在算法 2中给出。

备注1。 当存在多个抽样率相近的层时，算法 *1* 对于 p 给出的置信区间比算法 *2* 更宽，在隐私预算相同的情况下。然而，算法 *1* 还会为 $\hat{p}_h$ 生成隐私置信区间，这些区间可能对发布有所帮助。在第*4.2.1*节中，我们将对这两种算法进行定量比较。

备注 2。 比例始终介于 *0* 和 *1* 之间。可以将比例估计值（$\tilde{p}_h$ 算法 *1* 和 $\tilde{p}$ 算法 *2*中的）裁剪到区间 *[0,1]* ，而不损害隐私。当隐私预算非常小时，带噪声的比例估计值很可能落在 *[0,1]*之外。因此，裁剪会使置信区间更接近真实值，并提高覆盖率

算法 2 在总体层级添加噪声并使用公开样本量，PopNz-PubSz

输入： $\hat{p}, \hat{p}_h, n_h, N_h, w_h, \rho, \alpha$. **输出：** 总体比例的 ρ-零集中差分隐私（$1-\alpha$）置信区间。1：拆分预算 $\rho = \rho_1 + \rho_2$。2：生成噪声 $e \sim \mathcal{N}(0, \frac{\Delta_p^2}{2\rho_1})$，其中 $\Delta_p = \max_h \frac{w_h}{n_h}$，并令

$$\tilde{p} \leftarrow \hat{p} + e.$$

3：生成噪声 $e_V \sim \mathcal{N}(0, \frac{\Delta_V^2}{2\rho_2})$，其中 $\Delta_V = \max_h \left(\frac{C_h}{n_h}\left(1 - \frac{1}{n_h}\right)\right)$，且 $C_h = w_h^2 \frac{N_h - n_h}{N_h} \frac{1}{n_h - 1}$。令

$$\widetilde{V} \leftarrow \sum_{h=1}^{H} w_h^2 \left(\frac{N_h - n_h}{N_h}\right) \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1} + \frac{\Delta_p^2}{2\rho_1} + e_V.$$

4：返回

$$\tilde{p} \pm z_{1-\alpha/2}\sqrt{\widetilde{V}},$$

其中 $z_{1-\alpha/2}$ 是标准正态分布的 $(1-\alpha/2)$ 分位数。

将观察到该比率。预算适中或较大时，裁剪不会造成明显差异。

最后，还可以始终将输出的置信区间裁剪到 $[0,1]$，而不会造成隐私损失。

3.2. 使用私有样本量进行估计

当样本量属于公开信息时，保护比例的隐私实质上只需保护分子，即具有$y = 1$的个体数量（即计数）。在某些还需要估计子总体比例的情况下，使用公开样本量的算法1 和2 可能导致隐私泄露，因为这些计数会成为分母。例如，可以提出以下查询：（1）美国女性占总体的比例是多少；（2）美国女性中的失业人员占多少比例。女性人数是查询（1）中的分子，却是查询（2）中的分母。采用算法1 或2 可以保护查询（1）中的女性人数，却会在查询（2）中将其泄露。因此，子总体分析需要一种构造比例置信区间的方法，以同时保护计数和样本量的隐私。我们通过向样本量添加噪声来保护它们。因此，样本量并非固定值，所以我们需要采用具有邻接关系 $\sim_r$的无界差分隐私概念。

下面，我们通过在层级添加噪声，扩展算法1 以满足保护样本量隐私的需求。（如何扩展算法2并不明显。该算法在总体层级添加噪声，需要更复杂的机制；我们将在第6节简要讨论。）

首先，我们先考虑简单随机抽样的情形。其思路是对每一层的分子和分母分别添加相互独立的高斯噪声

。为简化记号，我们先考虑计数为 $c = \sum_{i=1}^{n} x_i$ 的单个层。我们知道

$$c \sim \text{Hypergeometric}(N, K, n),$$

其中， K 是具有感兴趣属性的个体总数。计数 c的均值为 $n\frac{K}{N} = np$，方差为 $n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1} = n^2\widehat{\text{Var}}(\hat{p})$。分别对 c 和 $\bar{n}$应用高斯机制，并使用隐私预算 ρ_1 和 ρ_2，可得到隐私计数 $\tilde{c}$ 和样本量 $\tilde{n}$:

$$\tilde{c} \mid c \sim \mathcal{N}(c, \frac{1}{2\rho_1})$$

and

$$\tilde{n} \sim \mathcal{N}(n, \frac{1}{2\rho_2}).$$

c 的无条件均值和方差为

$$\mathbb{E}(\tilde{c}) = \mathbb{E}[\mathbb{E}(\tilde{c} \mid c)] = \mathbb{E}(c) = np$$

and

$$\text{Var}(\tilde{c}) = \mathbb{E}\,\text{Var}(\tilde{c} \mid c) + \text{Var}\,\mathbb{E}(\tilde{c} \mid c) = \frac{1}{2\rho_1} + n^2\,\text{Var}(\hat{p}). \tag{6}$$

根据零集中差分隐私的组合性质，我们得到比例p的隐私估计量，记为 $\tilde{p}$，其隐私级别为 $\rho = \rho_1 + \rho_2$。由于 $\tilde{c}$ 和 $\tilde{n}$ 是相互独立的变量，原则上，

$$\mathbb{E}(\tilde{p}) = \mathbb{E}\left(\frac{\tilde{c}}{\tilde{n}}\right) = \mathbb{E}(\tilde{c})\mathbb{E}\left(\frac{1}{\tilde{n}}\right), \tag{7}$$

and

$$\text{Var}(\tilde{p}) = \mathbb{E}\left(\frac{\tilde{c}}{\tilde{n}}\right)^2 - \left(\mathbb{E}\left(\frac{\tilde{c}}{\tilde{n}}\right)\right)^2 = \mathbb{E}\tilde{c}^2\mathbb{E}\left(\frac{1}{\tilde{n}^2}\right) - (\mathbb{E}\tilde{c})^2\left(\mathbb{E}\frac{1}{\tilde{n}}\right)^2. \tag{8}$$

然而， $\frac{1}{\tilde{n}}$ 的矩不存在，因此 $\tilde{p}$的矩也不存在。一般而言，两个独立正态随机变量之比具有重尾分布，且不存在矩 [34, 18]。该分布的形状可能是单峰、双峰、对称或非对称的，主要取决于分母变量的变异系数 CV。当 CV 足够小时，正态分布近似可能效果良好。理论上已经证明，在以两个正态随机变量的均值之比为中心的区间内，正态分布可以与比率变量任意接近 [18]。实验给出了何时可以使用正态近似的指导原则。例如，一个简单的规则是，只要 CV 小于 0。1 [30]，这种近似就是合理的。其他实用规则见 [26, 34]。

我们利用正态近似，为比例构造一个 ρ-零集中差分隐私置信区间。我们提出以下估计策略

算法 3 使用私有样本量在层级添加噪声，

StrNz-PrivSz

输入： $N_h,\ w_h,\ n_h,\ c_h,\ \rho,\ \alpha$。

输出： 总体比例的 ρ-zCDP $(1-\alpha)$ 置信区间。

1: 划分预算 $\rho = \rho_1 + \rho_2$。

2: **对于** $h = 1$ 到 H **执行**

3: 生成 $e_h^{(1)} \sim \mathcal{N}(0,\ \frac{1}{2\rho_1})$ 和 $e_h^{(2)} \sim \mathcal{N}(0,\ \frac{1}{2\rho_2})$，并令

$$\begin{cases} \tilde{c}_h \leftarrow c_h + e_h^{(1)} \\ \tilde{n}_h \leftarrow \max\left(n_h + e_h^{(2)}, 2\right) \end{cases} \tag{9}$$

4: Let

$$\tilde{p}_h \leftarrow \frac{\tilde{c}_h}{\tilde{n}_h} \tag{10}$$

5: Let

$$\widetilde{V}_h \leftarrow \left(\frac{N_h - \tilde{n}_h}{N_h - 1}\right) \frac{\tilde{p}_h(1-\tilde{p}_h)}{\tilde{n}_h} + \frac{1}{2\rho_1 \tilde{n}_h^2} + \frac{\tilde{p}_h^2}{2\rho_2 \tilde{n}_h^2}. \tag{11}$$

6: </style)结束循环

7: 通过 $\tilde{p} \leftarrow$ 估计 p $\sum_{h=1}^{H} {}_{\sim} w_h \tilde{p}_h$ 并令 $V \leftarrow \sum_{h=1}^{H} w_h^2 \widetilde{V}_h$.

8: 返回

$$\tilde{p} \pm z_{1-\alpha/2} \sqrt{\widetilde{V}},$$

其中 $z_{1-\alpha/2}$ 是标准正态分布的 $(1-\alpha/2)$ 分位数。

算法3，**StrNz-PrivSz**。在该算法中，我们对 $\tilde{n}_h$ 在（9）中进行截断，以确保分母不会过小。否则，比率可能会变得任意大。这种后处理步骤保留了相同的隐私保证。在理论分析中，我们不对 $\tilde{n}_h$进行截断，而是考虑在事件 $\tilde{c}_h/\tilde{n}_h$给定的条件下的比率变量 $S_h = \{1 \le \tilde{n}_h \le 2n_h - 1\}$ （即 $\tilde{n}_h$均值周围的对称区域）。这更便于分析。算法中的 $\tilde{p}_h$与 $\tilde{c}_h/\tilde{n}_h \mid S_h$ 的渐近行为本质上相同，因为当 $\Pr(\tilde{n}_h \ge 2) \to 1$和$\Pr(S_h) \to 1$ 时，$n \to \infty$。我们将看到，根据对$\tilde{p}_h$ 的分析推导出的方差隐私估计量 $\tilde{c}_h/\tilde{n}_h \mid S_h$ 表现良好，而且该算法确实达到了所需的覆盖水平。

我们考虑两个独立正态变量的比值。根据独立性，尚不清楚的是正态分布倒数的性质。（需要说明的是，逆高斯分布与我们在此讨论的倒数分布是不同的分布。）在定理3.1中，我们给出了倒数正态分布的条件均值和方差的泰勒级数的一般形式。据我们所知，这是首个给出完整泰勒级数结果并明确指定余项的研究。我们在证明部分证明该定理。我们使用 $k = 2$ 推导 $\tilde{p}$ 算法3中11式所示的方差估计量。

定理 3.1（倒数正态分布的条件均值与方差）。对于随机变量 $X \sim N(\mu, \sigma^2)$，其中 $\mu > 1$ 且 $\sigma^2 > 0$，给定事件 $S = \{1 \le X \le 2\mu - 1\}$，对于任意整数 $k > 0$，其前两个矩

— $\frac{1}{X} \mid S$ 具有如下展开式：

$$\mathbb{E}\left(\frac{1}{X} \mid S\right) = \frac{1}{\mu}\sum_{j=0}^{k}\frac{(2j-1)!!\sigma^{2j}}{\mu^{2j}} + O\left(\frac{\sigma^{2k+2}}{\mu^{2k+2}}\right) \tag{12}$$

and

$$\mathbb{E}\left(\frac{1}{X^2} \mid S\right) = \frac{1}{\mu^2}\sum_{j=0}^{k}\frac{(2j+1)!!\sigma^{2j}}{\mu^{2j}} + O\left(\frac{\sigma^{2k+2}}{\mu^{2k+2}}\right). \tag{13}$$

4. 理论结果

本节给出隐私与渐近覆盖保证方面的理论结果。此外，还将从方差和宽度比的角度比较这三种算法。

4.1. 隐私与覆盖保证

我们的理论结果分为两部分。首先，所提出的算法在相应的邻接关系下满足所需的隐私级别，相关内容见定理 4.1。

定理 4.1 (隐私保证**)。** 算法 *1* 和 *2* 满足 ρ-零集中差分隐私，邻接关系为 $\sim_{ss}$；算法 *3* 满足 ρ-零集中差分隐私，邻接关系为 $\sim_{r}$。

证明在证明部分中给出。

另一方面，为了使这些置信区间具有实用价值，我们给出保证每种算法渐近覆盖率的定理。中心极限定理（中心极限定理）断言（大致而言），无论原始分布如何，样本均值渐近服从正态分布。因此，可以使用样本均值构造总体均值的置信区间。在我们考虑的有限总体抽样设计中，其他文献中也可以找到中心极限定理的变体。 [19, 25, 32] 我们在定理A.2 中重述了简单随机抽样下有限总体中心极限定理的一般形式，并在以下定理中给出渐近覆盖保证。

定理 4.2 (算法 1**)。** 对于大小为 N 的总体，令 p 表示总体中具有感兴趣属性的比例。在每层样本量为 n_h、该层大小为 N_h 的分层随机抽样中，$h = 1$、..、H，令

$\widetilde{V} = \sum_{h=1}^{H} w_h^2 \widetilde{V}_h$ *where*

$$\widetilde{V}_h = \left(\frac{N_h - n_h}{N_h}\right)\frac{\tilde{p}_h(1-\tilde{p}_h) + \frac{1}{2\rho n_h^2}}{n_h - 1} + \frac{1}{2\rho n_h^2}. \tag{14}$$

对于 $\rho > 0$，按算法 1所述进行。如果 $\rho = \omega(1/n_h)$ 对所有 h 均成立，那么当 $N_h - n_h$和 n_h 对每一层都趋于无穷大时，

(i) $\widetilde{V} \xrightarrow{p} \mathrm{Var}(\tilde{p})$，更具体地说，对于所有 h，

$$\widetilde{V}_h - \mathrm{Var}(\tilde{p}_h) = \widehat{\mathrm{Var}}(\hat{p}_h) - \mathrm{Var}(\hat{p}_h) + O_P\left(\frac{1}{\sqrt{\rho} n_h^2}\right) = O_P\left(\frac{1}{n_h^{3/2}}\right),$$

$\widehat{\ }$其中$\mathrm{Var}(\hat{\ }p_h)$ 是$\mathrm{Var}(\hat{\ }p_h)$的非隐私估计量；

(ii) 对于 $0 < \alpha < 1$,

$$\Pr\left(p \in \left(\tilde{p} - z_{1-\alpha/2}\sqrt{\widetilde{V}}, \tilde{p} + z_{1-\alpha/2}\sqrt{\widetilde{V}}\right)\right) \to 1 - \alpha. \tag{15}$$

定理 4.3 (算法 2)。 对于大小为 N 的总体，令 p 表示总体中具有感兴趣属性的比例。在每层样本量为 n_h、层大小为 N_h， $h = 1, .., H$ 时，令

$$\widetilde{V} = \sum_{h=1}^{H} w_h^2 \left(\frac{N_h - n_h}{N_h}\right) \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h - 1} + \frac{\Delta_p^2}{2\rho_1} + e_V \tag{16}$$

其中 $e_V \sim \mathcal{N}(0, \frac{\Delta_V^2}{2\rho_2})$对于 ρ_1， $\rho_2 > 0$如算法 2所述。如果 $\rho_1 = \omega(1/n_h)$ 且 $\rho_2 = \omega(1/n_h)$ 对于所有 h，那么随着 $N_h - n_h$ 和 n_h 在每一层都趋于无穷大，

(i) $\widetilde{V} \xrightarrow{p} \mathrm{Var}(\tilde{p})$，更具体地说，

$$\widetilde{V} - \mathrm{Var}(\tilde{p}) = \widehat{\mathrm{Var}}(\hat{p}) - \mathrm{Var}(\hat{p}) + O_P\left(\frac{1}{\max_h \sqrt{\rho_2} n_h^2}\right) = O_P\left(\frac{1}{\max_h n_h^{3/2}}\right);$$

$\widehat{\ }$其中，$\mathrm{Var}(\hat{\ }p)$ 是$\mathrm{Var}(\hat{\ }p)$的非隐私估计量；

(ii) 对于 $0 < \alpha < 1$,

$$\Pr\left(p \in \left(\tilde{p} - z_{1-\alpha/2}\sqrt{\widetilde{V}}, \tilde{p} + z_{1-\alpha/2}\sqrt{\widetilde{V}}\right)\right) \to 1 - \alpha. \tag{17}$$

上述定理的证明使用有限总体中心极限定理，并见证明部分。

对于算法 3， $\tilde{p}$ 的渐近性质除了依赖中心极限定理外，还建立在比率变量的正态近似之上。我们在此重温 [18] 于定理 A.4中的正态近似结果。基于该近似，我们证明了在简单随机抽样情况下， $\tilde{p}$ 具有一致性。

定理 4.4。 在简单随机抽样下，设 c 为具有感兴趣属性的个体数， n 为样本量。真实的总体比例记为 p。设 $\tilde{p}=\tilde{c}/\tilde{n}$，其中 $\tilde{c}\sim N(c, \frac{1}{2\rho_1})$ 且 $\tilde{n}\sim N(n, \frac{1}{2\rho_2})$，对 $\rho_1,\rho_2>0$均成立。在满足条件 $\rho_2=\omega(1/n)$、 $\rho_1=\omega(1/n)$时， $\tilde{p}$ 是 p 的相合估计量。</output>

基于上述一致性，我们建立了以下渐近性质：

定理 4.5 (算法 3)。 对于大小为 N 的总体，设 p 为总体中具有感兴趣属性的比例。在分层随机抽样中，若层大小为 n_h，各层的样本量为 N_h, $h=1,..,H$，令

$$\widetilde{V}=\sum_{h=1}^{H} w_h^2 \widetilde{V}_h \ where$$

$$\widetilde{V}_h=\left(\frac{N_h-\tilde{n}_h}{N_h-1}\right)\frac{\tilde{p}_h(1-\tilde{p}_h)}{\tilde{n}_h}+\frac{1}{2\rho_1\tilde{n}_h^2}+\frac{\tilde{p}_h^2}{2\rho_2\tilde{n}_h^2} \tag{18}$$

对于 ρ_1， $\rho_2>0$ 如算法 3所述。若 $\rho_1=\omega(1/n_h)$ 且 $\rho_2=\omega(1/n_h)$对所有 h均成立，则当 N_h-n_h 和 n_h 对每一层都趋于无穷大时，

(i) $\widetilde{V}\xrightarrow{p}\mathrm{Var}(\tilde{p}\mid S)$ 其中 S 是一个事件，且 $\Pr(S)\to 1$，更具体地说，对所有 h，

$$\tilde{V}_h-\mathrm{Var}(\tilde{p}_h\mid S)=\widehat{\mathrm{Var}}(\hat{p}_h)-\mathrm{Var}(\hat{p}_h)+O_p\left(\frac{1}{\rho_1 n_h^2}+\frac{1}{\rho_2 n_h^2}\right)=o_p\left(\frac{1}{n_h}\right),$$

其中 S_h 是一个事件，且 $\Pr(S_h)\to 1$;

(ii) 对于 $0<\alpha<1$,

$$\Pr\left(p\in\left(\tilde{p}-z_{1-\alpha/2}\sqrt{\widetilde{V}},\tilde{p}+z_{1-\alpha/2}\sqrt{\widetilde{V}}\right)\right)\to 1-\alpha. \tag{19}$$

事件 S_h 在第3.2 节中讨论， S 可以设为 $\cap_h S_h$。由于 $\tilde{p}$ 和 $\tilde{p}_h$ 的方差都不存在，我们转而考虑高概率事件下的条件方差。为证明定理4.5，我们先从单个层入手。我们使用正态分布（记为 p_h^*）来近似比例估计量 $\tilde{p}_h$的分布，在一个区间内，两者分布之间的距离趋于零。接着，对于多个层，我们证明正态变量的线性组合（记为 p^*）能够准确近似 $\tilde{p}$。最后但同样重要的是，由于定理4.4中所述的一致性，~含噪估计量 V 是 p^*方差的相合估计量。随后，我们可以使用~和 $\tilde{p}$ 构造沃尔德 V置信区间。详情见证明部分。

请注意，除方差估计的一致性外，上述结果还给出了收敛速率。与非隐私设置下的方差估计相比

在非隐私设置下，鉴于关于 ρ、 ρ_1 和 ρ_2 的条件，额外的偏倚仅是细微差异。事实上，我们设定这些条件，是为了确保估计方差时引入的噪声不会占主导地位。原则上，这些速率可用于实践中相应地调整置信区间的长度，不过本文不探讨这一方向。

4.2 方差比较

第4.1节介绍的定理保证，在适当条件下，可以渐近地达到所需的覆盖率。因此，为了比较所提出的不同置信区间的性能，我们比较它们的宽度，而宽度由其方差估计决定。本节将分析我们的方差估计，并将由此得到的宽度与非隐私置信区间的宽度进行比较。

4.2.1. 外在方差

为探究添加噪声会造成多少额外的不确定性，我们将隐私估计量的方差分解为两部分：（1）来自基于抽样数据进行估计的内在分量，即 $\mathrm{Var}(\hat{p})$；（2）由添加的噪声引入的外在分量，记作

$$V_{\mathrm{ex}} \overset{def}{=} \mathrm{Var}(\tilde{p}) - \mathrm{Var}(\hat{p}).$$

表1 给出了 $\tilde{p}$ 在三种算法下的（近似）方差，其中$w_h = \frac{N_h}{N}$ 为层权重。这些方差是在定理4.2、4.3和4.5的证明中推导得出的。附加方差项 V_{ex}可以改写为关于 $u_h \overset{def}{=} \frac{N_h}{n_h}$ 而非 w_h的形式，如表格第二行所示。事实上， u_h 在调查抽样文献中称为抽样权重。样本权重定义为样本中的每位受访者在总体中所代表的个体数量。它是抽样率的倒数 $\frac{n_h}{N_h}$ ，在调查数据的统计推断中发挥着重要作用[35, 12]。了解抽样权重与含噪估计量方差之间的关系，有助于实践者进行调查设计并选择算法。

表*1* $\tilde{p}$的（近似）方差。

算法	StrNz-PubSz	PopNz-PubSz	StrNz-PrivSz（近似）
$\mathrm{Var}(\tilde{p})$	$\mathrm{Var}(\hat{p}) + \frac{1}{2\rho}\sum_{h=1}^{H}\frac{w_h^2}{n_h^2}$	$\mathrm{Var}(\hat{p}) + \frac{1}{2\rho_1}\max_h \frac{w_h^2}{n_h^2}$	$\mathrm{Var}(\hat{p}) + \frac{1}{2\rho_1}\sum_{h=1}^{H}\frac{w_h^2}{n_h^2} + \frac{1}{2\rho_2}\sum_{h=1}^{H}\frac{w_h^2 p_h^2}{n_h^2}$
V_{ex}	$\frac{1}{2N^2}\sum_{h=1}^{H}\frac{u_h^2}{\rho}$	$\frac{1}{2N^2}$最大值$_h$ $\frac{u_h^2}{\rho_1}$	$\frac{1}{2N^2}\sum_{h=1}^{H}u_h^2(\frac{1}{\rho_1}+\frac{p_h^2}{\rho_2})$

在总体规模 N 固定且所选隐私级别为 ρ 的情况下，添加的噪声所引入的额外方差 V_{ex} 主要由 u_h 决定。在 PopNz-PubSz 中，我们在总体层级添加噪声， V_{ex} 仅由

所有层中最大的样本权重决定。如果在每一层中注入噪声，那么所有层的抽样权重会共同影响 V_{ex}。具体而言，对于 StrNz-PrivSz，V_{ex} 还会受到 p_h 的影响。对于这三种算法，抽样权重越小，额外方差越低。

为了进行比较，我们考察 PopNz-PubSz 和 StrNz-PrivSz 在预算分配 $\rho_1 = \rho_2 = \rho/2$下的 V_{ex} 比值。StrNz-PubSz 相对于 PopNz-PubSz 的 V_{ex} 比值为

$$\frac{\sum_{h=1}^{H} u_h^2}{2\max_h u_h^2}. \tag{20}$$

粗略地说，当层数较多时，在总体层级添加噪声会产生较小的方差。为了比较 StrNz-PrivSz 和 StrNz-PubSz，V_{ex} 的比值为

$$\frac{2\sum_{h=1}^{H} u_h^2(1+p_h^2)}{\sum_{h=1}^{H} u_h^2}, \tag{21}$$

由于在 StrNz-PrivSz 中保护样本量需要付出代价，该比值始终大于 2，且至多为 4。

4.2.2. 与非隐私置信区间比较：单层情形

为了从理论上评估宽度，我们还将置信区间宽度比与非隐私置信区间进行比较。由于在分层设置中，参数 N_h、n_h、p_h、ρ_h 会共同发挥作用，因此分析仅含一层的特殊情形更为实际。

令理论宽度比（TWR）为

$$\text{TWR} = \sqrt{\frac{\text{Var}(\tilde{p})}{\text{Var}(\hat{p})}}.$$

在实际应用中，实际宽度比（WR）定义为 $\sqrt{\widetilde{V}/\operatorname{Var}(\hat{p})}$，将~~非常接近~~ TWR，因为 V 是 $\text{Var}(\tilde{p})$的相合估计量。表 2列出了一些相关量。注意，$\frac{N-1}{N-n}$ 始终小于 1，但当总体规模远大于样本量时~~趋近~~于 1。

通过去掉因子 $\frac{N-1}{N-n}$ 并对 p取最小值，可以得到 TWR 的下界。~~可以看~~出，宽度比主要取决于 p 以及 n 与 ρ之间的相对大小。如果 p 趋于极端（趋近于 0 或 1），TWR 会显著增大；当 p 约为 0.5 时，TWR 接近下界。此外，添加的噪声会引入一个涉及 ρ的项。例如，在$\rho = 1/n$这一条件下，三种算法得到的区间长度至少分别为 $\sqrt{3} \approx 1.73$、$\sqrt{5} \approx 2.24$和 $\sqrt{3+2\sqrt{2}} \approx 2.41$ 倍。显然，在只有一层时，StrNz-PubSz 得到的置信区间比 PopNz-PubSz 更窄，因为式（20）中的 V_{ex} 之比为 1/2。然而，一旦层数足够多，使得式（20）大于 1，PopNz-PubSz 的表现就会优于 StrNz-PubSz。

表2 理论宽度比和下界。预算分配 $\rho_1 = \rho_2 = \rho/2$ 用于*PopNz-PubSz*和*StrNz-PrivSz*。

算法	StrNz-PubSz	PopNz-PubSz	StrNz-PrivSz
$\tilde{p}$	$\mathcal{N}(0, \frac{\hat{p}+\frac{1}{2\rho n^2}}{})$	$\mathcal{N}(0, \frac{\hat{p}+\frac{1}{\rho n^2}}{})$	$(c+\mathcal{N}(0, \frac{1}{\rho}))/(n+\mathcal{N}(0, \frac{1}{\rho}))$
$\mathrm{Var}(\tilde{p})$	$\frac{\mathrm{Var}(\hat{p})+\frac{1}{2\rho n^2}}{}$	$\frac{\mathrm{Var}(\hat{p})+\frac{1}{\rho n^2}}{}$	$\frac{\mathrm{Var}(\hat{p})+\frac{1+p}{\rho n^2}}{}$
TWR	$1+\frac{N-1}{N-n}\frac{2p}{p(1-p)n\rho}$	$1+\frac{N-1}{N-n}\frac{p}{p(1-p)n\rho}$	$1+\frac{N-1}{N-n}\frac{1+p^2}{p(1-p)n\rho}$
TWR 的下界	$\sqrt{1+\frac{2}{n\rho}}$	$\sqrt{1+\frac{4}{n\rho}}$	$\sqrt{1+\frac{2(1+\sqrt{2})n\rho}{}}$

5. 数值结果

本节通过模拟研究和应用来评估并展示所提出算法的数值性能。PopNz-PubSz 和 StrNz-PrivSz 使用预算 $\rho_1 = \rho_2 = \rho/2$ 。我们按照备注 2中的说明，将比例 $\tilde{p}_h$ 截断到 $[0, 1]$ 。

5.1. 模拟

我们设计了一组实验，以（1）检验含噪估计量的正态性，并（2）通过改变层的数量 H、真实总体比例 p和隐私级别ρ来评估所提出的置信区间的性能。为了生成数据，我们需要指定层大小 N_h 和抽样率 r_h。这些参数的设置见表3。我们为每一层生成一个比例，以形成层间异质性。随后计算真实总体比例，并在每次实验中报告。

表3 参数设置。由此得到的样本量介于*60*和*160*之间。

固定参数	值 / 分布	变化参数	值 / 分布
α	0.1	H	1 或 20
N_h	均匀分布(1500, 2000)	p_h	0.5，均匀分布(0.4, 0.6) 或 均匀分布(0.05, 0.15)
r_h	均匀分布(0.04, 0.08)	ρ	$1/\max(n_h)$ 或在图的坐标轴中指定

5.1.1. 正态性检验

我们首先检查三个算法中的 $\tilde{p}$ 分布是否与具有相应均值和方差的理论正态分布足够接近。图1 展示了理论分布$\tilde{p}$ 与其样本分布的 Q-Q 图：

- 非隐私：$\mathcal{N}(p,\ \mathrm{Var}(\hat{p}))$；
- StrNz-PubSz：$\mathcal{N}(p,\ \mathrm{Var}(\tilde{p}))$ 在$\mathrm{Var}(\tilde{p})$ 中（3）；
- PopNz-PubSz：$\mathcal{N}(p,\ \mathrm{Var}(\tilde{p}))$ 作为 $\mathrm{Var}(\tilde{p})$ 见式（5）；

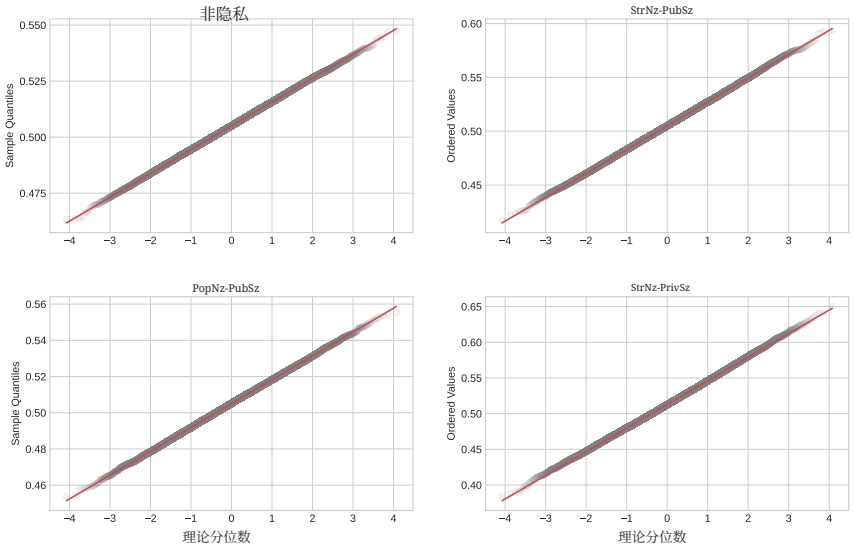


图 *1*：*Q-Q* 图：$\tilde{p}$ 具有 *20* 个层且 $p = 0.505$（由 $p_h \sim$ 均匀分布$(0.4,\ 0.6)$产生）的理论分布与样本分布，每种情况均基于 *10,000* 次重复。

- StrNz-PrivSz: $\mathcal{N}\left(p + \sum_{h=1}^{H} \frac{w_h p_h}{2\rho_2 n_h^2}, \sum_{h=1}^{H} w_h^2 V_h\right)$ with V_h specified in (51).

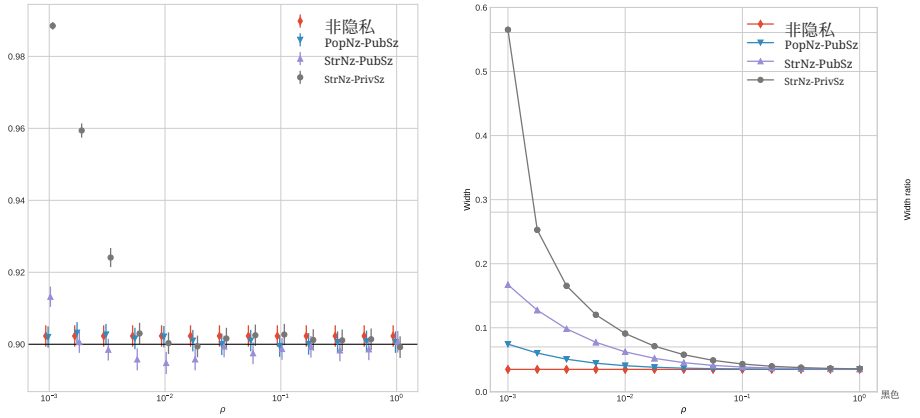
注意，$\tilde{p}$ 算法 StrNz-PubSz 和 PopNz-PubSz 中的估计量是 ω $\rho_2 =$ 的无偏估计，而 $\tilde{p}$ StrNz-PrivSz 中的估计量则不是。然而，在满足条件 $(1/n_h)$ 的定理 3中的情况下，偏差项 $\sum_{h=1}^{H} \frac{w_h p_h}{2\rho_2 n_h^2}$ 可以忽略，因此我们不会在算法 3中进行偏差校正。我们观察到理论分布与实验分布高度吻合，这表明这三种算法中的私有估计量确实非常接近正态分布。

5.1.2. 变化的关键参数

正态性检验的结果令人确信后，我们针对广泛的隐私预算、不同数量的层以及真实总体比例进行实验。

我们考察了 ρ 对三种隐私估计量性能的影响。模拟进行了 10,000 次重复，因此经验覆盖率落在$90\% \pm 0.006$（偏离两个标准差）范围内被认为是合适的。在图2a中，除 StrNz-PrivSz 在 ρ 小于约 0.005 时给出不必要的较高覆盖率外，经验覆盖率都较为合理。这是因为该方法的预算非常小，经过裁剪后，其覆盖真实值的频率高于所需水平。在这种情况下，置信区间过宽，实用性较低，如图2b所示。对于这三种方法，随着 ρ 变小，宽度都会增大。然而，宽度增长的速率

有所不同：在我们模拟的多层情形中，PopNz-PubSz 的宽度增长最慢，StrNz-PrivSz 增长最快，而 StrNz-PubSz 居中。因此，应综合考虑方法、宽度和覆盖率来选择最优隐私级别。例如，如果我们希望置信水平为 90%，且宽度小于 0.1，则可以将 ρ 选得尽可能小：对于 PopNz-PubSz，取 (1) 0.001；对于 StrNz-PubSz，取 (2) 0.003；对于 StrNz-PrivSz，取 (3) 0.01。



(A) 经验覆盖率　　(B) 宽度和比率

图*2*：设置：*20* 个层，且 $p = 0.505$ *(*$p_h \sim$ 均匀分布*(0.4, 0.6))*，重复 *10,000* 次。图 *(a)* 展示经验覆盖率，黑色实线表示 *90%* 的名义置信水平。图中显示了覆盖率的一个标准差误差条。*(b)* 展示平均宽度和宽度比，并以非隐私情形作为基准。宽度的误差条在图中不可见，因此未予显示。

此外，表 4展示了三个实验的数值结果，这些实验采用了不同的层数与总体真实比例组合。中间面板中的模拟与图 2 所示实验采用相同设置，但隐私级别固定为： $1/\max(n_h)$。这与多层情形下的 $\rho = 1/n$ （针对简单随机抽样）类似。在差分隐私文献中，通常会考察简单随机样本的 $\rho = 1/n$ ，以了解随着样本量增加， ρ 可以变得多小。请注意，较小的 ρ 意味着更高的隐私级别。

如上文所述，将 $\tilde{p}_h$ （或 $\tilde{p}$）裁剪到 $[0,1]$ 在某些情况下会取得更好的结果。这些结论与第4节中的分析一致。在所有模拟中，三个隐私方法的经验覆盖率均达到90%的名义置信水平，这由定理 4.2、4.3和4.5保证。当底部面板中的 StrNz-PrivSz 给出91.9%的置信水平时，这是由于裁剪造成的。（当层比例接近极端值时，裁剪现象会更加明显。）

平均宽度和宽度比（WR）会有所变化。只有一个层时，WR接近第4.2.2节给出的理论宽度比（TWR）下界，这表明下界几乎是紧的。只有一个层时，StrNz-PubSz的置信区间比PopNz-PubSz更窄。但是，当

表*4*基于*10,000*次重复次数，在 $\rho = 1/n$ （或 $\rho = 1/\max(n_h)$）情形下的模拟结果。层大小和抽样率按照表 *3*中的描述抽取。对于多层情形，所得的 n_h 样本量范围为*72*至*152*，且 ρ 被设为 $1/152 \approx 6.58 \times 10^{-3}$。对于单层情形，我们将样本量设为*152*，从而具有相同程度的隐私。

	非隐私	StrNz-PubSz	PopNz-PubSz	StrNz-PrivSz
	1 层， $p = 0.5$			
覆盖率	0.893	0.901	0.894	0.901
覆盖率标准差	3.09×10^{-3}	2.99×10^{-3}	3.08×10^{-3}	2.99×10^{-3}
宽度	0.127	0.228	0.295	0.327
宽度标准差	5.47×10^{-4}	9.85×10^{-4}	8.89×10^{-3}	3.15×10^{-2}
CI	(0.436, 0.564)	(0.386, 0.614)	(0.352, 0.648)	(0.34, 0.667)
WR	1	1.786	2.318	2.567
	20 层， $p = 0.505$ ($p_h \sim$ 均匀分布(0.4,0.6))			
覆盖率	0.902	0.895	0.902	0.902
覆盖率标准差	2.97×10^{-3}	3.07×10^{-3}	2.97×10^{-3}	2.97×10^{-3}
宽度	0.035	0.073	0.043	0.111
宽度标准差	1.08×10^{-4}	1.58×10^{-4}	5.87×10^{-4}	5.22×10^{-3}
CI	(0.488, 0.523)	(0.469, 0.542)	(0.483, 0.527)	(0.457, 0.568)
WR	1	2.074	1.239	3.168
	20 个层， $p = 0.103$ ($p_h \sim$ 均匀分布(0.05,0.15))			
覆盖率	0.902	0.919	0.904	0.899
覆盖率标准差	2.97×10^{-3}	2.73×10^{-3}	2.95×10^{-3}	3.01×10^{-3}
宽度	0.021	0.067	0.033	0.096
宽度标准差	6.17×10^{-4}	5.21×10^{-4}	8.71×10^{-4}	3.94×10^{-3}
CI	(0.092, 0.113)	(0.073, 0.143)	(0.086, 0.119)	(0.072, 0.168)
WR	1	3.189	1.571	4.563

层数增加时，PopNz-PubSz在宽度比方面优于StrNz-PubSz。层数增加意味着将总隐私预算分成更小的份额，从而整体上引入更多噪声。要达到相同的置信水平，置信区间就需要更宽。然而，对于StrNz-PrivSz，由于它还需要额外付出代价来同时保护样本量，因此它始终会产生最宽的置信区间。另一方面，在层数相同（此处为20）时，我们看到，更极端的 p_h 会比处于中间范围的 p_h 产生更大的宽度比。这是因为，因子 $p_h(1-p_h)$ 会发挥作用，就像 $p(1-p)$ 在表2 的单层情形理论宽度比中发挥作用一样。

我们还提供宽度的样本标准差（宽度标准差）。总体而言，非隐私方法的标准差小于隐私方法。在某些情况下，裁剪有助于降低隐私算法的宽度标准差。在隐私级别相同的情况下，与 StrNz-PubSz 相比，PopNz-PubSz 的宽度波动更大。这是因为我们使用一半的隐私预算，并直接向方差估计添加噪声。正如预期，StrNz-PrivSz 的宽度标准差最大，因为其宽度本身最大，而且比率变量经过设计具有重尾特征。

尽管如此，与宽度相比，所有方法的宽度标准差都很小，不会影响置信区间的有效性。

5.2. 应用

在本节中，我们将所提出的方法应用于来自 IPUMS USA 的 1940年人口普查完整普查数据 [36] ，并评估三种不同的差分隐私置信区间的性能。为对该数据集进行分层随机抽样，使用州级地理变量“STATEICP”（49 个类别，包括当时的 48 个州和华盛顿特区）进行分层。在具有 $H = 49$ 个层的分层随机抽样下，我们在第一个应用中估计全国失业率。在第二个应用中，我们关注黑人和白人男性之间收入水平的差异。

5.2.1. 失业率的置信区间

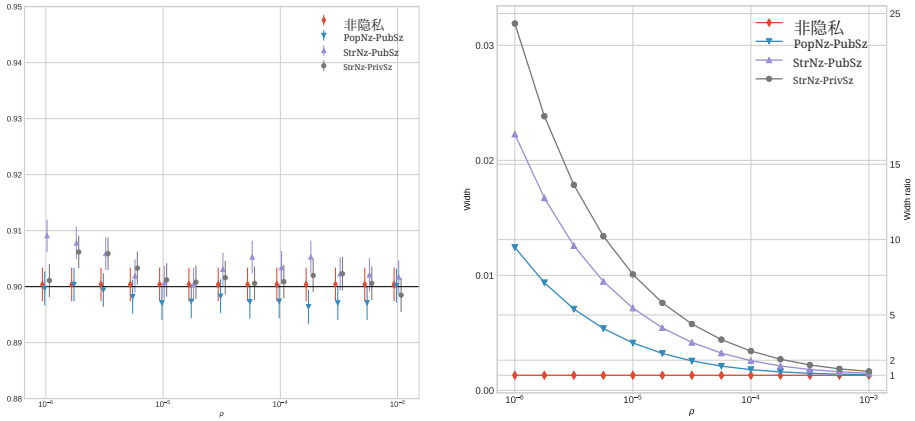
作为国民经济状况的重要指标，失业率是失业工人在由就业人员和失业人员组成的劳动力中所占的比例。因此，我们将所有就业人员或失业人员视为总体。在 1940年人口普查数据集中，二元特征“EMPSTAT”表示就业状况。完整普查被视为真实情况，而真实总体比例为 $p = 9.346\%$。为实施分层随机抽样，我们为全部 49 个层选择样本量或抽样率。为体现现代相关性，我们以当前美国社区调查（ACS）中采用的典型设计为参考进行模拟，选择 ACS 使用的典型抽样率范围，即 $[0.5\%、15\%]$。详情请参见表 5 。

表 5 抽样率

层大小	抽样率
$n_h \leq 5 \times 10^4$	15%
$5 \times 10^4 < n_h \leq 10^5$	10%
$10^5 < n_h \leq 5 \times 10^5$	5%
$5 \times 10^5 < n_h \leq 10^6$	2%
$10^6 < n_h \leq 5 \times 10^6$	1%
$n_h > 5 \times 10^6$	0.5%

为应用和评估所提出的算法，我们针对广泛的小额隐私预算进行实验：$\rho \in [10^{-6}, 10^{-3}]$。每种方法重复进行 10,000 次，并计算经验覆盖率、平均置信区间宽度和平均置信区间宽度比（WR）。如图3a所示，经验覆盖率始终接近名义水平；对于

整个隐私级别范围。在图3b中，以非隐私置信区间作为基准时，置信区间宽度和置信区间宽度比呈现相同的形状。即使 StrNz-PrivSz 给出的置信区间宽度是非隐私置信区间宽度的 8 倍，由于样本量较大，置信区间宽度也仅为 0.01。在选择最优隐私级别时，应同时考虑置信区间宽度和宽度比。



(A) 经验覆盖率 (B) 宽度和比率

图*3*：失业率差分隐私置信区间的经验覆盖率（含误差条）、平均宽度和宽度比。

5.2.2. 收入水平差异的置信区间

在第二个应用中，我们想考察白人男性（总体1）与黑人男性（总体2）的收入水平是否存在差异。请注意，我们只考虑1940年人口普查中收入数值有效的人。由于贫困线直到20世纪60年代才制定，因此1940年的数据中没有贫困线可用，我们改用全国平均收入作为阈值。我们感兴趣的是考察收入水平超过这一阈值的人所占子总体比例之间的差异。

地理特征“STATEICP”用于分层，形成 49 个层，其层大小范围为 $(41838, 4621442)$（白人男性总体）和 $(50, 309214)$（黑人男性总体）。抽样率根据层大小自适应选择。对于白人男性总体，抽样率范围为 $[0.5\%, 15\%]$, ；而由于黑人男性总体的层大小较小，其抽样率范围为 $[0.5\%, 30\%]$ 。此外，为了基于渐近结果进行可靠的近似，我们规定，如果抽样率得出的样本量小于 50，则将样本量调整为 50。详情见表6 。

表6 两个总体的抽样率。白人男性总体的层大小 $n_h \in (4.1 \times 10^4, 4.7 \times 10^6)$，黑人男性总体的层大小 $n_h \in (50,\ 3.1 \times 10^5)$。*如果上述抽样率得出的样本量小于50，则样本量将调整为50。

白人男性的层大小 N_h	抽样率
$N_h \leq 5 \times 10^4$	15%
$5 \times 10^4 < N_h \leq 10^5$	10%
$10^5 < N_h \leq 5 \times 10^5$	5%
$5 \times 10^5 < N_h \leq 10^6$	2%
$10^6 < N_h \leq 4 \times 10^6$	1%
$N_h > 4 \times 10^6$	0.5%

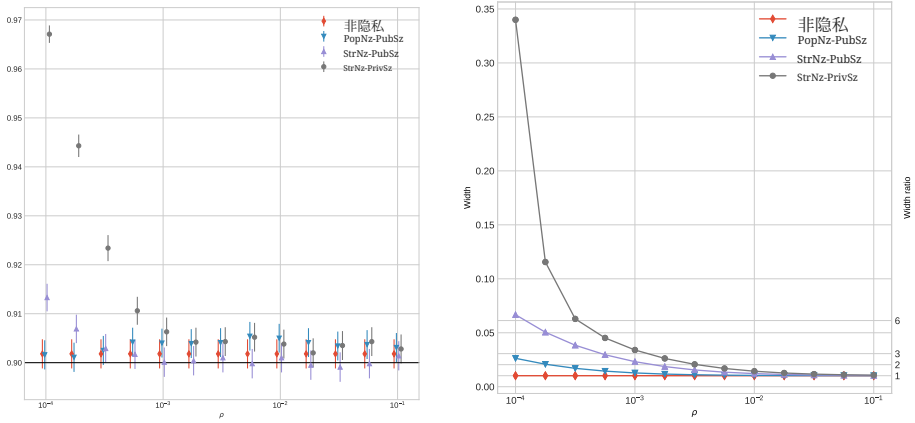
黑人男性的层大小 N_h	抽样率*
$N_h \leq 500$	30%
$500 < N_h \leq 5 \times 10^3$	15%
$5 \times 10^3 < N_h \leq 10^4$	5%
$10^4 < N_h \leq 2 \times 10^4$	2%
$2 \times 10^4 < N_h \leq 3 \times 10^4$	1%
$n_h > 3 \times 10^4$	0.5%

令 p_1 和 p_2 分别表示收入高于全国平均收入水平 $442.12 的符合条件个体所占的比例。这两个比例的真实值分别为 $p_1 = 49.0223\%$ 和 $p_2 = 29.5152\%$。令 $p_{\text{diff}} = p_1 - p_2$，则这两个比例的真实差值为 $p_{\text{diff}} = 19.5071\%$。根据两个独立正态分布的可加性，我们自然采用以下差分隐私置信区间：

$$\tilde{p}_{\text{diff}} + z_{1-\alpha/2}\sqrt{\widetilde{\mathrm{V}}(\tilde{p}_{\text{diff}})}, \tag{22}$$

其中 $\widetilde{V}(\cdot)$ 表示方差的隐私估计量，$\tilde{p}_{\text{diff}}$ 定义为 $\tilde{p}_1 - \tilde{p}_2$，且

$$\widetilde{\mathrm{V}}(p_{\text{diff}}) = \widetilde{\mathrm{V}}(p_1) + \widetilde{\mathrm{V}}(p_2).$$



(A) 经验覆盖率 (B) 宽度和比率

图4：在具有有效收入值的黑人和白人男性中，高于全国收入水平的比例之差的差分隐私置信区间的经验覆盖率（含误差条）、平均宽度和宽度比。

图 4中也呈现出与第一个应用类似的模式。所有置信区间的经验覆盖率均接近或高于名义置信水平，这与

第5.1.2节的模拟研究中一致。覆盖率较高的现象源于较小的 ρ 和有效的裁剪。当层大小的范围较大时（本应用中为（50，309214））， 也就是说，当各层的大小差异很大时，应选择较大的隐私预算 ρ。选择较小的ρ 会损害小层的估计效果。我们建议根据对宽度和/或宽度比的不确定性容忍度，选择尽可能小的 ρ。例如，如果准确性要求是宽度小于 0.05 或宽度比小于 5，那么图 ρ 中的实验结果4b里，0.0001 是 PopNz-PubSz 的最佳选择，0.0018 是 StrNz-PubSz 的最佳选择，0.0056 是 StrNz-PrivSz 的最佳选择。

6. 讨论

我们设计了三种算法，用于在分层随机抽样下构造总体比例的置信区间，并提供零集中差分隐私保证。我们同时考虑样本量公开和样本量属于隐私信息这两种情形。我们建立了包括隐私保证和渐近性质在内的理论结果。正如定理所述，在隐私预算与样本量之间满足适当关系的条件下，所得置信区间渐近地达到预期覆盖率；当样本量趋于无穷时，其宽度趋近于非隐私置信区间的宽度。

在模拟研究和两个应用中，我们针对各种参数设置试验了广泛范围的隐私预算。就经验覆盖率而言，这三种算法始终表现良好。即使在 $\rho = 1/\max_h n_h$ 的严格情形下，宽度和宽度比也处于合理范围内。实践中通常选择 0.001 到 10 之间的常数作为隐私预算。根据我们的实验，即使选择该范围内最小的预算 0.001，当最小层的样本量仅为 50 时（如第二个应用所示），这三种算法仍能取得相当好的结果。

第4.2节对三种算法的比较分析为实践者提供了切实可行的指导。当唯一目标是发布总体比例且层数足够多时（即关于样本权重的式(20)大于1），PopNz-PubSz是更好的选择。然而，如果还应发布层比例，或者层数很少，则StrNz-PubSz更可取。另一方面，当必须同时保护总体比例和样本量时，StrNz-PrivSz是本文提出的唯一算法。与公开样本量的情况相比，StrNz-PrivSz需要更大的预算才能满足相同的宽度要求，因为保护样本量会带来额外成本。

未来研究中有几个值得探讨的开放问题。本文讨论的是经典情形：层数固定，且样本量趋于无穷。原则上，在样本量有限的其他情形下，渐近正态性也成立。例如，已有研究表明

在非隐私情形下，当总样本量 $N \to \infty$ 由许多小样本或少数几个大样本组成，或由两者以某种方式组合而成时，在特定（复杂）条件下中心极限定理成立 [2]。在差分隐私的约束下，我们证明了差分隐私置信区间的隐私与准确性（亦称效用）之间的权衡取决于最小样本；也就是说，请回顾第 $\rho = \omega(1/n_h)$ 节中对所有 h的条件：4.1。总体隐私损失由各层中最大的隐私损失决定。然而，当层大小 N_h 保持有限时，这些层的权重（$w_h = N_h/N$）会随着 $N \to \infty$趋于 0。因此，只要 N 足够大，注入这些小层的噪声就不应损害区间的整体准确性。

更有意思的是，对于私有样本量情形，我们没有提供“PopNz-PrivSz”——一种与 PopNz-PubSz 类似的算法。为了同时保护总体比例和样本量，直接向非隐私聚合估计量添加噪声并不可行。除了直接向统计量添加噪声之外，还应考虑更复杂的机制。如果提出“PopNz-PrivSz”，我们预计它能得到更窄的置信区间，因为我们只需公布私有总体比例，而无需同时提供层比例的隐私置信区间。

未来研究的另一个方向是实现最优预算分配。我们没有讨论如何为 PopNz-PubSz 或 StrNz-PrivSz 最好地分配总预算。算法组合应用的预算分配也可能值得研究，例如第5.2.2节中，我们将算法应用两次，分别用于两个独立总体。

最后，一个广泛的研究方向是为基本 Wald 区间的其他替代方法开发差分隐私版本，例如 Wilson 区间、Jeffreys 区间等（参见 [22]，其中对七种此类比例置信区间进行了比较总结）。其中许多方法专为小样本量情形设计，而本文不考虑这种情形；我们预计，这种情形下可能需要采用与差分隐私相关的根本不同的方法。

附录A：证明

A.1. 定理 *3.1* 的证明

引理 A.1. 设 $X \sim N(\mu, \sigma^2)$ 且 $S = \{\mu - a \le X \le \mu + a\}$。对于任意 $a > 0$ 和整数 $k \ge 1$，条件偶数阶矩

$$\mathbb{E}[(X-\mu)^{2k} \mid S] = \sigma^{2k}(2k-1)!! - O\left(e^{-\frac{a^2}{2\sigma^2}} a^{2k-1}\right), \tag{23}$$

其中，大 O 项隐藏了一个依赖于 σ 和 k 的常数。

证明。 不失一般性，假设 $\mu = 0$。我们通过

归纳法证明。令 $k=1$，通过分部积分，

$$\begin{aligned}\mathbb{E}[X^2 I_S] &= \int_{-a}^{a} x^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}\, dx \\ &= \frac{\sigma}{\sqrt{2\pi}} \left(-x e^{-\frac{x^2}{2\sigma^2}} \Big|_{-a}^{a} + \int_{-a}^{a} e^{-\frac{x^2}{2\sigma^2}}\, dx \right).\end{aligned}$$

通过换元积分，第二项中的积分变为

$$\int_{-a}^{a} e^{-\frac{x^2}{2\sigma^2}}\, dx = \sigma\sqrt{2\pi}\,\mathrm{erf}\left(\frac{a}{\sigma\sqrt{2}}\right)$$

其中

$$\mathrm{erf}(z) = \sigma \int_0^z e^{-t^2}\, dt$$

是误差函数。那么，

$$\mathbb{E}[X^2 I_S] = \sigma^2\,\mathrm{erf}\left(\frac{a}{\sigma\sqrt{2}}\right) - O\left(e^{-\frac{a^2}{2\sigma^2}} a\right).$$

假设

$$\mathbb{E}[X^{2k} I_S] = \sigma^{2k}(2k-1)!!\,\mathrm{erf}\left(\frac{a}{\sigma\sqrt{2}}\right) - O\left(e^{-\frac{a^2}{2\sigma^2}} a^{2k-1}\right), \tag{24}$$

那么，对 $k+1$ 情形进行分部积分，

$$\begin{aligned}\mathbb{E}[X^{2(k+1)} I_S] &= \int_{-a}^{a} x^{2k+2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}\, dx \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_{-a}^{a} x^{2k+1} \cdot \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}\, dx \\ &= \frac{\sigma}{\sqrt{2\pi}} \left(-x^{2k+1} e^{-\frac{x^2}{2\sigma^2}} \Big|_{-a}^{a} + (2k+1) \int_{-a}^{a} x^{2k} e^{-\frac{x^2}{2\sigma^2}}\, dx \right) \\ &= \sigma^2(2k+1)\mathbb{E}[X^{2k} I_S] - O\left(e^{-\frac{a^2}{2\sigma^2}} a^{2k+1}\right).\end{aligned}$$

将（24）代入，得到

$$\mathbb{E}[X^{2(k+1)} I_S] = \sigma^{2k+2}(2k+1)!!\,\mathrm{erf}\left(\frac{a}{\sigma\sqrt{2}}\right) - O\left(e^{-\frac{a^2}{2\sigma^2}} a^{2k+1}\right).$$

至此，我们已证明（24）。注意到

$$\Pr(S) = \int_{-a}^{a} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}\, dx = \mathrm{erf}\left(\frac{a}{\sigma\sqrt{2}}\right),$$

并且 $\mathrm{erf}(z)$ 的值介于 $(-1,\ 1)$ 之间。因此，

$$\mathbb{E}[X^{2k} \mid S] = \mathbb{E}[X^{2k} I_S] / \Pr(S) = \sigma^{2k}(2k-1)!! - O\left(e^{-\frac{a^2}{2\sigma^2}} a^{2k-1}\right).$$

□

定理 3.1 中式（12）的证明。考虑 $\frac{1}{x}$ 在 $x=\mu$ 处的泰勒级数:

$$\frac{1}{x}=\sum_{j=0}^{\infty}\frac{(-(x-\mu))^j}{\mu^{j+1}}=\frac{1}{\mu}-\frac{x-\mu}{\mu^2}+\frac{(x-\mu)^2}{\mu^3}-\frac{(x-\mu)^3}{\mu^4}+\dots$$

令 y_m 为上述级数的部分和，即 $y_m(x)=\sum_{k=0}^{m}\left(-\frac{(x-\mu)^k}{\mu^{k+1}}\right)$。那么 $y_m(x)$ 收敛于包含 $[1,2\mu-1]$ 的 $(0,2\mu)$ 中的 $\frac{1}{x}$。令

$$g(x)=\sum_{k=0}^{\infty}\frac{|x-\mu|^k}{\mu^{k+1}}=\begin{cases}\frac{1}{x}, & \text{if } 1\le x\le\mu\\ \frac{1}{2\mu-x} & \text{if } \mu<x\le 2\mu-1\end{cases}$$

那么 g 可积，因为

$$\int_1^{2\mu-1}|g(x)|d\nu=\int_1^{\mu}\frac{1}{x}d\nu+\int_{\mu}^{2\mu-1}\frac{1}{2\mu-x}d\nu=2\int_1^{\mu}\frac{1}{x}d\nu<\infty,$$

其中 $d\nu=f(x)dx$ 是以事件 S 为条件，从 $\mathcal{N}(\mu,\sigma^2)$ 诱导得到的。另请注意，对于任意自然数 m 和 $x\in[1,2\mu-1]$，$|y_m(x)|\le g(x)$。根据控制收敛定理，对于$y_m(x)$，极限与积分运算可以交换。

$$\begin{aligned}\int_1^{2\mu-1}\frac{1}{x}d\nu&=\int_1^{2\mu-1}\lim_{m\to\infty}y_m(x)d\nu\\&=\lim_{m\to\infty}\int_1^{2\mu-1}y_m(x)d\nu\\&=\lim_{m\to\infty}\int_1^{2\mu-1}\left(\sum_{j=0}^{m}\frac{(-(x-\mu))^j}{\mu^{j+1}}\right)d\nu\\&=\lim_{m\to\infty}\left(\sum_{j=0}^{m}\int_1^{2\mu-1}\frac{(-(x-\mu))^j}{\mu^{j+1}}d\nu\right)\end{aligned}\tag{25}$$

那么，

$$\begin{aligned}\mathbb{E}\left(\frac{1}{X}\mid S\right)&=\sum_{j=0}^{\infty}\frac{1}{\mu^{j+1}}\mathbb{E}\left[(-(X-\mu))^j\mid S\right]\\&=\sum_{j=0}^{\infty}\frac{1}{\mu^{2j+1}}\mathbb{E}\left[(X-\mu)^{2j}\mid S\right]\\&=\sum_{j=0}^{k}\frac{1}{\mu^{2j+1}}\mathbb{E}\left[(X-\mu)^{2j}\mid S\right]+\frac{1}{\mu}\sum_{j=k+1}^{\infty}\mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2j}\mid S\right].\end{aligned}\tag{26}$$

第二个等式成立是因为对称性使奇数阶矩为零。

注意，给定事件 S, $|\frac{X-\mu}{\mu}| \leq \frac{\mu-1}{\mu} < 1$后，

$$\mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2k+2} \mid S\right] \leq \left(\frac{\mu-1}{\mu}\right)^{2} \mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2k} \mid S\right]. \tag{27}$$

由此可得

$$\begin{aligned}\sum_{j=k+1}^{\infty} \mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2j} \mid S\right] &\leq \sum_{j=0}^{\infty} \left(\frac{\mu-1}{\mu}\right)^{2j} \mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2k+2} \mid S\right] \\ &= \frac{\mu^2}{2\mu-1}\mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2k+2} \mid S\right] \\ &= O\left(\frac{1}{\mu^{2k+1}}\right) \cdot \mathbb{E}[(X-\mu)^{2k+2} \mid S].\end{aligned}$$

应用引理A.1，根据 $a = \mu - 1$的选取，式（26）变为

$$\mathbb{E}\left(\frac{1}{X} \mid S\right) = \frac{1}{\mu}\sum_{j=0}^{k} \frac{(2j-1)!!\sigma^{2j}}{\mu^{2j}} + O\left(\frac{\sigma^{2k+2}}{\mu^{2k+2}}\right).$$

□

定理*13*）中式（*3.1*的证明。我们对 $X \mid S$的二阶矩采用类似的处理过程。根据泰勒展开

$$\frac{1}{x^2} = \sum_{j=0}^{\infty} \frac{(j+1)(-(x-\mu))^j}{\mu^{j+2}} = \frac{1}{\mu^2} - \frac{2(x-\mu)}{\mu^3} + \frac{3(x-\mu)^2}{\mu^4} - \frac{4(x-\mu)^3}{\mu^5} + \cdots,$$

我们有

$$\begin{aligned}\mathbb{E}\left(\frac{1}{X^2} \mid S\right) &= \sum_{j=0}^{\infty} \frac{j+1}{\mu^{j+2}} \mathbb{E}\left[(-(X-\mu))^j \mid S\right] \\ &= \sum_{j=0}^{\infty} \frac{2j+1}{\mu^{2j+2}} \mathbb{E}\left[(X-\mu)^{2j} \mid S\right] \\ &= \sum_{j=0}^{k} \frac{2j+1}{\mu^{2j+2}} \mathbb{E}\left[(X-\mu)^{2j} \mid S\right] + \frac{1}{\mu^2} \sum_{j=k+1}^{\infty} (2j+1)\mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2j} \mid S\right].\end{aligned} \tag{28}$$

由于（27），因此可得

$$\begin{aligned}
&\sum_{j=k+1}^{\infty}(2j+1)\mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2j}\mid S\right]\\
&\le \mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2k+2}\mid S\right]\cdot\sum_{j=0}^{\infty}(2k+3+2j)\left(\frac{\mu-1}{\mu}\right)^{2j}\\
&=\mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2k+2}\mid S\right]\cdot\left[(2k+3)\sum_{j=0}^{\infty}\left(\frac{\mu-1}{\mu}\right)^{2j}+2\sum_{j=1}^{\infty}j\left(\frac{\mu-1}{\mu}\right)^{2j}\right]\\
&=\mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2k+2}\mid S\right]\cdot\left[\frac{(2k+3)\mu^2}{2\mu-1}+2\frac{\mu^2(\mu-1)^2}{(2\mu-1)^2}\right]\\
&=\mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2k+2}\mid S\right]\cdot O(\mu^2)\\
&=O\left(\frac{1}{\mu^{2k}}\right)\cdot\mathbb{E}[(X-\mu)^{2k+2}\mid S],
\end{aligned}\tag{29}$$

其中，术语 $\sum_{j=1}^{\infty}j\left(\frac{\mu-1}{\mu}\right)^{2j}$ 是一个等差几何数列的和。根据引理 A.1，(28)可化为

$$\mathbb{E}\left(\frac{1}{X^2}\mid S\right)=\frac{1}{\mu^2}\sum_{j=0}^{k}\frac{(2j+1)!!\sigma^{2j}}{\mu^{2j}}+O\left(\frac{\sigma^{2k+2}}{\mu^{2k+2}}\right).\tag{30}$$

□

A.2. 定理 *4.1*的证明

算法 *1*的证明。 在邻接关系 $\sim_{ss}$下，只有一个层内的一条记录发生变化，且样本量保持不变。对每一层应用参数为 ρ 的高斯机制，可得 ρ-零集中差分隐私保证。根据后处理性质，该置信区间也满足 ρ-零集中差分隐私。

□

算法 *2*的证明。 $\hat{p}$的敏感度以及 $\mathrm{Var}(\hat{p})$ 分别为 Δ_p 和 Δ_V。应用高斯机制可得， $\tilde{p}$ 满足 ρ_1-零集中差分隐私，且$\widetilde{V}$满足 ρ_2-零集中差分隐私。根据基本组合，置信区间 $\tilde{p}\pm z_{1-\frac{\alpha}{2}}\sqrt{\widetilde{V}}$满足$(\rho_1+\rho_2)$-零集中差分隐私。

□

算法 *3*的证明。 根据高斯机制和零集中差分隐私的基本组合性质，我们知道 $\tilde{p}_h$ 是 ρ-零集中差分隐私。在邻接关系$\sim_r$下，在一个层内只有一条记录发生变化。因此，根据后处理，置信区间是 ρ-零集中差分隐私。

□

A.3. 定理 4.2 的证明

在证明该定理之前，我们先回顾有限总体中心极限定理：

定理 A.2 （定理 1, [33]）。考虑一个有限总体 $\Pi = \{X_1, ..., X_N\}$，其大小为 N。令 μ 为总体均值，X_n 为从 Π中抽取的大小为 n 的简单随机样本的均值，且$\mathrm{Var}(X_n)$ 为 X_n的方差。Π 的有限总体方差记为

$$v = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu)^2.$$

当 $N \to \infty$ 时，如果

$$\frac{1}{\min(n, N-n)} \cdot \frac{\max_{1\le i\le N}(X_i-\mu)^2}{v} \to 0, \tag{31}$$

我们有

$$\frac{\bar{X}_n - \mu}{\sqrt{\mathrm{Var}(\bar{X}_n)}} \xrightarrow{d} \mathcal{N}(0,1). \tag{32}$$

$\bar{X}_n$ 的方差由总体方差 v 决定，而该方差未知。然而，样本方差 $\mathrm{Var}(X_n)$可用于估计 v。为确保将 $\mathrm{Var}(X_n)$替换为 $\mathrm{Var}(X_n)$后中心极限定理仍然成立，$\mathrm{Var}(X_n)$的一致性至关重要，正如下述引理所述。

引理 A.3. 令 $\mathrm{Var}(X_n)$ 为样本方差。$\mathrm{Var}(X_n)$ 是一个无偏估计量，用于估计 $\mathrm{Var}(X_n)$。此外，在满足定理 A.2中的条件时，当 $N \to \infty$,

$$\widehat{\mathrm{Var}}(\bar{X}_n)/\mathrm{Var}(\bar{X}_n) \xrightarrow{p} 1.$$

现在证明定理 4.2：

定理 4.2的证明。只需证明 $\frac{\tilde{p}_h - p_h}{\sqrt{\tilde{V}_h}} \xrightarrow{d} \mathcal{N}(0,1)$对所有 h都成立。根据定理 A.2中的有限总体中心极限定理，我们知道

$$\frac{\hat{p}_h - p_h}{\sqrt{\mathrm{Var}(\hat{p}_h)}} \xrightarrow{d} \mathcal{N}(0,1).$$

由于 $\tilde{p}_h = \hat{p}_h + e_h$其中 $e_h \sim \mathcal{N}(0, \frac{1}{2\rho n_h^2})$，因此我们有

$$\frac{\tilde{p}_h - p_h}{\sqrt{\mathrm{Var}(\tilde{p}_h)}} \xrightarrow{d} \mathcal{N}(0,1) \tag{33}$$

其中

$$\mathrm{Var}(\tilde{p}_h) = \mathrm{Var}(\hat{p}_h) + \frac{1}{2\rho n_h^2}.$$

Let

$$\begin{aligned}\widetilde{V}_h &= \left(\frac{N_h - n_h}{N_h}\right)\frac{\tilde{p}_h(1-\tilde{p}_h) + \frac{1}{2\rho n_h^2}}{n_h - 1} + \frac{1}{2\rho n_h^2} \\ &= \widehat{\mathrm{Var}}(\hat{p}_h) + \left(\frac{N_h - n_n}{N_h}\right)\frac{e_h - 2\tilde{p}_h e_h - e_h^2 + \frac{1}{2\rho n_h^2}}{n_h - 1} + \frac{1}{2\rho n_h^2}.\end{aligned} \tag{34}$$

由于 $e_h \sim \mathsf{N}(0, \frac{1}{2\rho n_h^2})$，我们有 $e_h = O_P(\frac{1}{\sqrt{\rho} n_h})$。于是，(34) 的第二项为 $O_P(\frac{1}{\sqrt{\rho} n_h^2})$，因此，$\widetilde{V}_h - \mathsf{Var}(\tilde{p}_h) = \widehat{\mathsf{Var}}(\hat{p}_h) - \mathsf{Var}(\hat{p}_h) + O_P\left(\frac{1}{\sqrt{\rho} n_h^2}\right)$。注意，$\widehat{\mathsf{Var}}(\hat{p}_h)$ 的阶为 $\frac{1}{n_h}$，且根据引理 A.3，$\widehat{\mathsf{Var}}(\hat{p}_h) \xrightarrow{p} \mathsf{Var}(\hat{p}_h)$。因此，$\widetilde{V}_h \xrightarrow{p} \mathsf{Var}(\tilde{p}_h)$，从而 $\widetilde{V} \xrightarrow{p} \mathsf{Var}(\tilde{p})$。

结合 V 与（33），我们有

$$\frac{\tilde{p}_h - p_h}{\sqrt{\widetilde{V}_h}} \xrightarrow{d} \mathcal{N}(0,1) \tag{35}$$

根据斯卢茨基定理。于是，$\frac{\tilde{p}-p}{\sqrt{\widetilde{V}}} \xrightarrow{d} \mathcal{N}(0,1)$。因此，由 $p \pm z_{1-\alpha/2}\sqrt{\widetilde{V}}$给出的置信区间具有渐近覆盖水平 $1-\alpha$。

□

A.4. 定理*4.3*的证明

证明。由于$\frac{\hat{p}-p}{\sqrt{\mathrm{Var}(\hat{p})}} \xrightarrow{d} \mathcal{N}(0,1)$且 $\tilde{p} = \hat{p} + \mathcal{N}(0, \Delta_p^2/2\rho_1)$，并且 $\Delta_p = \max_h \frac{w_h}{n_h}$，因此

$$\frac{\tilde{p} - p}{\sqrt{\mathrm{Var}(\tilde{p})}} \xrightarrow{d} \mathcal{N}(0,1),$$

and

$$\mathrm{Var}(\tilde{p}) = \mathrm{Var}(\hat{p}) + \frac{\Delta_p^2}{2\rho_1}.$$

在算法2中，我们设

$$\widetilde{V} = \widehat{\mathrm{Var}}(\hat{p}) + \frac{\Delta_p^2}{2\rho_1} + e_V, \tag{36}$$

其中 $e_V \sim \mathcal{N}(0, \frac{\Delta_V^2}{2\rho_2})$，且 $\Delta_V = \max_h \left(\frac{C_h}{n_h}\left(1 - \frac{1}{n_h}\right)\right)$ 和 $C_h = w_h^2 \frac{N_h - n_h}{N_h}\frac{1}{n_h - 1}$。由于 $\Delta_V = O(\frac{1}{\max_h n_h^2})$，我们有 $e_V = O_P(\frac{1}{\max_h \sqrt{\rho_2} n_h^2})$。因此，$\widetilde{V} - \mathsf{Var}(\tilde{p}) = \widehat{\mathsf{Var}}(\hat{p}) - \mathsf{Var}(\hat{p}) + O_P\left(\frac{1}{\max_h \sqrt{\rho} n_h^2}\right)$。由有限总体中心极限定理可知，$\widehat{\mathsf{Var}}(\hat{p}) \xrightarrow{p} \mathsf{Var}(\hat{p})$，因此我们有 $\widetilde{V} \xrightarrow{p} \mathsf{Var}(\tilde{p})$。

因此，根据斯卢茨基定理，

$$\frac{\tilde{p} - p}{\sqrt{\widetilde{V}}} \xrightarrow{d} \mathcal{N}(0,1). \tag{37}$$

那么， $p \pm z_{1-\alpha/2}\sqrt{\widetilde{V}}$给出的置信区间具有渐近覆盖率 $1-\alpha$。

□

A.5. 定理 *4.4* 的证明

证明。 对于 $\tilde{n} \sim \mathcal{N}(n, \frac{1}{2\rho_2})$，根据命题 3.1，我们推导出条件期望的 k 阶泰勒级数，其中 $\tilde{p}$ 给定 $S = \{1 \le \tilde{n} \le 2n-1\}$：

$$\mathbb{E}\left(\tilde{p} \mid S\right) = p \sum_{j=0}^{k} \frac{(2j-1)!!}{n^{2j}(2\rho_2)^j} + O\left(\frac{1}{n^{2k+1}\rho_2^{k+1}}\right). \tag{38}$$

例如，当 $k = 2$ 时，

$$\mathbb{E}\left(\tilde{p} \mid S\right) = p\left(1 + \frac{1}{2n^2\rho_2} + \frac{3}{4n^4\rho_2^2}\right) + O\left(\frac{1}{n^5\rho_2^3}\right). \tag{39}$$

为了得到条件方差的泰勒展开，我们代入

$$\mathbb{E}\left(\frac{1}{\tilde{n}} \mid S\right) = \frac{1}{n}\sum_{j=0}^{k} \frac{(2j-1)!!}{n^{2j}(2\rho_2)^j} + O\left(\frac{1}{n^{2k+2}\rho_2^{k+1}}\right)$$

and

$$\mathbb{E}\left(\frac{1}{\tilde{n}^2} \mid S\right) = \frac{1}{n^2}\sum_{j=0}^{k} \frac{(2j+1)!!}{n^{2j}(2\rho_2)^j} + O\left(\frac{1}{n^{2k+2}\rho_2^{k+1}}\right)$$

into

$$\mathrm{Var}(\tilde{p} \mid S) = \mathbb{E}(\tilde{p}^2 \mid S) - (\mathbb{E}(\tilde{p} \mid S))^2 = \mathbb{E}\tilde{c}^2\mathbb{E}\left(\frac{1}{\tilde{n}^2} \mid S\right) - (\mathbb{E}(\tilde{p} \mid S))^2,$$

由此我们推导出条件方差的一般展开式：

$$\begin{aligned}\mathrm{Var}(\tilde{p} \mid S) = \mathrm{Var}(\hat{p}) \sum_{j=0}^{k} \frac{(2j+1)!!}{n^{2j}(2\rho_2)^j} + p^2 \left(\sum_{j=0}^{k} \frac{(2j+1)!!}{n^{2j}(2\rho_2)^j} - \left(\sum_{j=0}^{k} \frac{(2j-1)!!}{n^{2j}(2\rho_2)^j}\right)^2\right) \\ + \frac{1}{2\rho_1}\sum_{j=0}^{k} \frac{(2j+1)!!}{n^{2j+2}(2\rho_2)^j} + O\left(\frac{1}{n^{2k}\rho_2^{k+1}}\right) + O\left(\frac{1}{n^{2k+2}\rho_1\rho_2^{k+1}}\right).\end{aligned} \tag{40}$$

当 $k = 2$,

$$\begin{aligned}\mathrm{Var}(\tilde{p} \mid S) = \mathrm{Var}(\hat{p})\left(1 + \frac{3}{2n^2\rho_2} + \frac{15}{4n^4\rho_2^2}\right) + p^2\left(\frac{1}{2n^2\rho_2} + \frac{2}{n^4\rho_2^2} - \frac{6}{8n^6\rho_2^3} - \frac{9}{16n^8\rho_2^4}\right) \\ + \frac{1}{2\rho_1}\left(\frac{1}{n^2} + \frac{3}{2n^4\rho_2} + \frac{15}{4n^6\rho_2^2}\right) + O\left(\frac{1}{n^4\rho_2^3}\right) + O\left(\frac{1}{n^6\rho_1\rho_2^3}\right).\end{aligned} \tag{41}$$

基于泰勒展开，并使用 $k=2$ 式（39）和（41）中给出的条件均值和方差，在条件 $\frac{1}{\rho_1 n}=o(1)$ 和 $\frac{1}{\rho_2 n}=o(1)$ 下，我们有

$$\mathbb{E}(\tilde{p}\mid S)=p+o\left(\frac{1}{n}\right)$$

and

$$\mathrm{Var}(\tilde{p}\mid S)=\mathrm{Var}(\hat{p})+o\left(\frac{1}{n}\right).$$

因此，$\tilde{p}\mid S$ 是渐近无偏的。注意，Var(ˆp)的阶为$\frac{1}{n}$，因此$\tilde{p}\mid S$ 的方差趋于零。因此，$\tilde{p}\mid S$ 依概率收敛于 p。还要注意，当 $n\to\infty$时，Pr(S) $\to 1$；那么对于任意 $\epsilon>0$,

$$\Pr(|\tilde{p}-p|>\epsilon)=\Pr(|\tilde{p}-p|>\epsilon\mid S)+\Pr(|\tilde{p}-p|>\epsilon\mid S^{\mathrm{c}})\to 0.$$

也就是说，$\tilde{p}$ 是 p的相合估计量。

□

A.6. 定理 *4.5* 的证明

为了证明定理 4.5，我们需要以下定理和引理。

定理 A.4（定理 1, [18]）。设 X 为一个均值 μ_x为正、方差为 σ_x^2、变异系数为 $\delta_x=\sigma_x/\mu_x$ 的正态随机变量，满足$0<\delta_x<\lambda\le 1$，其中 λ 为已知常数。对于每个 $\epsilon>0$，存在$\gamma(\epsilon)\in(0,\sqrt{\lambda^2-\delta_x^2})$，并且还存在一个与 X独立的正态随机变量 Y，其均值 μ_y为正、方差为 σ_y^2、变异系数为 $\delta_y=\sigma_y/\mu_y$，满足以下条件，

$$0<\delta_y\le\gamma(\epsilon)\le\sqrt{\lambda^2-\delta_x^2}<\lambda \tag{42}$$

满足以下结果的任意 z，只要其属于该区间

$$I=\left[\beta-\frac{\sigma_z}{\lambda},\beta+\frac{\sigma_z}{\lambda}\right],$$

其中 $\beta=\mu_x/\mu_y$、$\sigma_z=\beta\sqrt{\delta_x^2+\delta_y^2}$ 满足

$$|G(z)-F_Z(z)|<\epsilon,$$

其中 $G(z)$ 是$N(\beta,\sigma_z^2)$的累积分布函数，而 F_Z是 $Z=X/Y$的累积分布函数。注意，一旦给定的 Y满足相应的 G 与 F_Z 之间的接近性，那么任何变异系数更小的其他随机变量也都满足这一结果。

引理 A.5。对于大小为 N 的总体，设 p 为总体中具有感兴趣属性的真实比例。考虑样本量为 n 的简单随机抽样。令 $Z^*\sim N(p,V)$，其中 $V=\binom{N-n}{N-1}\frac{p}{n}(1-p+\frac{1}{2\rho_1 n^2}+\frac{p^2}{2\rho_2 n^2})$。如果

$\rho_1 = \omega(1/n^2)$ 且 $\rho_2 = \omega(1/n)$，当 $N-n$ 和 n 均趋于无穷大时，则对于任意 $z \in (0,\ 2p)$，

$$|F_{\tilde{p}}(z) - F_{Z^*}(z)| \to 0. \tag{43}$$

证明。根据定理 A.2 中的中心极限定理，我们知道 $\hat{p} \sim \mathcal{AN}(p,\ \mathrm{Var}(\hat{p}))$。回顾 $\tilde{c} = n\hat{p} + \mathcal{N}(n,\ \frac{1}{2\rho_1})$，则 $\tilde{c} \sim \mathcal{AN}(np,\ n^2\mathrm{Var}(\hat{p}) + \frac{1}{2\rho_1})$。

令 $\tilde{X} \sim \mathcal{AN}(np,\ n^2\mathrm{Var}(\hat{p}) + \frac{1}{2\rho_1})$ 和 $X \sim \mathcal{N}(np,\ n^2\mathrm{Var}(\hat{p}) + \frac{1}{2\rho_1})$。因此，对于任意 $\epsilon > 0$，存在某个 $n_0 = n_0(\epsilon)$，使得对于任意 x 和 $n > n_0$，

$$|F_{\tilde{X}}(x) - F_X(x)| < \epsilon, \tag{44}$$

其中 F表示累积分布函数。令 $Y \sim \mathcal{N}(n,\ \frac{1}{2\rho_2})$、$\tilde{Z} = \tilde{X}/Y$ 和 $Z = X/Y$，则

$$F_{\tilde{Z}}(z) = \Pr\left(\frac{\tilde{X}}{Y} < z\right) = \Pr(\tilde{X} < Yz) = \int_{-\infty}^{\infty} F_{\tilde{X}}(yz) f_y(y) dy,$$

其中 $f_y(y)$是 Y的密度函数。由 (44)，$|F_{\tilde{X}}(yx) - F_X(yx)| < \epsilon$。因此，

$$\int_{-\infty}^{\infty} (F_X(yx) - \epsilon) f_y(y) dy < \int_{-\infty}^{\infty} F_{\tilde{X}}(yx) f_y(y) dy < \int_{-\infty}^{\infty} (F_X(yx) + \epsilon) f_y(y) dy,$$

这等价于

$$\left| \int_{-\infty}^{\infty} F_{\tilde{X}}(yx) f_y(y) dy - \int_{-\infty}^{\infty} F_X(yx) f_y(y) dy \right| < \epsilon,$$

即，

$$|F_{\tilde{Z}}(z) - F_Z(z)| < \epsilon. \tag{45}$$

设 δ_x和 δ_y分别为 X和 Y的变异系数，则 $\delta_x^2 = (\mathrm{Var}(\hat{p}) + \frac{1}{2\rho_1 n^2})/p^2$ 且 $\delta_y^2 = \frac{1}{2\rho_2 n^2}$。在条件 $\frac{1}{\rho_1 n} = o(1)$ 下，由于 $\mathrm{Var}(\hat{p}) = O(\frac{1}{n})$，我们有 $\delta_x^2 = O(\frac{1}{n})$。在条件 $\frac{1}{\rho_2 n} = o(1)$ 下，我们知道 $\delta_y^2 = o(\frac{1}{n})$，进而有 $\delta_y = o(\delta_x)$。当 n足够大时，δ_y足够小。设 $\lambda = \sqrt{\delta_x^2 + 2\delta_y^2}$和 $F_{Z^*}(z)$为 $Z^* \sim \mathrm{N}(p,\ \mathrm{Var}(\hat{p}) + \frac{1}{2\rho_1 n^2} + \frac{p^2}{2\rho_2 n^2})$ 的分布函数。根据引理 A.4，对于与 X独立的正态随机变量 Y，当 δ_y足够小时，条件 (42) 成立，且我们有

$$|F_Z(z) - F_{Z^*}(z)| < \epsilon, \tag{46}$$

对于任意满足 $\sigma_{z^*} = p\sqrt{\delta_x^2 + \delta_y^2}$的 $z \in I = [p - \frac{\sigma_{z^*}}{\lambda}, p + \frac{\sigma_{z^*}}{\lambda}]$。因此，对于 $z \in I$，

$$|F_{\tilde{Z}}(z) - F_{Z^*}(z)| < |F_{\tilde{Z}}(z) - F_Z(z)| + |F_Z(z) - F_{Z^*}(z)| < 2\epsilon. \tag{47}$$

还要注意，当 $n \to \infty$、$\frac{\sigma_{z^*}}{\lambda} \longrightarrow p$ 时，I 的极限为 $(0, 2p)$。

到目前为止，我们已证明，当 n 趋于无穷大时，在条件 $\frac{1}{\rho_1 n} = o(1)$ 和 $\frac{1}{\rho_2 n} = o(1)$ 下，对于 $z \in I_h$，

$$|F_{\tilde{Z}}(z) - F_{Z^*}(z)| \to 0. \tag{48}$$

□

引理 A.6. 设 $Z_1, ..., Z_H$ 和 $Z_1^*, ..., Z_H^*$ 是依赖于 n 的相互独立的连续随机变量。令 F 表示分布函数。当 $n \to \infty$ 时，若

$$|F_{Z_h}(z) - F_{Z_h^*}(z)| \to 0$$

对区间 (a_h, b_h) 中的任意 $h = 1, ..., H$ 和 z 均成立，且 $\Pr(Z_h \in (a_h, b_h)) \to 1$，$\Pr(Z_h^* \in (a_h, b_h)) \to 1$。那么，

$$\left|F_{\sum_{h=1}^H c_h Z_h}(z) - F_{\sum_{h=1}^H c_h Z_h^*}(z)\right| \to 0$$

对任意 $z \in \left(\sum_{h=1}^H c_h a_h, \sum_{h=1}^H c_h b_h\right)$ 均成立，其中 c_h's 为常数。

证明。 只需证明，对于任意 $z \in (a_1 c_1 + a_2 c_2, b_1 c_1 + b_2 c_2)$，

$$\left|F_{c_1 Z_1 + c_2 Z_2}(z) - F_{c_1 Z_1^* + c_2 Z_2^*}(z)\right| \to 0$$

当 $n \to \infty$ 时。我们有

$$\begin{aligned}
&F_{c_1 Z_1 + c_2 Z_2}(z) \\
&= \Pr(c_1 Z_1 + c_2 Z_2 < z) \\
&= \Pr\left(Z_1 < \frac{z - c_2 Z_2}{c_1}\right) \\
&= \int_{\mathbb{R}} F_{Z_1}\left(\frac{z - c_2 x}{c_1}\right) f_{Z_2}(x) dx \\
&= \int_{\mathbb{R}} \left[F_{Z_1}\left(\frac{z - c_2 x}{c_1}\right) - F_{Z_1^*}\left(\frac{z - c_2 x}{c_1}\right)\right] f_{Z_2}(x) dx + \int_{\mathbb{R}} F_{Z_1^*}\left(\frac{z - c_2 x}{c_1}\right) f_{Z_2}(x) dx \\
&= \int_{\mathbb{R}} \left[F_{Z_1}\left(\frac{z - c_2 x}{c_1}\right) - F_{Z_1^*}\left(\frac{z - c_2 x}{c_1}\right)\right] f_{Z_2}(x) dx + F_{c_1 Z_1^* + c_2 Z_2}(z).
\end{aligned} \tag{49}$$

当 $a_1 < (z - c_2 x)/c_1 < b_1$ 时，我们知道

$$\left|F_{Z_1}\left(\frac{z - c_2 x}{c_1}\right) - F_{Z_1^*}\left(\frac{z - c_2 x}{c_1}\right)\right| \to 0. \tag{50}$$

由于 $F_{Z_1}(b_1) - F_{Z_1}(a_1) \to 1$ 且 $F_{Z_1^*}(b_1) - F_{Z_1^*}(a_1) \to 1$，对于任意 $a < a_1$，均有 $F_{Z_1}(a) \to 0$ 且 $F_{Z_1^*}(a) \to 0$；对于任意 $b > b_1$，$F_{Z_1}(b) \to 1$且 $F_{Z_1^*}(b) \to 1$。因此，当 $(z - c_2 x)/c_1$ 位于 (a_1, b_1) 之外时，(50) 仍然成立。于是，(49) 右端的第一项收敛于 0。然后

$$\left|F_{c_1 Z_1 + c_2 Z_2}(z) - F_{c_1 Z_1^* + c_2 Z_2}(z)\right| \to 0.$$

同样地，我们有

$$\left|F_{c_1 Z_1^* + c_2 Z_2}(z) - F_{c_1 Z_1^* + c_2 Z_2^*}(z)\right| \to 0.$$

根据三角不等式，

$$\left|F_{c_1 Z_1 + c_2 Z_2}(z) - F_{c_1 Z_1^* + c_2 Z_2^*}(z)\right| \to 0.$$

□

定理 *4.5* 的证明。根据引理 A.5，对于每个层，在条件$\rho_1 = \omega(1/n_h)$ 和 $\rho_2 = \omega(1/n_h)$下，$\tilde{p}_h$ 的分布函数在区间 $\mathcal{N}(p_h, V_h)$ 上收敛于其对应的分布函数，其中区间为 $(0, 2p_h)$，其中

$$V_h = \left(\frac{N_h - n_h}{N_h - 1}\right)\frac{p_h(1-p_h)}{n_h} + \frac{1}{2\rho_1 n_h^2} + \frac{p_h^2}{2\rho_2 n_h^2}. \tag{51}$$

令 $p^* \sim \mathcal{N}(p, V)$ 其中 $V = \sum_{h=1}^{H} w_h^2 V_h$。根据引理 A.6，在区间 $(0, 2p)$内，有

$$|F_{\tilde{p}}(z) - F_{p^*}(z)| \to 0. \tag{52}$$

其中 $F_{\tilde{p}}$ 表示由算法 $\tilde{p}$ 设计的 3 的分布函数，而F_{p^*} 是 p^*的分布函数。

令 $L = p - z_{1-\alpha/2}\sqrt{V}$，$U = p + z_{1-\alpha/2}\sqrt{V}$，$\tilde{L} = p - z_{1-\alpha/2}\sqrt{\tilde{V}}$和 $\tilde{U} = p + z_{1-\alpha/2}\sqrt{\tilde{V}}$。注意，$L$和 U是常数，而 $\tilde{L}$和 $\tilde{U}$是随机变量。假设 n_h 足够大，U 和 L 位于以下结论成立的区间内，这是由式（52）可得，

$$|F_{\tilde{p}}(U) - F_{p^*}(U)| \to 0 \tag{53}$$

and

$$|F_{\tilde{p}}(L) - F_{p^*}(L)| \to 0. \tag{54}$$

另一方面，根据定理 3.1 和 4.4，我们知道 $\tilde{p}_h \xrightarrow{p} p_h$ 和 $\frac{1}{\tilde{n}_h} \xrightarrow{p} \frac{1}{n_h}$ 在条件$\rho_1 = \omega(1/n_h)$ 和 $\rho_2 = \omega(1/n_h)$下成立。根据连续映射定理，$\widetilde{V}_h \xrightarrow{p} V_h$ 当 $n_h \to \infty$时，因此 $\widetilde{V} \xrightarrow{p} V$。因此，$\tilde{U} \xrightarrow{p} U$和 $\tilde{L} \xrightarrow{p} L$。由于 $F_{\tilde{p}}$ 是连续的，我们有

$$|F_{\tilde{p}}(\tilde{U}) - F_{\tilde{p}}(U)| \xrightarrow{p} 0 \tag{55}$$

and

$$|F_{\tilde{p}}(\tilde{L}) - F_{\tilde{p}}(L)| \xrightarrow{p} 0. \tag{56}$$

因此，

$$\begin{aligned}
&\Pr\left(p \in \left(\tilde{p} - z_{1-\alpha/2}\sqrt{\widetilde{V}}, \tilde{p} + z_{1-\alpha/2}\sqrt{\widetilde{V}}\right)\right) \\
&= \Pr\left(p - z_{1-\alpha/2}\sqrt{\widetilde{V}} < \tilde{p} < p + z_{1-\alpha/2}\sqrt{\widetilde{V}}\right) \\
&= \left(F_{\tilde{p}}(\tilde{U}) - F_{\tilde{p}}(U)\right) + \left(F_{\tilde{p}}(U) - F_{p^*}(U)\right) \\
&\quad - \left(F_{\tilde{p}}(\tilde{L}) - F_{\tilde{p}}(L)\right) - \left(F_{\tilde{p}}(L) - F_{p^*}(L)\right) + \left(F_{p^*}(U) - F_{p^*}(L)\right).
\end{aligned}$$

综合（53）至（56）以及$F_{p^*}(U) - F_{p^*}(L) = 1 - \alpha$，我们得到

$$\lim_{n\to\infty} \Pr\left(p \in \left(\tilde{p} - z_{1-\alpha/2}\sqrt{\widetilde{V}}, \tilde{p} + z_{1-\alpha/2}\sqrt{\widetilde{V}}\right)\right) \to 1 - \alpha.$$

由于$\frac{1}{\tilde{n}_h} \xrightarrow{p} \frac{1}{n_h}$，在条件 $\rho_1 = \omega(1/n_h)$和 $\rho_2 = \omega(1/n_h)$下，有 $\frac{1}{2\rho_1\tilde{n}_h^2} = o_P(\frac{1}{n_h})$和 $\frac{\tilde{p}_h^2}{2\rho_2\tilde{n}_h^2} = o_P(\frac{1}{n_h})$。因此，注入噪声导致的 $\tilde{p}$ 条件方差估计额外误差为$O_p\left(\frac{1}{\rho_1 n_h^2} + \frac{1}{\rho_2 n_h^2}\right) = o_p\left(\frac{1}{n_h}\right)$。

□

致谢

我们感谢 Rolando Rodriguez、Brian Finley、Jörg Drechsler、Gary Benedetto、Michael Freiman 和 Justin Doty（排名不分先后）与我们进行了有益的交流并提出了意见。本项目由美国人口普查局合作协议 CB20ADR0160001 资助。

参考文献

[1] Abowd., J. M.（2016）。科学可重复性挑战与统计机构的隐私保护。[2] Bickel, P. J. 和 Freedman, D. A.（1984）。分层抽样中的渐近正态性与自助法。《统计年鉴》 **12** 470 – 482。[3] Brawner, T. 和 Honaker, J.（2018）。自助法推断与差分隐私：免费的标准误。未发表手稿。[4] Bun, M. 和 Steinke, T.（2016）。集中差分隐私：简化、扩展与下界。*arXiv* **abs/1605.02065**。[5] 美国人口普查局（2020）。2020年人口普查的披露规避：简介。[6] Cai, T. T., Wang, Y. 和 Zhang, L.（2021）。隐私成本：差分隐私下参数估计的最优收敛速率。《统计年鉴》 **49**2825–2850。[7] Covington, C., He, X., Honaker, J. 和Kamath, G.（2021）。差分隐私下的无偏统计估计与有效置信区间。*arXiv***abs/2110.14465**。[8] Ding, B., Kulkarni, J. 和 Yekhanin, S.（2017）。以隐私方式收集遥测数据。载于神经信息处理系统进展 *30*：*2017*年神经信息处理系统年会，*2017*年*12*月*4*日至*9*日，美国加利福尼亚州长滩3571–3580。[9] D'Orazio, V., Honaker, J. 和 King, G.（2015）。社会科学推断中的差分隐私。*ERN*：其他计量经济学：计量经济学与统计方法（专题）。

[10] Drechsler, J., Globus-Harris, I., Mcmillan, A., Sarathy, J. 和 Smith, A.（2022）。中位数的非参数差分隐私置信区间。《调查统计与方法学杂志》10 804- 829。[11] Du, W., Foot, C., Moniot, M., Bray, A. 和 Groce, A.（2020）。差分隐私置信区间。arXiv abs/2001.02285。[12] DuMouchel, D. G. J. William H.（1983）。在分层样本的多元回归分析中使用抽样调查权重。《美国统计协会杂志》78 535–543。[13] Dwork, C. 和 Lei, J.（2009）。差分隐私与稳健统计。载于第四十一届 ACM 计算理论专题讨论会论文集 371–380。[14] Dwork, C., McSherry, F., Nissim, K. 和 Smith, A.（2006）。为隐私数据分析中的敏感度校准噪声。载于密码学理论会议 265–284。Springer。[15] Dwork, C. 和 Roth, A.（2014）。差分隐私的算法基础。《理论计算机科学基础与趋势》9 211–407。[16] Dwork, C. 和 Rothblum, G. N.（2016）。集中差分隐私。CoRR abs/1603.01887。[17] Dwork, C. 和 Smith, A.（2009）。统计学中的差分隐私：已知与待学。《隐私与保密杂志》1。[18] Díaz-Francés, E. 和 Rubio, F.（2013）。关于两个独立正态随机变量之比的分布是否存在正态近似。《统计论文》54。[19] Erdös, P. 和 Rényi, A.（1959）。有限总体抽样的中心极限定理。《匈牙利科学院数学研究所刊物》4 49–61。[20] Erlingsson, U., Pihur, V. 和 Korolova, A.（2014）。RAPPOR：随机化可聚合隐私保护序数响应。载于2014年 ACM SIGSAC 计算机与通信安全会议论文集。CCS '14 1054–1067。计算机协会，纽约，美国纽约州。[21] Ferrando, C., Wang, S. 和 Sheldon, D.（2022）。差分隐私置信区间的参数自助法。载于 AISTATS。[22] Franco, C., Little, R. J. A., Louis, T. A. 和 Slud, E. V.（2019）。复杂抽样调查中比例置信区间的比较研究。《调查统计与方法学杂志》7 3 334-364。[23]Gaboardi, M., Rogers, R. M. 和 Sheffet, O.（2019）。局部隐私均值估计：Z 检验与紧置信区间。arXiv abs/1810.08054。[24] Groshen, E. L. 和 Goroff, D.（2022）。披露规避与2020年人口普查：研究人员需要了解什么？《哈佛数据科学评论》特刊 2。[25] Hájek, J.（1960）。简单随机抽样中的极限分布，来自

有限总体。《Magyar Tud. Akad. Mat. Kutató Int. Közl.》5，361–374。[26] Hayya, J., Armstrong, D. 和 Gressis, N.（1975）。关于两个正态分布变量之比的注记。《管理科学》21，1338-1341。[27] Karwa, V. 和 Vadhan, S.（2017）。有限样本差分隐私置信区间。CoRR abs/1711.03908。[28] Kifer, D. 和 Machanavajjhala, A.（2011）。数据隐私没有免费午餐。载于 ACM SIGMOD会议。[29] Kroll, M.（2021）。局部差分隐私下定点密度估计。《电子统计学杂志》15，1783–1813。[30] Kuethe, D., Caprihan, A., Gach, H., Lowe, I. 和 Fukushima, E.（2000）。使用惰性氟化气体通过核磁共振成像观察通气受阻。《应用生理学杂志》（贝塞斯达，马里兰州：1985）88，2279-86。[31] Lam-Weil, J., Laurent, B. 和 Loubes, J.-M.（2022）。局部差分隐私约束下密度与多项分布拟合优度检验的极小极大最优方法。《伯努利》28，579–600。[32] Lehmann, E. L.（编）（1999）。《大样本理论要素》。施普林格纽约，纽约州纽约市。[33] Li, X. 和 Ding, P.（2017）。有限总体中心极限定理的一般形式及其在因果推断中的应用。《美国统计学会会刊》112，1759-1769。[34] Marsaglia, G.（2006）。《正态变量之比》。《统计软件杂志》16，1–10。[35] Pfeffermann, D.（1993）。抽样权重在调查数据建模中的作用。《国际统计评论》61，317–37。[36]Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J. 和 Sobek, M.（2018）。IPUMS USA：版本8.0，美国人口普查局披露规避研究所用的1940年人口普查数据提取 [数据集]。明尼阿波利斯，明尼苏达州：IPUMS。[37] Smith, A. D.（2009）。渐近最优且保护隐私的统计估计。载于《密码学与网络安全》，第8届国际会议 CANS 2009，日本金泽，2009年12月12日至14日。会议论文集（J. A. Garay, A. Miyaji 和 A. Otsuka 编）。计算机科学讲义5888，53–57。Springer。[38] Team, A. D. P.（2017）。《大规模隐私保护学习》。[39]Thompson, M. E.（1997）。《抽样调查理论》。《统计学与应用概率专著》。施普林格美国。[40]van Erven, T. 和 Harremos, P.（2014）。Rényi散度与Kullback-Leibler散度。《IEEE信息论汇刊》60，3797-3820。[41] Wang, Y., Kifer, D. 和 Lee, J.（2019）。经验风险最小化的差分隐私置信区间。《隐私与保密杂志》9。[42] Wasserman, L. 和 Zhou, S.（2010）。差分隐私的统计框架。《美国统计学会会刊》105，375-389。