The Korean Journal of Applied Statistics
2022, Vol. 35, No. 1, 1–17

DOI: https://doi.org/10.5351/KJAS.2022.35.1.001
Print ISSN1225-066X / Online ISSN 2383-5818

A parametric bootstrap test for comparing differentially private histograms

Juhee Son[a], Min-Jeong Park[b], Sungkyu Jung[1,a]

[a]Department of Statistics, Seoul National University [b]Statistics Korea

Abstract

We propose a test of consistency for two differentially private histograms using parametric bootstrap. The test can be applied when the original raw histograms are not available but only the differentially private histograms and the privacy level α are available. We also extend the test for the case where the privacy levels are different for different histograms. The resident population data of Korea and U.S in year 2020 are used to demonstrate the efficacy of the proposed test procedure. The proposed test controls the type I error rate at the nominal level and has a high power, while a conventional test procedure fails. While the differential privacy framework formally controls the risk of privacy leakage, the utility of such framework is questionable. This work also suggests that the power of a carefully designed test may be a viable measure of utility.

Keywords: parametric bootstrap, differential privacy, consistency of two histograms

1. 서론

히스토그램은 표본을 이용하여 모집단의 분포를 살펴보는 대표적인 방법이며, 이를 이용하여 두 집단의 모분포가 동일한지에 대한 통계적 검정을 시행할 수 있다. 히스토그램의 동질성 검정은 다양한 분야에서 사용될 수 있다. 예를 들면, 임상 실험 전후에 관심이 있는 특성치의 분포에 차이가 있는지 등, 두 집단의 분포 동질성을 검정할 수 있다. 히스토그램에서 각 구간 도수들의 벡터를 다항분포 표본으로 본다면 구간의 길이와 개수가 같은 두 히스토그램은 다항분포에서의 두 표본이 된다. 이 때 두 히스토그램의 동질성 검정은 두 다항분포 표본이 동일한 모수(proportion parameter)에서 추출된 것인지를 검정하는 것과 같으므로, 다항분포의 카이제곱 검정과 같은 방법을 이용할 수 있다.

한편, 최근 많은 데이터 정보들에 대한 접근 가능해지면서, 서로 다른 정보들을 결합했을 때 개인 정보의 노출 가능성에 대한 연구가 이루어지고 있다 (Homer 등, 2008). 이러한 가능성을 정량적으로 설명하면서 (formal privacy) 유계를 정해 놓은 것이 차등정보보호의 개념이다. 차등정보보호를 만족하도록 설계된 메커니즘은 이용자가 요구하는 값에 잡음을 추가하여 약간 다른 값을 생성하며 차등정보보호는 참값을 제공할 확률(가능도 함수)을 이용해 정의된다. α-차등정보보호란 개체 하나만 다른 두 개의 데이터베이스가 있을 때 차등정보보호 메커니즘을 각 데이터베이스에 적용하여 특정 값이 얻어질 가능도 함수의 비율이 언제나 α 이하가

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C2002256).
This paper is part of the first author's master thesis from the Seoul National University.
[1] Corresponding author: Department of Statistics, Seoul National University, 1 Gwanak-Ro, Gwanak-Gu, Seoul 08826, Korea. E-mail: sungkyu@snu.ac.kr

Published 28 February 2022 / journal homepage: http://kjas.or.kr

되는 것을 말한다. 차등정보보호 메커니즘은 차등정보보호를 만족하도록 원자료를 변형하는 방식을 의미하고 이 메커니즘을 적용한 자료를 α-차등정보보호자료라고 한다 (Dwork 등, 2006a). 미국과 호주 등에서는 개인 정보 보호를 위해 원자료 대신에 이러한 차등정보보호 자료를 공개하는 것을 고려하고 있다.

히스토그램에 차등정보보호를 적용하는 메커니즘은 대부분 원자료를 통해 얻은 원 히스토그램에 잡음(noise)을 더하는 것이다. 그러나 차등정보보호 히스토그램의 동질성 검정에 기존에 사용하던 카이제곱 검정을 적용할 경우, 검정통계량이 더 이상 카이제곱 분포를 따르지 않아 검정 오류가 증가하는 문제가 발생한다. 따라서, 차등정보보호가 적용된 히스토그램들의 동질성 검정을 위해서는 새로운 검정 방법이 요구된다.

본 논문에서는 새로운 검정 방법으로 모수적 부트스트랩(parametric bootstrap)을 이용할 것을 제안하고 미국과 한국의 연령별 인구분포 자료에 이를 적용하여 그 실용성(utility)을 보인다. 특히, 차등정보보호 자료의 동질성 검정을 위한 기존의 연구들과의 차별성은 원자료에 대한 정보가 전혀 주어지지 않고 차등정보보호의 수준 α와 차등정보보호 자료만이 주어져 있는 상황에서 사용할 수 있다는 점에 있다. 또한, 동질성 검정에 사용되는 두 히스토그램에 서로 다른 차등정보보호의 수준이 적용되었을 때에도 사용할 수 있다.

본 논문은 총 5장으로 구성되어 있다. 2장에서는 전통적인 히스토그램 동질성 검정, 차등정보보호, 그리고 히스토그램에 차등정보보호를 적용하는 방법에 대해서 다룬다. 또한, 차등정보보호가 적용된 히스토그램에 기존의 동질성 검정이 사용될 수 없음을 확인한다. 3장에서는 차등정보보호 히스토그램의 동질성 검정을 다룬 기존의 논문들을 살펴본다. 4장에서는 본 논문에서 제안하고 있는 모수적 부트스트랩 검정 방법을 설명하고, 미국과 한국의 연령별 인구 분포 자료를 이용하여 그 성능과 유용성을 보인다. 마지막으로 5장에서 결론을 맺는다.

2. 동질성 검정과 차등정보보호

이 장에서는 전통적으로 사용되어 온 히스토그램에 대한 동질성 검정을 살펴본다. 그리고 차등정보보호의 확률적 정의와 차등정보보호가 적용된 히스토그램을 생성하는 방법을 설명한다. 마지막으로 전통적인 카이제곱 동질성 검정이 차등정보보호 히스토그램에 적용될 수 없음을 보인다.

2.1. 히스토그램의 동질성 검정

본 논문에서는 두 히스토그램의 구간의 길이와 개수가 동일함을 가정한다. 따라서, 각 구간을 범주로 보면 아래와 같이 구간이 M개인 각각의 히스토그램을 동일한 범주에 대한 다항분포에서의 표본으로 생각할 수 있다.

$$C_i = (C_{i1}, \ldots, C_{ij}, \ldots, C_{iM})^t \sim \text{Multinomial}\,(N_i, p_i)\,, \quad i = 1, 2, \quad j = 1, \ldots, M. \tag{2.1}$$

여기서 C_{ij}는 i번째 히스토그램의 j번째 구간의 도수(count)를 의미하고 N_i와 p_i는 i번째 히스토그램의 총 도수 $\sum_{j=1}^{M} C_{ij}$ 와 모수(proportion parmameter)를 의미한다. 즉, 두 히스토그램의 동질성 검정은 아래와 같은 두 다항분포 표본에 대한 모수의 동질성 검정과 동일하다.

$$\begin{aligned} &H_0 : p_1 = p_2 \quad \textbf{vs} \quad H_1 : p_1 \neq p_2 \\ &\text{for} \quad C_1 \sim \text{Multinomial}\,(N_1, p_1)\,, \quad C_2 \sim \text{Multinomial}\,(N_2, p_2)\,. \end{aligned} \tag{2.2}$$

또한, 해당 검정에 사용되는 검정통계량은 아래와 같다.

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{M} \frac{\left(C_{ij} - E_{ij}\right)^2}{E_{ij}} \quad \text{where} \quad E_{ij} = \frac{N_i \left(C_{1j} + C_{2j}\right)}{N_1 + N_2}. \tag{2.3}$$

귀무가설(null hypothesis) 하에서 주어진 검정 통계량은 근사적으로 자유도 $M-1$의 카이제곱 분포를 따른다. 따라서, 두 히스토그램의 동질성 검정은 주어진 히스토그램들을 사용하여 검정통계량을 계산한 후 카이제곱 분포 하에서의 P-value를 계산하여 주어진 유의수준보다 작은 경우 귀무가설을 기각한다.

2.2. 차등정보보호

α-차등정보보호란 하나만 다른 두 개의 데이터베이스 중 한 데이터베이스의 정보를 가진 외부인이 다른 데이터베이스를 공격한다는 상당히 엄격한 가정하에서 공개된 자료에서 개인의 정보가 노출되는 위험을 제어하기 위하여 해당 자료에 적용하는 기술적 조치이다. α-차등정보보호에 대한 정의는 다음과 같다 (Dwork, 2006b; Park 등, 2018).

Definition 1. *데이터베이스 D_1과 D_2를 고려하고, 각각의 데이터베이스는 하나의 개체만 다르고 동일하다고 하자. 임의의 개인정보 노출 제한 방법을 확률적 함수 K라고 표현할 때, $K(D)$는 노출 제한 방법이 적용된 공개용 자료가 된다. 집합 S는 노출제한 방법이 적용된 모든 가능한 자료의 집합을 나타낸다.*

이러한 가정 하에서, α-차등정보보호란 다음과 같은 조건을 만족하는 것이다.

$$\log \frac{P[K(D_1)\in S]}{P[K(D_2)\in S]} \le e^{\alpha} \approx 1+\alpha \quad \textit{for all } S \subseteq \textit{range}(K).$$

해당 정의에서 D_1과 D_2를 한 개의 구간에만 대하여 도수가 1만큼 차이가 있는 거의 같은 두 히스토그램으로 생각하면 $K(D_1)$과 $K(D_2)$는 K라는 함수를 이용하여 각각에 차등정보보호를 적용한 두 자료로 볼 수 있다. 이러한 맥락에서 히스토그램에 대한 차등정보보호란 한 사람의 포함 여부가 다른 두 히스토그램에 차등정보보호를 적용했을 때 각각으로부터 얻을 수 있는 정보의 차이가 정해진 수준 이하여야 함을 의미한다. 이 때, α가 작을수록 두 차등정보보호 자료의 분포에는 차이가 없으므로 작은 α 값은 더 강한 개인 정보 보호를 의미한다.

이러한 차등정보보호가 개인정보 노출을 제어하는 데에 도움이 됨은 다음 예시에서 확인할 수 있다. 한국에서 가장 부유한 자산가의 이름과 한국인들의 자산의 히스토그램 자료가 공개되었다고 가정하자. 이 경우 두 자료를 결합하면 해당 자산가의 자산 수준이 노출된다. 그러나 히스토그램에 차등정보보호를 적용하여 잡음이 더해진 자산 자료를 배포하는 경우 해당 개인의 자산 정보가 보호될 수 있다.

기존의 여러 연구에서 히스토그램에 적용할 수 있는 차등정보보호 방법을 다양하게 제시하고 있는데, 그 중에서 본 논문에서 사용한 방법은 '보정된 교란 히스토그램'이다 (Wasserman과 Zhou, 2010; Park 등, 2019). 해당 방법은 기본적으로 히스토그램의 도수에 라플라스 잡음을 첨가하는 '교란' 방법에 해당한다. 그런데 도수가 너무 작거나 잡음으로 절대값이 큰 음수가 더해졌을 때 차등정보보호 히스토그램의 도수가 음수가 될 수 있다. 이 경우 음의 도수를 0으로 보정 하는 것이 '보정된 교란 히스토그램' 방법이며 아래와 같이 표현된다.

$$X_j = \max\left\{0, C_j + Z_j\right\}, \quad Z_j \sim \text{Lap}\left(\frac{2}{\alpha}\right), \quad j = 1,\ldots,M. \tag{2.4}$$

$$C = (C_1,\ldots,C_M)^t, \quad X = (X_1,\ldots,X_M)^t \in \mathbb{R}^M.$$

여기서 C_j 는 원자료 히스토그램의 j번째 구간의 도수를, X_j는 차등정보보호 히스토그램의 j번째 도수를 의미한다. Z_j는 첨가된 라플라스 잡음을 의미한다. Wasserman과 Zhou (2010)는 이 방법을 적용한 히스토그램이 임의의 α에 대하여 α-차등정보보호 자료임을 증명하였다. 한편 라플라스 잡음 대신에 가우시안(Gaussian) 잡음을 사용할 수도 있다 (Dwork 등, 2006a,b). 그러나, 더 약한 정보보호 수준인(α,δ) 차등정보보호를 하기 위해 사용되는 가우시안 잡음의 분산이 α 차등정보보호를 위해 사용되는 라플라스 잡음의 분산보다 더 크기 때문에 차등정보보호 자료의 유용성이 더 낮아지게 된다 (Gaboardi 등, 2016). 본 논문에서 이후에 차등정보보호를 적용하였다는 것은 라플라스 잡음을 이용한 보정된 교란 방법을 적용하였음을 의미한다.

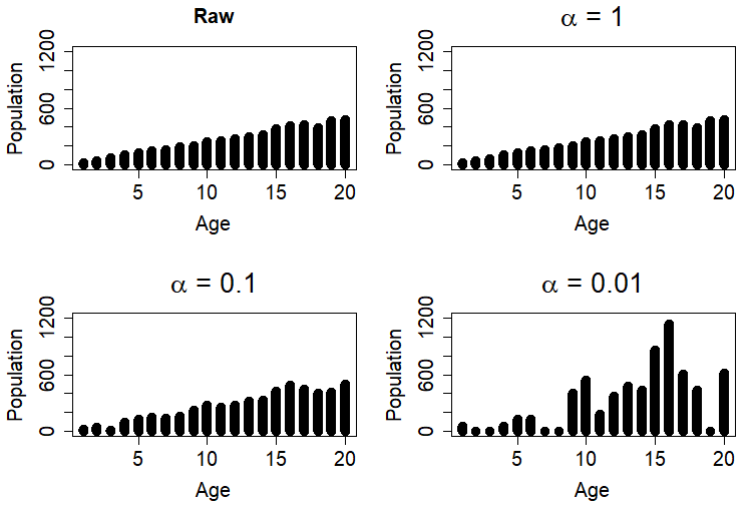


Figure 1: *Original and Differentially private histograms.*

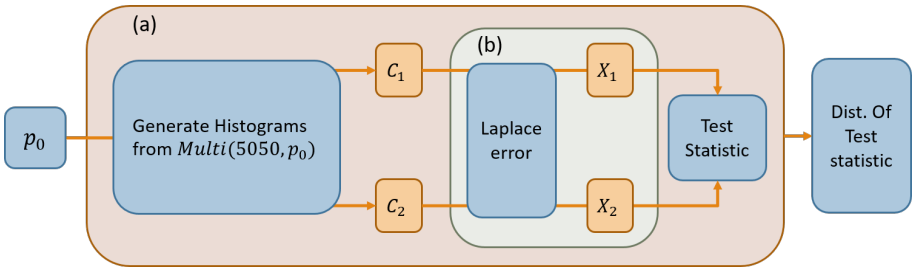


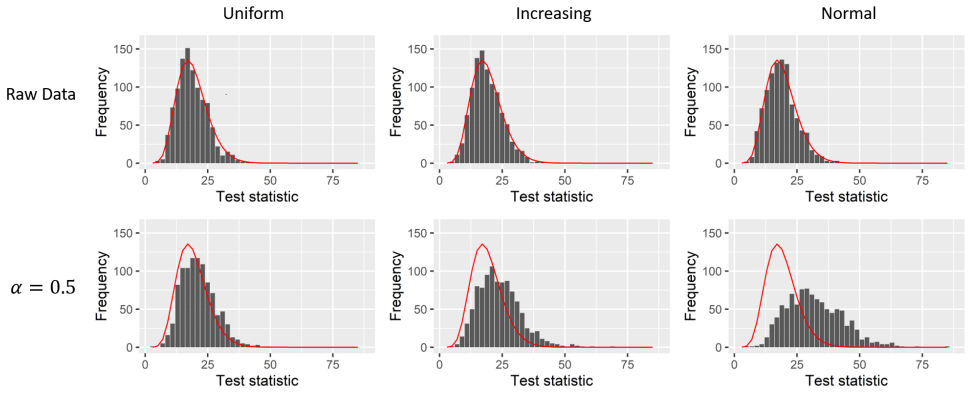
Figure 2: *Framework to check the distribution of test statistic of the traditional chi-square test.*

2.3. 차등정보보호 히스토그램에의 전통적 동질성 검정의 적용

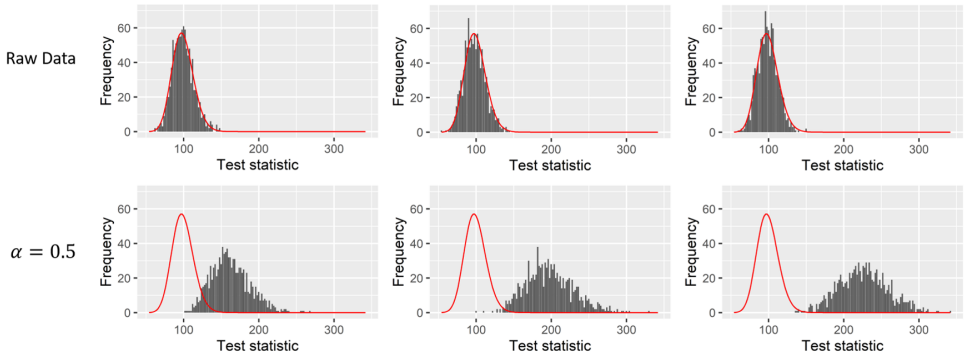
이 절에서는 새로운 동질성 검정 방법이 필요함을 확인하기 위해, 차등정보보호 히스토그램의 동질성 검정에 전통적인 카이제곱 통계량을 사용할 수 없음을 보이고자 한다.

먼저, 차등정보보호를 적용하였을 때 히스토그램의 분포가 어떻게 달라지는지 살펴보자. 이를 위하여 2020년 한국의 연령별 인구분포 데이터를 사용하였으며 이 자료는 4.2절에서도 사용된다. Figure 1은 해당 데이터의 원 히스토그램과 각기 다른 수준의 차등정보보호를 적용하였을 때의 히스토그램을 보여준다. 약한 수준의 차등정보보호($\alpha = 1$)에서는 원 히스토그램과 차등정보보호 히스토그램의 분포가 거의 같은 것을 볼 수 있다. 그러나 강한 차등정보보호를 적용하면 차등정보보호 히스토그램이 원 히스토그램과 상당히 다른 분포를 따르게 된다. 또한, 차등정보보호를 적용하였을 때 히스토그램의 총 도수(N) 역시 달라짐을 알 수 있다.

이제, 차등정보보호가 적용된 히스토그램에 대한 동질성 검정에서 검정통계량의 분포를 살펴보자. 검정통계량은 귀무가설 하에서 일정한 분포를 따르는 성질을 가져야 한다. 해당 성질을 갖지 않으면 검정통계량에 기반한 P-value를 안정적으로 계산할 수 없기 때문이다. 즉, 각 히스토그램이 생성된 두 다항분포의 모수가 동일하다면 검정통계량은 모수 값과 관계없이 항상 분포가 일정해야 한다. 그러나 다양한 모수에서 생성된 두 히스토그램에 차등정보보호를 적용하였을 때 전통적인 카이제곱 통계량이 더 이상 카이제곱 분포를 따르지



(a) $M = 20$



(b) $M = 100$

Figure 3: *Null distributions of test statistics.*

않을 뿐만 아니라 일정한 분포를 따르지 않음을 Figure 2와 같은 모의실험을 통해 확인할 수 있다.

모의실험은 길이가 같은 M개의 구간을 가진 두 히스토그램에 대해서 시행되었다. M개의 구간에 대해서, 주어진 $p_0 \in \mathbb{R}^M$ 를 모수로 하는 다항분포를 생각하자. 이러한 다항분포에서 도수의 합이 5,050이 되도록 두 개의 히스토그램을 생성하고, $\alpha = 0.5$ 수준의 차등정보보호를 적용한 뒤 앞에서 다룬 카이제곱 통계량을 계산한다. 이러한 과정을 (a)라고 할 때 이를 1000번을 반복하면 귀무가설 하에서의 1000개의 검정통계량을 얻게 되는데, 이 값들에 대하여 히스토그램을 그려서 검정통계량의 분포를 확인하였다. 관심 있는 특성치가 연속형 변수인 경우 정의역의 일부를 M개의 일정한 구간으로 나누어서, j번째 구간에 해당하는 확률 $p_{0j}, j = 1, \ldots, M$ 을 계산하였다. $p_0 = (p_{01}, \ldots, p_{0M})'$에 대하여 다음과 같은 3가지 경우를 고려하였다.

- Uniform : $p_{0j} = \frac{1}{M}, \quad j = 1, \ldots, M.$
- Increasing : $p_{0j} = \frac{j}{\sum_{i=1}^{M} i}, \quad j = 1, \ldots, M.$
- Normal : $p_{0j} = \frac{\phi\left(-3+\frac{6}{M-1}\times(j-1)\right)}{\sum_{i=1}^{M}\phi\left(-3+\frac{6}{M-1}\times(i-1)\right)}$, ϕ는 $N(0,1)$의 밀도함수, $\quad j = 1, \ldots, M.$

Uniform은 모수를 설정할 때 모든 구간의 기대 도수가 일정하도록 한 것을, Increasing은 각 구간의 기대

도수가 점차 증가하는 모양이 되도록 한 것을 나타낸다. Normal의 경우에는 –3과 3을 양 끝점으로 하여 거리가 균등한 M개의 점을 생각하고, 해당 점에서의 $N(0,1)$의 밀도함수 값을 가중치로 이용하여 p_0를 정하였다. 그 결과, 검정통계량의 분포는 Figure 3과 같다.

Figure 3의 (a)와 (b)는 총 도수를 5,050으로 고정한 후 구간의 수 M을 각각 20과 100으로 다르게 했을 때의 히스토그램이다. 각 열은 왼쪽부터 모수가 Uniform, Increasing, Normal인 경우 검정통계량의 분포이다. (a)와 (b) 각각의 첫번째 행은 Figure 2의 (b)부분을 적용하지 않아 차등정보보호가 적용되지 않은 원 히스토그램에 대한 검정통계량의 분포를 보여준다. (a)와 (b)에서 붉은 선은 각각 자유도 19와 99의 카이제곱 분포를 나타내는데 각 첫번째 행에서는 히스토그램이 해당 분포를 따르고 있음을 알 수 있다. 반면, 차등정보보호 히스토그램에 대한 검정통계량의 분포를 보여주는 두번째 행에서는 분포가 전반적으로 오른쪽으로 이동하고 산포가 커졌으므로 더 이상 카이제곱 분포를 따르지 않음을 알 수 있다. 또한 통계량의 분포 변화 정도가 모수에 따라서 달라지고 있으므로 차등정보보호 히스토그램에 대해서는 기존의 카이제곱 동질성 검정을 그대로 사용하기 어렵다고 판단된다.

한편, 구간의 수가 늘어남에 따라서 검정통계량의 분포에 더 큰 변화가 나타난 것을 알 수 있다. 이는 총 도수가 고정되었을 때 구간의 수가 증가하면 각 구간 내의 도수가 줄어들어 잡음의 영향이 증가하기 때문이다. 따라서 같은 수준의 차등정보보호임에도 불구하고 구간의 수가 증가함에 따라 검정통계량에 더 큰 변화가 나타나게 된다.

3. 기존 연구

앞 절에서는 차등정보보호 히스토그램의 동질성 검정에 대한 새로운 검정 방법의 필요성을 논의하였다. 이 절에서는, 기존의 다른 문헌에서도 이러한 논의를 찾아볼 수 있어 여기서 간단히 소개한다. Wang 등 (2017)에서는 차등정보보호를 적용할 때 두 다항분포 표본에 대한 독립성과 동질성 검정 및 하나의 다항분포 표본에서 적합성 검정(goodness of fit test)을 시행하는 방안을 다루고 있다. 먼저 차등정보보호를 적용한 다항분포 표본을 사용하여 전통적인 카이제곱 통계량을 계산하면 적절한 조건 하에서 이 통계량이 라플라스 분포를 따르는 확률 변수와 특정 모수의 다항분포 확률 변수에 관한 식으로 정리될 수 있음을 보였다. 다음으로 라플라스 분포와 해당 모수의 다항분포 확률 변수를 생성하여 통계량을 계산하는 과정을 반복하여 주어진 차등정보보호 표본의 통계량의 경험적 P-value를 주어진 유의수준과 비교하여 동질성 검정을 수행할 것을 제안하고 있다.

한편, Gaboardi 등 (2016)에서는 차등정보보호 다항분포 표본의 독립성 검정과 적합성 검정을 제안하고 있다. 제안된 검정 중에 동질성 검정과 유사한 독립성 검정을 살펴보면 L_1 loss와 L_2 loss를 조합한 형태의 목적함수 (Lee 등, 2015)를 사용하여 차등정보보호 표본으로부터 모수를 추정한 후 추정된 모수를 이용하여 전통적인 독립성 검정통계량을 계산한다. 그 후에, 추정된 모수로부터 새로 다항분포 표본을 생성하여 앞과 동일하게 새로 모수를 추정하고 검정통계량을 계산하는 과정을 반복한다. 이렇게 계산된 검정통계량들을 사용하여 (1– 주어진 유의수준) × 100(%) 분위수를 계산하고 처음에 계산한 통계량이 이 분위수보다 크면 귀무가설을 기각한다.

기본적으로 두 논문은 주어진 차등정보보호 다항분포 표본에 대하여 원 표본의 총합($N_i = \sum_{j=1}^{M} C_{ij}$) 을 알고 있다고 가정한다. 그러나 차등정보보호를 적용한 자료의 총합 $n_i = \sum_{j=1}^{M} X_{ij}$는 원자료의 총합 N_i와 같지 않으며, $\sum_{j=1}^{M} \max\{0,\ C_{ij} + Z_{ij}\}$으로부터 얻어짐을 주지할 필요가 있다. 따라서, 원자료를 공개하지 않고 차등정보보호 자료 만을 공개한다고 할 때 원자료의 총합을 안다는 가정은 비현실적이다. 또한, Wang 등 (2017)은 차등정보보호 다항분포 표본의 일부가 5 미만인 경우를 가정하지 않고 있으며 Gaboardi 등 (2016)은 해당 경우에 대한 별도의 설명없이 귀무가설을 기각하지 않는다는 결론을 내리고 있다. 따라서 원 다항분포 표본의 일부 구간의 도수가 매우 작거나 도수에 더해지는 잡음이 큰 음의 값을 가진다면 검정을 할 수 없거나 귀무가설을 항상 기각하지 않는 문제가 발생할 수 있다.

본 논문에서는 이러한 문제를 해결하기 위하여 원자료에 대한 정보를 전혀 사용하지 않고 차등정보보호 히스토그램만을 가지고 수행하는 동질성 검정을 제안한다. 즉, 차등정보보호가 적용된 히스토그램과 차등정보보호 수준 α만이 접근 가능한 정보라고 가정한다. 또한, 제안하는 동질성 검정은 도수가 5 미만인 것과 관계없이 항상 검정 가능하며 차등정보보호 자료를 이용하여 실제 도수를 참값과 가깝게 추정한다.

4. 모수적 부트스트랩을 이용한 동질성 검정

4.1. 이론적 구조

본 논문에서 검정하고자 하는 가설과, 전통적으로 해당 가설을 검정하기 위해 사용하는 검정통계량을 정리하면 다음과 같다.

$$H_0 : p_1 = p_2 \quad \textbf{vs} \quad H_1 : p_1 \neq p_2 \tag{2.2}$$
$$\text{for} \quad C_1 \sim \text{Multinomial}\,(N_1, p_1)\,, \quad C_2 \sim \text{Multinomial}\,(N_2, p_2)\,.$$
$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{M} \frac{(C_{ij} - E_{ij})^2}{E_{ij}} \quad \text{where} \quad E_{ij} = \frac{N_i(C_{1j} + C_{2j})}{N_1 + N_2}. \tag{2.3}$$

위의 귀무가설 하에서의 통계량을 계산하고자 할 때 필요한 값은 N_i 와 C_{ij}이다. 그러나 차등정보보호 자료만을 제공받는 경우 주어지는 정보는 차등정보보호 수준 α_i와 아래와 같이 계산된 X_{ij} 뿐이다.

$$X_{ij} = \max\left\{0, C_{ij} + Z_{ij}\right\}, \quad Z_{ij} \sim \text{Lap}\left(\frac{2}{\alpha_i}\right), \quad i = 1, 2, \quad j = 1, \ldots, M. \tag{2.4}$$
$$X_i = (X_{i1}, \ldots, X_{iM})^t \in \mathbb{R}^M, \quad i = 1, 2.$$

이러한 가정 하에서 본 논문에서 제안하고 있는 모수적 부트스트랩을 이용한 검정 방법의 구조는 다음과 같다. 먼저, 주어진 차등정보보호 자료인 X_1과 X_2로부터 귀무가설 하에서의 각 구간의 비율에 대응되는 다항분포의 모수 p_0와 N_i를 추정한다. 그리고 이 $\widehat{p_0}$과 $\widehat{N}_i$으로부터 두 히스토그램 U_1과 U_2를 생성하고 차등정보보호를 적용해서 귀무가설 하에서의 검정통계량을 계산하는 과정을 B번 반복한다. 이렇게 얻은 B개의 검정통계량에 대하여 $(1 - 0.05) \times 100(\%)$ 분위수를 계산하고, 처음에 주어진 X_1과 X_2를 이용하여 계산한 통계량이 이 분위수보다 크면 귀무가설을 기각한다.

이때 핵심이 되는 부분은 p_0와 N_i를 추정하고 U_1과 U_2를 생성하는 과정이므로 이를 상세히 설명하면 다음과 같다. 앞서 언급한 바와 같이 귀무가설 하에서의 p_0를 계산하고자 할 때 필요한 값은 N_i 와 C_{ij}이나, 필요한 값이 모두 교란되어 알 수 없는 상황이다. 즉, 주어진 자료 중 C_{ij}에 대한 정보를 가진 값은 차등정보보호 자료인 X_{ij}가 유일하다. 따라서, 적률추정량의 정의를 X_{ij}에 적용하여 C_{ij}와 N_i를 추정하고 이를 이용하여 p_0를 추정하는 것이 자연스럽다. 먼저, 차등정보보호가 적용된 i번째 히스토그램의 j번째 구간의 도수인 X_{ij}의 조건부 기댓값을 계산해보면 아래와 같다.

$$E(X_{ij}|C_{ij}) = E\left(\max\left\{0, C_{ij} + Z_{ij}\right\}\middle|C_{ij}\right) = C_{ij} + \frac{1}{\alpha}\exp\left(-\frac{\alpha C_{ij}}{2}\right) =: g(C_{ij}). \tag{4.1}$$

만약 원자료의 빈도인 C_{ij}가 충분히 크다면 뒤의 오차항이 0에 가까워서 X_{ij}를 C_{ij}의 추정값으로 사용할 수 있다. 그러나 C_{ij}가 충분히 크지 않다면 오차항이 유의미하게 0보다 크게 된다. 따라서 X_{ij}를 C_{ij}의 추정값으로 쓰게 되면 실제 C_{ij}보다 큰 값으로 과대추정을 하게 된다. 이를 면밀히 살펴보기 위하여 Figure 4는 위에 정의한 함수 $g(C_{ij})$의 그래프를 각 차등정보보호 정도에 따라 나타내고 있다.

Figure 4는 왼쪽 위부터 Z방향으로 α가 1, 0.1, 0.05, 0.01인 경우이며, 각 그래프에서 검정색 선과 빨간색 선은 $g(C_{ij})$와 C_{ij}의 값을 각각 나타낸다. 앞서 언급한 바와 같이 C_{ij}가 작을 때, X_{ij}의 기댓값은 실제 C_{ij}값보다

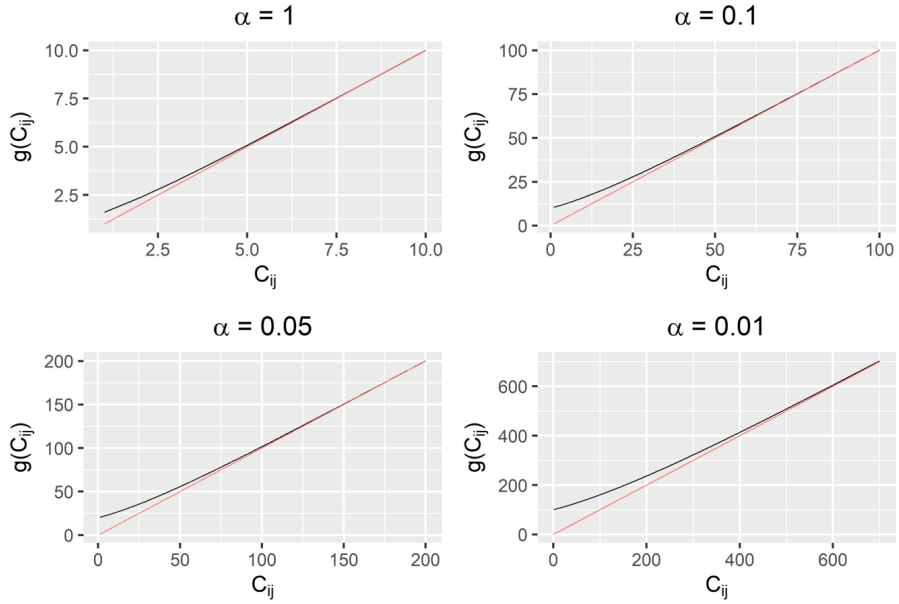


Figure 4: *Conditional Expectation of X_{ij}. Black curve represents $g(C_{ij})$; red curve represents C_{ij}.*

유의미하게 큰 것을 알 수 있다. 본 논문에서는 이러한 오차를 줄이기 위하여 $g(C_{ij})$의 역함수를 이용하고자 한다. 우선, 아래와 같이 $\widetilde{C_{ij}}$를 정의한다.

$$\widetilde{C_{ij}} = \begin{cases} 0, & \text{if } X_{ij} \le \frac{1}{\alpha_i} \\ g^{-1}(X_{ij}), & \text{if } \frac{1}{\alpha_i} < X_{ij} < k_{ij}(\alpha_i) \\ X_{ij}, & \text{otherwise} \end{cases} \tag{4.2}$$

$$\text{where} \quad k_{ij}(\alpha_i) = \arg\min_{C_{ij}} \left| C_{ij} - g(C_{ij}) \right| < 5.$$

식 4.1을 보면, C_{ij} 가 0 이하인 경우, X_{ij}의 기대값은 $1/\alpha_i$보다 작다. C_{ij}는 0 보다 작을 수 없으므로 X_{ij}가 $1/\alpha_i$ 보다 작은 경우에는 $\widetilde{C_{ij}}$를 0으로 정의한다. 그리고 실제 C_{ij}와 X_{ij}의 기댓값의 차이가 5보다 작으면 X_{ij}를 $\widetilde{C_{ij}}$의 값으로 사용하도록 하고 그 외의 경우에는 $g^{-1}(X_{ij})$의 값을 사용한다.

이와 같은 $\widetilde{C_{ij}}$를 이용하여 최종적으로 N_i, C_{ij}, p_0는 아래와 같이 추정할 수 있다.

$$\widehat{N_i} = \left\lfloor \sum_{j=1}^{M} \widetilde{C_{ij}} \right\rfloor, \quad \widehat{C_{ij}} = \widetilde{C_{ij}} \times \frac{\widehat{N_i}}{\sum_{j=1}^{M} \widetilde{C_{ij}}}, \quad \widehat{p_0} = \frac{\widehat{C_1} + \widehat{C_2}}{\widehat{N_1} + \widehat{N_2}}. \tag{4.3}$$

여기서 $\lfloor x \rceil$는 x에서 가장 가까운 정수를 의미한다. $\widetilde{C_{ij}}$를 바로 사용하지 않고 위의 과정을 거치는 이유는 $\widehat{N_i}$를 자연수로 추정하기 위해서이다.

마지막으로, Multinomial$(\widehat{N_1}, \widehat{p_0})$과 Multinomial$(\widehat{N_2}, \widehat{p_0})$에서 U_1과 U_2를 생성하게 된다. Algorimthm 1과 Figure 5는 지금까지 설명한 과정을 나타낸다.

한편, 이와 같은 $\widehat{C_{ij}}$을 사용하여 기존의 카이제곱 검정을 수행하는 것을 생각해볼 수 있다. 구간 내의 빈도가 충분히 큰 경우에는 차등정보보호 히스토그램의 도수인 X_{ij}도 크기 때문에 $X_{ij} = \widetilde{C_{ij}} = \widehat{C_{ij}}$가 될 것이므로 Figure 3과 같은 결과를 얻는다. 즉, 제 1종 오류 확률이 통제되지 않는다. 또한, 작은 표본 수나 구간 수의 증가

Algorithm 1 Framework for parametric bootstrap test

procedure ParaBoot $(X_1, X_2, \alpha_1, \alpha_2, B)$

$T = \sum_{i=1}^{2}\sum_{j=1}^{M} \frac{(X_{ij}-E_{ij})^2}{E_{ij}}$ where $E_{ij} = \frac{n_i(X_{1j}+X_{2j})}{n_1+n_2}$, $n_i = \sum_{j=1}^{M} X_{ij}$ $i = 1, 2$, $j = 1, \ldots, M$

$\widetilde{C_{ij}} = \begin{cases} 0, & \text{if } X_{ij} \le \frac{1}{\alpha_i} \\ g^{-1}(X_{ij}), & \text{if } \frac{1}{\alpha_i} < X_{ij} < k_{ij}(\alpha_i) \\ X_{ij}, & \text{otherwise} \end{cases}$ where $k_{ij}(\alpha_i) = \arg\min_{C_{ij}} \left| C_{ij} - g\left(C_{ij}\right) \right| < 5$

$\widehat{N_i} = \left\lfloor \sum_{j=1}^{M} \widetilde{C_{ij}} \right\rceil$, $\lfloor x \rceil$ = (nearest integer from x)

$\widehat{C_{ij}} = \widetilde{C_{ij}} \times \frac{\widehat{N_i}}{\sum_{j=1}^{M} \widetilde{C_{ij}}}$

$\widehat{p_0} = \frac{\widehat{C_1}+\widehat{C_2}}{\widehat{N_1}+\widehat{N_2}}$

for $b \in [B]$ **do**

Generate U_1 and U_2 from Multinomial$\left(\widehat{N_i}, \widehat{p_0}\right)$

$Y_{ij} = \max\left\{0, U_{ij} + Z_{ij}\right\}$, where $Z_{ij} \overset{i.i.d.}{\sim} \text{Lap}\left(\frac{2}{\alpha_i}\right)$

$T_b = \sum_{i=1}^{2}\sum_{j=1}^{M} \frac{(Y_{ij}-E_{ij})^2}{E_{ij}}$, where $E_{ij} = \frac{\widehat{N_i}(Y_{1j}+Y_{2j})}{\widehat{N_1}+\widehat{N_2}}$

P-value $= \sum_{b=1}^{B} \frac{I(T_b > T)}{B}$

return P-value

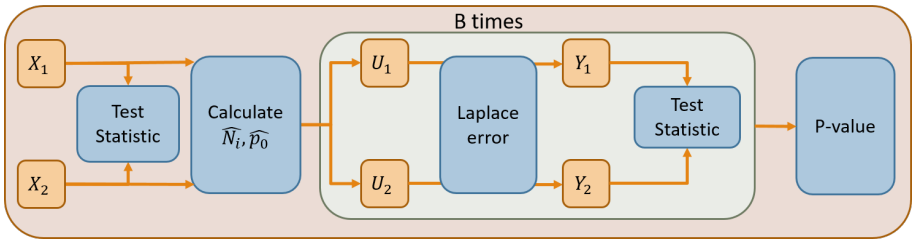


Figure 5: *Framework for parametric bootstrap test.*

등으로 인해 개별 구간 내의 빈도가 작은 경우에도 X_{ij} 보다 좋은 추정값인 $\widehat{C_{ij}}$을 사용할 때의 제 1종 오류 확률은 여전히 통제되지 않는다.

4.2. 한국과 미국의 인구 데이터에의 적용

이 절에서는 한국과 미국의 연령별 인구 분포 데이터를 이용하여 제안한 검정 방법의 성능을 확인하고자 한다. 성능평가 기준은 제 1종의 오류가 일어나는 확률의 통제와 검정력이다. 제안한 검정 방법의 유의수준 5%에서의 제 1종 오류 확률과 검정력은 Algorithm 2 *RejectRate* 알고리즘의 방법과 Figure 6으로 추정할 수 있다.

길이가 같은 구간이 M개인 두 히스토그램, A와 B에 대한 동질성 검정을 생각하자. 먼저, 각각에서 총 도수가 N_i가 되도록 다항분포 표본을 생성한 결과를 C_1과 C_2라고 하고, 각각에 차등정보보호를 적용한 자료를 X_1과 X_2라고 정의한다. 그러면, 앞의 절에서 설명한 바와 같이 X_1과 X_2만을 이용하여 N_i와 귀무가설 하에서의 다항분포 모수인 p_0를 추정하고 P-value를 계산할 수 있다. 이 때, 만약 A와 B가 분포가 같았다면, P-value가 0.05보다 작은 경우는 제 1종의 오류를 범한 것이므로 1,000개의 P-value 중 0.05보다 작은 비율을 계산하여 제 1종 오류의 확률을 계산할 수 있다. 역으로, 사실은 A와 B는 분포가 달랐다면 P-value가 0.05보다 큰 경우는 제 2종의 오류를 범한 것이다. 따라서 전체 과정을 1000번 반복하여 얻은 1,000개의 P-value 중에서 0.05보다

Algorithm 2 Calculate the error rate

procedure RejectRate $(A, B, \alpha_1, \alpha_2, N, B)$
  for $k \in [1000]$ **do**
    Generate C_1 and C_2 from Multinomial(N, p_i) where p_1 and p_2 are the proportion of A and B.
    $X_{ij} = \max\left\{0, C_{ij} + Z_{ij}\right\}$, where $Z_{ij} \overset{i.i.d.}{\sim} \text{Lap}\left(\frac{2}{\alpha_i}\right)$
    $P - value_k = P_{ARA}B_{OOT}\,(X_1, X_2, \alpha_1, \alpha_2, B)$
  Rate $= \sum_{k=1}^{1000} \frac{I(P-value_k<0.05)}{1000}$ ▷ Type 1 error under H_0 or Power under H_1
  return Rate

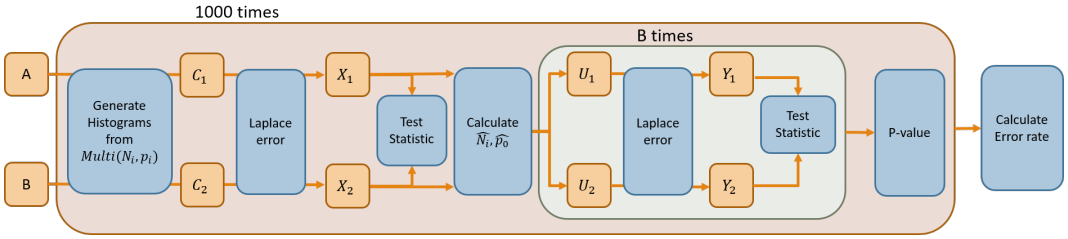


Figure 6: *Framework to measure performance of the test.*

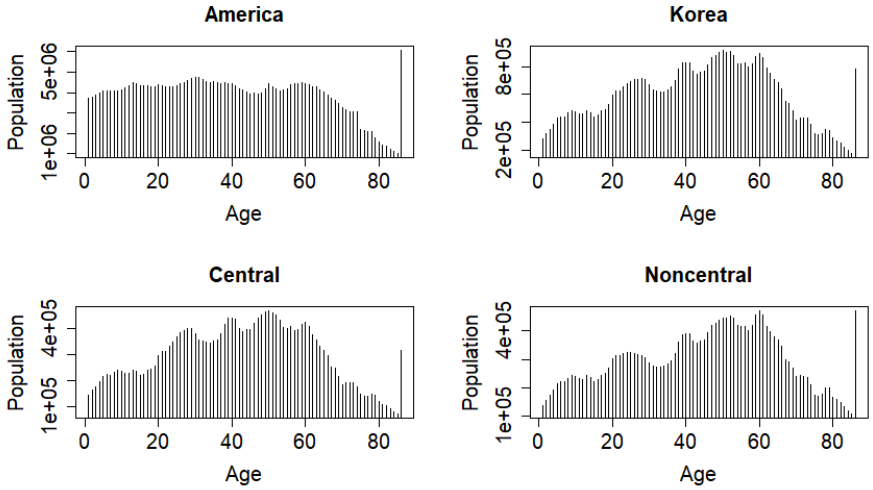


Figure 7: *Distribution of populations.*

작은 비율을 계산하면 1에서 제 2종 오류의 확률을 뺀 것이므로 검정력을 계산할 수 있다.

검정 방법의 성능을 평가하는 데에 사용된 자료는 통계청과 U.S. Census Bureau (2020)에서 발표한 한국과 미국의 2020년 연령별 인구분포 데이터이다. 먼저 Figure 7과 Table 1은 한국과 미국, 한국의 수도권과 비수도권의 인구분포를 정리한 자료이다. 여기서 수도권이란, 서울과 경기, 인천을 포함하고, 비수도권은 그 외의 지역을 포괄한다.

Figure 7에서 각각의 구간은 1세 간격으로 0세부터 84세까지는 각 구간의 빈도수이고 85세 구간은 85세 이상을 의미한다. 그래프에서 알 수 있듯이 한국과 미국의 인구분포는 눈에 띄게 다른 것을 알 수 있다. 반면에,

Table 1: The size of population

	America	Korea	Central	Noncentral
N	332,599,000	51,349,259	25,675,740	25,673,519

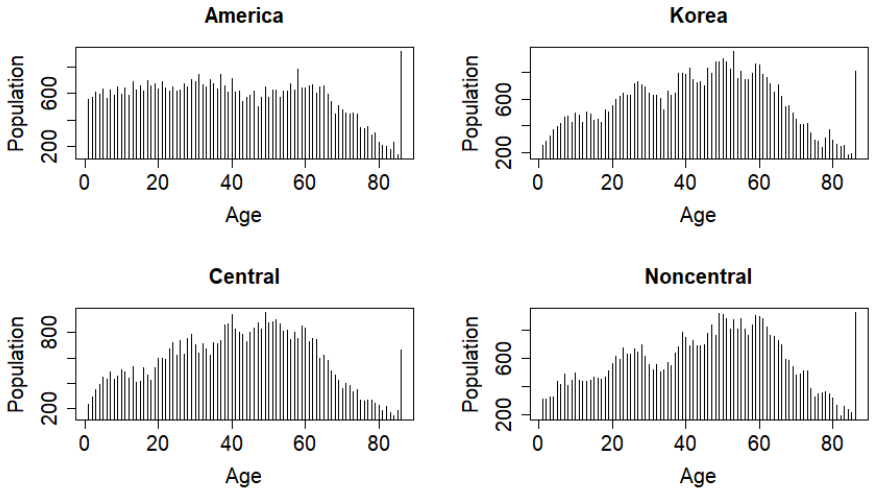


Figure 8: *Differentially private Distribution of populations.*

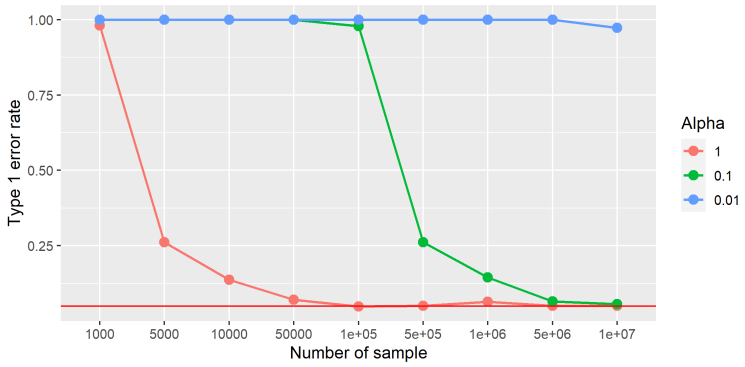
한국의 수도권과 비수도권의 경우 약간의 차이는 있으나 한국과 미국에 비하면 그 정도가 미미하다. 따라서 한국과 미국을 비교하는 경우를 효과 크기(effect size)가 큰 경우로, 한국의 수도권과 비수도권을 비교하는 경우를 효과 크기가 중간인 경우로 나누어 검정력을 평가하였다. 성능평가에 사용된 표본 수와 차등정보보호 수준, 히스토그램 구간의 수는 다음과 같다. 구간의 수는 히스토그램을 그릴 때 순서대로 1세, 5세, 10세 단위를 선택하여 얻어진 값들이다.

- 표본 수 : $N \in \left\{10^k, 5\times 10^k : k \in \{3,4,5,6,7\}\right\}$
- 차등정보보호 수준 : $\alpha \in \{1, 0.1, 0.01\}$
- 구간의 수 : $M \in \{86, 18, 9\}$

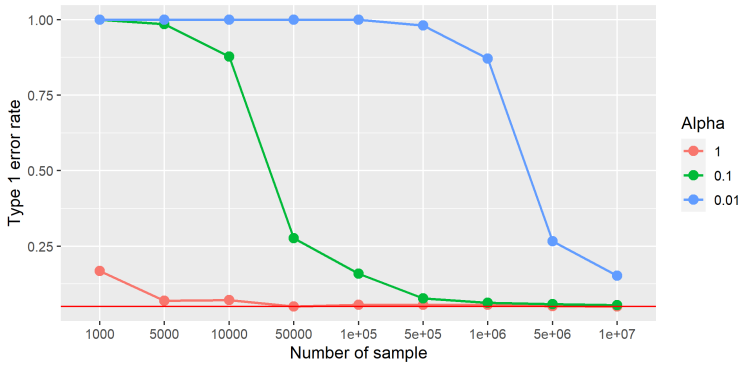
Figure 8은 Figure 7의 데이터로부터 50000개의 표본을 뽑아 히스토그램을 그린 후 중간 차등정보보호($\alpha = 0.1$)를 적용한 그림이다.

검정력을 살펴보기 전에 우선 한국의 인구분포로부터 두 개의 히스토그램을 생성해서 제 1종 오류 확률이 잘 통제되는지 확인해보았다. 먼저, X_1과 X_2에 전통적인 동질성 검정을 적용했을 때의 결과를 살펴보면 Figure 9와 같다. 그래프의 x축은 표본 수 N이고 y축은 제 1종 오류 확률이다. 이 경우 제 1종 오류가 잘 통제되지 않음을 알 수 있는데 이는 앞에서 차등정보보호를 적용했을 때에는 검정통계량의 분포가 카이제곱 분포보다 오른쪽으로 치우쳤던 것과 맥락을 같이한다.

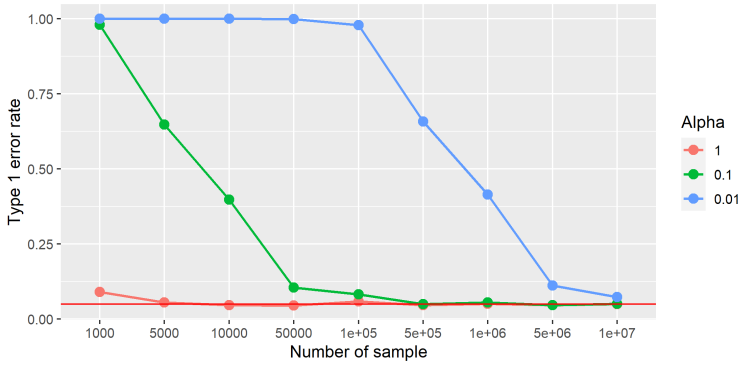
또한, 구간의 개수가 작아짐에 따라서 각 구간 내의 도수가 증가하여 잡음의 영향이 작아지는데 그 결과로서 1종 오류 확률이 낮아지는 것을 알 수 있다. 그럼에도 불구하고 1종 오류 확률을 통제하기 위해서는 히스토그램 도수의 수가 천만 이상이 되어야 함을 알 수 있다. 정리하자면, 전통적인 동질성 검정은 사용할 수 없다.



(a) $M = 86$



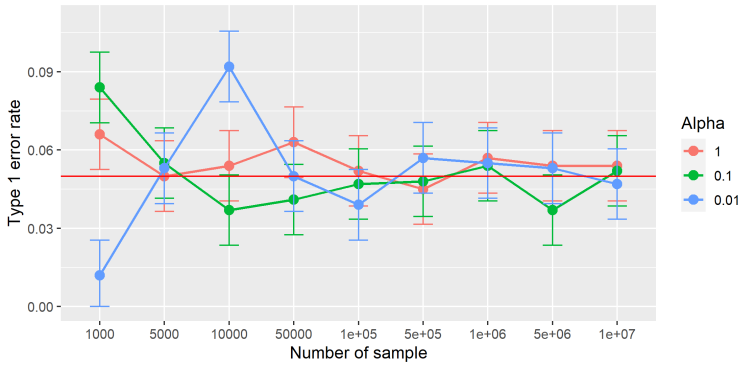
(b) $M = 18$



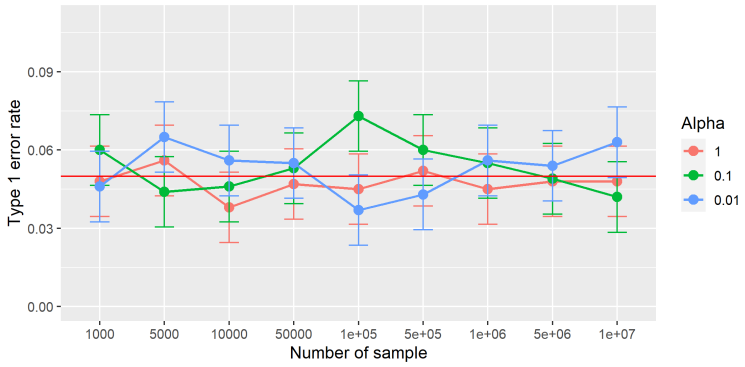
(c) $M = 9$

Figure 9: *Type 1 error rate for chi-square test.*

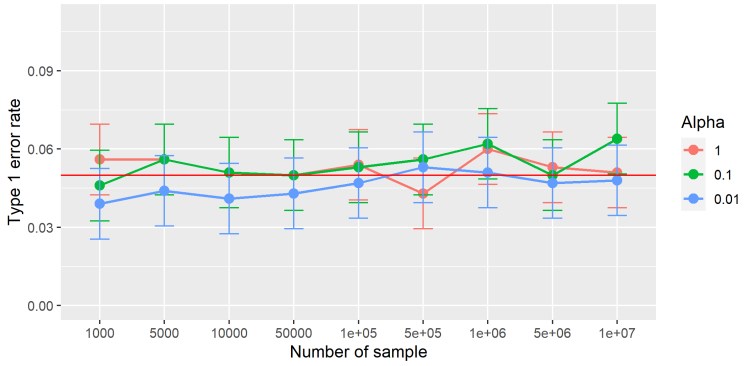
Figure 10은 새롭게 제안한 모수적 부트스트랩을 이용한 경우의 제 1종 오류 확률을 나타낸 자료이다. 각 점에서 표시된 구간은 유의수준 5%에서의 제 1종 오류 확률의 신뢰구간을 표기한 것이다. Figure 9과 비교했을 때 눈에 띄게 제 1종 오류가 잘 통제되고 있음을 알 수 있다. 특히, 구간의 갯수가 86개로 많고 강한



(a) $M = 86$



(b) $M = 18$

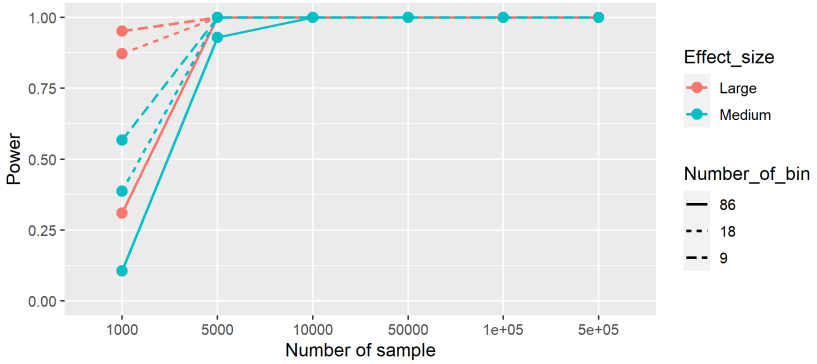


(c) $M = 9$

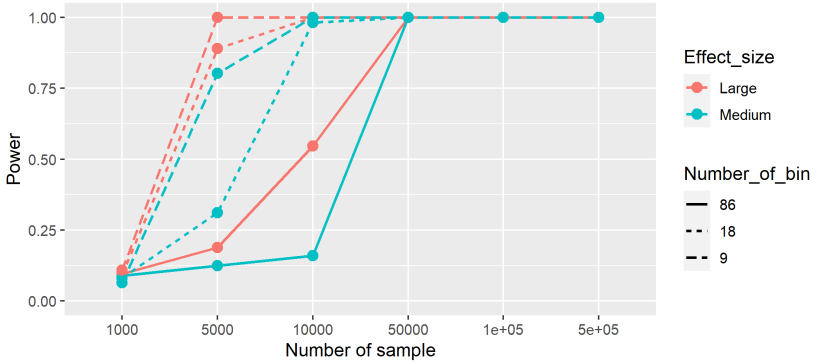
Figure 10: *Type 1 error rate for parametric bootstrap test.*

차등정보보호가 적용된 경우에도 표본수가 50,000개 이상이면 제 1종 오류가 잘 통제되고 있다.

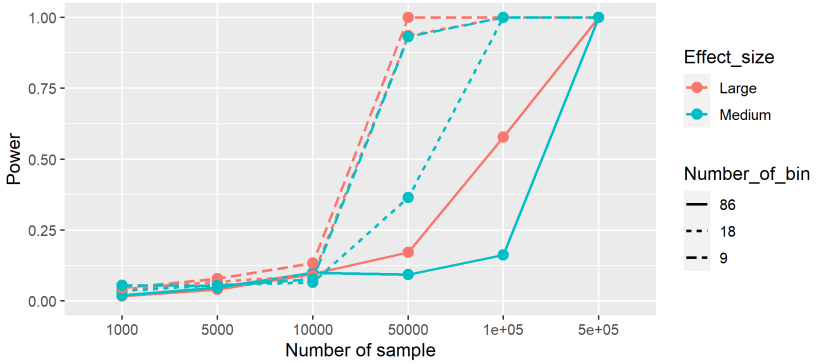
다음으로 Figure 11을 통해 새로운 검정 방법의 검정력을 확인하자. Figure 11은 위에서부터 차례대로 약한 차등정보보호($\alpha = 1$), 중간 차등정보보호($\alpha = 0.1$), 강한 차등정보보호($\alpha = 0.01$)를 적용했을 때의 검



(a) $\alpha = 1$



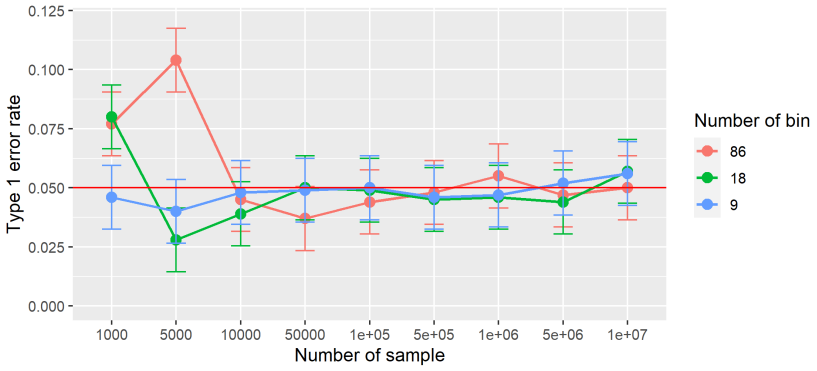
(b) $\alpha = 0.1$



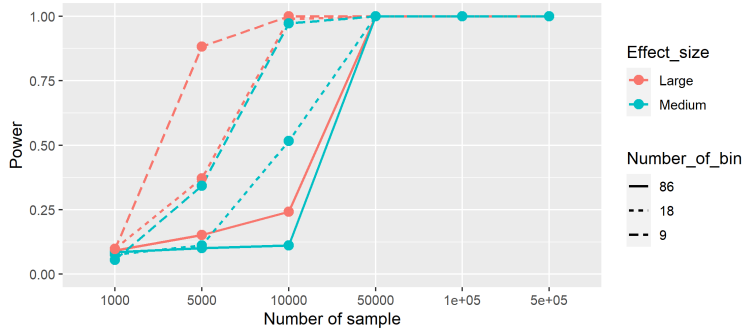
(c) $\alpha = 0.01$

Figure 11: *Power for parametric bootstrap test.*

정력이다. 약한 차등정보보호를 적용했을 때는 표본 수가 5,000개로 작을 때도 검정력이 매우 좋은 것을 알 수 있다. 또한, 정보보호 수준을 높일수록(즉, α가 작을수록), 구간의 개수가 많을수록, 효과 크기가 작을수록, 검정력을 보장하기 위한 표본 수 N이 증가하는 것을 알 수 있다.



(a) Type 1 error rate



(b) Power

Figure 12: *Error rate for parametric bootstrap test with different level of differential privacy.*

제안된 동질성 검정의 또 다른 장점은, 두 개의 히스토그램이 서로 다른 수준의 차등정보보호 자료일 때도 사용할 수 있다는 점이다. 이를 확인하기 위하여 한 히스토그램은 $\alpha = 0.1$의 정보보호를, 다른 히스토그램은 $\alpha = 0.05$의 정보보호를 적용한 후, Figure 12와 같이 제 1종 오류의 통제 여부와 검정력을 확인해보았다.

Figure 12의 (a)는 서로 다른 수준의 정보보호 처리를 한 한국의 인구분포의 두 히스토그램 표본을 가지고 모수적 부트스트랩 방법을 이용한 경우의 제 1종 오류 확률을 나타낸 그림이다. 전체적으로 유의수준 5%에서 잘 통제되고 있음을 알 수 있다. Figure 12의 (b)는 서로 다른 수준의 정보보호 처리를 적용한 미국과 한국, 한국의 수도권과 비수도권의 인구분포 표본을 가지고 모수적 부트스트랩 방법을 이용한 경우의 검정력을 나타낸 그림이다. 검정력 역시 표본 수가 10,000개 또는 50,000개 이상에서 매우 좋은 것을 확인할 수 있다.

5. 결론

본 논문에서는 모수적 부트스트랩을 이용한 두 차등정보보호 히스토그램의 동질성 검정을 제안하였다. 해당 검정 방법은 원자료에 대한 정보가 전혀 없을 때와 두 히스토그램에 적용된 차등정보보호의 수준이 다를 때에도 사용할 수 있다는 장점이 있다. 또한, 구간의 길이와 개수가 같은 두 히스토그램은 두 개의 다항분포 표본으로 볼 수 있으므로, 해당 검정 방법은 범주형 자료의 동질성 검정에 바로 적용될 수 있다. 이러한 검정 방법의 성능을 평가하기 위해 본 논문에서는 미국과 한국의 연령별 인구분포 자료를 통해 제 1종 오류의 확률과 검정력을 추정하였다. 그 결과, 작은 표본 수에서도 제 1종 오류가 잘 통제되고, 검정력 역시 표본 수가 증가

함에 따라서 빠르게 증가했다. 한편, 본 논문의 의의는 차등정보보호를 구현하는 매커니즘의 유용성(utility) 척도의 관점에서도 찾을 수 있다. Geng과 Viswanath (2015)에서는 차등정보보호를 구현하기 위해 더해진 오차의 분산이 작을수록 해당 매커니즘의 유용성이 크다고 보았다. 그러나 본 논문에서는, 부트스트랩을 이용한 동질성 검정을 수행했을 때의 높은 검정력이 좋은 유용성의 또 다른 척도로 사용될 수 있음을 보였다. 따라서, 후속 연구에서는 차등정보보호 히스토그램을 생성하는 매커니즘들을 비교함에 있어, 이와 같은 검정력을 비교 기준으로 사용할 수 있다.

References

Dwork C, McSherry F, Nissim K, and Smith A (2006a). Calibrating noise to sensitivity in private data analysis, *Theory of Cryptography Conference*, 265–284.

Dwork C (2006b). Differential privacy, *International Colloquium on Automata, Languages, and Programming*, 1–12.

Gaboardi M, Lim H, Rogers R, and Vadhan S (2016). Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *Proceedings of Machine Learning*, **48**, 2111–2120.

Geng Q and Viswanath P (2015). The optimal noise-adding mechanism in differential privacy, *IEEE Transactions on Information Theory*, **62**, 925–951.

Homer N, Szelinger S, Redman M, *et al.* (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays, *Plos Genetics*, **4**, e1000167.

Lee J, Wang Y, and Kifer D (2015). Maximum likelihood postprocessing for differential privacy under consistency constraints. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 635–644.

Park M, Lee Y, and Kwon S (2018). An study on differential privacy, *Statistics Research Institute*.

Park M, Kwon S, and Jung J (2019). An experimental study on applying differential privacy, *Statistics Research Institute*.

Wang Y, Lee J, and Kifer D (2017). Revisiting differentially private hypothesis tests for categorical data, Retrieved June 22nd, 2021 from: arXiv:1511.03376.

Wasserman L and Zhou S (2010). A statistical framework for differential privacy, *Journal of the American Statistical Association*, **105**, 375-–389.

US Census Bureau (2020). 2020 Demographic analysis estimates press kit, Retrieved feb 9th 2022 from https://www.census.gov/newsroom/press-kits/2020/2020-demographic-analysis.html

Received August 11, 2021; Revised September 28, 2021; Accepted October 25, 2021

모수적 부트스트랩을 이용한 차등정보보호 히스토그램의 동질성 검정

손주희[a], 박민정[b], 정성규[1,a]

[a]서울대학교 통계학과; [b]통계청

요 약

본 논문에서는 모수적 부트스트랩을 이용한 두 차등정보보호 히스토그램의 동질성 검정을 제안한다. 제안된 검정 방법은 차등정보보호 히스토그램과 적용된 차등정보보호 수준 정보만 있을 때에도 사용 가능하며, 비교하고자 하는 두 히스토그램에 적용된 차등정보보호의 수준이 다를 때에도 사용할 수 있다는 장점이 있다. 검정 방법의 성능을 평가하기 위해 미국과 한국의 연령별 인구분포 자료를 사용하고, 제 1종 오류의 확률이 잘 통제됨과 높은 검정력을 확인한다.

주요용어: 모수적 부트스트랩, 차등정보보호, 히스토그램의 동질성 검정

이 논문은 한국연구재단(NRF)과 과학기술정보통신부(MSIT)의 지원을 받아 연구되었음(No. 2019R1A2C2002256).
본 연구는 제 1저자 손주희의 서울대학교 석사학위논문의 일부로 활용된 논문임.

[1]교신저자: (08826) 서울시 관악구 관악로 1, 서울대학교 자연과학대학 통계학과. E-mail: sungkyu@snu.ac.kr