Received: 15 September 2023 | Revised: 11 November 2023 | Accepted: 16 January 2024

DOI: 10.1002/sta4.658

ORIGINAL ARTICLE

Highly private large-sample tests for contingency tables

Sungkyu Jung | Seung Woo Kwak

Department of Statistics, Seoul National University, Seoul, Republic of Korea

Correspondence
Sungkyu Jung, Department of Statistics, Seoul National University, 25-320, 1, Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea.
Email: sungkyu@snu.ac.kr

Funding information
This work was partially supported by the NRF Grants RS-2023-00218231, RS-2023-00275839 and RS-2023-00301976 and IITP Grant 2022-0-00937, funded by the Korean government (MSIT and MOE).

Abstract

Differential privacy is a foundational concept for safeguarding sensitive individual information when releasing data or statistical analysis results. In this study, we concentrate on the protection of privacy in the context of goodness-of-fit (GOF) and independence tests, utilizing perturbed contingency tables that adhere to Gaussian differential privacy within the high-privacy regime, where the degrees of privacy protection increase as the sample size increases. We introduce private test procedures for GOF, independence of two variables and the equality of proportions in paired samples, similar to McNemar's test. For each of these hypothesis testing situations, we propose private test statistics based on the χ^2 statistics and establish their asymptotic null distributions. We numerically confirm that Type I error rates of the proposed private test procedures are well controlled and have adequate power for larger sample sizes and effect sizes. The proposal is demonstrated in private inferences based on the American Time Use Survey data.

KEYWORDS
differential privacy, Gaussian differential privacy, goodness-of-fit test, independence test

1 | INTRODUCTION

Official administrative data are frequently distributed in the form of contingency tables, as they are a practical and effective means of summarizing, analysing and communicating categorical data and providing relationships among categorical variables. These relationships are often identified and confirmed through hypothesis tests. However, there have been concerns regarding whether the tables used for these tests have the risk of violating privacy, as they may contain sensitive private information. When releasing data or results from statistical analyses, data users and publishers often believed that anonymization, which removes personal information, would be sufficient to protect privacy. However, several studies(Lin et al., 2004; Samarati & Sweeney, 1998; Sweeney, 2002) have shown that this is not the case due to auxiliary information. To address this issue, differential privacy (DP) has gained attention across various fields dealing with private sensitive individual data, as it quantifies the risk of privacy violations formally and mathematically (Dwork, 2006).

In the DP framework, each individual's sensitive information is often protected by adding a random perturbation to the data (Bun & Steinke, 2016; Dwork & Roth, 2014; Dong et al., 2022; Mironov, 2017). Particularly for contingency tables, the cell counts are perturbed by adding random noises, while the distribution and variance of the noise are determined by the prespecified level of privacy protection (Dolera & Favaro, 2021; Gaboardi & Rogers, 2018; Gaboardi et al., 2016; Johnson & Shmatikov, 2013; Kim et al., 2023; Rogers & Kifer, 2017; Rinott et al., 2018; Smith, 2011; Sheffet, 2018; Son et al., 2022; Uhler et al., 2013; Vu & Slavkovic, 2009; Wang et al., 2015). To be specific, write $\mathbf{x}=(x_{ij})$ for the original $r\times c$ contingency table. By adding independent standard normal random noises z_{ij} to each cell count, the perturbed table is $\mathbf{u}=(u_{ij})$, $u_{ij}=x_{ij}+z_{ij}$. This table $\mathbf{u}$ and any further analysis based on $\mathbf{u}$ are said to satisfy '$\sqrt{2}$-GDP', where the term 'GDP' stands for a particular measure of DP and $\sqrt{2}$ is the level of privacy protection; see Section 1.1, in which we provide necessary background on DP.

In this paper, we investigate large-sample test procedures for randomly perturbed contingency tables, specifically when substantial perturbations are applied to achieve higher levels of privacy protection. We introduce novel *differentially private* procedures for conducting goodness-of-fit

(GOF) and independence (homogeneity) tests on general $r \times c$ contingency tables, along with a private test for equality of proportions in 2×2 tables. Conventional test procedures for unperturbed tables, such as the well-known χ^2 tests and McNemar's test (McNemar, 1947), serve as the basis for our proposal. Unlike some of the previous attempts at differentially private contingency table analyses (Johnson & Shmatikov, 2013; Smith, 2011; Vu & Slavkovic, 2009; Uhler et al., 2013) that assume asymptotically negligible perturbation, which limits their applicability in finite-sample-size scenarios, we take a different approach. We explicitly model the amount of perturbation to increase proportionally with the sample size n, ensuring that the noise neither overwhelms the original data nor becomes negligible. Our research reveals that the optimal balance is achieved when the noises z_{ij} follow a normal distribution with a standard deviation proportional to $\sqrt{n}$, striking the perfect trade-off between privacy protection and accurate analysis.

In each hypothesis test scenario, namely, GOF, independence and equality of proportions, we establish differentially private test statistics represented as χ^2 statistics. These test statistics solely rely on the perturbed cell counts in $\mathbf{u}$. Additionally, we derive their asymptotic sampling distributions under the corresponding null hypothesis. Since we assumed a high-privacy regime (in which the privacy protection becomes stronger as n increases), the private test statistics asymptotically follow a weighted sum of χ^2 distributions each with one degree of freedom. Although this null distribution may involve unknown true parameters, we conduct empirical analysis demonstrating that substituting the parameters with natural sample counterparts yields satisfactory results, even with moderate sample sizes. This approach ensures the validity and practicality of our proposed differentially private test procedures.

It is important to note that our study adopts a specific modelling approach for perturbing each cell count, using the normal distribution commonly known as the *Gaussian mechanism* in the DP literature. While the amount of noise can be calibrated to satisfy specific levels of privacy, adhering to normal distribution enables us to derive the asymptotic null distribution, even when the perturbation is non-negligible. This is in contrast to previous approaches of DP contingency table analysis (Dolera & Favaro, 2021; Gaboardi & Rogers, 2018; Gaboardi et al., 2016; Kim et al., 2023; Johnson & Shmatikov, 2013; Rogers & Kifer, 2017; Sheffet, 2018; Rinott et al., 2018; Smith, 2011; Son et al., 2022; Uhler et al., 2013; Vu & Slavkovic, 2009; Wang et al., 2015), which use Laplace or truncated exponential mechanisms, as well as Gaussian mechanisms. Those approaches are thus limited to negligible noises (Johnson & Shmatikov, 2013; Smith, 2011; Uhler et al., 2013; Vu & Slavkovic, 2009) or rely on resampling-based test procedures (Gaboardi & Rogers, 2018; Kim et al., 2023; Rogers & Kifer, 2017; Son et al., 2022; Wang et al., 2015). Our use of the Gaussian mechanism allows for a more flexible and accurate analysis, especially in scenarios where the perturbation is significant.

The remaining sections of the paper are structured as follows. Section 1.1 provides the necessary background on DP. In Section 2, we propose test procedures of differentially private GOF and independence tests for $r \times c$ contingency tables and differentially private test of equality of proportions in paired samples. In Section 3, through simulation studies, we confirm that the type I error rates of the proposed tests are controlled at the specified significance level, and the power increases under various alternative situations. We also demonstrate their practical application and effectiveness in real-world situations using American Time Use Survey (ATUS) data in Section 3.3. Section 4 presents a brief discussion summarizing the key findings of our research, potential implications and suggestions for further exploration. Technical details and proofs are contained in Appendix A.

1.1 | DP and related notions

DP is defined to protect individual privacy by disguising the possible change of an algorithm's output caused by changing one input value. Let $\mathcal{X}$ be a data collection. When two datasets $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ differ in one observation, we say that $\mathbf{x}$ and $\mathbf{y}$ are neighbouring datasets and denote as $\mathbf{x} \sim \mathbf{y}$.

A randomized algorithm $\mathcal{M} : \mathcal{X} \to \mathbb{R}^d$ is (ϵ, δ)-DP if for all $S \subseteq \text{Range}(\mathcal{M})$ and for all neighbouring datasets $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, $Pr[\mathcal{M}(\mathbf{x}) \in S] \leq e^{\epsilon} Pr[\mathcal{M}(\mathbf{y}) \in S] + \delta$ holds (Dwork et al., 2006). If $\delta = 0$, $\mathcal{M}$ is ϵ-differentially private (Dwork, 2006).

This definition of DP provides us with means to measure the privacy guarantees of algorithms, but the interpretation of ϵ is not easy. By viewing the DP framework as hypothesis testing, the interpretation of DP becomes easier. The hypotheses related to this view of DP are whether a random variable M, an output of randomized algorithm $\mathcal{M}$, follows a distribution P or Q, where $M = \mathcal{M}(\mathbf{x}) \sim P$ and $M = \mathcal{M}(\mathbf{y}) \sim Q$. That is, the null hypothesis is $H_0 : M \sim P$ (or the dataset at hand is $\mathbf{x}$), and the alternative hypothesis is $H_1 : M \sim Q$ (or the dataset at hand is $\mathbf{y}$). The Gaussian differential privacy (GDP) (Dong et al., 2022) is based on this point of view, and the difficulty (or the type I and II errors) of the most powerful test for the hypotheses are measured in terms of two Gaussian distributions. To measure the privacy guarantees of algorithms with respect to the GDP, a trade-off function is used. For a test procedure $\phi : M \to [0,1]$ for the hypotheses, let $\alpha_\phi = \mathrm{E}_{M \sim P}[\phi(M)]$ and $\beta_\phi = 1 - \mathrm{E}_{M \sim Q}[\phi(M)]$, which stand for the type I and II error rates of ϕ, respectively. The trade-off function $T(P,Q) : [0,1] \to [0,1]$ between two distribution P and Q in the hypotheses is defined as $T(P,Q)(\alpha) = \inf\{\beta_\phi : \alpha_\phi \leq \alpha\}$. When the P and Q are $N(0,1)$ and $N(\mu,1)$ for some $\mu > 0$, the trade-off function is $G_\mu := T(N(0,1), N(\mu,1))$ and $G_\mu(\alpha) = \Phi(\Phi^{-1}(1-\alpha) - \mu)$ where Φ is the standard normal cumulative distribution function. In the context of neighbouring datasets $\mathbf{x}$ and $\mathbf{y}$, if a privacy mechanism $\mathcal{M}$ has a trade-off function that exceeds the trade-off function G_μ, then $\mathcal{M}$ is referred to as μ-GDP. It is important to note that a smaller value of $\mu > 0$ corresponds to a higher level of privacy since $P \sim N(0,1)$ and $Q \sim N(\mu,1)$ are more difficult to discern when μ is small.

We primarily focus on an additive mechanism that adds random noise to the given statistic. The mechanism can be expressed as

$$\mathcal{M}(\mathbf{x}) = f(\mathbf{x}) + \mathbf{z}, \tag{1}$$

where $f : \mathcal{X} \to \mathbb{R}^d$ represents a function summarizing the data or estimator of a particular parameter. A simple Gaussian mechanism uses $\mathbf{z} \sim N(0, \sigma^2 \mathbf{I}_d)$, where $\mathbf{I}_d$ is the d-dimensional identity matrix. For any privacy level $\mu > 0$ and an algorithm f, the μ-GDP satisfying mechanism can be tightly obtained by calibrating the variance σ^2. For this purpose, one needs to measure the ℓ_2 sensitivity $\Delta_f = \max_{\mathbf{x} \sim \mathbf{y} \in \mathcal{X}} \|f(\mathbf{x}) - f(\mathbf{y})\|_2$, which is the maximum ℓ_2 difference of the algorithm output $f(\mathbf{x})$ over exchanging one observation in $\mathbf{x} \in \mathcal{X}$.

Theorem 1 Gaussian mechanism, Dong et al. (2022). The mechanism $\mathcal{M}$ in (1) satisfies μ-GDP if $\mathbf{z} \sim N\left(0, \sigma^2_{\Delta_f,\mu} \mathbf{I}_d\right)$ where $\sigma_{\Delta_f,\mu} = \Delta_f / \mu$ and Δ_f is ℓ_2 sensitivity of f.

Let f be a function that outputs a d-category contingency table. The ℓ_2 sensitivity $\Delta = \Delta_f$ of f is $\sqrt{2}$ (Kim et al., 2023). This is because contingency tables from two neighbouring sets, differing only by one observation, can have at most one count difference in exactly two categories in the contingency table. For any $\mu > 0$, setting $\sigma^2 = \Delta^2/\mu^2 = 2/\mu^2$ in the Gaussian mechanism (1) leads that the resulting perturbed contingency table satisfies μ-GDP. Conversely, for any variance $\sigma_0^2 > 0$, the resulting Gaussian mechanism is μ_0-GDP for $\mu_0 = \sigma_0/\Delta$. This seamless translation between the variance and the privacy parameter μ is difficult when (ϵ, δ)-DP is used. In particular, for a given variance $\sigma_0^2 > 0$, the mechanism is (ϵ, δ)-DP for infinitely many pairs of (ϵ, δ), as long as they satisfy $\sigma_0 = \Delta\sqrt{2\log(1.25/\delta)}/\epsilon$. (Typically, δ is prespecified to be smaller than $1/n$.) In this work, we choose to work with the GDP framework for its simplicity. We remark that it is the Gaussian perturbation that makes the use of GDP more preferable. However, as it will be clearer in the next section, the calculation of the null distributions for the private tests we propose is also made possible by assuming the Gaussian perturbation.

2 | DIFFERENTIALLY PRIVATE TESTS FOR CONTINGENCY TABLES

In this section, we define test statistics based on perturbed data and derive their asymptotic null distributions for each hypothesis test scenario. The original unperturbed $r \times c$ contingency table is denoted by $\mathbf{x}$, consisting of cell counts corresponding to each category. Let n denote the total cell count or the sample size. An application of ordinary Gaussian mechanism leads to an additively perturbed contingency table $\mathbf{u} = \mathbf{x} + \mathbf{z}$, where each element of $\mathbf{z}$ independently follows $N(0, \sigma_n^2)$. The amount of perturbation, or the standard deviation $\sigma_n = \sigma(\mu, \Delta) = \Delta/\mu$, depends on the specified level μ of the GDP framework and $\Delta = \sqrt{2}$, the ℓ_2 sensitivity of the contingency table.

In this work, we assume that the privacy parameter μ depends on the sample size n. Specifically, we set $\mu = \mu_n = \mu_0/\sqrt{n}$ for a constant $\mu_0 > 0$. This modelling choice allows us to achieve higher privacy protection for datasets with larger sample sizes, as a smaller value of μ corresponds to stronger privacy protection. As a consequence, the variance σ_n^2 of the Gaussian mechanism is also influenced by the sample size and can be expressed as follows:

$$\sigma_n^2 = \frac{\Delta^2}{\mu^2} = \frac{2}{\mu_0^2} n.$$

For simplicity, we denote $\sigma^2 := 2/\mu_0^2$.

Additionally, we assume that both the sample size n and σ (or equivalently μ_0, or μ) are publicly known information. This approach aligns with the methodology employed by the U.S. Census Bureau in their disclosure of the 2020 Census Redistricting File, wherein certain parameters are regarded as publicly accessible information (U.S. Census Bureau, 2021).

2.1 | Differentially private GOF test

To formally define our private GOF test procedure, we model the unperturbed contingency table $\mathbf{x}$ to follow a multinomial distribution. Treating $\mathbf{x} = (x_1, \ldots, x_d)$ as a random vector with d entries[1] (for $d \geq 2$), we model $\mathbf{x} \sim M_d(n, \mathbf{p})$, where M_d stands for the d-dimensional multinomial distribution, and $\mathbf{p} = (p_1, \ldots, p_d) \in \mathcal{S}_d$ is the unknown parameters representing the population proportions. Here, $\mathcal{S}_d := \left\{\mathbf{p} \in \mathbb{R}^d : \mathbf{p}^\top \mathbf{1}_d = \sum_{i=1}^d p_i = 1, p_i > 0, \forall i = 1, \ldots, d\right\}$.

We assume that one observes a perturbed contingency table $\mathbf{u} = \mathbf{x} + \mathbf{z}$, where $\mathbf{z} \sim N_d(\mathbf{0}, \sigma_n^2 \mathbf{I}_d)$, but not the original table $\mathbf{x}$. Based on the single (perturbed) observation $\mathbf{u}$, and further assuming that σ and n is known, a test of GOF aims to compare the null hypothesis $H_0 : \mathbf{p} = \mathbf{p}_0$ against general alternative $H_1 : \mathbf{p} \neq \mathbf{p}_0$, for prespecified null probabilities $\mathbf{p}_0 = (p_1^0, \ldots, p_d^0) \in \mathcal{S}_d$. A natural test statistic for the private GOF test is obtained by treating the perturbed table $\mathbf{u}$ as if it were an unperturbed table $\mathbf{x}$ and is defined as the 'χ^2-statistic'

$$T_1(\mathbf{u};\mathbf{p}_0) := \sum_{i=1}^{d} \frac{\left(u_i - np_i^0\right)^2}{np_i^0}. \tag{2}$$

We note that this form of private statistics has also been used in Wang et al. (2015) and Gaboardi et al. (2016). If there were no perturbation, in which case $\mathbf{u} = \mathbf{x} \sim M_d(n, \mathbf{p}_0)$, it is well-known that the distribution of $T_1(\mathbf{u};\mathbf{p}_0)$ converges to the χ^2_{d-1} distribution, as $n \to \infty$. We show in Theorem 2 that the asymptotic distribution of the statistic (2) is no longer χ^2_{d-1} but is a weighted sum of χ^2_1 distributions. For a d-vector $\mathbf{p} = (p_1, \ldots, p_d)$, define $\sqrt{\mathbf{p}} := (\sqrt{p_1}, \ldots, \sqrt{p_d})$, and write $D_{\mathbf{p}}$ for the diagonal matrix whose ith diagonal element is p_i. For a $d \times d$ symmetric matrix Σ, $\lambda_i(\Sigma)$ stands for the ith largest eigenvalue of Σ.

> **Theorem 2.** Let $\mathbf{p} \in \mathcal{S}_d$ and $\sigma > 0$ be given. For each $n = 1, 2, \ldots$, define the random vector $\mathbf{u} := \mathbf{x} + \mathbf{z}$, where $\mathbf{x} \sim M_d(n, \mathbf{p})$ and $\mathbf{z} \sim N_d(\mathbf{0}, \sigma_n^2 \mathbf{I}_d)$, $\sigma_n^2 = \sigma^2 n$, are independent. Then, as $n \to \infty$, $T_1(\mathbf{u}, \mathbf{p})$ converges in distribution to $\sum_{i=1}^{d} \lambda_i(\Sigma_1) Z_i^2$, where Z_i^2s are independent and each follows χ^2_1 distribution, and $\Sigma_1 = \mathbf{I}_d - \sqrt{\mathbf{p}}\sqrt{\mathbf{p}}^\top + \sigma^2 D_{\mathbf{p}}^{-1}$.

Thanks to Theorem 2, the asymptotic null distribution of the test statistic (2) is specified exactly, as the matrix $\Sigma_1 = \mathbf{I}_d - \sqrt{\mathbf{p}_0}\sqrt{\mathbf{p}_0}^\top + \sigma^2 D_{\mathbf{p}_0}^{-1}$ is given by the null proportions $\mathbf{p}_0$. For an observed test statistic T_1, the (asymptotic) p value of the GOF test is $p(T_1) = P\left(\sum_{i=1}^{d} \lambda_i(\Sigma_l) Z_i^2 > T_1\right)$. At significance level $\alpha \in (0,1)$, the null hypothesis is rejected if $p(T_1) \leq \alpha$. The computation of the p value $p(T_1)$ requires an evaluation of the cdf of the scale mixture of chi-squared random variables $\sum_{i=1}^{d} \lambda_i(\Sigma) Z_i^2$. Such a task is well studied in the literature; see (Davies, 1973, 1980; Liu et al., 2009). We used the `R` package `mgcv` (Wood, 2023) in our numerical studies.

We remark that the convergence of the distribution of $T_1(\mathbf{u};\mathbf{p})$ to $\sum_{i=1}^{d} T_1(\Sigma_1) Z_i^2$ heavily relies on the convergence of the scaled and shifted multinomial random vector $\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}) := \sqrt{n}(\frac{\mathbf{x}}{n} - \mathbf{p})$ to the normal distribution $N_d(\mathbf{0}, \Sigma_0)$ where $\Sigma_0 = D_{\mathbf{p}} - \mathbf{p}\mathbf{p}^\top$. This convergence is also crucial for the usual, non-private χ^2-statistic to converge to the χ^2 distribution. The χ^2 approximation of the distribution of the non-private test statistic is known to be appropriate when each np_i (the expected cell count) is at least five. Since the random perturbation $\mathbf{z}$ is independent of $\mathbf{x}$, it contributes to the statistic $T_1(\mathbf{u}, \mathbf{p})$ as a weighted sum of χ^2 distributions, even for finite-sample sizes. This characteristic allows for the inclusion of random perturbations while preserving the convergence properties, resulting in the fact that the asymptotic null distribution is quite accurate even for small sample sizes. We note that private GOF test procedures proposed in Gaboardi et al. (2016) and Rogers and Kifer (2017) are equivalent to our proposal (despite the seemingly different formulations). We emphasize that a rigorous theoretical justification of the null distribution is first established in Theorem 2.

In many practical applications of the proposed GOF test procedure, the level μ of privacy protection or the variance σ_n^2 may be fixed and not varied as the sample size n changes from a dataset to another. However, one can still apply the test procedure by specifying σ^2 to be σ_n^2/n. This pragmatic approach allows for more accurate control of type I error, thereby avoiding the false assumption of negligible perturbation (e.g., $\sigma_n^2 = O(1)$) and leading to more accurate and reliable results in practice.

2.2 | Differentially private independence tests

We now consider the situation where two categorical random variables Y_1 and Y_2 are observed simultaneously, where Y_1 has r categories and Y_2 has c categories. The pairs of (Y_1, Y_2) observed over n subjects can be summarized in the typical $r \times c$ contingency table, arranged in a $r \times c$ matrix $\mathbb{X}$, whose (i,j)th element x_{ij} is the count of subjects belonging to the ith category for Y_1 and the jth category for Y_2. Deploying the typical vec() operation (stacking the columns), we write

$$\mathbf{x} = \mathrm{vec}(\mathbb{X}) \sim M_{rc}(n, \boldsymbol{\pi}),$$

for some $\boldsymbol{\pi} \in \mathcal{S}_{rc}$. The elements of $\mathbf{x} = (x_{ij})$ are indexed by $i = 1, \ldots, r$ and $j = 1, \ldots, c$.

Now, suppose the $r \times c$ contingency table is released privately. In particular, let $\mathbb{Z}$ represent the $r \times c$ noise matrix to be added, which is independent of the matrix $\mathbb{X}$ and its (i,j)th element Z_{ij} independently follows $N(0, \sigma_n^2)$. The privatized table $\mathbb{U} := \mathbb{X} + \mathbb{Z}$ is released. The null and alternative hypotheses of the differentially private independence test are

$$H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}^{(2)} \otimes \boldsymbol{\pi}^{(1)} \text{ vs } H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}^{(2)} \otimes \boldsymbol{\pi}^{(1)},$$

where $\boldsymbol{\pi}$ is the cell probabilities of the given contingency table, for some $\boldsymbol{\pi}^{(2)} \in \mathcal{S}_c$ and $\boldsymbol{\pi}^{(1)} \in \mathcal{S}_r$, and $\otimes$ represents the Kronecker product. $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$ are row and column marginal probabilities, respectively.

We begin with various ways of defining the estimator of cell frequencies under the null hypothesis. Naive estimators of $\pi_i^{(1)}$ for $i=1,\ldots,r$ and $\pi_j^{(2)}$ for $j=1,\ldots,c$ are

$$\tilde{\pi}_i^{(1)} = \frac{\sum_{j=1}^{c} u_{ij}}{n}, \text{ and } \tilde{\pi}_j^{(2)} = \frac{\sum_{i=1}^{r} u_{ij}}{n}.$$

A drawback of this naive approach is that $\tilde{\boldsymbol{\pi}}^{(1)} = (\tilde{\pi}_i^{(1)})_{i=1,\ldots,r}$ does not lie in $\mathcal{S}_r$. In particular, the requirement that the sum of the proportion to one, that is, $\mathbf{1}_r^\top \tilde{\boldsymbol{\pi}}^{(1)} = 1$ and $\mathbf{1}_c^\top \tilde{\boldsymbol{\pi}}^{(2)} = 1$, is violated. Alternatively, we consider the following estimators. The first estimator is based on the perturbed total counts, $n^U = \mathbf{1}_{rc}^\top \mathbf{u}$, and is

$$\hat{\boldsymbol{\pi}}^{(1)U} = \frac{\mathbb{U}\mathbf{1}_c}{n^U} \text{ and } \hat{\boldsymbol{\pi}}^{(2)U} = \frac{\mathbb{U}^\top \mathbf{1}_r}{n^U}. \tag{3}$$

The estimators in (3) have been also used in Rogers and Kifer (2017), in which the (limiting) null distribution of the test statistic based on these estimators is *misspecified* as a χ^2 distribution; see Theorem 3. Alternatively, one can define

$$\hat{\boldsymbol{\pi}}^{(1)G} := \operatorname{argmin}_{\mathbf{a} \in \mathbb{R}^r, \mathbf{1}_r^\top \mathbf{a} = 1} \|\tilde{\boldsymbol{\pi}}^{(1)} - \mathbf{a}\|_2^2 = \tilde{\boldsymbol{\pi}}^{(1)} - \frac{\mathbf{1}_r^\top \tilde{\boldsymbol{\pi}}^{(1)} - 1}{r}\mathbf{1}_r = \tilde{\boldsymbol{\pi}}^{(1)} - \frac{z_{..}}{nr}\mathbf{1}_r \text{ and } \hat{\boldsymbol{\pi}}^{(2)G} = \tilde{\boldsymbol{\pi}}^{(2)} - (nc)^{-1} z_{..}\mathbf{1}_c. \tag{4}$$

These estimators are given by orthogonally projecting the raw probabilities $\tilde{\boldsymbol{\pi}}^{(1)}$ onto the affine hyperplane defined by $\mathbf{1}_r^\top \tilde{\boldsymbol{\pi}}^{(1)} = 1$. Note that Gaboardi et al. (2016) also considered a projection-based estimator, but has utilized the L_2-distance minimizing projection onto $\overline{\mathcal{S}}_r$. Our approach is simpler and thus allows for an application of the central limit theorem and the delta method.

Based on each of the estimators, we define two versions of chi-squared statistics: For $\hat{\boldsymbol{\pi}}^U = \hat{\boldsymbol{\pi}}^{(2)U} \otimes \hat{\boldsymbol{\pi}}^{(1)U}$ and $\hat{\boldsymbol{\pi}}^G = \hat{\boldsymbol{\pi}}^{(2)G} \otimes \hat{\boldsymbol{\pi}}^{(1)G}$, the test statistics are

$$\chi_n^2(\mathbb{U}) := \sum_{i,j} \frac{\left(u_{ij} - n\hat{\pi}_{ij}^U\right)^2}{n\hat{\pi}_{ij}^U} \text{ and} \tag{5}$$

$$\chi_G^2(\mathbb{U}) := \sum_{i,j} \frac{\left(u_{ij} - n\hat{\pi}_{ij}^G\right)^2}{n\hat{\pi}_{ij}^G}. \tag{6}$$

Theorem 3. Let $\hat{\boldsymbol{\pi}}^{(1)U}$, $\hat{\boldsymbol{\pi}}^{(2)U}$, $\hat{\boldsymbol{\pi}}^{(1)G}$ and $\hat{\boldsymbol{\pi}}^{(2)G}$ be defined in (4) and (3), and $\sigma > 0$ be given. For each $n = 1,2,\ldots$, define the random vector $\mathbf{u} = \mathbf{x} + \mathbf{z}$ where $\mathbf{x} \sim M_{rc}(n, \boldsymbol{\pi})$ and $\mathbf{z} \sim N\left(\mathbf{0}, \sigma_n^2 \mathbf{I}_{rc}\right)$, $\sigma_n^2 = n\sigma^2$, are independent. Then, under the null hypothesis that Y_1 and Y_2 are independent when $Y_1 \sim M_r\left(n, \boldsymbol{\pi}^{(1)}\right)$, $Y_2 \sim M_c\left(n, \boldsymbol{\pi}^{(2)}\right)$ (i.e., $\boldsymbol{\pi} = \boldsymbol{\pi}^{(2)} \otimes \boldsymbol{\pi}^{(1)}$ for some $\boldsymbol{\pi}^{(2)} \in \mathcal{S}_c$ and $\boldsymbol{\pi}^{(1)} \in \mathcal{S}_r$),

$$\chi_n^2(\mathbb{U}) \Rightarrow \sum_{i=1}^{rc} \lambda_i(\Sigma_U) Z_i^2 \text{ and} \tag{7}$$

$$\chi_G^2(\mathbb{U}) \Rightarrow \sum_{i=1}^{rc} \lambda_i(\Sigma_G) Z_i^2, \tag{8}$$

as $n \to \infty$, where

$$\Sigma_U = \mathbf{I}_{rc} - \sqrt{\boldsymbol{\pi}}\sqrt{\boldsymbol{\pi}}^\top - \mathbf{A}\left(\mathbf{A}^\top \mathbf{A}\right)^{-1}\mathbf{A}^\top + D_{\boldsymbol{\pi}}^{-1/2} \Sigma_{\sigma,\boldsymbol{\pi}}^U D_{\boldsymbol{\pi}}^{-1/2} \text{ and} \tag{9}$$

$$\Sigma_G = \mathbf{I}_{rc} - \sqrt{\boldsymbol{\pi}}\sqrt{\boldsymbol{\pi}}^\top - \mathbf{A}\left(\mathbf{A}^\top \mathbf{A}\right)^{-1}\mathbf{A}^\top + D_{\boldsymbol{\pi}}^{-1/2} \Sigma_{\sigma,\boldsymbol{\pi}}^G D_{\boldsymbol{\pi}}^{-1/2}. \tag{10}$$

Here, $\mathbf{A} = D_{\boldsymbol{\pi}}^{-\frac{1}{2}}\nabla$, and ∇ is the $rc \times (r+c-2)$ matrix consisting of partial derivatives of $\boldsymbol{\pi}$ with respect to $\pi_i^{(1)}$, $i = 1,\ldots,r-1$ and $\pi_j^{(2)}$, $j = 1,\ldots,c-1$, and is

$$\nabla = \left[\boldsymbol{\pi}^{(2)} \otimes \begin{pmatrix} \mathbf{I}_{r-1} \\ -\mathbf{1}_{r-1}^{\top} \end{pmatrix}, \begin{pmatrix} \mathbf{I}_{c-1} \\ -\mathbf{1}_{c-1}^{\top} \end{pmatrix} \otimes \boldsymbol{\pi}^{(1)}\right].$$

The $rc \times rc$ matrices $\Sigma_{\sigma,\boldsymbol{\pi}}^{U}$ and $\Sigma_{\sigma,\boldsymbol{\pi}}^{G}$ in (9) and (10) depend on $\boldsymbol{\pi}^{(1)}$, $\boldsymbol{\pi}^{(2)}$ and σ and the exact expression are given in Appendix A.0.3.

Since $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$ are unknown, we can obtain the estimates $\hat{\Sigma}_U$ and $\hat{\Sigma}_G$ by replacing $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$ in Σ_U and Σ_G with $\hat{\boldsymbol{\pi}}^{(1)U}$ and $\hat{\boldsymbol{\pi}}^{(2)U}$ and $\hat{\boldsymbol{\pi}}^{(1)G}$ and $\hat{\boldsymbol{\pi}}^{(2)G}$, respectively. For observed test statistics $\chi_n^2(\mathbb{U})$ and $\chi_G^2(\mathbb{U})$, the p values of the independence tests are $p(\chi_n^2) = P\left(\sum_{i=1}^{rc}\lambda_i(\Sigma_U)Z_i^2 > \chi_n^2(\mathbb{U})\right)$ and $p(\chi_G^2) = P\left(\sum_{i=1}^{rc}\lambda_i(\Sigma_G)Z_i^2 > \chi_G^2(\mathbb{U})\right)$, respectively. We reject the null hypothesis at level α when $p(\chi_n^2) < \alpha$ and $p(\chi_G^2) < \alpha$, respectively.

Since the magnitude of the added noise can sometimes exceed the observed counts of a cell or the sample size n, the perturbed counts or the perturbed total sample size can be negative values. This is more likely to occur under the high-privacy regime with small sample sizes. Since the estimates of marginal probabilities are calculated based on these perturbed counts, negative marginal probabilities can be obtained, but these values cannot be used for the test. To resolve this issue, the negative estimates are replaced with small probabilities. This is reasonable since the negative estimates arise when the marginal values are not sufficiently large to outweigh the added noise.

2.3 | Test of equality of proportions in paired samples

In this subsection, we consider testing for equality of proportions in paired samples, based on a private contingency table. Let a 2×2 table represent the binary responses of two questions where x_{ij} and π_{ij} stand for the counts and probabilities of the (i,j)th cell, respectively. The totals of the ith row and the jth column in the given table are denoted by $x_{i.}$ and $x_{.j}$, correspondingly. Similarly, the marginal probabilities are presented by $\pi_{i.}$ and $\pi_{.j}$ for $i = 1,2$ and $j = 1,2$.

When $\pi_{1.}$ and $\pi_{.1}$ are similar, the same holds for π_{12} and π_{21} due to the relationships $\pi_{1.} = \pi_{11} + \pi_{12}$ and $\pi_{.1} = \pi_{11} + \pi_{21}$. If one wants to test the homogeneity of approval for the two questions, the null hypothesis is $H_0 : \pi_{i.} = \pi_{.i}$, which is equivalent to $H_0 : \pi_{12} = \pi_{21}$.

Now, suppose that the perturbed table is transformed to the vector, and the vector is expressed as $\mathbf{u} = \mathbf{x} + \mathbf{z}$ where $\mathbf{x} = (x_{12}, x_{21}, x_{11} + x_{22}) \sim \mathrm{Mult}(n, \boldsymbol{\pi})$, $\mathbf{z} = (z_{12}, z_{21}, z_{11} + z_{22})$, and $z_{ij} \sim N(0, \sigma_n^2)$. Then,

$$\frac{\mathbf{u} - n\mathbf{p}}{\sqrt{n}} \Rightarrow N\left(\mathbf{0}, \Sigma_0 + \frac{\sigma_n^2}{n}\mathbf{C}\right), \text{ where } \mathbf{C} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

The test statistic for the test of equality of proportions is defined as

$$t_n(\mathbf{u}) = \frac{u_{12} - u_{21}}{\sqrt{\hat{n}^* + 2\sigma_n^2}}, \tag{11}$$

where $\hat{n}^* = u_{12} + u_{21}$ is an estimate of n^* for the unknown π_1 under the null hypothesis.

Theorem 4. Let $\sigma > 0$ be given and $\sigma_n = \sigma\sqrt{n}$. For each $n = 1,2,\ldots$, define the random vector $\mathbf{u} = \mathbf{x} + \mathbf{z}$ where $\mathbf{x} = (x_{12}, x_{21}, x_{11} + x_{22}) \sim \mathrm{Mult}(n, \boldsymbol{\pi})$, and $z_{ij} \sim N(0, \sigma_n^2)$ given that $\boldsymbol{\pi} = (\pi_{12}, \pi_{21}, \pi_0)$ and $\mathbf{z} = (z_{12}, z_{21}, z_{11} + z_{22})$. Under the null hypothesis that $\pi_{12} = \pi_{21}$, the limiting distribution of the private test statistic $t_n(\mathbf{u})$ in (11) is $N(0,1)$.

For an observed test statistic t_n, the p value of the test of equality of proportions in paired samples is $p(t_n) = 1 - \Phi(t_n(\mathbf{u}))$. The null hypothesis is rejected at level α when $p(t_n) < \alpha$.

3 | SIMULATIONS

In this section, we investigate the type I error rate and the power of the private tests discussed in Section 2. The performances of our proposal are compared with close competitors including Gaboardi et al. (2016) and Rogers and Kifer (2017). We employ the continuous Gaussian distribution to generate the noise although one may consider using the discrete noises from the discrete Laplace distribution (Ghosh et al., 2009) or the discrete Gaussian distribution (Canonne et al., 2022) as used in Nikolov et al. (2016) and Haney et al. (2021). However, the discrete Laplace noise cannot be used for μ-GDP (Kim et al., 2023), and the results of the discrete Gaussian noise are presented in the Supporting Information. In short, the simulation results do not show a significant difference between test performances under the (continuous) Gaussian and discrete Gaussian perturbations.

The type I error rate and power of the tests depend on various factors, including the sample size n, privacy parameter μ (or the amount σ of the perturbation) and the effect size, which measures a difference between the null distribution and the alternative distribution. To measure the effect size, Cohen's ω (Cohen, 1988) is used, which is defined as

$$\omega = \sqrt{\sum_{i=1}^{d} \frac{(p_i - p_i^0)^2}{p_i^0}}$$

where p_i and p_i^0 represent the probability of the ith cell in the alternative and null distribution, respectively. Figure 1 provides examples of multinomial distributions for $d = 20$ cells and their effect sizes $\omega = 0, 0.3, 0.6$, compared to the global null distribution, putting equal probabilities to each cell.

Throughout, we will investigate the performances of the proposed private test procedures under two asymptotic scenarios:

- High-privacy regime, in which for a given σ, the variance $\sigma_n^2 = \sigma^2 n$ increases as the sample size increases, thus resulting in higher privacy ($\mu \asymp 1/\sqrt{n}$) as n increases.
- Fixed level of privacy. For a fixed level μ, the variance of the mechanism is calibrated to satisfy μ-GDP, for each sample size n.

3.1 | Type I error

Figure 2 displays the type I error rates of the private GOF test, independence test and test of equality of proportions, proposed in Section 2, under various scenarios. Overall, our proposal controls the occurrence of type I errors under the given significance level.

The outcomes are from the fixed σ scenario (upper plots in Figure 2) and the fixed μ scenario (lower plots in Figure 2). Based on the results of the fixed σ case, stronger privacy (indicated by smaller μ values) is guaranteed as the sample size increases. For instance, when the sample size is 100, μ is 0.14, while μ becomes less than 0.1 when the sample size exceeds 500. The lower plots in Figure 2 show that smaller σ is needed to achieve the same privacy guarantee μ with a larger sample size.

In Figure 2 and throughout, $\mathtt{T}_1$ denotes the results of private GOF test based on (2) while $\mathtt{Priv}_G$ is the results of the test proposed by Gaboardi et al. (2016), and $\mathtt{Unprj}$ and $\mathtt{Prj}$ are the results of the test suggested by Rogers and Kifer (2017). χ_G^2 and χ_n^2 represent the results of private independence tests (7) and (8), respectively. t_n is the result of the private test of equality of proportions in paired samples using test statistic in (11).

For the private GOF test, the test statistics of $\mathtt{T}_1$ and $\mathtt{Priv}_G$ are the same as expressed in the quadratic form of vectors of perturbed counts asymptotically following multivariate normal distribution. The limiting distributions of both test statistics rely on eigenvalues of the covariance

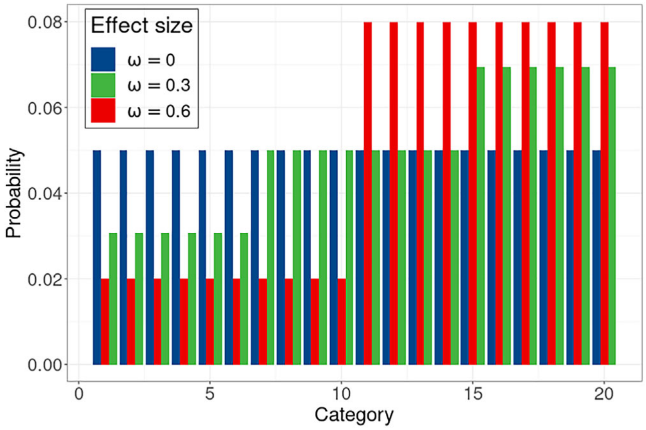

FIGURE 1 Example of the population distributions when the number of categories is 20. Blue, red and green bars indicate the multinomial distributions with effect size ω 0, 0.3 and 0.6, respectively.

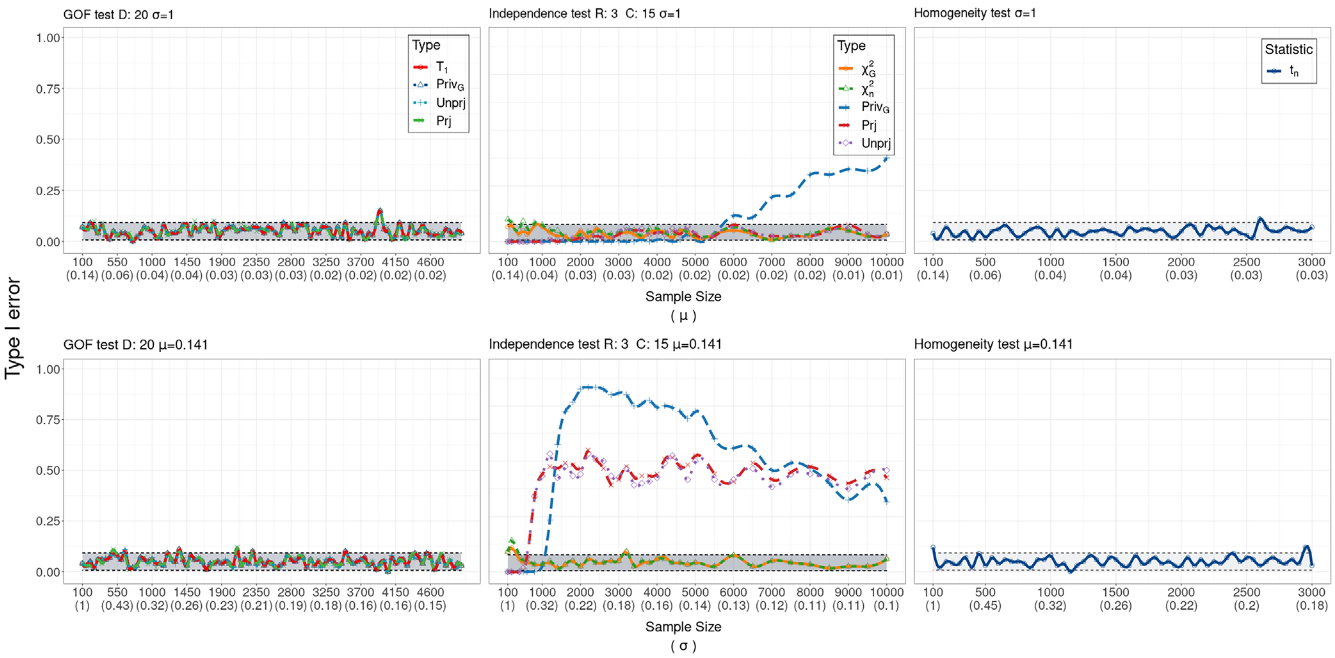

FIGURE 2 Type I error of the private tests according to the sample size increases are depicted. From the first column to the last, the results of GOF, independence and test of equality of proportions in paired samples are shown. In the first row, $\sigma = 1$, and in the second row, $\mu = 0.141$ ($\sigma_n = 10$). For the GOF test, there are 20 categories, and for the independence test, $r \times c = 3 \times 15$. The shaded band represents two standard errors of the empirical type I error.

matrices of the perturbed counts. That is, the test statistics and the limiting null distribution of $\texttt{Priv}_\texttt{G}$ and $\texttt{T}_1$ are equivalent, as noted in Section 2.1. As a result, the test results are the same in our simulation setting. $\texttt{Unprj}$ defines test statistic as $n^{-1}(\mathbf{u}-n\mathbf{p})^\top M(\mathbf{u}-n\mathbf{p})$ where M is the inverse of the covariance matrix of the asymptotic distribution of $n^{-1/2}(\mathbf{u}-n\mathbf{p})$ which is the multivariate normal distribution with mean zero and covariance matrix $D_{\sqrt{\mathbf{p}}}\Sigma_1 D_{\sqrt{\mathbf{p}}}$. $\texttt{Prj}$ projects $(\mathbf{u}-n\mathbf{p})$ by $P = \mathbf{I}_d - d^{-1}\mathbf{1}\mathbf{1}^\top$ and uses the test statistic $n^{-1}(\mathbf{u}-n\mathbf{p})^\top PMP(\mathbf{u}-n\mathbf{p})$. $\texttt{T}_1$ converges to the quadratic of multivariate normal random vector and the test statistic of $\texttt{Unprj}$ is a scaled version of $\texttt{T}_1$. Thus, the limiting distribution of the statistic $\texttt{Unprj}$ is χ^2_d, and it is the same as $\texttt{T}_1$. For a very small σ value, that is, very weak privacy protection is assumed, $\texttt{Prj}$ is suggested. Each test statistic follows χ^2_d and χ^2_{d-1} since it is known that $\mathbf{y}^\top V\mathbf{y}$ follows χ^2_k when $\mathbf{y} \sim N(\mathbf{0}, \Sigma)$, $V\Sigma$ is idempotent, and $\text{rank}(V) = k$ by Driscoll (1999). As a result, in the private GOF test, the type I error rates of these tests are the same.

Plots in the second column of Figure 2 show the type I error of two differentially private independence tests based on two test statistics $\chi^2_n(\mathbb{U})$ and $\chi^2_G(\mathbb{U})$ in Theorem 3. Other results $\texttt{Priv}_\texttt{G}$, $\texttt{Unprj}$ and $\texttt{Prj}$ are obtained from Gaboardi et al. (2016) and Rogers and Kifer (2017), respectively. The test statistics for these methods are the same with the private GOF test, except that $\mathbf{p}$ is replaced by estimates. When the sample size is small the results are not provided as they do not draw conclusions when the noisy cell count or the expected count of any cell is too small. We treat no decision as not rejecting the null hypothesis which results in small type I error rates. In Figure 2, there are zero type I error rates when the sample size is smaller than 2000. Moreover, under the fixed σ setting, the type I error of $\texttt{Priv}_\texttt{G}$ increases as the sample size increases. The $\texttt{Prj}$ and $\texttt{Unprj}$ have zero type I errors like $\texttt{Priv}_\texttt{G}$ does, but the rates get stable with more samples. However, for fixed μ scenario, the type I error rates of $\texttt{Priv}_\texttt{G}$, $\texttt{Prj}$ and $\texttt{Unprj}$ are not controlled even when the sample size is 10,000. The differences can be caused by the estimation of marginal probabilities and the estimation of the covariance matrix. For $\texttt{Priv}_\texttt{G}$ independence test, the test statistic looks similar with χ^2_G, but the guessed covariance of the limiting distribution is not the same as in Theorem 3. As a result, the type I error of the $\texttt{Priv}_\texttt{G}$ test is not controlled. Moreover, the estimator obtained from the optimization is not easy to show the characteristics under the asymptotic situation. For the private independence tests of $\texttt{Unprj}$, the test statistic is the same as that of the GOF test, except for replacing the probabilities with their estimates. However, in this case, estimated M is not the inverse of the covariance matrix of the limiting distribution of $n^{-1/2}(\mathbf{u}-n\mathbf{p})$, leading to a failure to converge to the χ^2 distribution. The same issue occurs with the test statistic of $\texttt{Prj}$. Incorrect estimation of the covariance matrix prevents convergence to the χ^2 distributions with the corresponding degrees of freedom.

In the simulation results, we observe that the type I error rates of our methods are controlled for every σ we have tested, for all four test procedures proposed unless the n is very small. In particular, the privacy independence tests, using χ^2_G and χ^2_n, require $n \geq 1000$ as they require estimation of unknown true proportions in the limiting null distribution; the other two tests (T_1 and t_n) control type I error rates for all sample sizes and all values of privacy parameters.

3.2 | Power

To assess the power of the private tests, we simulate different effect sizes. For the GOF and independence tests, we consider moderate ($\omega = 0.3$) and large ($\omega = 0.6$) effect sizes while the test of equality of proportions in paired samples uses small ($\omega = 0.1$) and moderate effect sizes are used.

The results of the private GOF test are presented in Figure 3, the independent test results are shown in Figure 4 and the test of equality of proportions in paired samples results are displayed in Figure 5. We observe that the power of these tests increases with larger sample sizes and effect sizes. However, when compared to non-private tests, it becomes evident that the private tests require a larger sample size to approach the power of the non-private counterparts.

The private GOF test results based on $\mathtt{T}_1$, $\mathtt{Priv}_G$, `Unprj` and `Prj` are presented in Figure 3. These results are indistinguishable, as the type I error results shown in Figure 2. Even in cases with a moderate effect size, differentially private GOF tests exhibit comparable power to the non-private test when the sample size is around 2000. The last plot in Figure 3 represents the fixed μ case, and it shows a similarity to the fixed σ case.

For the power of the private independence tests (in Figure 4), the higher number of cells, compared to the private GOF tests, and the influence of noise on estimated marginal probabilities cause the power to converge more slowly to the non-private tests as the sample size increases. Other test results are omitted from the figure as they do not control the type I error rate under the given σ (or μ) and sample sizes. Furthermore, there is no significant difference between χ^2_G and χ^2_n in the simulations.

For the private test of equality of proportions for paired samples (in Figure 5), the power converges to that of the non-private tests with a much smaller sample size compared to the private GOF test under the same conditions (in the second and third plots). Therefore, the power of the test, assuming a smaller effect size $\omega = 0.1$ is provided in the first plot. Note that the scale of the x-axis in the first plot is different from that in the second or last plots in the figure. The non-private test also requires a larger sample size when the effect size is small, as opposed to the moderate effect size $\omega = 0.3$ in the second and last plots in Figure 5. The private test converges to the non-private test power when the sample size is about 8000 while others achieve this at a sample size of approximately 500.

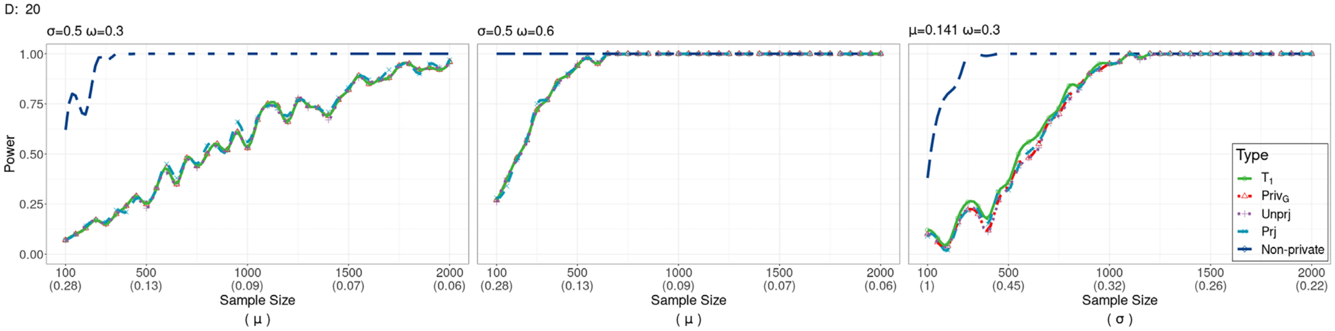

FIGURE 3 Power of the GOF test results when there are 20 categories. It illustrates the impact of increasing the sample size with $\sigma = 0.5$ (in the first and second plots) and $\mu = 0.141$ (in the last plot) with a large effect size of $\omega = 0.6$ (in the second plot) and moderate effect size of $\omega = 0.3$ (in the first and last plots)

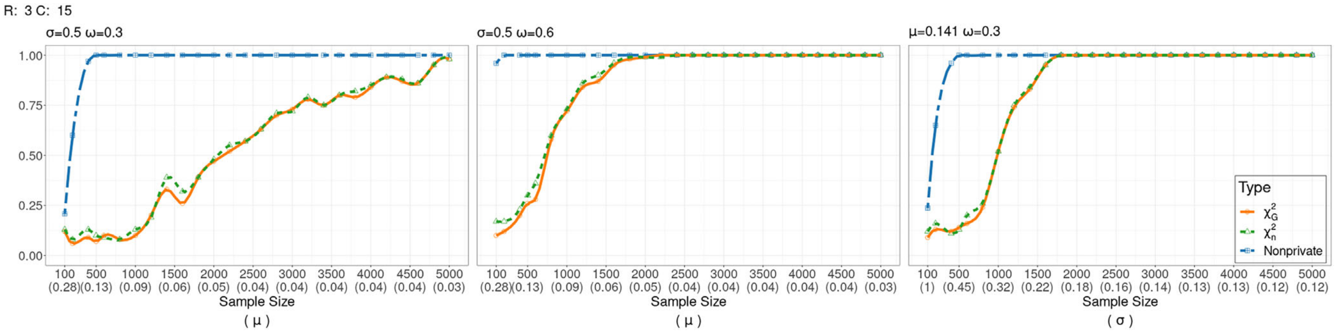

FIGURE 4 Power of the independence test under the null hypothesis of uniform distribution is displayed for a dimension of $r = 3$ and $c = 15$ as sample size increases. The plots show the power of the test with $\sigma = 0.5$ (in the first and second plots) and $\mu = 0.141$ (in the last plot), considering both a large effect size of $\omega = 0.6$ (in the second plot) and a moderate effect size of $\omega = 0.3$ (in the first and last plots)

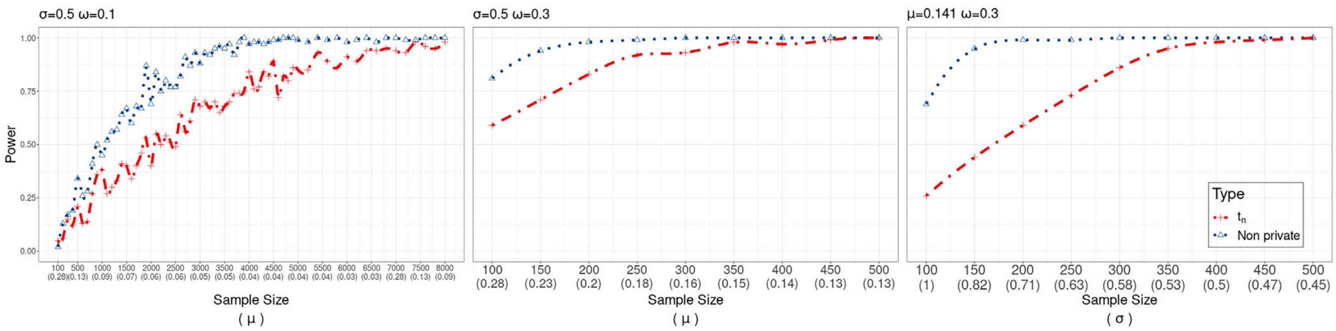

FIGURE 5 Power of the differentially private test of equality of proportions in paired samples is depicted as the sample size increases. Each plot illustrates the power of the test when $\sigma = 0.5$ (in the first and second plots) and $\mu = 0.141$ (in the last plot) with a small effect size of $\omega = 0.1$ (in the first plot) and a moderate effect size of $\omega = 0.3$ (in the second and last plots)

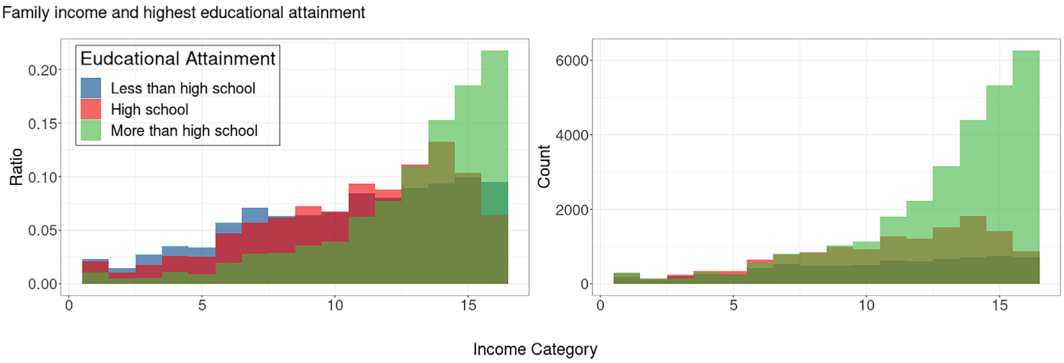

FIGURE 6 Distributions of household income in the American Time Use Survey 2021 based on the highest educational attainment. The left plot displays the ratios of each category across corresponding education levels while the right plot shows the frequencies of each category within corresponding education levels.

The simulation results indicate that the private tests require larger sample sizes to achieve the desired type I error rate and match the power of non-private tests. This need for a larger sample size is particularly evident with increased dimensions, as the total noise level rises.

Additionally, Figures 2 and 4 reveal that when the sample size is small, the private test results may lack reliability, with a low proportion of rejections regardless of whether the null hypothesis is true or not. However, as the sample size increases, the test results are similar to those of non-private tests. Determining an exact sample size for reasonable power solely based on privacy guarantees is challenging, as it depends on both known factors (the dimensions of data and privacy level) and unknown factors (effect size and population distribution).

3.3 | Real data analysis

In this section, we apply the private independence test to a real dataset from the ATUS. The dataset comprises two variables: `PEEDUCA` (highest level of school) and `HEFAMINC` (household income) from the ATUS Current Population Survey (ATUS-CPS) for the year 2021. The `HEFAMINC` variable contains 16 categories, with Category 1 representing households with less than $5000 of income and Category 16 representing households with over $150,000 of income. The `PEEDUCA` variable also has 16 categories, covering educational attainment from less than first grade to a doctoral degree. However, for the purpose of the independence test, it has been grouped into three broader categories based on diploma achievement: 'less than high school', 'high school' and 'more than high school'.

As depicted in Figure 6, the differences become more significant at higher income levels (category greater than 14) when the education level is higher. The population ratios for less than high school, high school and more than high school are 15%, 26% and 59%, respectively. The ratios serve as the basis for generating samples for each sample size in the independence test. The sample data are randomly extracted from the dataset, and the power of the test is computed based on the generated samples. Additionally, we consider test results without noise to compare the private and non-private tests. The null hypothesis is that the `HEFAMINC` is independent of `PEEDUCA`.

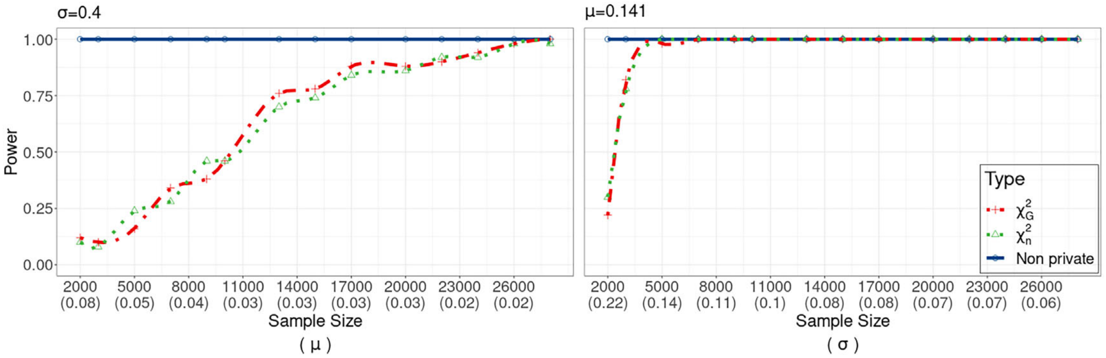

FIGURE 7 Power of the private independence tests for the household income (HEFAMINC) and educational attainment (PEEDUCA) variables in ATUS-CPS data when $\sigma = 0.4$ and $\mu = 0.141$, with varying sample sizes.

Figure 7 shows the results of applying differentially private independence tests to the ATUS dataset. The non-private test result concludes that the house income is not independent of the area. When $\sigma = 0.4$, the private test requires more than 26,000 samples to yield a similar result to the non-private test. However, for $\mu = 0.141$ (right plot in Figure 7), which uses smaller σ than 0.4 for any sample size greater than 2000, the private test results are the same as the non-private test when the sample size is greater than 5000.

4 | DISCUSSION

In this study, we investigated the asymptotic null distribution of the private test statistics based on χ^2 statistic under the high-privacy regime. Under the assumption that the curator provides perturbed tables, our goal is to find the null distribution of the differentially private GOF and independence tests based on χ^2 statistics. We demonstrated that the private test statistics can be expressed in quadratic forms of the random vectors that follow a multivariate normal distribution as $n \to \infty$, and the limiting null distribution for the private test statistics is the weighted sum of the χ^2 distributions. The weights are determined by the eigenvalues of the covariance matrix of the random vectors and prespecified σ^2.

We believe that the impact of the noise is not significantly reduced as the sample size increases. It would be better to acknowledge the impact of the noise by adjusting the noise variance σ_n depending on the sample size n. Then, the privacy parameter μ for GDP, derived as Δ/σ_n, becomes inherently linked to the sample size n. Both simulation experiments and application to real data reveal that type I errors and test powers become stable and approach the performance of non-private tests as the sample size grows. These results indicate that by embracing the presence of noise and setting noise variance accordingly, we can attain a stronger privacy guarantee with larger sample sizes, facilitating the successful execution of private tests. Since the asymptotic test requires a large-sample size, the test results may not work well on smaller sample sizes. However, as demonstrated in simulation experiments, the test performs appropriately for moderately large-sample sizes $n > 200$. This fast convergence is in part due to the use of Gaussian noise, resulting in the χ^2 mixture for the null distribution.

For this study, we assume that n, σ and a perturbed table are provided by the curators (or data distributors). However, it's worth noting that they may prefer perturbed tables to exhibit a total of n instead of the sum of noises and n. Under this consideration, conducting private GOF and independence tests remains an intriguing challenge and direction for future exploration.

AUTHOR CONTRIBUTIONS

Sungkyu Jung guided in conceptualization and formulation in this study and edited manuscript. Seung Woo Kwak wrote the manuscript, designed and performed the numerical experiments and prepared figures. All authors read and approved the final version of the manuscript.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

ORCID

Sungkyu Jung https://orcid.org/0000-0002-6023-8956

Seung Woo Kwak https://orcid.org/0000-0001-5951-5728

ENDNOTE

[1] If $\mathbf{x}$ is a $r \times c$ table, a suitable vectorization can be applied, in which case $d = rc$.

REFERENCES

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (2007). *Discrete multivariate analysis: Theory and practice*: Springer Science & Business Media.

Bun, M., & Steinke, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. arXiv, https://arxiv.org/abs/1605.02065

Canonne, C., Kamath, G., & Steinke, T. (2022). The discrete Gaussian for differential privacy. *Journal of Privacy and Confidentiality*, *12*(1), 1–4. https://journalprivacyconfidentiality.org/index.php/jpc/article/view/784

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.).: Routledge.

Davies, R. B. (1973). Numerical inversion of a characteristic function. *Biometrika*, *60*(2), 415–417. https://doi.org/10.1093/biomet/60.2.415

Davies, R. B. (1980). Algorithm as 155: The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *29*(3), 323–333. http://www.jstor.org/stable/2346911

Dolera, E., & Favaro, S. (2021). The power of private likelihood-ratio tests for goodness-of-fit in frequency tables. arXiv, https://arxiv.org/abs/2109.09630

Dong, J., Roth, A., & Su, W. J. W. (2022). Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *84*(1), 3–37.

Driscoll, M. F. (1999). An improved result relating quadratic forms and chi-square distributions. *The American Statistician*, *53*, 273–275. https://api.semanticscholar.org/CorpusID:119704835

Dwork, C. (2006). Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming, Part II (ICALP 2006)*, Lecture Notes in Computer Science, *4052*, Springer-Verlag, pp. 1–12. https://www.microsoft.com/en-us/research/publication/differential-privacy/

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Halevi, S., & Rabin, T. (Eds.), *Theory of cryptography*: Springer Berlin Heidelberg, pp. 265–284.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, *9*(3–4), 211–407. https://doi.org/10.1561/0400000042

Gaboardi, M., Lim, H., Rogers, R., & Vadhan, S. (2016). Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *International Conference on Machine Learning*, PMLR, pp. 2111–2120.

Gaboardi, M., & Rogers, R. (2018). Local private hypothesis testing: Chi-square tests. In Dy, J., & Krause, A. (Eds.), *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 80: PMLR, pp. 1626–1635. https://proceedings.mlr.press/v80/gaboardi18a.html

Ghosh, A., Roughgarden, T., & Sundararajan, M. (2009). Universally utility-maximizing privacy mechanisms. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC '09, Association for Computing Machinery, pp. 351–360. https://doi.org/10.1145/1536414.1536464

Haney, S., Sexton, W., Machanavajjhala, A., Hay, M., & Miklau, G. (2021). *Differentially private algorithms for 2020 census detailed DHC race & ethnicity*. https://www2.census.gov/about/partners/cac/sac/meetings/2022-03/dhc-attachment-1-safetab-dp-algorithms.pdf (U.S. Census Bureau), https://arxiv.org/abs/2107.10659 (arXiv).

Johnson, A., & Shmatikov, V. (2013). Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, Association for Computing Machinery, pp. 1079–1088. https://doi.org/10.1145/2487575.2487687

Kim, M., Lee, J., Kwak, S. W., & Jung, S. (2023). Differentially private multivariate statistics with an application to contingency table analysis. https://arxiv.org/abs/2211.15019

Lin, Z., Owen, A. B., & Altman, R. B. (2004). Genomic research and human subject privacy. *Science*, *305*(5681), 183–183.

Liu, H., Tang, Y., & Zhang, H. H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*, *53*(4), 853–856. https://www.sciencedirect.com/science/article/pii/S0167947308005653

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157. https://doi.org/10.1007/bf02295996

Mironov, I. (2017). Renyi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, IEEE, pp. 263–275.

Nikolov, A., Talwar, K., & Zhang, L. (2016). The geometry of differential privacy: The small database and approximate cases. *SIAM Journal on Computing*, *45*(2), 575–616. https://doi.org/10.1137/130938943

Rinott, Y., O'Keefe, C. M., Shlomo, N., & Skinner, C. (2018). Confidentiality and differential privacy in the dissemination of frequency tables. *Statistical Science*, *33*(3), 358–385. https://doi.org/10.1214/17-STS641

Rogers, R., & Kifer, D. (2017). A new class of private chi-square hypothesis tests. In Singh, A., & Zhu, J. (Eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, Vol. 54: PMLR, pp. 991–1000. https://proceedings.mlr.press/v54/rogers17a.html

Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression: Computer Science Laboratory, SRI International (Tech. Rep.) http://www.csl.sri.com/papers/sritr-98-04/

Sheffet, O. (2018). Locally private hypothesis testing. In Dy, J., & Krause, A. (Eds.), *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 80: PMLR, pp. 4605–4614. https://proceedings.mlr.press/v80/sheffet18a.html

Smith, A. (2011). Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, STOC '11, Association for Computing Machinery, pp. 813–822. https://doi.org/10.1145/1993636.1993743

Son, J., Park, M., & Jung, S. (2022). A parametric bootstrap test for comparing differentially private histograms. *The Korean Journal of Applied Statistics*, *35*(1), 1–17.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 557–570. https://doi.org/10.1142/S0218488502001648

U.S. Census Bureau (2021). *Disclosure avoidance for the 2020 census: An introduction*: U.S. Department of Commerce.

Uhler, C., Slavkovic, A. B., & Fienberg, S. E. (2013). Privacy-preserving data sharing for genome-wide association studies. *Journal of Privacy and Confidentiality*, *5*(1), 137–166.https://journalprivacyconfidentiality.org/index.php/jpc/article/view/629

Vu, D., & Slavkovic, A. B. (2009). Differential privacy for clinical trial data: Preliminary evaluations. In *2009 IEEE International Conference on Data Mining Workshops*, IEEE, pp. 138–143. https://api.semanticscholar.org/CorpusID:12371233

Wang, Y., Lee, J., & Kifer, D. (2015). Revisiting differentially private hypothesis tests for categorical data. arXiv, https://arxiv.org/abs/1511.03376

Wood, S. (2023). mgcv: Mixed GAM computation vehicle with automatic smoothness estimation. R package version 1.9-0.

AUTHOR BIOGRAPHIES

Sungkyu Jung is a professor of Statistics at the Seoul National University. His research interest lies in the theoretical study and applications of modern statistics and data science in the analysis of data that lie on non-standard spaces and data privacy, including differential privacy and synthetic data generation.

Seung Woo Kwak is a post-doctorate researcher of LAMP (Learning and Academic research institution for Master's, PhD student and Post-docs) at Seoul National University. His research interest focuses on differential privacy, synthetic data generation and statistical machine learning.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Jung, S., & Woo Kwak, S. (2024). Highly private large-sample tests for contingency tables. *Stat*, *13*(1), e658. https://doi.org/10.1002/sta4.658

APPENDIX A: PROOFS AND TECHNICAL DETAILS

A.0.1 | Proof of Theorem 2

Proof of Theorem 2. The result of $T_1(\mathbf{u};\mathbf{p})$ can be derived as

$$\begin{aligned} T_1(\mathbf{u};\mathbf{p}) &= \sum_{i=1}^{d} \frac{(u_i - np_i)^2}{np_i} = \sum_{i=1}^{d} \frac{n\left(\hat{p}_i - p_i + \frac{z_i}{n}\right)^2}{p_i} \\ &= \mathbf{y}^\top \left(D_{\mathbf{p}}^{-1} + O_p\left(n^{-\frac{1}{2}}\right)\right)\mathbf{y} = \mathbf{y}^\top D_{\mathbf{p}}^{-1}\mathbf{y}, \end{aligned}$$

where $\hat{p}_i = \frac{x_i}{n}$ and $\mathbf{y} = \sqrt{n}\{\hat{\mathbf{p}} - \mathbf{p} + \frac{\mathbf{z}}{n}\}$. Then, $\mathbf{y} \Rightarrow N_d(\mathbf{0}, D_{\mathbf{p}} - \mathbf{p}\mathbf{p}^\top + \sigma^2 \mathbf{I}_d)$.

$T_1(\mathbf{u},\mathbf{p})$ converges in distribution to $\sum_{i=1}^{d} \lambda_i(\Sigma_1) Z_i^2$, where Z_i^2s are independent and each follows χ_1^2 distribution. The proof is completed by using a classical result on the convergence of distribution and the quadratic forms; see Theorems 14.3-6 and 14.3-7 in Bishop et al. (2007). □

A.0.2 | Proof of Theorem 1

For the proof of the limiting distribution of the test statistics, we will exploit the fact that the data $\mathbf{x} = \mathrm{vec}(\mathbb{X})$ and the noise $\mathbf{z} = \mathrm{vec}(\mathbb{Z})$ are independent. Write for $i = 1,\ldots,r$, $j = 1,\ldots,c$, $z_{i\cdot} = \sum_{j=1}^{c} z_{ij}$, $z_{\cdot j} = \sum_{i=1}^{r} z_{ij}$, $z_{\cdot\cdot} = \sum_{i=1}^{r}\sum_{j=1}^{c} z_{ij}$, and observe that

$$n^{-\frac{1}{2}} z_{ij} \sim N(0,\sigma^2), n^{-\frac{1}{2}} z_{i\cdot} \sim N(0, c\sigma^2), n^{-\frac{1}{2}} z_{\cdot j} \sim N(0, r\sigma^2), \tag{A1}$$

and

$$n^{-\frac{1}{2}} z_{\cdot\cdot} \sim N(0, rc\sigma^2).$$

We next list some classical results on the multinomially distributed random matrix $\mathbb{X}$ and its associated χ^2-statistic. Note that for $\hat{\boldsymbol{\pi}}^{(1)} = \mathbb{X}\mathbf{1}_c/n$ and $\hat{\boldsymbol{\pi}}^{(2)} = \mathbb{X}^\top \mathbf{1}_r/n$, the central limit theorem gives

$$\sqrt{n}\left(\hat{\boldsymbol{\pi}}^{(1)} - \boldsymbol{\pi}^{(1)}\right) \Rightarrow N_r\left(\mathbf{0}, D_{\boldsymbol{\pi}^{(1)}} - \boldsymbol{\pi}^{(1)}\left(\boldsymbol{\pi}^{(1)}\right)^\top\right) \text{ and} \tag{A2}$$

$$\sqrt{n}\left(\hat{\boldsymbol{\pi}}^{(2)} - \boldsymbol{\pi}^{(2)}\right) \Rightarrow N_c\left(\mathbf{0}, D_{\boldsymbol{\pi}^{(2)}} - \boldsymbol{\pi}^{(2)}\left(\boldsymbol{\pi}^{(2)}\right)^\top\right), \tag{A3}$$

as $n \to \infty$. In particular, for $i = 1,\ldots,r$, $j = 1,\ldots,c$,

$$\hat{\pi}_i^{(1)} = \pi_i^{(1)} + O_p\left(n^{-\frac{1}{2}}\right) \text{ and } \hat{\pi}_j^{(2)} = \pi_j^{(2)} + O_p\left(n^{-\frac{1}{2}}\right). \tag{A4}$$

The following result is excerpted from the proof of Theorem 14.8-4 of Bishop et al. (2007).

Lemma 1 Bishop et al. (2007). For $\hat{\mathbf{p}} = \mathbf{x}/n$ and $\hat{\boldsymbol{\pi}} = \hat{\boldsymbol{\pi}}^{(2)} \otimes \hat{\boldsymbol{\pi}}^{(1)}$, $\sqrt{n}D_{\boldsymbol{\pi}}^{-\frac{1}{2}}(\hat{\mathbf{p}} - \hat{\boldsymbol{\pi}}) \Rightarrow N_{rc}(\mathbf{0}, \mathbf{D}_{\boldsymbol{\pi},\mathbf{A}})$ as $n \to \infty$, where $\mathbf{D}_{\boldsymbol{\pi},\mathbf{A}} = \mathbf{I}_{rc} - \sqrt{\boldsymbol{\pi}}\sqrt{\boldsymbol{\pi}}^\top - \mathbf{A}\left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top$. Here, $\mathbf{A} = D_{\boldsymbol{\pi}}^{-\frac{1}{2}}\nabla$, and ∇ is the $rc \times (r+c-2)$ matrix consisting of partial derivatives of $\boldsymbol{\pi}$ with respect to $\pi_i^{(1)}$, $i = 1,\ldots,r-1$ and $\pi_j^{(2)}$, $j = 1,\ldots,c-1$, and is expressed as

$$\nabla = \left[\boldsymbol{\pi}^{(2)} \otimes \begin{pmatrix} \mathbf{I}_{r-1} \\ -\mathbf{1}_{r-1}^\top \end{pmatrix}, \begin{pmatrix} \mathbf{I}_{c-1} \\ -\mathbf{1}_{c-1}^\top \end{pmatrix} \otimes \boldsymbol{\pi}^{(1)}\right],$$

as defined in Theorem 3.

To further simplify, we collect the terms with order $O_p(1)$ and $O_p\left(n^{-\frac{1}{2}}\right)$ and use the O_p notation for the terms smaller than $n^{-\frac{1}{2}}$. For this purpose, the results of (A1) and (A4) are extensively used in the following calculation. The cross-product term rewrites to

$$\begin{aligned}\left(\hat{\pi}_i^{(1)} + n^{-1}z_{i\cdot}\right)\left(\hat{\pi}_j^{(2)} + n^{-1}z_{\cdot j}\right) &= \hat{\pi}_{ij} + \left(\pi_i^{(1)} + O_p\left(n^{-\frac{1}{2}}\right)\right)n^{-1}z_{\cdot j} + n^{-1}z_{i\cdot}\left(\pi_j^{(2)} + O_p\left(n^{-\frac{1}{2}}\right)\right) + O_p\left(n^{-1}\right), \\ &= \hat{\pi}_{ij} + \pi_i^{(1)}n^{-1}z_{\cdot j} + n^{-1}z_{i\cdot}\pi_j^{(2)} + O_p\left(n^{-1}\right),\end{aligned} \tag{A5}$$

and the ‘error term’ becomes

$$\frac{n}{n + z_{\cdot\cdot}} = \frac{1}{1 + n^{-1}z_{\cdot\cdot}} = 1 - n^{-1}z_{\cdot\cdot} + O_p\left(n^{-1}\right). \tag{A6}$$

Moreover, we can write

$$\hat{\pi}_{ij}n^{-1}z_{\cdot\cdot} = \left(\pi_{ij} + O_p\left(n^{-\frac{1}{2}}\right)\right)n^{-1}z_{\cdot\cdot} = \pi_{ij}n^{-1}z_{\cdot\cdot} + O_p\left(n^{-1}\right). \tag{A7}$$

Proof. Note that

$$\chi_n^2(\mathbb{U}) = \sum_{i,j} \frac{n\left(\hat{p}_{ij} + n^{-1}z_{ij} - \hat{\pi}_{ij}^U\right)^2}{\hat{\pi}_{ij}^U},$$

where

$$\hat{\pi}_{ij}^U = \left(\hat{\pi}_i^{(1)} + n^{-1}z_{i\cdot}\right)\left(\hat{\pi}_j^{(2)} + n^{-1}z_{\cdot j}\right)\frac{1}{\left(1 + n^{-1}z_{\cdot\cdot}\right)^2}.$$

Utilizing the expansion $\frac{1}{(1+x)^2} = 1 - 2x + O(x^2)$ and (A5)–(A7),

$$\hat{\pi}^U_{ij} = \hat{\pi}_{ij} + \pi_i^{(1)} n^{-1} z_{\cdot j} + n^{-1} z_{i\cdot} \pi_j^{(2)} - 2\pi_{ij} n^{-1} z_{\cdot\cdot} + O_p(n^{-1}). \tag{A8}$$

We have

$$\hat{p}^U_{ij} - \hat{\pi}^U_{ij} = (\hat{p}_{ij} - \hat{\pi}_{ij}) + n^{-1} \hat{\mathbf{z}}^{\boldsymbol{\pi}}_{ij} + O_p\left(n^{-1}\right), \tag{A9}$$

where

$$\hat{\mathbf{z}}^{\boldsymbol{\pi}}_{ij} = z_{ij} + \pi_i^{(1)} z_{\cdot j} + \pi_j^{(2)} z_{i\cdot} - 2\pi_{ij} z_{\cdot\cdot}$$

In addition, by (A8),

$$\hat{\pi}^U_{ij} = \pi_{ij} + O_p\left(n^{-\frac{1}{2}}\right).$$

From (A9),

$$\begin{aligned} \chi^2_n(\mathbb{U}) &= \sum_{i,j} \frac{\left(u_{ij} - n\hat{\pi}^U_{ij}\right)^2}{n\hat{\pi}^U_{ij}} \\ &= \sum_{i,j} \frac{n\left(\hat{p}_{ij} + n^{-1} z_{ij} - \hat{\pi}^U_{ij}\right)^2}{\hat{\pi}^U_{ij}} \\ &= \sum_{i,j} \frac{\left(\sqrt{n}(\hat{p}_{ij} - \hat{\pi}_{ij}) + n^{-\frac{1}{2}} \hat{\mathbf{z}}^{\boldsymbol{\pi}}_{ij} + O_p\left(n^{-\frac{1}{2}}\right)\right)^2}{\pi_{ij} + O_p(n^{-\frac{1}{2}})} \\ &= \mathbf{w}_n^{G\top} \left(D_{\boldsymbol{\pi}}^{-1} + O_p\left(n^{-\frac{1}{2}}\right)\right) \mathbf{w}_n^G, \end{aligned} \tag{A10}$$

where $\mathbf{w}_n^G = \sqrt{n}(\hat{\mathbf{p}} - \hat{\boldsymbol{\pi}}) + n^{-\frac{1}{2}} \hat{\mathbf{z}}^{\boldsymbol{\pi}} + O_p(n^{-\frac{1}{2}})$. One can check that

$$n^{-\frac{1}{2}} \hat{\mathbf{z}}^{\boldsymbol{\pi}} = ((\mathbf{I}_c - D_{\boldsymbol{\pi}^{(2)}} \mathbf{J}_c) \otimes (\mathbf{I}_r - D_{\boldsymbol{\pi}^{(1)}} \mathbf{J}_r) + (D_{\boldsymbol{\pi}^{(2)}} \mathbf{J}_c) \otimes (D_{\boldsymbol{\pi}^{(1)}} \mathbf{J}_r)) n^{-\frac{1}{2}} \mathbf{z},$$

and $n^{-\frac{1}{2}} \hat{\mathbf{z}}^{\boldsymbol{\pi}} \sim N\left(0, \Sigma^U_{\sigma,\boldsymbol{\pi}}\right)$ where

$$\Sigma^U_{\sigma,\boldsymbol{\pi}} = \sigma^2 [\mathbf{A}_1 \otimes \mathbf{A}_2 + \mathbf{A}_3 \otimes \mathbf{A}_4][\mathbf{A}_1 \otimes \mathbf{A}_2 + \mathbf{A}_3 \otimes \mathbf{A}_4]^\top, \tag{A11}$$

and $\mathbf{A}_1$–$\mathbf{A}_4$ are defined as

$$\mathbf{A}_1 = (\mathbf{I}_c - D_{\boldsymbol{\pi}^{(2)}} \mathbf{J}_c), \tag{A12}$$

$$\mathbf{A}_2 = (\mathbf{I}_r - D_{\boldsymbol{\pi}^{(1)}} \mathbf{J}_r), \tag{A13}$$

$$\mathbf{A}_3 = D_{\boldsymbol{\pi}^{(2)}} \mathbf{J}_c, \tag{A14}$$

$$\mathbf{A}_4 = D_{\boldsymbol{\pi}^{(1)}} \mathbf{J}_r. \tag{A15}$$

Thus, together with Lemma 1, as $n \to \infty$,

$$D_{\boldsymbol{\pi}}^{-1/2} \mathbf{w}_n^G \Rightarrow N_{rc}(\mathbf{0}, \Sigma_U),$$

where

$$\Sigma_U = \mathbf{I}_{rc} - \sqrt{\boldsymbol{\pi}} \sqrt{\boldsymbol{\pi}}^\top - \mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top + D_{\boldsymbol{\pi}}^{-1/2} \Sigma^U_{\sigma,\boldsymbol{\pi}} D_{\boldsymbol{\pi}}^{-1/2}.$$

So, we have

$$\chi_n^2(\mathbb{U}) \Rightarrow \sum_{i=1}^{rc} \lambda_i(\Sigma_U) Z_i^2.$$

□

Proof. Now,

$$\chi_G^2(\mathbb{U}) = \sum_{i,j} \frac{\left(u_{ij} - n\hat{\pi}_{ij}^G\right)^2}{n\hat{\pi}_{ij}^G},$$

where

$$\hat{\boldsymbol{\pi}}^{(1)G} = \tilde{\boldsymbol{\pi}}^{(1)} - \frac{\mathbf{1}_r^\top \tilde{\boldsymbol{\pi}}^{(1)} - 1}{r}\mathbf{1}_r = \tilde{\boldsymbol{\pi}}^{(1)} - \frac{z_{..}}{nr}\mathbf{1}_r, \tag{A16}$$

$$\begin{aligned} \hat{\boldsymbol{\pi}}^{(2)G} &= \tilde{\boldsymbol{\pi}}^{(2)} - \frac{\mathbf{1}_c^\top \tilde{\boldsymbol{\pi}}^{(2)} - 1}{c}\mathbf{1}_c = \tilde{\boldsymbol{\pi}}^{(2)} - \frac{z_{..}}{nc}\mathbf{1}_c \text{ and} \\ \hat{\boldsymbol{\pi}}^G &= \hat{\boldsymbol{\pi}}^{(2)G} \otimes \hat{\boldsymbol{\pi}}^{(1)G}. \end{aligned} \tag{A17}$$

From (A16) and (A17),

$$\begin{aligned} \hat{\pi}_{ij}^G &= \hat{\pi}_i^{(1)G}\hat{\pi}_j^{(2)G} \\ &= \left(\tilde{\pi}_i^{(1)} - \frac{z_{..}}{nr}\right)\left(\tilde{\pi}_j^{(2)} - \frac{z_{..}}{nc}\right) \\ &= \left(\hat{\pi}_i^{(1)} + n^{-1}z_{i.} - \frac{z_{..}}{nr}\right)\left(\hat{\pi}_j^{(2)} + n^{-1}z_{.j} - \frac{z_{..}}{nc}\right) \\ &= \left(\hat{\pi}_i^{(1)} + n^{-1}z_{i.} - \frac{z_{..}}{nr}\right)\hat{\pi}_j^{(2)} + \left(\hat{\pi}_i^{(1)} + n^{-1}z_{i.} - \frac{z_{..}}{nr}\right)n^{-1}z_{.j} - \left(\hat{\pi}_i^{(1)} + n^{-1}z_{i.} - \frac{z_{..}}{nr}\right)\frac{z_{..}}{nc} \\ &= \left(\hat{\pi}_i^{(1)}\hat{\pi}_j^{(2)} + n^{-1}z_{i.}\hat{\pi}_j^{(2)} - \frac{z_{..}}{nr}\hat{\pi}_j^{(2)}\right) + \left(\hat{\pi}_i^{(1)}n^{-1}z_{.j} + n^{-1}z_{i.}n^{-1}z_{.j} - \frac{z_{..}}{nr}n^{-1}z_{.j}\right) - \left(\hat{\pi}_i^{(1)}\frac{z_{..}}{nc} + n^{-1}z_{i.}\frac{z_{..}}{nc} - \frac{z_{..}}{nr}\frac{z_{..}}{nc}\right) \\ &= \hat{\pi}_{ij} + n^{-1}z_{i.}\hat{\pi}_j^{(2)} - (nr)^{-1}z_{..}\hat{\pi}_j^{(2)} + \hat{\pi}_i^{(1)}n^{-1}z_{.j} + n^{-2}z_{i.}z_{.j} - n^{-2}r^{-1}z_{..}z_{.j} - (nc)^{-1}\hat{\pi}_i^{(1)}z_{..} - n^{-2}c^{-1}z_{i.}z_{..} + n^{-2}(rc)^{-1}z_{..}^2 \\ &= \hat{\pi}_{ij} + n^{-1}(z_{i.} - r^{-1}\sum_{i=1}^{r} z_{i.})\hat{\pi}_j^{(2)} + n^{-1}(z_{.j} - c^{-1}\sum_{j=1}^{c} z_{.j})\hat{\pi}_i^{(1)} + n^{-2}z_{i.}z_{.j} - n^{-2}r^{-1}z_{..}z_{.j} - n^{-2}c^{-1}z_{i.}z_{..} + n^{-2}(rc)^{-1}z_{..}^2 \\ &= \hat{\pi}_{ij} + n^{-1}(z_{i.} - r^{-1}\sum_{i=1}^{r} z_{i.})\hat{\pi}_j^{(2)} + n^{-1}(z_{.j} - c^{-1}\sum_{j=1}^{c} z_{.j})\hat{\pi}_i^{(1)} + O_p(n^{-1}) \\ &= \hat{\pi}_{ij} + n^{-1}(z_{.j} - r\overline{Z})\hat{\pi}_i^{(1)} + n^{-1}(z_{i.} - c\overline{Z})\hat{\pi}_j^{(2)} + O_p(n^{-1}), \end{aligned}$$

where $\overline{Z} = \frac{\sum_{i,j} z_{ij}}{rc} = \frac{z_{..}}{rc}$. From (A4), we have

$$\begin{aligned} \hat{\pi}_{ij} &+ n^{-1}(z_{.j} - r\overline{Z})\hat{\pi}_i^{(1)} + n^{-1}(z_{i.} - c\overline{Z})\hat{\pi}_j^{(2)} + O_p(n^{-1}) \\ &= \hat{\pi}_{ij} + n^{-1}(z_{.j} - r\overline{Z})\left(\pi_i^{(1)} + O_p(n^{-\frac{1}{2}})\right) + n^{-1}(z_{i.} - c\overline{Z})\left(\pi_j^{(2)} + O_p(n^{-\frac{1}{2}})\right) + O_p(n^{-1}) \\ &= \hat{\pi}_{ij} + n^{-1}(z_{.j} - r\overline{Z})\pi_i^{(1)} + n^{-1}(z_{i.} - c\overline{Z})\pi_j^{(2)} + O_p(n^{-1}). \end{aligned}$$

Now,

$$\begin{aligned} n^{-1}u_{ij} - \hat{\pi}_{ij}^G &= \hat{p}_{ij} + n^{-1}z_{ij} - \left(\hat{\pi}_{ij} + n^{-1}(z_{.j} - r\overline{Z})\pi_i^{(1)} + n^{-1}(z_{i.} - c\overline{Z})\pi_j^{(2)} + O_p(n^{-1})\right) \\ &= (\hat{p}_{ij} - \hat{\pi}_{ij}) + n^{-1}\left(z_{ij} - z_{.j}\pi_i^{(1)} - z_{i.}\pi_j^{(2)} + r\overline{Z}\pi_i^{(1)} + c\overline{Z}\pi_j^{(2)}\right) + O_p(n^{-1}) \\ &= (\hat{p}_{ij} - \hat{\pi}_{ij}) + n^{-1}\mathbf{z}_{ij}^{\pi,G} + O_p(n^{-1}), \end{aligned}$$

where

$$\mathbf{z}_{ij}^{\pi,G} = z_{ij} - \pi_i^{(1)}z_{.j} - \pi_j^{(2)}z_{i.} + r\pi_i^{(1)}\overline{Z} + c\pi_j^{(2)}\overline{Z}.$$

Then,

$$\mathbf{u} - n\hat{\boldsymbol{\pi}}^{G} = n(\hat{\mathbf{p}} - \hat{\boldsymbol{\pi}}) + n^{-1}\hat{\mathbf{z}}^{\pi,G} + O_p(n^{-\frac{1}{2}})),$$

where

$$\hat{\mathbf{z}}^{\pi,G} = \left((\mathbf{I}_c - D_{\boldsymbol{\pi}^{(2)}}\mathbf{J}_c) \otimes (\mathbf{I}_r - D_{\boldsymbol{\pi}^{(1)}}\mathbf{J}_r) - (D_{\boldsymbol{\pi}^{(2)}}\mathbf{J}_c) \otimes (D_{\boldsymbol{\pi}^{(1)}}\mathbf{J}_r) + (\mathbf{1}_c \otimes \boldsymbol{\pi}^{(1)})\mathbf{1}_{rc}^{\top}\frac{1}{c} + (\boldsymbol{\pi}^{(2)} \otimes \mathbf{1}_r)\mathbf{1}_{rc}^{\top}\frac{1}{r} \right)\mathbf{z}.$$

So,

$$\begin{aligned} \chi^2_G(\mathbb{U}) &= \sum_{i,j} \frac{(u_{ij} - n\hat{\pi}^{G}_{ij})^2}{n\hat{\pi}^{G}_{ij}} \\ &= (\mathbf{w}^{G}_n)^{\top}\left(D_{\boldsymbol{\pi}}^{-1} + O_p\left(n^{-\frac{1}{2}}\right)\right)\mathbf{w}^{G}_n, \end{aligned}$$

where $\mathbf{w}^{G}_n = \sqrt{n}((\hat{\mathbf{p}} - \hat{\boldsymbol{\pi}}) + n^{-\frac{1}{2}}\hat{\mathbf{z}}^{\pi,G} + O_p(n^{-\frac{1}{2}})$. Thus, together with Lemma 1, as $n \to \infty$,

$$D_{\boldsymbol{\pi}}^{-\frac{1}{2}}\mathbf{w}^{G}_n \Rightarrow N_{rc}(\mathbf{0}, \Sigma_G),$$

where

$$\Sigma_G = \mathbf{I}_{rc} - \sqrt{\boldsymbol{\pi}}\sqrt{\boldsymbol{\pi}}^{\top} - \mathbf{A}(\mathbf{A}^{\top}\mathbf{A})\mathbf{A}^{\top} + D_{\boldsymbol{\pi}}^{-\frac{1}{2}}\Sigma^{G}_{\sigma,\boldsymbol{\pi}}D_{\boldsymbol{\pi}}^{-\frac{1}{2}},$$

$$\Sigma^{G}_{\sigma,\boldsymbol{\pi}} = \sigma^2[\mathbf{A}_1 \otimes \mathbf{A}_2 - \mathbf{A}_3 \otimes \mathbf{A}_4 + \mathbf{A}_5 + \mathbf{A}_6][\mathbf{A}_1 \otimes \mathbf{A}_2 - \mathbf{A}_3 \otimes \mathbf{A}_4 + \mathbf{A}_5 + \mathbf{A}_6]^{\top}. \tag{A18}$$

$\mathbf{A}_1$–$\mathbf{A}_4$ are defined in (A12)–(A15), and

$$\mathbf{A}_5 = (\mathbf{1}_c \otimes \boldsymbol{\pi}^{(1)})\mathbf{1}_{rc}^{\top}c^{-1}, \tag{A19}$$

$$\mathbf{A}_6 = (\mathbf{1}_r \otimes \boldsymbol{\pi}^{(2)})\mathbf{1}_{rc}^{\top}r^{-1}. \tag{A20}$$

Thus,

$$\chi^2_G(\mathbb{U}) \Rightarrow \sum_{i=1}^{rc}\lambda_i(\Sigma_G)Z_i^2.$$

□

A.0.3 | Details of $\Sigma^{U}_{\sigma,\boldsymbol{\pi}}$ and $\Sigma^{G}_{\sigma,\boldsymbol{\pi}}$

The expansions of $\Sigma^{U}_{\sigma,\boldsymbol{\pi}}$ (A11) and $\Sigma^{G}_{\sigma,\boldsymbol{\pi}}$ (A18) in Appendix A.0.2 share common terms which facilitate the computation of the covariance matrices. Let $\mathbf{A}_1$–$\mathbf{A}_6$ be from (A12)–(A15) and (A19) and (A20). We have $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ and $(A \otimes B)^{\top} = A^{\top} \otimes B^{\top}$.

1. $\Sigma^{U}_{\sigma,\boldsymbol{\pi}}$ in (A11) can be expressed as

$$\begin{aligned} \Sigma^{U}_{\sigma,\boldsymbol{\pi}} &= \sigma^2[(\mathbf{I}_c - D_{\boldsymbol{\pi}^{(2)}}\mathbf{J}_c) \otimes (\mathbf{I}_r - D_{\boldsymbol{\pi}^{(1)}}\mathbf{J}_r) + (D_{\boldsymbol{\pi}^{(2)}}\mathbf{J}_c) \otimes (D_{\boldsymbol{\pi}^{(1)}}\mathbf{J}_r)] \\ &\quad [(\mathbf{I}_c - D_{\boldsymbol{\pi}^{(2)}}\mathbf{J}_c) \otimes (\mathbf{I}_r - D_{\boldsymbol{\pi}^{(1)}}\mathbf{J}_r) + (D_{\boldsymbol{\pi}^{(2)}}\mathbf{J}_c) \otimes (D_{\boldsymbol{\pi}^{(1)}}\mathbf{J}_r)]^{\top} \\ &= \sigma^2[\mathbf{A}_1 \otimes \mathbf{A}_2 + \mathbf{A}_3 \otimes \mathbf{A}_4][\mathbf{A}_1 \otimes \mathbf{A}_2 + \mathbf{A}_3 \otimes \mathbf{A}_4]^{\top} \\ &= \sigma^2\left[(\mathbf{A}_1 \otimes \mathbf{A}_2)(\mathbf{A}_1 \otimes \mathbf{A}_2)^{\top} + (\mathbf{A}_1 \otimes \mathbf{A}_2)(\mathbf{A}_3 \otimes \mathbf{A}_4)^{\top} + (\mathbf{A}_3 \otimes \mathbf{A}_4)(\mathbf{A}_1 \otimes \mathbf{A}_2)^{\top} + (\mathbf{A}_3 \otimes \mathbf{A}_4)(\mathbf{A}_3 \otimes \mathbf{A}_4)^{\top}\right] \\ &= \Sigma_{\sigma,\boldsymbol{\pi}} + \sigma^2\left[(\mathbf{A}_1 \otimes \mathbf{A}_2)(\mathbf{A}_3 \otimes \mathbf{A}_4)^{\top} + (\mathbf{A}_3 \otimes \mathbf{A}_4)(\mathbf{A}_1 \otimes \mathbf{A}_2)^{\top} + (\mathbf{A}_3 \otimes \mathbf{A}_4)(\mathbf{A}_3 \otimes \mathbf{A}_4)^{\top}\right] \\ &= \Sigma_{\sigma,\boldsymbol{\pi}} + \sigma^2[(\mathbf{A}_1\mathbf{A}_3^{\top}) \otimes (\mathbf{A}_2\mathbf{A}_4^{\top}) + (\mathbf{A}_3\mathbf{A}_1^{\top}) \otimes (\mathbf{A}_4\mathbf{A}_2^{\top}) + (\mathbf{A}_3\mathbf{A}_3^{\top}) \otimes (\mathbf{A}_4\mathbf{A}_4^{\top})], \end{aligned}$$

where $\Sigma_{\sigma,\boldsymbol{\pi}} = \sigma^2(\mathbf{A}_1 \otimes \mathbf{A}_2)(\mathbf{A}_1 \otimes \mathbf{A}_2)^{\top} = \sigma^2(\mathbf{A}_1\mathbf{A}_1^{\top}) \otimes (\mathbf{A}_2\mathbf{A}_2^{\top})$.

2. $\Sigma^{U}_{\sigma,\boldsymbol{\pi}}$ in (A18) can be expressed as

$$\begin{aligned}
\Sigma^{G}_{\sigma,\boldsymbol{\pi}} &= \sigma^2\big[(\mathbf{I}_c - D_{\boldsymbol{\pi}^{(2)}}\mathbf{J}_c)\otimes(\mathbf{I}_r - D_{\boldsymbol{\pi}^{(1)}}\mathbf{J}_r) - D_{\boldsymbol{\pi}^{(2)}}\mathbf{J}_c\otimes D_{\boldsymbol{\pi}^{(1)}}\mathbf{J}_r + (\mathbf{1}_c\otimes\boldsymbol{\pi}^{(1)})\mathbf{1}_{rc}^{\top}c^{-1} \\
&\quad +(\mathbf{1}_r\otimes\boldsymbol{\pi}^{(2)})\mathbf{1}_{rc}^{\top}r^{-1}\big]\big[(\mathbf{I}_c - D_{\boldsymbol{\pi}^{(2)}}\mathbf{J}_c)\otimes(\mathbf{I}_r - D_{\boldsymbol{\pi}^{(1)}}\mathbf{J}_r) - D_{\boldsymbol{\pi}^{(2)}}\mathbf{J}_c\otimes D_{\boldsymbol{\pi}^{(1)}}\mathbf{J}_r \\
&\quad +(\mathbf{1}_c\otimes\boldsymbol{\pi}^{(1)})\mathbf{1}_{rc}^{\top}c^{-1} + (\mathbf{1}_r\otimes\boldsymbol{\pi}^{(2)})\mathbf{1}_{rc}^{\top}r^{-1}\big]^{\top} \\
&= \sigma^2[\mathbf{A}_1\otimes\mathbf{A}_2 - \mathbf{A}_3\otimes\mathbf{A}_4 + \mathbf{A}_5 + \mathbf{A}_6][\mathbf{A}_1\otimes\mathbf{A}_2 - \mathbf{A}_3\otimes\mathbf{A}_4 + \mathbf{A}_5 + \mathbf{A}_6]^{\top} \\
&= \Sigma_{\sigma,\boldsymbol{\pi}} \\
&\quad +\sigma^2\big[(\mathbf{A}_3\mathbf{A}_3^{\top})\otimes(\mathbf{A}_4\mathbf{A}_4^{\top}) - (\mathbf{A}_1\mathbf{A}_3^{\top})\otimes(\mathbf{A}_2\mathbf{A}_4^{\top}) - (\mathbf{A}_3\mathbf{A}_1^{\top})\otimes(\mathbf{A}_4\mathbf{A}_2^{\top})) \\
&\qquad +\mathbf{A}_5(\mathbf{A}_1^{\top}\otimes\mathbf{A}_2^{\top}) + \mathbf{A}_6(\mathbf{A}_1^{\top}\otimes\mathbf{A}_2^{\top}) - \mathbf{A}_5(\mathbf{A}_3^{\top}\otimes\mathbf{A}_4^{\top}) - \mathbf{A}_6(\mathbf{A}_3^{\top}\otimes\mathbf{A}_4^{\top}) \\
&\qquad +(\mathbf{A}_1\otimes\mathbf{A}_2)\mathbf{A}_5^{\top} - (\mathbf{A}_3\otimes\mathbf{A}_4)\mathbf{A}_5^{\top} + (\mathbf{A}_1\otimes\mathbf{A}_2)\mathbf{A}_6^{\top} - (\mathbf{A}_3\otimes\mathbf{A}_4)\mathbf{A}_6^{\top} \\
&\qquad +\mathbf{A}_5\mathbf{A}_6^{\top} + \mathbf{A}_6\mathbf{A}_5^{\top} + \mathbf{A}_5\mathbf{A}_5^{\top} + \mathbf{A}_6\mathbf{A}_6^{\top}\big].
\end{aligned}$$

A.0.4 | Proof of Theorem 4

Let the 2×2 table be perturbed and vectorized as $\mathbf{u}=\mathbf{x}+\mathbf{z}$, where $\mathbf{x}=(x_{12},x_{21},x_{11}+x_{22})\sim \text{Mult}(n,\boldsymbol{\pi})$, $\mathbf{z}=(z_{12},z_{21},z_{11}+z_{22})$, $z_{ij}\sim N(0,\sigma_n^2)$ for $i=1,2$ and $j=1,2$. Based on the limiting distribution of the multinomial random variables, we have

$$\frac{\mathbf{u}-n\mathbf{p}}{\sqrt{n}} \Rightarrow N\left(\mathbf{0},\Sigma_0+\frac{\sigma_n^2}{n}\mathbf{C}\right), \text{ where } \mathbf{C}=\begin{bmatrix}1&0&0\\0&1&0\\0&0&2\end{bmatrix}.$$

Under the null hypothesis that $x_{12}=x_{21}=n\boldsymbol{\pi}$, the private test statistic t_n is defined as

$$t_n(\mathbf{u})=\frac{\mathbf{b}^{\top}\mathbf{u}}{\sqrt{\mathbf{b}^{\top}(n\Sigma+\sigma_n^2\mathbf{C})\mathbf{b}}}=\frac{u_{12}-u_{21}}{\sqrt{2n\pi_1+2\sigma_n^2}}=\frac{u_{12}-u_{21}}{\sqrt{n^*+2\sigma_n^2}},$$

where $\mathbf{b}=(1,-1,0)$. We have $t_n\Rightarrow N(0,1)$ since

$$\frac{\mathbf{b}^{\top}\mathbf{u}}{\sqrt{n}}\Rightarrow N\left(0,\mathbf{b}^{\top}\left(\Sigma_0+\frac{\sigma_n^2}{n}\mathbf{C}\right)\mathbf{b}\right)=N\left(0,2\pi_1+\frac{2\sigma_n^2}{n}\right).$$