연구보고서 2019-20

차등정보보호 적용 실험 연구

박민정 · 권성훈 · 정재호

Statistics Korea
Statistics Research Institute

발간사

21세기 데이터 혁명시대에 빅데이터와 공공데이터의 홍수 속에서 통계는 더 이상 전문가들만의 전유물이 아닌 국민의 삶에 밀접한 영향을 미치는 중요한 요소로 보편화되고 있습니다. 특히 국가통계는 개인과 기업의 의사결정뿐만 아니라 정부의 정책결정에 필수적인 도구로서 그 중요성이 증대되고 있습니다. 이러한 변화에 따라 통계청은 빅데이터의 활용, 조사자료와 행정자료 간의 연계 등과 같은 통계생산 방식의 혁신을 통해서 응답자 부담은 최소화하면서 동시에 보다 정확하고 사용자 친화적인 통계를 만들고자 끊임없이 노력하고 있습니다.

이러한 통계청의 싱크탱크로서 전략적인 연구를 수행하고 있는 통계개발원은 국내의 유일한 「국가통계 전문연구기관」입니다. 통계개발원은 2006년에 창립한 이래 국가통계방법론과 증거기반 정책연구 분야에서 창의적인 발전을 해오고 있습니다. 특히 최근에는 기존의 조사통계를 보다 효율적으로 작성하기 위한 각종 추계기법과 관련된 통계방법론적 연구는 물론 데이터에 기반한 국가정책이 수립될 수 있도록 경제·사회현상에 대한 심층 분석 연구를 강화하고 있습니다. 또한 저출산·고령사회 현상 등으로 인해 대내외적으로 관심이 높아지고 있는 인구집단 및 인구동향에 관한 분석연구 및 인구동태 관련 방법론 연구를 밀도 깊게 수행하고 있습니다. 이러한 연구의 구체적인 결과를 중심으로 통계개발원은「2019년 연구보고서」를 발간하게 되었습니다.

이번 「2019년 연구보고서」는 주택임대차정보시스템 자료를 이용한 집세 물가지수 작성방안 연구 등 「행정자료 활용」 관련 연구를 비롯하여 세대별 니스(NEES) 특성 분석 등 「인구·사회통계」 기반 정책연구는 물론이고 통계기반정책평가 표준화 연구 등 「정책활용방법」 관련 통섭적인 연구를 수록하고 있습니다. 또한 통계방법론 분야에 있어서 동 연구보고서는 국민법의식조사 조사표 개선연구 등 「인지실험 기반 조사방법론 연구」 및 차등정보보호 적용 실험 연구 등 「개인정보보호」 관련 혁신적인 연구 등을 제시하고 있습니다.

통계개발원은 본 연구보고서가 데이터 사용자의 통계활용에 실제적인 도움이 되고 국가통계생산자의 혁신적인 통계개발에 유용한 자료로 활용될 수 있기를 기대합니다. 내년이면 15주년을 맞는 통계개발원이 “국가통계의 싱크탱크”로서 대내외적으로 선도적인 역할을 할 수 있도록 독자 여러분의 지속적인 관심을 부탁드립니다. 국가통계의 개선과 혁신적 개발을 위하여 제언이 있다면 언제든지 말씀하여 주십시오. 겸허히 귀 기울이겠습니다. 아울러 실용적이고 품질 높은 연구 결과를 도출하기 위해 최선을 다한 연구진에게 따스한 감사를 전합니다. 끝으로 본 연구에서 제시된 내용 및 결과는 저자 개인의 의견이며, 통계청 또는 통계개발원의 공식적인 견해가 아님을 밝혀둡니다.

2020년 5월

통계청 통계개발원장



전 영 일

목 차

제1장 서론 ····· 1

제2장 차등정보보호 이해 ····· 4

1. 노출위험과 정보손실 ····· 5

2. 차등정보보호와 통계학 ····· 13

제3장 차등정보보호 히스토그램 생성 알고리즘 ····· 25

1. 원자료의 히스토그램 ····· 25

2. 차등정보보호 히스토그램: 교란 방식 ····· 34

3. 차등정보보호 히스토그램: 재현 방식 ····· 40

제4장 차등정보보호 히스토그램 구현 및 평가 ····· 42

1. 차등정보보호 히스토그램: 모의실험 ····· 42

2. 차등정보보호 히스토그램: 실제자료 적용실험 ····· 53

3. 재현자료로 활용 가능성 ····· 76

제5장 결론 ····· 79

참고문헌 ····· 83

부록 ····· 84

요 약

데이터 이용 활성화와 개인정보보호 문제는 언제나 서로 충돌하는 관계에 있다. 이를 해결하기 위하여 개인정보의 노출위험을 관리하면서 자료제공 범위를 넓히기 위한 전문적이고 기술적인 노력이 필요하다. 최근 통계개발원에서는 자료를 제공할 때 개인정보 노출위험을 제어하기 위한, 즉 데이터 정보보호(data privacy)를 위한 통계적 기법들을 꾸준히 연구해왔으며, 본 연구는 차등정보보호 적용 실험을 다룬다. 본 연구에서는 선행연구를 숙지한 독자들이 차등정보보호 적용을 이해할 수 있도록 관련 이론을 정리하였고, 차등정보보호 히스토그램을 생성하는 몇몇 알고리즘을 소개했다. 또한 차등정보보호 히스토그램 생성 모의실험을 실시하고, 실제자료에 적용한 결과를 제시하였다. 더불어 재현자료 활용 가능성을 간략히 검토하였다. 본 연구에서 수행된 기술적인 노력이 사회 전반의 데이터 공유 및 개인정보보호 활동에 더하여 우리나라의 발전에 도움이 되기를 기대한다.

주요 용어 : 차등정보보호, 재현자료, 노출위험, 정보손실, 데이터 정보보호

제1장 서 론

4차 산업혁명 시대의 주도권을 확보하기 위한 노력들이 활발하다. 플랫폼 생태계를 구성하는 다수의 기업들은 경쟁우위를 확보하기 위해 다양한 활동을 하고 있다. 무엇보다 인공지능 등 기반기술 역량을 강화하기 위해 노력하고 있으며, 다른 한편으로는 그러한 기술을 응용하려는 분야에 대해 전문지식과 노하우를 확보하기 위해 애쓰고 있다. 그러나 전문지식과 기술력을 갖추고 플랫폼을 구축하여 사업을 진행한다고 하더라도 유용한 데이터 없이는 우위적인 플랫폼을 운영할 수 없다.

정부의 경우도 마찬가지이다. 공공서비스를 혁신하거나 경제발전을 촉진하는 정책수행에 데이터 활용은 필수적이다. 예를 들어 복지 정책 등, 정부 본연의 역할을 더욱 효율적으로 수행하기 위해 데이터를 활용한다면 제반 행정 비용을 파격적으로 감소시킬 수 있다. 뿐만 아니라 국가가 보유한 데이터를 이용할 수 있도록 국민들에게 제공한다면, 데이터 사업 등의 경제활동을 촉진하고 사회를 발전을 시키는 데 있어 크게 기여할 수 있다. 이 모든 활동에서 요구되는 전제 조건은 데이터 공유이다.

사회 구성원들이 데이터를 활발하게 사용하고자 할 때, 가장 큰 걸림돌은 개인정보보호이다. 공공 데이터의 경우를 생각해보자. 통계기관이 경제총조사나 사업체 조사 자료를 모두 공개하여 기업들이 그 자료들을 분석할 수 있다고 가정하자. 기업들은 바람직한 경제활동에 데이터를 이용할 수도 있지만, 경쟁자들의 자원 분배 방식을 엿보거나 약점을 잡기 위하여 데이터를 분석할 수도 있다. 창업에 도움을 줄 목적으로 인구주택총조사 자료를 손쉽게 이용할 수 있는 플랫폼을 제공하였는데, 개인들이 호기심으로 이웃의 혼인상태나 교육수준을 알아내거나 험담의 소재로 이용할 수도 있다. 불만이 있는 개인들은 특정 기관을 당황하게 하거나 타격을 줄 목적으로 자료를 분석할 수도 있다.

데이터 이용 활성화와 개인정보보호 문제는 언제나 서로 충돌하는 관계에 있다. 어느 하나를 이루려면 다른 하나를 포기해야 한다. <표 1>은 센서스 자료제공에서 이러한 딜레마를 해결하기 위한 미국의 정책적 노력을 보여주고 있다(Ruggles et al., 2018). 시대상황에 따라 자료 이용을 장려하는 모습을 보여주기도 하고, 반대로 개인정보보호를 우선하는 정책을 추진하기도 한다. 다만 최근에 와서 매스킹 등의 기술을 도입하여 적절한 해결책을 모색하고 있음을 알 수 있다. 이는 노출위험을 관리하면서 자료제공 범위를 넓히는 과정 속에 적절한 타협점을 찾기 위한 전문적이고 기술적인 노력이 지속되고 있음을 의미한다.

〈표 1〉 미국의 센서스 자료에 대한 개인정보보호 정책 변화

시기	주요 내용	R↓	U↑	T
1790년	센서스법에서 방문조사 완료 후 사본을 공적 장소에 게시할 것을 요구		√	
1850년	센서스 결과는 정부에 의해서 독점적으로 이용되어야 하고, 사적인 목적으로 사용될 수 없음을 선언	√		
1880년	조사자에게 다른 사람에게 조사정보를 공개하지 않을 것에 대한 선서를 요구	√		
1910년	대통령은 수집된 모든 센서스 정보에 대한 기밀성(confidentiality) 확보를 선언	√		
1921년	문맹 현황 조사를 위해 센서스 자료의 사용 허용으로 비밀보호 완화		√	
1929년	센서스법에서 제공된 정보는 통계목적으로만 사용되어야 하고, 어떤 특정 시설이나 개인이 제공한 자료를 식별할 수 없으며 개인보고서 자료로 사용될 수 없음을 규정	√		
제2차 세계대전 중	1929년 센서스법에서 규정한 비밀보호 규정 폐지		√	
1947년	1929년 법령의 비밀보호 조항 복원	√		
1954년	센서스법 Title 13에서 개인정보 비밀보호 강화	√		
1962년	센서스법에서 센서스국의 타 정부기관과의 개인 또는 시설에 대한 정보 공유를 명백히 금지하고 센서스 자료를 사법절차의 증거로 사용할 수 없음을 규정	√		
1990년	가구자료 중 핵심변수에 있어 동일한 특성이 있고 지역이 다른 가구들 사이에 교환 기법 사용 등			√
2000년	센서스 국은 이전보다 훨씬 적은 항목이 포함된 마이크로데이터 샘플을 만들 것을 제안, 매스킹 기법 사용			√
2000년 이후	센서스 표본조사를 대체한 ACS에도 센서스에 활용한 비밀보호 기법을 적용, 비밀보호 기법을 지속적으로 개선			√

주: R↓: 노출위험 축소 정책 U↑: 자료유용성 확대 정책 T : 신기술 도입 정책

활발한 데이터 이용을 위해 데이터의 직접 배포뿐만 아니라 데이터 센터(Research Data Center), 데이터 거래소(Data Exchange) 등을 활용하는 방식도 충분히 검토할 수 있다. 다만 어느 방식이든 데이터를 공유하고자 할 때 노출위험 축소와 자료유용성 확보 사이의 딜레마를 해결하기 위해서는 발전된 최신 이론 및 기술을 적절히 이용하는 전략

이 필요하다.

최근 통계개발원에서는 자료를 제공할 때 개인정보 노출을 제어하기 위한, 즉 데이터 정보보호[1](data privacy)에 관한 기법들을 꾸준히 연구해왔다. 주요 방안으로는 매스킹 처리, 베이지안 재현자료 활용, 차등정보보호 적용 등이 있다. 이러한 기법들이 가지는 장단점을 올바르게 이해하기 위해서는 노출위험과 정보손실에 대한 지식이 선행되어야 한다. 최근 일련의 보고서 발간을 통해 통계개발원은 이러한 작업을 순차적으로 진행하였으며, 이번 연구에서는 차등정보보호 적용 실험을 다루었다.

자료를 제공하는 방식에 있어 차등정보보호는 쿼리(query)에 의존하는 특징이 있다. 본 연구에서는 실험 대상 쿼리로 히스토그램을 선택하여 차등정보보호를 실제로 적용한다면 어떠한 통계가 제공되는지 확인하고자 한다. 이를 위해 차등정보보호 히스토그램을 생성하는 여러 알고리즘을 비교, 검토하였다. 또한 적절한 자료제공, 즉 일정 수준으로 개인정보 노출제어가 보장되고 이용자가 수용할 수 있을 정도의 정보손실만 발생되게 하는데 필요한 조건들을 살펴보았다. 나아가 차등정보보호를 이용해 재현자료를 생성하고 사용할 수 있을지 간략히 검토하였다. 이는 정보손실이 작지만 노출위험 수준이 담보되지 않고 생성 과정이 복잡한 베이지안 재현자료를 대신할 수 있는지, 그 가능성을 타진하기 위함이다.

이를 위해 2절에서는 선행연구를 숙지한 독자들이 차등정보보호를 다른 기법들과 비교하며 이해할 수 있도록 노출위험과 정보손실의 의미를 정리하였다. 더불어 통계학 관점에서 차등정보보호를 이해하고, 차등정보보호 자료를 생성하는 교란 방식과 재현 방식을 설명하였다. 3절에서는 차등정보보호 히스토그램을 생성하는 여러 알고리즘을 정리하여 제시하였다. 4절에서는 차등정보보호 히스토그램을 생성하는 모의실험을 진행한 후, 실제 가계금융복지조사 보도자료 일부에 대하여 차등정보보호 적용 실험을 수행하고 그 결과를 정리하였다. 또한 재현자료로써 활용 가능성도 간략히 언급하였다.

유럽에서는 일반개인정보보호규정(GDPR)의 시행이 시작('18년 5월)되었고 이웃나라 일본은 자국의 개인정보보호 관리 조치를 인정받아 GDPR의 면제국으로 지정되었다. 우리나라는 아직 면제국에 포함되지 않는다. 통계적 비밀보호, 데이터 정보보호를 위한 본 연구의 기술적인 노력이 사회 전반의 데이터 공유 및 개인정보보호 활동과 더불어 우리나라의 발전에 도움이 되기를 기대한다.

---

1) 통계적 노출제어(statistical disclosure control), 통계적 노출제한(statistical disclosure limitation), 통계적 비밀보호(statistical data confidentialtiy), 데이터 정보보호(data privacy)를 모두 「자료를 제공할 때 자료 내에서 개인정보가 드러나는 것을 방지하기 위한 기술적 조치」를 의미하는 용어로 사용한다. 이러한 조치를 적용하면 데이터는 각종 교란(perturbation) 등으로 원자료와 달라질 수 있다. 다만 이들 용어를 자료 제공과 반대의 목적으로 사용되는 「자료의 유출을 방지하기 위한 데이터 보안(security)」과 구분하기 바란다.

제2장 차등정보보호 이해

차등정보보호를 이해하기 위한 연구가 2018년 진행된 바 있다. 박민정 외(2018)에서는 차등정보보호에 관한 개념적 이해, 이론적 이해, 활용 사례를 다루었으며, 이를 통해 독자들이 차등정보보호를 이해하고 나아가 활용할 아이디어를 찾는데 도움을 주고자 하였다. 후속 과제인 본 연구는 차등정보보호를 실제로 적용할 때 나타나는 결과를 살펴보고, 활용 전략을 세우는 방법을 생각해보려고 한다. 이를 위해 주요 실험 대상 쿼리로 히스토그램을 선정하였다. 쿼리 응답 결과물에 라플라스 잡음을 더하여 차등정보보호를 만족하는 히스토그램을 구하는 것은 단순하나, 특정 알고리즘이나 재현자료 생성 방식을 이용하는 것은 상당히 복잡하며 이론적인 설명이 필요하다. 한편, 차등정보보호 히스토그램을 구하기 위하여 생성했던 재현자료를 이용해 다른 통계량을 구할 수도 있다. 이러한 내용을 교란 방식과 재현 방식으로 나누어 설명한다.

이 절에서는 독자들이 선행 연구의 차등정보보호 개념, 이론 및 활용 사례를 충분히 숙지하였음을 가정하고, 먼저 노출위험과 정보손실 관계를 통해 차등정보보호를 다른 방법들과 비교한다. 다음으로 통계학 관점에서 차등정보보호를 이해하기 위해 필요한 내용을 설명한다.

참고로 차등정보보호를 논의할 때 대다수 문헌에서 노출제어 수준을 ϵ으로 표현한다. 반면 차등정보보호를 연구한 통계학 분야의 주요 문헌에서는 α를 사용한다. 아마도 통계학에서 주로 잡음을 표현할 때 ϵ을 사용하고, 유의수준을 표현할 때는 α를 사용하기 때문일 것으로 추측해 본다. 차등정보보호 모수 역시 통계학적으로 다시 정의되는 노출위험 수준을 표현하는 것이므로(Wasserman과 Zhou, 2010; Duchi et al., 2018), ϵ이 아닌 α를 사용한다고 생각하는 것이 차등정보보호를 이해하는데 도움이 될지도 모르겠다. 따라서 차등정보보호를 논할 때 ϵ을 사용하는 것이 독자들에게 친숙하겠지만, 본 연구는 통계학 기반에서 제시된 이론적 설명을 따르므로 ϵ을 대신하여 α를 사용하겠다.

1. 노출위험과 정보손실

(요약) 매스킹 기법 적용이나 베이지안 재현자료 생성으로는 노출위험 수준을 정량적으로(formally) 제시하기 어렵다. 반면에 차등정보보호는 주어진 수준 α로 노출위험을 제어하면서 자료를 생성한다. 한편 기존의 거리 기반 정보손실을 사용한다면, 차등정보보호 자료에서도 노출위험-정보손실 상충관계는 여전히 존재한다.

가. 매스킹 처리에서 노출위험-정보손실 관계

최근까지 자료를 제공할 때 주로 매스킹(masking) 기법을 이용해 개별 정보의 노출위험을 제어해 왔다. 매스킹 기법은 세부 종류가 다양하여 하나의 자료에 대해서도 수십 개 이상의 매스킹 처리 결과를 생성할 수 있다. 따라서 원래 자료를 대신할 수 있는 다양한 방안의 후보군 중에서 선택을 위한 적절한 판단 기준이 필요하다. 이러한 판단의 기준으로 사용하는 것이 노출위험과 정보손실 측도이다. 노출위험과 정보손실(자료 유용성) 사이의 관계를 종합적으로 판단하기 위해 각 측도의 계산 결과를 평면에 표현하는 것을 위험-유용성 지도(risk-utility map)라고 부른다. [그림 1]은 위험-유용성 지도의 한 예를 나타낸다(Hundepool et al., 2012). 가로축에는 자료 유용성을, 세로축에는 노출위험을 표시한다. 원자료가 가지는 자료 유용성은 유지하면서 노출위험만 주어진 기준 이하로 낮추는 것이 가장 이상적이다. 그러나 노출위험을 줄이기 위해 매스킹 기법을 적용하여 변형된 자료는 자료 유용성이 낮아질 수밖에 없다.

한편, 가로축에 노출위험을, 세로축에 자료의 정확성을 기재하여 정보보호 처리된 자료를 평가할 수도 있다. 또한 반대로 가로축에 정보손실을, 세로축에 노출위험을 측정한 결과를 표시할 수도 있다. [그림 2] 및 [그림 3]은 각 방식들을 나타낸다. 어느 방식이든 노출위험만을 줄이는 방안을 찾는 것이 가장 이상적이지만 정보손실의 증가 또는 유용성의 감소가 발생한다. 따라서 이상적인 결론과 가장 가까운 방안이 무엇인지 정량적인 수치를 제시하여 판단하기는 어렵다는 문제가 발생한다. 노출위험을 줄이려고 자료를 강하게 변형할수록 정보손실이 증가하는데, 노출위험과 정보손실은 언제나 상충(tradeoff) 관계에 있기 때문이다. 따라서 최적 방안 선택에 대한 적절한 기준이 필요하다.



[그림 1] 위험-유용성 지도(노출위험-자료유용성)



[그림 2] 위험-유용성 지도(정확성-노출위험)



[그림 3] 위험-유용성 지도(노출위험-정보손실)

한편, 각 축을 나타내는 노출위험 측도와 정보손실 측도는 다음과 같이 결정한다. 먼저 노출위험은 측정에 사용되는 변수 범위와 통계적 모형에 따라 구분할 수 있다. 키변수[2]만을 사용해서 유일성을 측정하는 모형들이 있는데, 이를 유일성에 근거한 노출위험 측도라 하자. 키변수들이 가지는 정보에 근거하여 각 레코드가 모집단에서 유일한 정도를 표현하는 조건부 확률의 값을 추정하고, 개체별 노출위험 측도로 사용할 수 있다. 파일 단위로 노출위험을 비교하기 위해서는 그러한 모형을 이용해 얻은 개체별 노출위험의 평균이나 합을 활용한다. 그러나 이러한 방식은 민감변수에 대한 매스킹 처리 효과를 감안하지 못하는 단점을 가진다.

이러한 단점을 극복하기 위해 키변수 뿐만 아니라 잡음 추가 등의 방식으로 변조한 민감변수의 값도 고려하여 노출위험을 측정하도록 제안된 이론이 있다. 이를 침입자 의사결정론 기반 노출위험 측도라 하자. 이러한 접근 방식에 의하면, 통계기관이 제공하는 파일에서 외부인(침입자)이 목표로 하는 레코드를 찾을 때 외부인은 통계적 의사결정론을 이용해 매칭 확률을 계산한다고 가정한다. 따라서 매스킹 기법을 적용한 파일들에 대하여 레코드의 유형별로 이러한 확률을 계산하고, 그 확률을 최소화하는 매스킹 방안을 선택하는 것이 노출위험을 최소화하는 방안이 된다. 침입자 의사결정에 근거한 노출위험 측정은 매우 설득력이 있으나, 파일마다 이러한 통계적인 모형을 구축하는 것이 대단히 방대하고 어렵다는 단점이 있다.

실무적으로 키변수와 (연속형인 경우가 많은) 민감변수를 모두 고려해서 노출위험을 계산하기 위해서는, 단순히 매스킹 전후의 거리를 고려하는 경우가 많다. 예를 들면, 먼저 원자료에서 레코드마다 순위에 근거해 어떤 구간을 만들고, 매스킹 처리 전후 각 레코드의 값이 그 구간 내에 있는지 없는지를 살펴본다. 다음으로 전체 레코드 중 각 구간 내에 원자료가 있는 비율을 이용해 파일 단위의 노출위험을 계산할 수 있다(Templ과 Meindl, 2008).

한편, 정보손실은 다양한 거리 함수들을 이용해 매스킹 기법 적용 전후 자료를 비교하여 측정한다. 자료값 사이의 거리, 자료 분포의 특징을 나타내는 공분산이나 주성분을 이용하여 계산한 거리, 분포를 비교하기 위한 쿨백-라이블러 거리(Kullback-Liebler divergence) 및 경험적 분포 등을 이용할 수 있다. 추정값의 신뢰구간 중복, 성향 점수 또는 군집 분석을 이용하는 아이디어들도 제시된 바 있다.

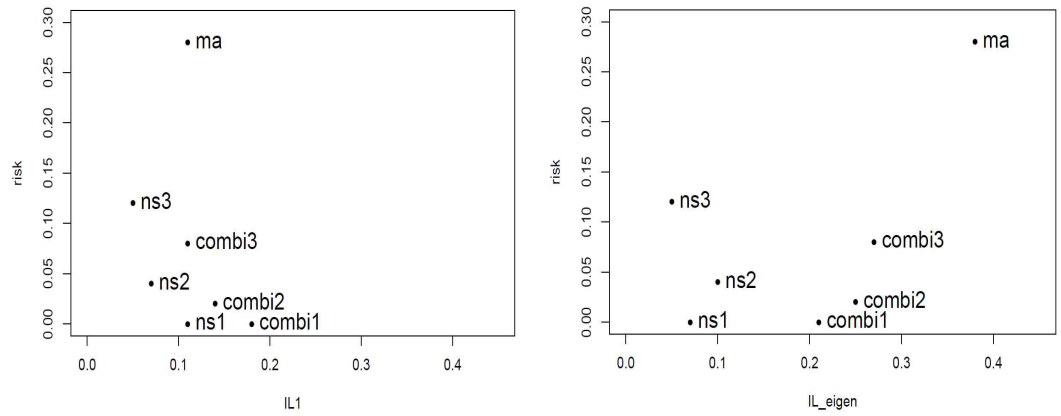
종합하면, 자료에 대해 노출위험과 정보손실을 실무적으로 측정하고자 할 때 매스킹 전후 자료 사이의 거리를 이용한다. 자료의 변화가 적어서 거리가 가까울수록 노출위험이 크고 정보손실이 작으며, 자료에 변화가 많아 거리가 멀어지면 노출위험이 작고 정보

2) 응답자(레코드)의 특성을 나타내는 변수

손실이 큰 결과를 얻는다. 즉, 노출위험과 정보손실 사이에 상충관계가 있으므로 다양한 매스킹 방안 중에서 좋은 선택을 할 수 있는 명확한 기준을 제시하기 어렵다. 이상적으로는 노출위험과 정보손실을 모두 줄이는 방안을 찾아야 하지만, 실제 적용에 있어서 두 가지 사항이 서로 상충관계를 발생시키기 때문이다. 또한 노출위험과 정보손실에 대하여 여러 종류의 측도들이 존재하여 어떤 측도를 사용하는 것이 적절한지 설명하기 어렵다는 단점도 있다.

나. 매스킹 처리에서 노출위험-정보손실 관계 예제

[그림 4]는 가계금융복지조사 마이크로데이터에 여러 매스킹 기법을 적용한 결과를 보여준다. 왼쪽 그림에서 가로축의 정보손실 측도는 자료값 사이의 거리를 이용했고, ns1 방안이 ns2 방안에 비해 노출위험이 작고 정보손실이 크다는 것을 알 수 있다. 그러나 둘 중 어떤 방안이 좋은지 결정하는데 명확한 기준을 찾는 것은 어렵다. 노출위험과 정보손실 중에서 어느 측도를 우선으로 판단해야 할지 결정하기 쉽지 않기 때문이다. 또한 [그림 4]의 오른쪽 그림에서는 같은 자료에 대해 변수들 사이의 구조에 관한 정보를 더 잘 잡아내도록, 자료값 대신에 고유값을 이용해 정보손실을 측정하였다. 이때는 ns1 방식이 ns2 방식에 비해서 절대적으로 더 나음을 알 수 있다. 즉, 정보손실 측도를 바꾸면 매스킹 각 방안에 대한 평가 결과가 달라진다.



[그림 4] 매스킹 방안에 대한 위험-유용성 지도 예제

이처럼 거리 기반 노출위험과 정보손실 측도를 이용하는 경우, 매스킹 기법 적용 결과를 평가하는 기준을 명확하게 하는 것은 어렵다. 나아가 측도를 바꾸는 경우 평가 결과가 일관되지 않다는 단점이 있다. 좀 더 자세한 내용은 박민정(2014)을 참고하라.

지금까지 살펴본 바와 같이 매스킹 기법을 사용하여 통계적 비밀보호를 고려할 때 여러 문제점이 있음이 잘 알려져 있어 이를 극복하기 위한 노력이 지속적으로 이루어져 왔다. 그러한 노력의 결과로 베이지안[3] 재현자료(synthetic data) 및 차등정보보호(differential privacy)가 제안되었으며, 최근 더욱 활발히 연구되고 있다.

---

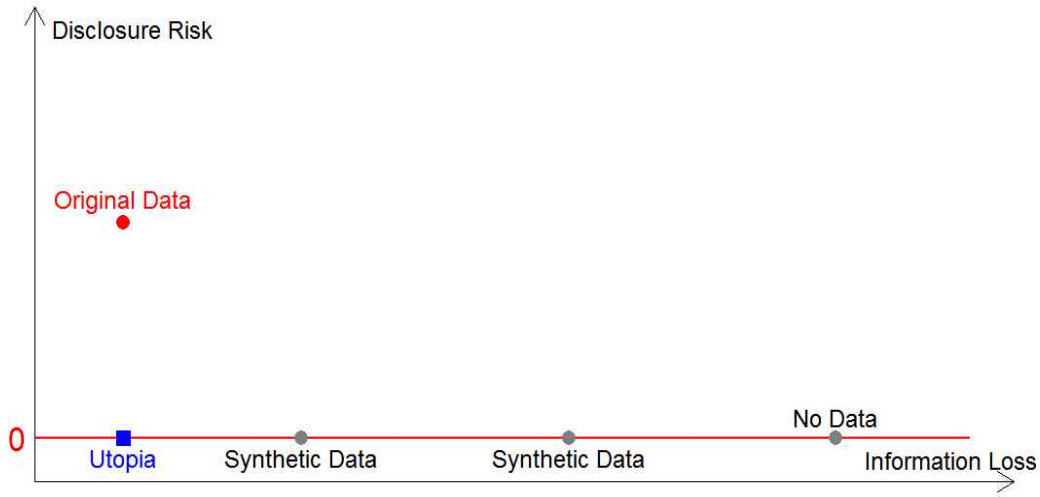
3) 베이지안 관점의 다중대체 방식으로 생성한 재현자료를 베이지안 재현자료라 부르도록 한다. 다음 소절의 <표 2> 용어 정리에서 관련 논의가 제시된다.

다. 베이지안 재현자료에서 노출위험-정보손실 관계

재현자료는 다중대체(multiple imputation) 기법을 이용해서 노출위험 없이 자료를 제공하기 위해 제안되었다. 모집단에서 조사되지 않은 레코드들을 마치 무응답 개체처럼 취급하고 다중대체하여 재현 모집단을 만든다. 생성된 재현 모집단에서 일부 자료를 단순임의추출해서 만든 자료를 완전(fully) 재현자료라고 부른다. 이상적으로 완전 재현자료에는 원자료가 전혀 포함되지 않아 노출위험이 없다고 볼 수도 있다. 때문에 재현자료를 생성하는 모형을 정밀하게 구축하여 정보손실을 줄이려는 노력이 있었다.

[그림 5]는 완전 재현자료의 위험-유용성 지도를 나타낸다. 모수적으로 모형을 적절히 구축하면 노출위험을 0이라고 생각할 수 있다. 또한 모형의 정확성이 높을수록 정보손실이 줄어들 수 있다. 그러나 실제자료에 대해서 모수적으로 적절한 모형을 구축하는 것은 대단히 어려운 작업이므로, [그림 5]의 원점에서 먼 재현자료를 생성하게 된다.

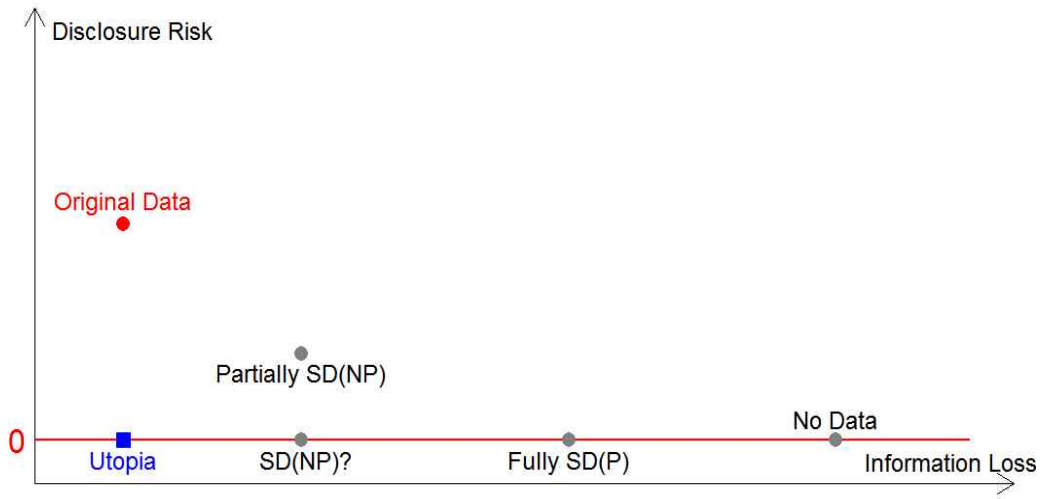
이를 개선하여 자료 유용성을 높이기 위해서는 붓스트랩과 같은 비모수적인 방법을 사용하여 완전 재현자료를 생성한다. 그러면 실제자료를 직접 사용하게 되므로, 실제자료가 재현자료에 나타날 수 있다. 따라서 완전 재현자료가 제안된 원래 의도와 다르게 노출위험이 없이 안전하다고 말하기 어렵다. 또한 충분한 자료 유용성을 확보하는 것도 쉽지 않으므로 실제자료에 완전 재현자료 생성 방식을 적용한 사례를 찾기 어렵다.



[그림 5] 이상적인 완전 재현자료의 위험-유용성 지도

한편, 모든 자료를 재현하는 경우 정보손실을 줄이기 어려우므로, 자료의 일부만을 결측값처럼 취급하고 다중대체 기법을 적용해서 재현자료를 생성할 수 있다. 이렇게 생성된 자료를 부분(partially) 재현자료라고 한다. 부분 재현자료는 안전하다고 판단된 레코드나 변수에 대해서는 원자료의 값을 그대로 유지한다. 완전 재현자료와 마찬가지로 실제자료에 대하여 모수적인 모형을 구축하기는 쉽지 않으므로 비모수적 기법을 사용하게 된다. 비모수적 기법으로 베이지안 붓스트랩 등을 사용한다는 것은 원래 자료의 값을 사용하여 재현자료를 생성하는 것을 의미한다. 따라서 부분 재현자료의 경우에도 노출위험이 어느 정도 발생한다고 할 수 있고, 발생한 노출위험을 정량적으로 측정하기 어렵다는 측면에서는 매스킹과 동일한 한계를 가진다.

[그림 6]은 재현자료에서 노출위험이 발생해서 매스킹 처리의 경우와 유사한 형태의 위험-유용성(정보손실)을 보이는 것을 나타낸다. 원래 재현자료를 제안했던 아이디어처럼 노출위험을 완전히 억제하는 것이 아니라, 정보손실을 줄이기 위해 다시 노출위험을 어느 정도 늘리게 된다. 게다가 매스킹 기법을 사용할 때처럼 노출위험이 증가한 규모의 측정 또한 복잡하다. 즉, 매스킹이나 재현자료 생성으로는 노출위험이나 정보손실 수준을 정량적으로 제시하기 어렵다는 것을 알 수 있다. 재현자료 생성 기법에 관한 더 자세한 내용을 알기 원한다면 박민정과 김정연(2017) 및 수록된 문헌을 참고하라.

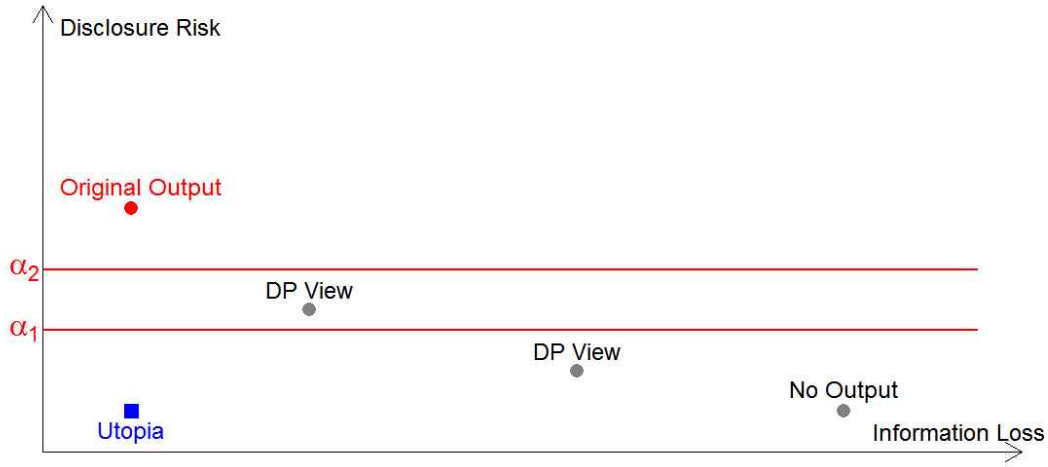


주: Fully SD(P)는 모수적으로 생성한 완전 재현자료를, Partially SD(NP)는 비모수적으로 생성한 부분 재현자료를 나타낸다.

[그림 6] 재현자료의 위험-유용성 지도

라. 차등정보보호 자료에서 노출위험-정보손실 관계

차등정보보호는 노출위험 개념을 새롭게 제시하고, 노출위험 측도를 정량적으로 제안하여 큰 주목을 받았다. 그동안 사람들의 생각을 지배했던 Dalenius의 정보보호[4] 정의를 만족시키는 것은 외부 정보의 존재로 불가능함을 수학적으로 증명하고, 나아가 새로운 노출위험 측도의 개념과 수학적인 정의를 제시했기 때문이다. 차등정보보호 개념이 제안되기 이전에는 노출위험 측정을 외부인의 자료 접근 중심으로 생각하였으나, 차등정보보호 제안 이후에는 응답자의 데이터 수집 참여 중심으로 생각하는 사고의 전환이 일어났다. 이를 통해 달성 불가능한 도전에 매달리지 않고, 새로운 방법을 찾는 노력이 시작될 수 있었다.[5]



[그림 7] 차등정보보호 자료의 위험-유용성 지도

차등정보보호는 공표하는 통계의 생산 과정에 대한 정보보호이며, 노출위험이 주어진 수준 α를 기준으로 제어되면서 자료를 생성한다는 장점을 가진다. 즉, 매스킹 기법 적용이나 재현자료 생성을 통해 사후적으로 노출위험을 측정하던 방식과 완전히 다르다. 다만 선행 연구인 박민정 외(2018)에서 소개한 활용사례들이 보여주는 것처럼, 정보보호 수준이 강력해지면(α가 작아지면) 정보손실이 커지게 된다. 따라서 매스킹 기법을

4) "Anything that can be learned from the database about a specific individual should be learned without access to the database."[Dalenius, T.(1977), Towards a methodology for statistical disclosure control, Statistik Tidskrift, 15, 429-444.] 이 명제는 이후에 Dwork(2006)에서 다음과 같이 표현되기도 하였다. "Access to a statistical database should not enable one to learn anything about an individual that could not be learned without access."

5) 관련 내용을 더 알아보기 원하는 독자는 박민정(2016)의 3절을 참고할 수 있다.

이용할 때처럼 노출위험과 정보손실 사이의 상충관계가 여전하나, 차등정보보호라는 노출위험 정의를 사용하므로 노출위험 측도에 수학적 엄밀함과 일관성이 있다는 장점이 있다. 차등정보보호를 만족시키도록 생성한 자료의 위험-유용성은 [그림 7]과 같이 표현된다. 한편, 차등정보보호를 만족시키도록 변조된 자료를 차등정보보호 자료(differentially private view, DP view)라고 부른다.

지금까지 노출위험과 정보손실 측도를 살펴보고, 매스킹 기법 활용, 베이지안 재현자료 생성, 차등정보보호 적용에 따른 노출위험-정보손실 관계를 정리하였다. 다음 소절에서는 차등정보보호에 관한 통계학적 관점 및 히스토그램 생성 알고리즘을 정리하여 본 연구에서 수행하는 차등정보보호 적용 실험의 의의를 이해하도록 한다.

2. 차등정보보호와 통계학

(요약) 차등정보호호를 적용하는 방식은 다음과 같이 교란 방식과 재현 방식으로 나눌 수 있으며, 통계학적 활용을 위해서는 재현 방식이 선호된다(Wasserman, 2012).

A. 교란 방식(the query-response model of privacy)

B. 재현 방식(the sanitized database model of privacy)

통계학적 관점(framework)에서 노출위험을 나타내는 차등정보보호 모수 α는 가설검정 오차의 개념으로 설명된다(Duchi et al., 2018). 또한, 최소최대 이론을 이용해 추론의 관점에서 정보손실을 설명할 수 있다(Wasserman과 Zhou, 2010).

가. 차등정보보호 적용 방식: 관점의 차이

차등정보보호는 공학 분야에서 처음 제안되었지만, 차등정보보호 정의는 통계학 지식을 기반으로 이루어져 있다. 다만 차등정보보호를 만족시키는 방법은 주로 요청받은 쿼리에 대한 최종 결과물(또는 특정 공표통계)에 단순히 라플라스 (또는 이중지수 분포의) 잡음을 추가하는 것이었다. 이러한 접근 방식에 대해 Wasserman과 Zhou(2010)는 통계학적인 관점(framework)을 가지고 차등정보보호를 이해하고 활용할 것을 제안하였다.

이를 논하기에 앞서, 차등정보보호를 처음 제안한 공학 분야에서 바라보는 통계학과, 통계학자들이 생각하는 통계학의 범위가 다르다는 Wasserman(2012)의 설명을 살펴보자. [그림 8]은 통계학 연구자와 공학 분야 연구자가 가지는 통계학에 대한 관점의 차이를 보여준다.

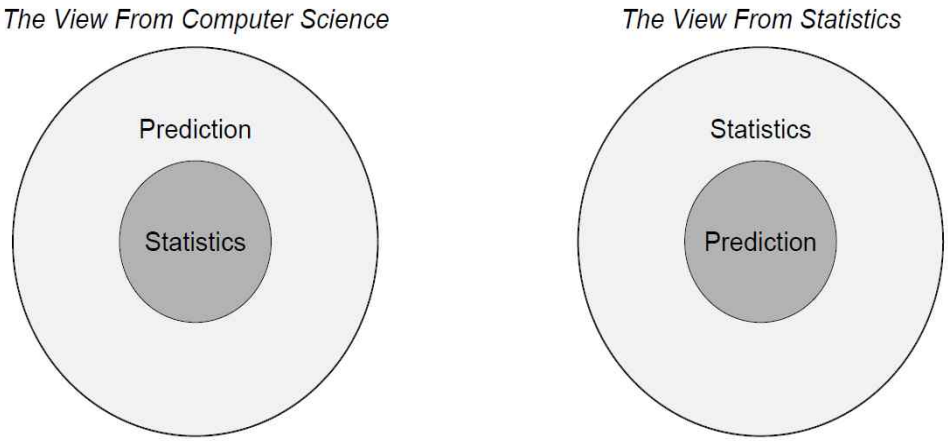


Figure 1: Statisticians view prediction (classification, regression, etc.) as a part of statistics. Some in CS see statistics as a subset of prediction.

[그림 8] 통계학을 바라보는 관점의 차이

Wasserman(2012)에 따르면 공학 분야의 많은 연구자들은 통계학을 예측을 위한 방법의 하나로 정의한다. 따라서 차등정보보호를 연구할 때도 분석(예측) 결과에 집중하며, 정보보호(privacy)를 위해 데이터 분석 결과에 어떻게 잡음을 추가할 것인지 다양한 알고리즘을 개발하는 것이 주 관심사이다. 즉, 예측 결과를 통해 개인정보침해가 일어나지 않도록 차등정보보호 구현 알고리즘을 연구하는데 관심사가 국한되어 있다. 또한 정보손실을 단순히 잡음이 추가된 결과물과 원래 결과물과의 차이로 측정한다.

반면 많은 통계학 연구자들은 예측을 포함한 다양한 자료 분석 과정을 다루는 것을 통계학이라고 생각한다. 따라서 분석 결과뿐만 아니라 데이터 자체에 집중한다. 통계학자는 예측 문제를 해결하기 위해서만 데이터를 사용하는 것이 아니라, 더 넓은 범위의 여러 종류의 분석에 데이터를 사용하고자 하기 때문이다. 따라서 주어진 쿼리 이외의 문제에도 자료를 사용하고자 하므로, 데이터 자체를 어떻게 제공할지에 관심을 가진다. 또한 정보손실 측정의 문제도 추론의 정확성을 기준으로 접근한다. 이에 Wasserman(2012)은 차등정보보호를 적용하는 방식을 다음과 같이 나누어 다루고 있다.

A. the query-response model of privacy - 교란 방식

B. the sanitized database model of privacy - 재현 방식

첫 번째 방식은 처음 차등정보보호가 제안된 개념에서 기인하였으며, 특정 쿼리에 대한 결과물에 잡음을 직접 추가하여 변조된 결과물(perturbed output)을 제공하는 것을 말한다. 본 연구에서는 이를 교란 방식이라고 하겠다. 두 번째 방식은 차등정보보호의 개념을 통계적인 방식으로 재해석한 개념을 기반으로 한다. 차등정보보호를 특정 쿼리에 응답하는 결과값을 주기 위해 적용하는 것이 아니라, 그 쿼리를 계산할 수 있도록 재현된 자료(sanitized database)를 제공하는 것을 말한다. 이를 재현 방식이라고 하겠다.

한편, 선행 연구에서는 차등정보보호 자료를 모두 비식별 데이터(sanitized data)라고 지칭했다(박민정 외, 2018). 이는 비식별이라는 용어가 국내에서 이미 익숙하고, 다중대체 기법을 활용한 재현자료와 용어의 중복을 피하기 위함이었다. 그러나 현재 해외에서 차등정보보호를 만족하는 자료를 지칭할 때 재현자료(synthetic data)라는 용어를 널리 사용하고 있다. 혼란을 피하기 위하여 본 연구에서는 다음과 같이 용어를 정리하도록 한다. 우선 다중대체 기법으로 생성한 재현자료를 베이지안 재현자료라고 부르도록 한다. 다음으로 차등정보보호 자료 중에서 재현 방식으로 생성한 자료를 차등정보보호 재현자료라고 부르겠다. 이는 현재 해외의 용어 사용 흐름을 따르기 위함이다.

〈표 2〉 용어 정리

<table>
<tr><th>데이터 정보보호 방식</th><th>기존 국내(국외) 문헌</th><th>신규 제안</th></tr>
<tr><td>매스킹 기법 적용</td><td>비식별 데이터
(masked data)</td><td>비식별 데이터
(masked data)</td></tr>
<tr><td>다중대체 기법으로 재현자료 생성</td><td>재현자료
(synthetic data)</td><td>베이지안 재현자료
(Bayesian synthetic data)</td></tr>
<tr><td>차등정보보호 적용 - 교란 방식</td><td rowspan="2">비식별 데이터
(sanitized database)
(synthetic data)

차등정보보호 자료
(DP view)</td><td>차등정보보호 자료
(DP view)</td></tr>
<tr><td>차등정보보호 적용 - 재현 방식</td><td>차등정보보호 재현자료
(synthetic data)</td></tr>
</table>

공학 분야에서 처음 제안된 차등정보보호 정의는 [정의 1]과 같다(Dwork, 2006; 박민정 외, 2018)[6]. 차등정보보호를 만족하는 가장 대표적인 랜덤화 함수 K는 요청받은 쿼리의 결과에 라플라스 잡음을 더하는 함수이며 이는 교란 방식에 해당한다. 반면, 통계

학적 관점에서 차등정보보호 정의는 [정의 2]와 같다(Duchi et al., 2018; 박민정 외, 2018). 차등정보보호를 랜덤화 함수뿐만 아니라 재현자료를 포함한 개념으로 확장하고 있음을 확인할 수 있다.

[정의 1] D 및 D' 를 개체(레코드) 하나만 다른 데이터베이스라고 하자. 만약 주어진 랜덤화 함수 K가 적당한 $\alpha > 0$ 에 대하여 다음과 같은 조건을 만족하면 K를 α-차등정보보호(α-differential privacy) 랜덤화 함수라고 한다.

$$P[K(D_1) \in S] \leq e^{\alpha} P[K(D_2) \in S] \ for\ all\ S \subseteq range(K).$$ ■

[정의 2] 원자료 $X = (X_1, ..., X_n)$ 에 대하여 재현자료 $Z = (Z_1, ..., Z_k)$의 조건부 밀도함수(conditional density function)를 $f_{Z|X}$라고 하자. 만약 $f_{Z|X}$가 적당한 $\alpha > 0$ 에 대하여 다음 부등식을 만족하면 $f_{Z|X}$를 X에 대한 α-차등정보보호 기법(α-differential privacy mechanism)이라고 하고 Z를 X에 대한 α-차등정보보호 자료(α-differentially private view)라고 한다.

$$sup_{\|u-v\|_0 = 1}\ sup_z \frac{f_{Z|X}(z|u)}{f_{Z|X}(z|v)} \leq exp(\alpha).$$

단, $\|u-v\|_0 = \sum_{i=1}^{n} I(u_i \neq v_i)$ 이며 I는 지시(indicator) 함수이다. ■

한편, 차등정보보호 자료는 [정리 1]과 같은 성질을 가진다는 것이 알려져 있다(Wasserman과 Zhou, 2010; 박민정 외, 2018). [정리 1]은 α-차등정보보호 자료를 측정 가능한(measurable) 함수를 사용하여 변환해도 여전히 α-차등정보보호 자료가 됨을 의미하며, 특히 α-차등정보보호를 만족하는 재현자료를 구축하는 방법을 설명하고 있다.

[정리 1] $e = (e_1, ..., e_m)$를 밀도함수 f_{ϵ}에서 추출한 임의표본(independent and identically distributed random variables)이라고 하자. 만약 e가 원자료 $X = (X_1, ..., X_n)$와 독립이면 e와 X로부터 얻은 확률벡터 $Z(X, e)$는 다음과 같은 성질을 가진다.

---

6) 원문 문헌의 내용을 국문 선행 연구 보고서에서 이미 번역하여 설명하였고, 기존 보고서에 설명된 내용을 차용했으므로 참고 문헌을 모두 표시하였다. 이하 선행 연구 보고서의 참고 문헌 중복 수록은 동일한 이유를 가진다.

① $Z(X,e)$가 X에 대한 α-차등정보보호 자료이면 임의의 함수[7] S에 대하여 $W=S(Z(X,e))$도 X에 대한 α-차등정보보호 자료이다.
② $Z(X,e)$가 X에 대한 α-차등정보보호 자료이고 f가 $Z(X,e)$의 밀도함수이면 f로부터 추출한 임의표본도 X에 대한 α-차등정보보호 자료이다. ■

나. 차등정보보호의 노출위험과 통계학의 가설검정

차등정보보호의 개념에서 모수 α의 역할은 직관적으로 이해하기 쉽지 않다. 최근 차등정보보호에 대한 이론이 통계학적 관점에서 연구되면서 모수 α를 [정리 2]에 제시된 가설에 대한 검정력의 상한(upper bound)으로 해석할 수 있음이 증명되었다(Wasserman과 Zhou, 2010; 박민정 외, 2018).

[정리 2] 원자료 $X=(X_1,...,X_n)$의 재현자료를 $Z=(Z_1,\ ...,Z_k)$라 하자. 다음과 같은 가설을 검정하기 위하여 재현자료 Z로부터 유의수준(significance level)이 γ인 검정법 T_i를 구성한다고 하자.

$$H_0 : X_i = u \ \text{ vs } \ H_1 : X_i = v, \quad u \neq v .$$

만약 Z가 X에 대한 α-차등정보보호 자료이면, 임의의 $i \in \{1,\ ...,n\}$에 대하여 T_i의 검정력(power)의 상한은 $\gamma exp(\alpha)$이다. ■

위 정리는 주어진 재현자료 Z를 사용하여 원자료에서 특정 레코드 X_i의 값을 식별(identification)할 수 있는가를 설명하고 있다. [정리 2]의 결과에 따르면 $\alpha \to 0$인 경우 $\gamma exp(\alpha) \to \gamma$가 되므로 검정법의 검정력이 유의수준과 거의 같아진다. 유의수준 γ는 0.1 이하인 경우가 대부분이므로, 결국 강한 차등정보보호($\alpha \to 0$)가 적용된 재현자료를 사용하면 특정 레코드의 값에 관하여 어떤 판단을 내리기가 어려워진다는 것을 알 수 있다.

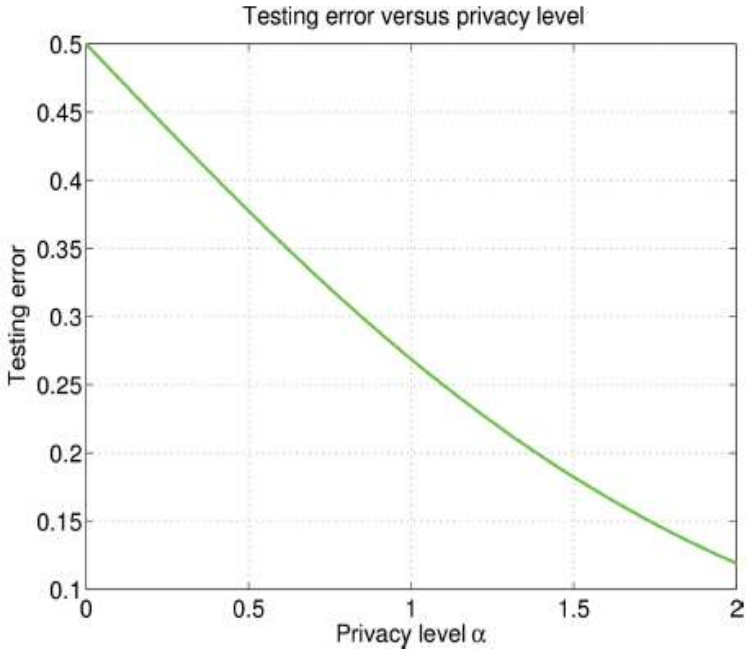
이러한 아이디어를 바탕으로 Duchi et al.(2018)은 다음과 같이 정의한 검정오차(testing error)를 사용하여 차등정보보호 수준 α를 설명하였다.

$$P_{err} = \frac{1}{2} inf_{\psi} \{ P(\psi(Z) \neq v | X = v) + P(\psi(Z) \neq u | X = u) \} .$$

---

7) 측도 가능한 함수(measurable function)

위 검정오차의 정의에서 $\psi(Z)$는 재현자료를 이용한 검정결과이며, 첫 항은 제2종의 오류를, 두 번째 항은 제1종의 오류를 의미한다. [그림 9]는 차등정보보호 수준 α와 검정오차 사이의 관계를 보여준다. 만약 차등정보보호를 강력하게 적용하면($\alpha \to 0$) 제안된 검정오차는 $1/2$에 가까워지게 된다. 이는 차등정보보호 자료를 사용한 어떠한 검정법도 무작위 추측(random guess)에 불과함을 의미하며 결국 노출위험이 매우 낮아진다는 것을 뜻한다.



출처 : Duchi et al.(2018)

[그림 9] 차등정보보호 수준에 따른 검정 오차

이상의 논의에서 차등정보보호 수준 α를 검정력 혹은 검정오차의 개념으로 이해할 수 있음을 확인하였다. 즉, 정보보호 수준이 α로 주어졌다는 것은 외부인이 특정 개체를 식별하고자 하는 가설에 대해 검정오차가 일정수준 이상이 된다는 것을 보장한다. 검정오차가 높을수록 노출위험이 낮아지므로 차등정보보호 자료를 사용하는 이용자가 차등정보보호 수준 α를 좀 더 유의미하게 해석할 수 있다.

이와 같이 통계적 가설검정의 관점에서 차등정보보호 수준 α를 이해할 수 있지만, 실제 상황에서 그 의미를 해석하기 위해서는 추가 연구가 필요하다. 또한 박민정 외(2018) 2절의 보험료 사례와 같이, 일반적으로 이해하기 쉬운 예제를 활용하여 차등정보보호 수준 α를 설명하기 위한 노력도 꾸준히 지속될 필요가 있다.

다. 차등정보보호의 정보손실과 통계학의 최소최대 추정량

Wasserman(2012)은 차등정보보호에서 최소최대 이론(statistical minimax theory)의 중요성을 강조하였다. 최소최대 위험(minimax risk)은 추정량(estimator)의 정확성(accuracy)을 측정하는 하나의 방법이며, 추정량이 최악의 상황에서 어느 정도의 위험을 수반하는가를 측정한다. 최소최대 위험은 추정량을 얻는 방법의 정확성 또는 정보손실로 이해해도 무방하다.[8] 먼저 다음과 같이 몇 가지 기호를 정의하자.

$X_1, ..., X_n$	원자료
χ	원자료의 표본공간(sample space)
$\boldsymbol{P}$	χ 위에서 정의된 분포의 집합
$Z_1, ..., Z_k$	α-차등정보보호 자료
$\aleph$	α-차등정보보호 자료의 표본공간
$\boldsymbol{Q}_\alpha$	$\aleph$ 위에서 정의된 α-차등정보보호 기법(조건부 분포)의 집합
Θ	모수공간(parameter space)
$\theta : \boldsymbol{P} \to \Theta$	P에서 정의된 모수(parameter)
ρ	Θ 위에서 정의된 거리(metric)
$\hat{\theta}_n : \chi^n \to \Theta$	추정량(estimator)
$\hat{\theta}_n : \aleph^k \to \Theta$	재현자료 추정량(estimator)
$\Phi : R_+ \to R_+$	$\Phi(0) = 0$ 이고 증가하는 함수

재현자료 추정량 $\hat{\theta}_k = \hat{\theta}(Z_1, ..., Z_k)$에 대하여 아래와 같이 정의한 함수 ℓ을 손실함수(loss function)라고 한다.

$$\ell(\hat{\theta}_k, \theta) = (\Phi \circ \rho)(\hat{\theta}_k, \theta)$$

8) 다만 최소최대 위험은 차등정보보호를 생성하는 메커니즘에 대한 정보손실 측도이며, 데이터마다 측정하는 고전적인 정보손실 측도가 아님을 주의할 필요가 있다. 또한 최소최대 기준은 상당히 보수적이기 때문에 최소최대 추정량을 구하였다고 하더라도 그 추정량이 실제적으로 효율적인지에 관하여는 추가적인 분석이 필요하다.

손실함수는 θ 대신 $\hat{\theta}_k$를 사용하는 과정에서 발생하는 손실의 크기를 측도하는 함수이며, 간단하게 θ와 $\hat{\theta}_k$ 사이의 차이(difference)라고 생각해도 좋다. $\hat{\theta}_k$를 평가하는 경우 손실함수를 측도로 직접 사용하지는 않으며, 다음과 같이 위험함수(risk function) R을 고려한다.

$$R_{P,Q}(\hat{\theta}_k, \theta) = E_{P,Q}(\ell(\hat{\theta}_k, \theta)), \ Q \in \boldsymbol{Q}_\alpha, \ P \in \boldsymbol{P}.$$

손실함수가 α-차등정보보호 자료를 사용하여 추정량을 만들었을 때 발생하는 손실의 크기를 나타낸다면, 위험함수는 수집 가능한 모든 자료를 사용하여 추정량을 만들었을 때 발생하는 손실의 크기를 나타낸다. 즉 위험함수는 분포 P와 Q에서 발생하는 손실의 기댓값을 의미한다. 만약 위험함수가 0으로 수렴하면 추정량 $\hat{\theta}_k$가 모수 θ에 대하여 일치성(consistency)을 가진다고 말하며, 이를 고정된 분포 P와 Q에서 모수 θ에 대한 추정량의 평가 측도로 사용할 수 있다.

추정량 $\hat{\theta}_k$와 조건부분포 Q가 고정된 경우, 위험함수를 분포집합 $\boldsymbol{P}$에서 모수공간 Θ를 거쳐 양의 실수 집합으로 가는 하나의 함수로 이해할 수 있으며, 다음과 같이 최대위험 R을 정의할 수 있다.

$$R_Q(\hat{\theta}_k, \boldsymbol{P}) = sup_{P \in \boldsymbol{P}} \ R_{P,Q}(\hat{\theta}_k, \theta).$$

최대위험 $R_Q(\hat{\theta}_k, \boldsymbol{P})$는 가능한 모든 원자료 분포 중에서 고려한 추정량 $\hat{\theta}_k$의 최대위험이며, 최악의 분포에서 발생하는 $\hat{\theta}_k$에 대한 손실의 기댓값을 의미한다. 최대위험은 고정된 하나의 추정량과 조건부분포에 대한 위험을 의미하므로, 어떠한 추정량이 최적인가를 평가할 때 주어진 최대위험을 추정량 $\hat{\theta}_k$과 조건부분포 Q의 함수로 간주하여 다음과 같이 최소최대 위험(minimax risk) $R_{n,k}(\boldsymbol{P})$를 정의한다.

$$R_{n,k}(\boldsymbol{P}) = inf_{\hat{\theta}_k, \boldsymbol{Q}_\alpha} \ R_Q(\hat{\theta}_k, \boldsymbol{P}).$$

최소최대 위험은 α-차등정보보호 자료를 사용한 모든 추정량에 대하여 발생할 수 있는 최대위험을 의미한다. 따라서 주어진 추정량 $\hat{\theta}_k$가 최소최대 위험보다 더 큰 최대위험을 갖는다면 다른 추정량을 구하여 최대위험을 줄일 수 있다는 것을 의미하며, 최소최대 위험과 같은 최대위험을 갖는다면 이를 최소최대 추정량(minimax estimator)이라고 한다.

최소최대 위험은 어떠한 분포집합을 고려하는가에 따라 결정되며, 거의 대부분의 문제에서 잘 알려져 있다. 따라서 보통의 경우 주어진 추정량 $\hat{\theta}_k$의 최대위험을 계산하여 최소최대 위험과 비교하는 과정을 통해 추정량을 평가한다.

α-차등정보보호 자료를 통해 얻은 추정량과 원자료에서 얻은 추정량의 최소최대위험이 동일한 수렴속도(convergence rate)를 가지면, 원래 자료와 α-차등정보보호 자료를 이용한 추정량은 동일한 정확도(accuracy)를 가진다고 이해할 수 있다. 이는 최소최대 위험이라는 측도로 정보손실을 측정할 수 있고, 또한 이 측도가 효율적인 α-차등정보보호 자료를 재현하는 기준이 될 수 있음을 의미한다.

정리하면 만약 주어진 정보보호 수준 α를 기준으로 노출위험을 제어하는 차등정보보호를 적용할 때, 최소최대 위험이 동일한 메커니즘을 찾을 수 있다면 원자료와 차등정보보호 자료는 모수를 추정하기 위한 동일한 정보를 가진다고 할 수 있다.[9] 다음 [그림 0-10]은 정보손실 측도로 최소최대 위험을 사용하고, 차등정보보호 수준 α를 기준으로 노출위험을 제어할 때, 차등정보보호 자료의 위험-유용성 지도를 나타낸다.



주: 정보손실 측도가 최소최대 위험일 때, 원자료와 동일한 최소최대 위험을 가지는 매커니즘에서 차등정보보호 자료를 생성한 경우

[그림 10] 차등정보보호 자료의 위험-유용성 지도

참고로 베이지안 재현자료 연구는 정보손실을 줄이는 모형을 개발하는 것이 주요 목적이다. 생성된 자료는 안전하다고 여기므로, 유용성 향상에 집중하기 때문이다. 다만, 실제자료에 대해서 정보손실을 줄이도록 모형을 만들려고 하면, 비모수적 기법을

9) 원자료도 정보손실을 가진다는 것을 생각하면, 본문의 내용은 차등정보보호 자료를 사용한 추정량의 정확도에 대한 위험이 원자료의 그것과 동일하면 두 자료의 정보손실이 같다는 것을 의미한다. 또한 이 정보손실을 최소최대 위험으로 측정한다. 차등정보보호 자료의 정보손실이 절대적으로 0이 라는 의미가 아님을 주의하기 바란다.

사용해야 하고 완전 재현방식보다는 부분 재현자료 생성을 선택할 수밖에 없다. 따라서 노출위험은 발생하지만 발생한 노출위험의 정도를 측정하는 것이 쉽지 않다. 즉, 베이지안 재현자료의 경우 차등정보보호와 달리 노출위험 제어를 보장하지 않으므로 매스킹과 유사한 딜레마에 도달하게 된다. 반면 차등정보보호는 노출위험 수준을 보장하며, 최소최대 위험을 기준으로 효율적인 알고리즘을 찾는다면 정보손실을 원래 자료와 동일하게 할 수 있다. 다만 정보손실 측도가 기존과 다르다는 것은 주지할 필요가 있다.

라. 차등정보보호 자료 생성: 교란 방식

원자료로부터 차등정보보호 자료를 얻는 가장 단순한 방법은 원자료를 이용해 구한 쿼리의 결과물에 적당한 오차를 더하여 교란(perturbation)하는 방법이다. 이때 교란 매커니즘이 생성하는 결과물은 차등정보보호 정의를 만족시켜야 한다. 다음의 예제는 Dwork et al.(2006)에서 제안하였으며 자료의 합 및 히스토그램에 대하여 α-차등정보보호를 구현하는 방법을 기술하고 있다.

[예제 1] 원자료를 $X = (X_1, ..., X_n)$, $X_i \in \{0,1\}$, $i \in \{1, ..., n\}$ 이라 하자. 오차 e 의 밀도함수가 $f_\epsilon(x) \propto exp(-|x-\mu|/\sigma)$ [10]이고 X와 독립이라고 하자. X와 e 을 사용하여 다음과 같이 원자료의 합에 대한 차등정보보호 자료를 구성할 수 있다.

$$Z = \sum_{i=1}^{n} X_i + e\,.$$

위 차등정보보호 자료가 X에 대한 α-차등정보보호 자료인가를 확인하기 위하여 간단한 계산 과정을 진행하면 $\sigma = 1/\alpha$ 인 경우 다음과 같은 부등식을 얻을 수 있으며, 이는 교란 매커니즘이 차등정보보호를 만족시킴을 나타낸다.

$$sup_{\|u-v\|_0 = 1}\ sup_z \frac{F_{Z|X}(z|u)}{F_{Z|X}(z|v)} = sup_{a,b \in \{0,1\}}\ exp(\alpha|a-b|) \le exp(\alpha). \quad \blacksquare$$

[예제 2] 원자료를 $X = (X_1, ..., X_n)$ 이라고 하면 X의 히스토그램은 다음과 같은 형태의 밀도함수로 정의된다.

---

10) 위치모수가 μ, 산포모수가 σ인 이중지수(double exponential) 혹은 라플라스(Laplace) 분포.

$$\hat{f}_m(x) = \sum_{j=1}^{m} \frac{C_j}{nh} I(x \in B_j).$$

단, h, B_j, C_j는 각각 히스토그램의 평활모수(bandwidth), 막대(bin), 막대 B_j에서의 자료의 빈도(frequency)이며, m은 막대의 개수이다. $V=(V_1,...,V_m)$을 아래와 같은 이중지수분포에서 추출한 임의표본이라고 하자.

$$g(v) = \frac{\alpha}{4} e^{-\alpha\frac{|v|}{2}}.$$

그러면 다음과 같이 생성한 자료 $D=(D_1, ..., D_m)$는 원자료 X에 대한 α-차등정보보호 자료이다.

$$D_j = C_j + V_j,\ j \in \{1, ..., m\}.\ \blacksquare$$

교란 매커니즘은 다양하게 만들 수 있으며, [정리 1]에 의해 차등정보보호 자료를 생성하는 알고리즘 중간에 적절한 변환 과정을 적용해도 여전히 차등정보보호를 만족한다. 따라서 주어진 α에 대해 정보손실을 줄일 수 있도록 효율적인 알고리즘을 고안하는 연구가 많이 수행되고 있으며, 하나의 주제에 대해 다양한 알고리즘이 제안되는 경우가 많다. 많은 경우, 교란 방식으로 얻은 차등정보보호 자료에 대해서는 고전적인[11] 정보손실 측도들을 사용한다.

마. 차등정보보호 자료 생성: 재현 방식

앞에서 설명한 [예제 2]는 Dwork et al.(2006)이 소개하였으며 원자료의 빈도 C_j 대신 이중지수분포 방법을 적용한 α-차등정보보호 자료 D_j를 사용하여 히스토그램을 생성할 수 있음을 보여준다. 다만 [예제 2]에서는 잡음을 추가하면서 막대가 가지는 빈도가 음수가 될 수 있다.

잡음을 직접 추가하는 교란 방법을 사용하는 대신에 Wasserman과 Zhou(2010)는 먼저 α-차등정보보호를 만족하는 재현자료를 생성하고, 재현자료를 사용하여 히스토그램을 작성하는 방법을 제안하였다. 이를 도식화하면 다음과 같다.

$$X_1, ..., X_n \xrightarrow[estimation]{density} \hat{p} \xrightarrow{privatize} \hat{\boldsymbol{p}}^* \xrightarrow{sample} Z_1, ..., Z_k.$$

11) 매스킹 기법을 평가할 때와 유사한 거리 기반 정보손실 측도들을 말한다. 자세한 내용은 박민정(2014)의 3절을 참조할 수 있다.

다음 [정리 3]은 [예제 2]의 α-차등정보보호 자료(히스토그램) D를 사용하여 또 다른 α-차등정보보호 자료(재현자료)를 생성할 수 있음을 보여주며, 이는 [정리 2]에 의한 결과이기도 하다.

[정리 3] 원자료 $X = (X_1, ..., X_n)$이 $X_i \in [0, 1]$, $i \in \{1, ..., n\}$이라고 하자. 임의의 k에 대하여 아래와 같은 밀도함수에서 추출한 임의표본 $Z = (Z_1, ..., Z_k)$는 원자료 X에 대한 α-차등정보보호 자료이다.

$$\tilde{f}_n(x) = \sum_{j=1}^{m} \frac{\hat{q}_j}{h} I(x \in B_j).$$

단, $\widetilde{D}_j = max\{D_j, 0\}$이고 $\hat{q}_j = \widetilde{D}_j / \sum_{s=1}^{m} \widetilde{D}_s$이다. ■

[정리 3]에서 소개한 차등정보보호 방법의 특징은 차등정보보호 자료의 개수 k를 임의로 조정할 수 있다는 점이다. F를 원자료의 분포함수라 하자. Wasserman과 Zhou (2010)는 $k \geq n$이고 L_2-거리함수를 손실함수(loss function)로 사용하는 경우, 적당한 정칙조건(regularity condition) 하에서 차등정보보호 자료 Z로 추정한 경험적 분포함수(empirical distribution) $\hat{F}_Z$의 위험함수(risk function)가 $E_{L_2}(F, \hat{F}_Z) = O(n^{-2/3})$을 만족하며, 따라서 최소최대 추정량(minimax risk estimator)이 된다는 것을 증명하였다. 또한 교란 히스토그램에서 재현자료를 얻는 것이 평활 히스토그램(smoothed histogram)에서 재현자료를 구성하는 방법보다 성능이 뛰어나다는 것을 추가로 증명하였다.

이상에서 소개한 방법 이외에도 기저함수를 이용하거나 일반화 이중지수분포를 활용하는 방법 등이 소개되어 있으나, 본 연구의 범위를 넘어선다고 판단되어 생략한다. 관련 내용은 선행연구를 참고하기 바란다. 다음 절에서는 차등정보보호를 적용하는 방식과 적용 결과에 대하여 논의하기 위해 α-차등정보보호를 만족하는 히스토그램 중 일부를 소개하고 각각의 특징을 살펴본다.

제3장 차등정보보호 히스토그램 생성 알고리즘

히스토그램은 사용자가 시스템에 요청하는 가장 대표적인 쿼리 중 하나이며, 자료 탐색 과정에서 가장 빈번하게 사용하는 통계이다. 히스토그램의 경우 쿼리 민감도(Dwork et al., 2006)가 1이므로, 정보보호 수준 α 와 자료의 유용성과의 관계를 파악하기에 가장 용이하다고 할 수 있다. 이러한 이유로 차등정보보호를 연구한 많은 논문들이 히스토그램을 주요 실험 대상으로 삼고 있다. 차등정보보호를 실제로 통계를 제공하는 시스템에 적용한다면 어떤 결과가 나타나는지를 파악하기 위해, 본 연구에서도 히스토그램을 실험 대상으로 선정하였다.

본 절에서는 차등정보보호 히스토그램(DP-compliant histogram)에 관한 설명을 교란 방식과 재현 방식으로 구분하여 진행하며, 각 히스토그램의 작성 절차와 이론적인 특징을 간략하게 정리하여 소개한다.

1. 원자료의 히스토그램

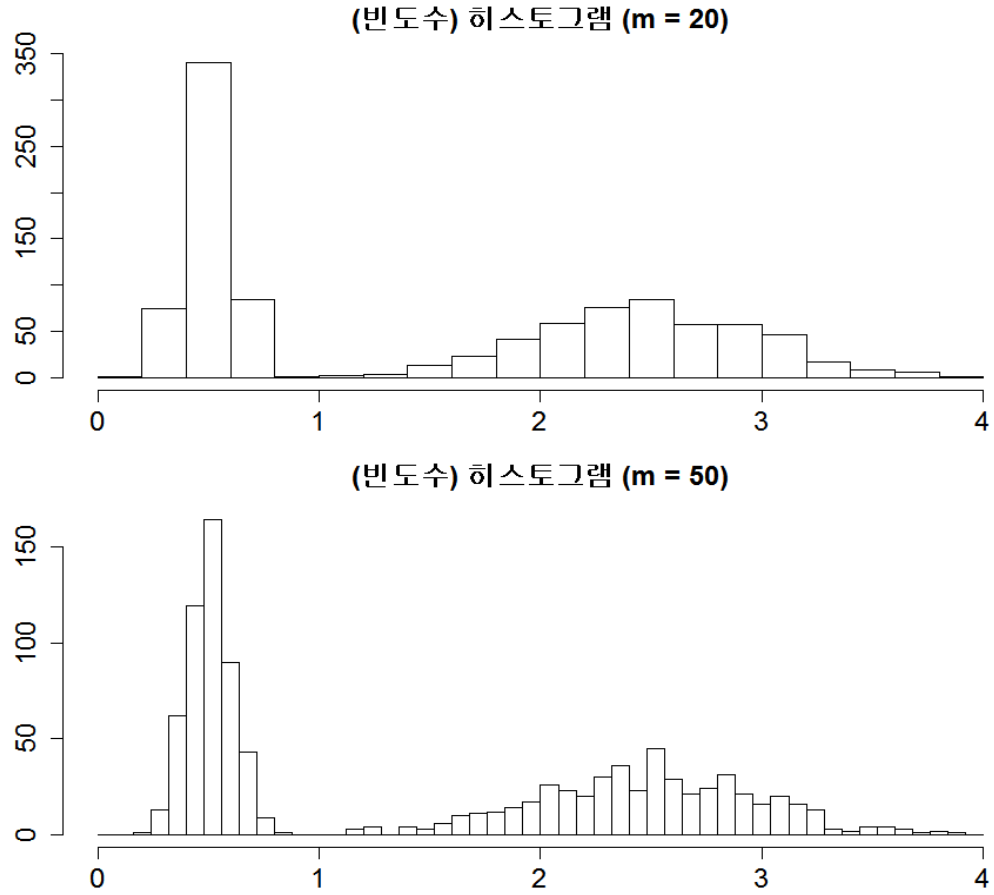
(요약) 통계학에서는 보통 동일한 크기의 막대를 사용하여 히스토그램을 작성한다. 적당한 조건을 만족시키는 히스토그램 추정량은 최소최대 추정량이며, 이는 작성된 히스토그램이 효율적이라는 것을 의미한다. 이제 차등정보보호를 만족하면서 동일한 효율성을 가지는 히스토그램을 생성하는 알고리즘을 찾는 것이 주요 논의 사항이 된다.

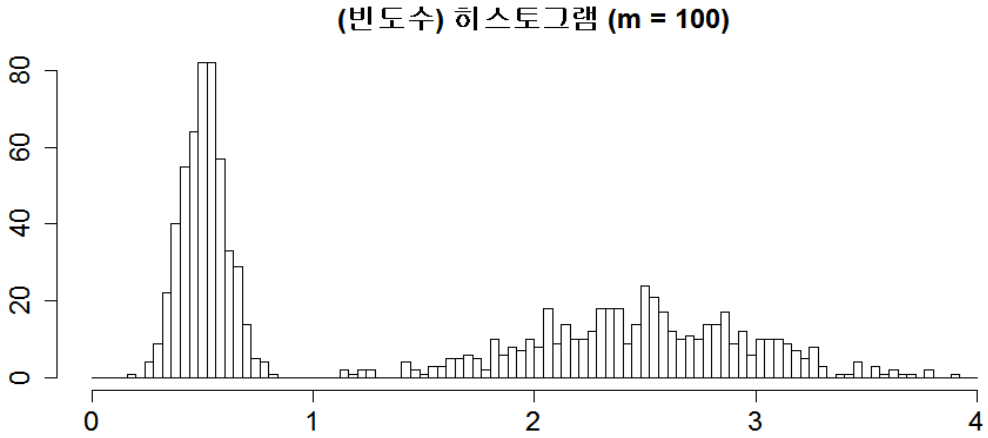
한편, 공학 분야에서는 동일한 크기의 막대로 그린 히스토그램에서 각 막대를 통합해 더욱 적절한 구조, 즉 최적 구조를 가지는 히스토그램을 생성하는 알고리즘들이 제안된 바 있다. 이런 최적 구조를 이용해 정보손실이 작은 차등정보보호 히스토그램을 생성하는 알고리즘을 고안할 수 있으므로, 최적 구조 히스토그램 생성 알고리즘도 함께 소개한다.

가. 히스토그램의 특징

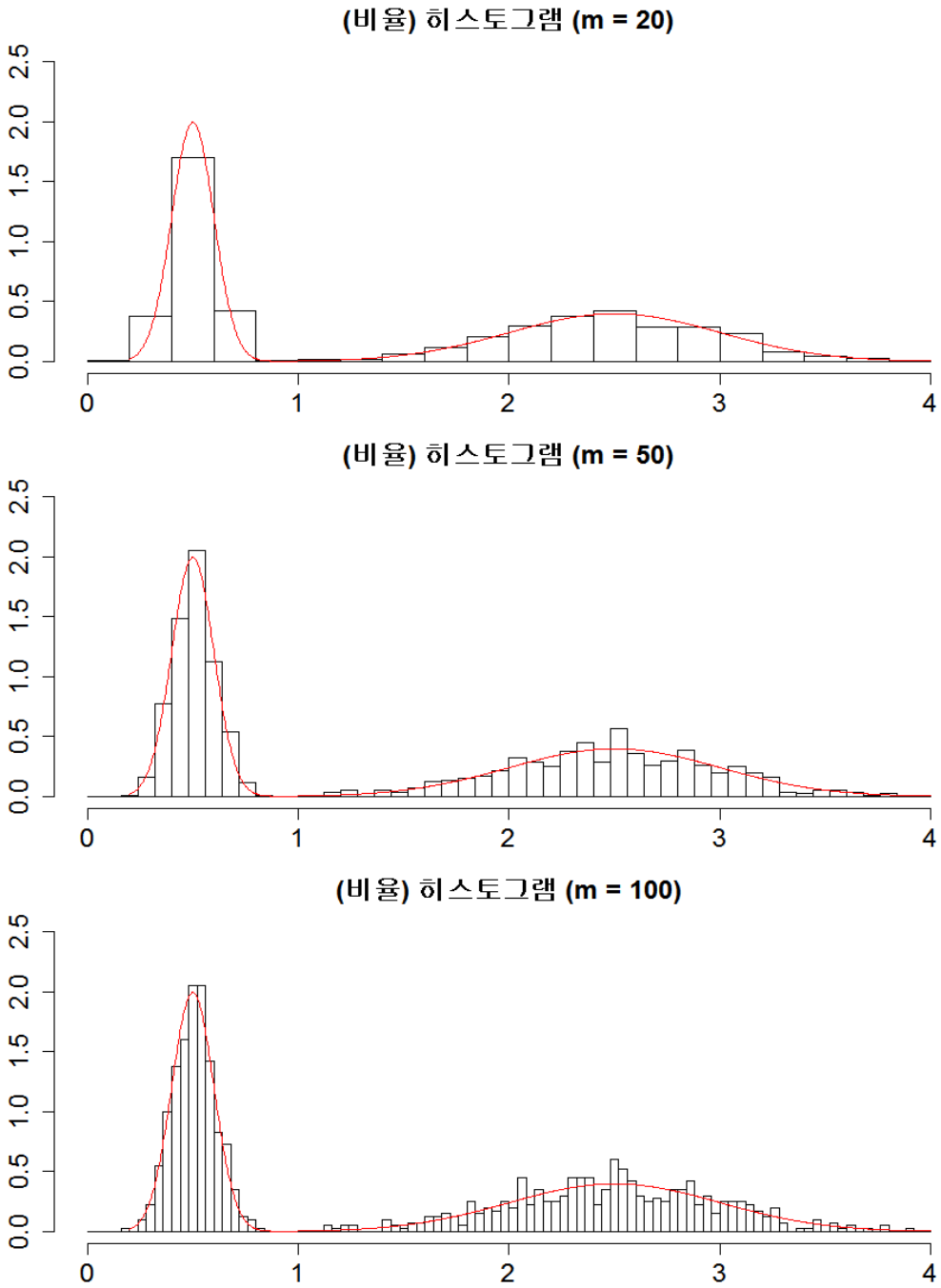
히스토그램의 특징을 설명하기 위해 두 정규분포 $N(0.5, 0.1^2)$ 과 $N(2.5, 0.5^2)$ 의 혼

합분포(normal mixture)를 고려하였다. [그림 11]은 이 혼합분포에서 임의표본을 추출하여 작성한 히스토그램이며 표본의 수는 $n = 1000$, 막대의 개수는 $m \in \{20, 50, 100\}$ 이다. 막대의 개수 m이 클수록 이용자는 자료 분포를 더 자세히 알 수 있으며, 동일한 개수의 막대를 사용하더라도 각 막대가 시작하는 위치에 따라 국소적으로 히스토그램의 모양이 달라질 수 있다.





[그림 11] 빈도수를 이용해 작성한 히스토그램: 막대 개수에 따른 변화



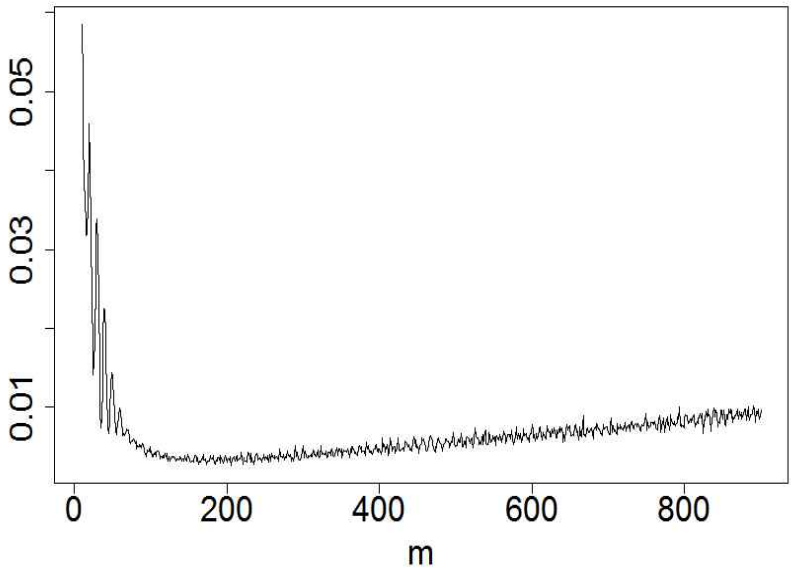
[그림 12] 비율을 이용해 작성한 히스토그램: 막대 개수에 따른 변화

빈도수 대신에 비율을 기준으로 히스토그램을 작성하면, [그림 12]와 같다. 붉은 색 곡선은 이 혼합분포의 밀도함수를 나타낸다. [그림 12]는 혼합분포의 밀도함수와 비교하기 위하여 [그림 11]에서 모든 막대의 면적의 합이 1이 되도록 조정된 것으로 이해할 수

있다.

막대의 개수 m이 크다는 것만으로 자료의 분포가 반드시 정확하게 표현되는 것은 아니다. 자료의 개수에 비해서 m이 지나치게 크면 오히려 자료의 분포를 적절하게 보여주는데 방해가 될 수 있다. 또한 주어진 m에 대하여 각 막대가 가지는 범위가 자료의 특성을 잘 표현하도록 설정되어야, 즉 막대의 위치가 적절하게 잡혀야 원래의 분포를 더욱 잘 표현할 수 있다.

[그림 13]은 1000개 자료에 대해 m개 막대를 가지는 히스토그램을 그릴 때 정보손실을 보여준다. 정보손실은 분포 참값과 m개 막대를 가지는 히스토그램 값 사이의 평균제곱오차(MSE)를 이용해 계산하였다. 막대의 개수 m이 지나치게 크면, 오히려 원래 분포의 값을 기준으로 정보손실이 커지는 것을 알 수 있다.



[그림 13] 막대의 개수에 따른 히스토그램의 정보손실

또한 유사한 크기의 m이라고 하더라도 막대의 범위가 어디에서 시작하게 되느냐에 따라 정보손실이 국소적으로 달라지는 것을 볼 수 있다. 많은 이용자들이 히스토그램을 쉬운 통계로 여길 수도 있지만, 사실 히스토그램은 상당히 복잡한 주제이다. 그러하기에 최적의 히스토그램을 구하는 알고리즘에 대해 오랜 기간 상당히 많은 연구 결과가 축적되어 있기도 하다.

나. 기본 히스토그램

원자료를 $X=(X_1,\ ...,\ X_n)$이라고 하고, 편의상 $X_i \in \chi = R^d,\ i \le n,\ d \ge 1$이라고 하자. 원자료 X의 히스토그램은 밀도함수의 추정량이며 다음과 같이 정의된다.

$$\hat{f}_m(x) = \sum_{j=1}^{m} \frac{C_j}{nh^d} I(x \in B_j) = \sum_{j=1}^{m} \frac{\hat{p}_j}{h^d} I(x \in B_j).$$

단, h, B_j, C_j는 각각 히스토그램의 평활모수(bandwidth), 막대[12](bin), 막대 B_j에서의 자료의 빈도(frequency)이며, m은 막대의 개수이다. 또한 $\hat{p}_j = C_j/n$이다.

임의의 $B \subset R^d$에 대하여 $V = Vol(B)$를 B의 부피(volume)라고 하자. $d=1$인 경우 V는 길이(length)가 된다. 다음 <표 3>은 원자료를 사용한 히스토그램의 작성 절차이며, 본 연구에서는 이 히스토그램을 "기본 히스토그램"이라고 한다.

〈표 3〉 기본 히스토그램

(1) 히스토그램 막대의 개수 m을 결정한다.

(2) 다음을 만족하는 m개의 막대(bin) B_j를 결정한다.

$$B_j \cap B_{j'} = \varnothing,\ \forall j \ne j',\ \cup_{j=1}^{m} B_j = R^d.$$

(3) 막대에 포함된 원자료의 빈도(frequency) C_j[13], $j \le m$을 계산한다.

$$C_j = \sum_{i=1}^{n} I(X_i \in B_j),\ j \le m.$$

(4) 막대의 부피 $V_j = Vol(B_j),\ j \le m$를 사용하여 히스토그램을 작성한다.

$$\hat{f}_m^{original}(x) = \frac{1}{n} \sum_{j=1}^{m} \frac{C_j}{V_j} I(x \in B_j),\ x \in R^d.$$

---

12) 히스토그램 bin을 번역한 용어로 '막대'를 사용하며, '구간'을 병행하여 이용한다.

13) 합이 n인 가중치 w_i, $i \le n$이 주어진 경우, 가중빈도 $C_j = \sum_{j=1}^{n} w_i I(X_i \in B_j)$, $j \le m$을 사용할 수 있다. 다만, 이른바 빅데이터(정밀하게 표본설계되어 얻어진 조사자료가 아닌, 행정자료 등의 전수자료)를 대상으로 차등정보보호를 적용하는 경우, 가중값에 대해 고려할 이유가 없다.

표본의 수 n이 무한히 증가하면, 밀도함수 공간에 정의된 적당한 거리(distance)에 대하여, 위 히스토그램이 밀도함수로 수렴한다는 사실이 잘 알려져 있다. 따라서 히스토그램은 밀도함수에 대한 일치(consistent) 추정량이다. 하지만 원자료를 사용한 히스토그램은 정보누출의 위험성이 존재하며[14] 이러한 위험성을 제거하기 위하여 다양한 α-차등정보보호 기법을 적용할 수 있다.

한편, 특정 조건(Lipschitz condition)을 만족하는 밀도함수 추정량은 [정리 4]와 같은 최소최대 수렴 속도(minimax rate of convergence)를 가지는 것으로 알려져 있다.

[정리 4] 적당한 상수 $L > 0$에 대하여 원자료의 밀도함수 p가 다음의 성질을 만족한다고 하자.

$$p \in \boldsymbol{P} = \{p : |p(x) - p(y)| \le L\|x - y\|\}$$

만약 $h \to 0$, $nh^d \to \infty$, $m = m_n \asymp n^{d/(2+d)}$ 이면 원자료의 기본 히스토그램 $\hat{f}_m^{original}$은 다음을 만족한다.

$$\int (p - \hat{f}_m^{original})^2 = O_p(n^{-2/(2+d)}).$$

단, $a_n \asymp b_n$은 충분히 큰 n에 대하여 a_n/b_n 및 b_n/a_n이 모두 유계임을 의미한다. ■

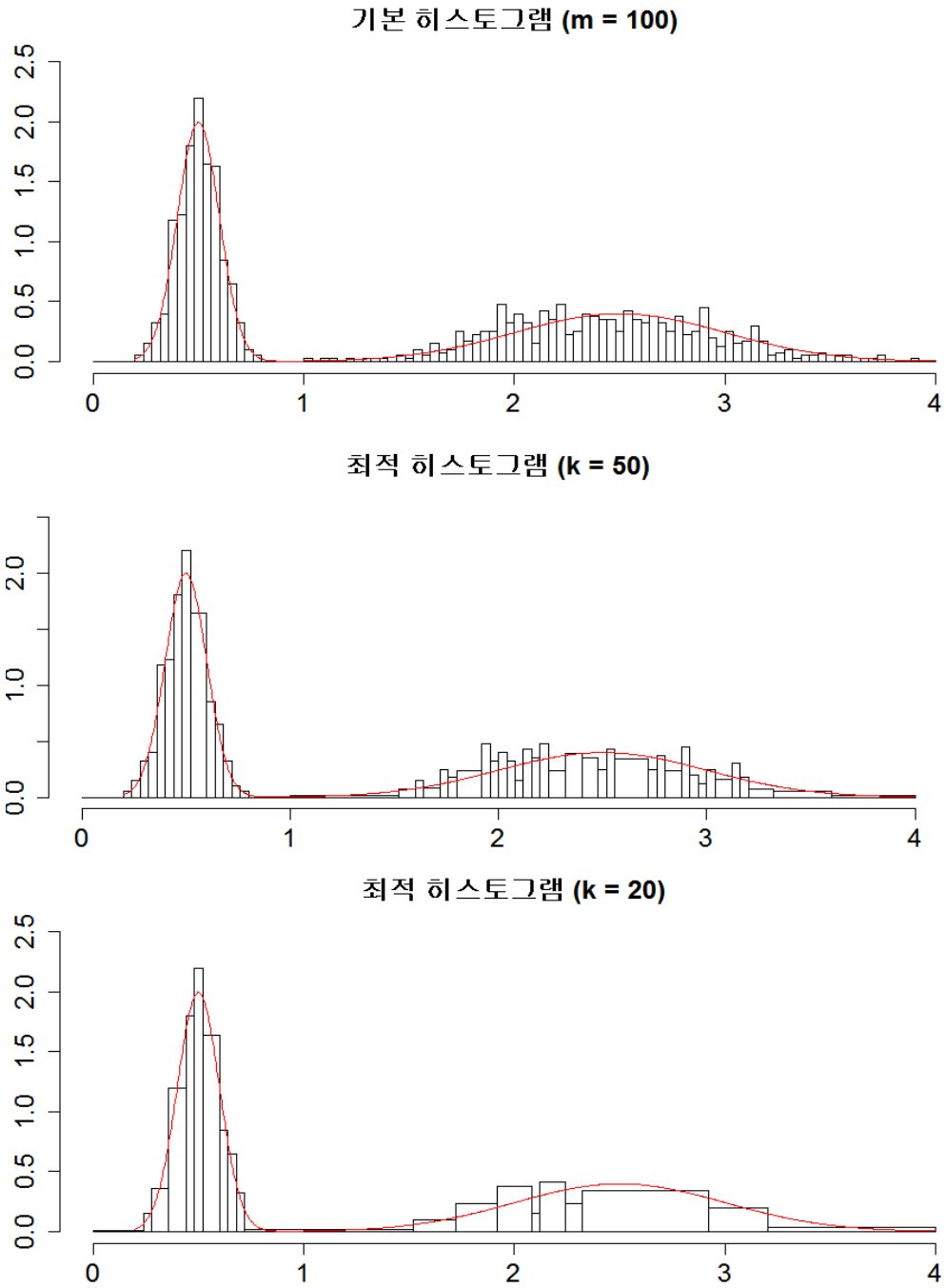
즉, L_2-거리 함수를 사용하여 최소최대 위험을 정의하는 경우 기본 히스토그램은 원자료 밀도함수의 최소최대 추정량이 된다. 따라서 기본 히스토그램은 α-차등정보보호 히스토그램을 평가하는 과정에서 하나의 비교 기준으로 사용 가능하다.

다. 최적 구조 히스토그램

동일한 평활모수 h를 사용하는 경우 기본 히스토그램은 모든 막대의 너비가 동일한 특징을 가진다. 하지만 막대의 너비가 다양하게 정의된 히스토그램도 꾸준히 연구되고 있다. Jagadish et al.(1998)은 이러한 히스토그램 중 하나를 제안하였다. 너비가 균일한 m개의 막대로 구성된 히스토그램이 주어져 있다고 하자. Jagadish et al.(1998)은 주어진

14) 이용자가 자료를 공격적으로 분석한다고 가정해보자. 이용자는 먼저 데이터베이스 D 전체에 대해 히스토그램을 그린다. 다음으로 자료 하나가 부족하거나 많은 데이터베이스 D'에 대한 히스토그램을 요청하여 둘을 비교할 것이다. D 및 D'가 하나의 자료가 있고 없고의 차이를 가진다면, 히스토그램 중 하나의 막대에서 값이 1만큼 차이날 것이다. 차이가 발생한 막대가 가지는 자료 값의 범위를 통해 이용자는 하나의 개체가 가지는 값을 유추할 수 있다. m의 값이 크면 좀 더 자세한 범위에 대한 정보를 얻을 수 있다. 따라서 경우에 따라 정확한 값이 노출될 수 있다.

히스토그램의 막대를 병합하여 $k(\leq m)$ 개의 막대를 가지는 새로운 히스토그램으로 구조를 바꾸는 방법을 연구하였으며, 가능한 방법 중 최적 구조를 가지는 히스토그램을 찾는 알고리즘을 제안했다. 본 연구에서는 Jagadish et al.(1998)의 알고리즘으로 얻는 히스토그램을 “최적 구조 히스토그램”이라 하겠다.



[그림 14] 기본 히스토그램 및 최적 구조 히스토그램

[그림 14]는 100 개의 막대로 구성된 기본 히스토그램($MSE = 0.2038$)과 기본 히스토그램으로부터 구한 두 개의 최적 구조 히스토그램을 보여준다. 새로 생성하는 히스토그램에서 막대의 개수를 $k = 50$ 으로 정하면, 결합하는 막대의 구성에 따라 많은 구조들이 존재한다. 그 중 Jagadish의 알고리즘을 이용해 최적 구조를 구하면 [그림 14]의 두 번째 그림이 되며, 정보손실은 $MSE = 0.2037$ 로 측정된다. 막대의 개수가 $k = 20$ 인 경우, 세 번째 그림이 보여주는 히스토그램이 최적 구조이며, 정보손실은 0.2032 로 측정된다. 최적 구조를 구하는 알고리즘의 세부 내용은 <부록> 1.을 참고하라. 이용자에게 원자료의 히스토그램을 제공할 때도 이러한 방식을 활용하는 것은 검토해 볼 만하다. 시각적으로 더 좋은 효과가 있을 수 있기 때문이다.

한편 차등정보보호 히스토그램을 구할 때 이런 방식을 활용하는 이유는 다음과 같다. 앞에서 설명했던 교란 방식, 즉 최종 통계량에 직접 잡음을 더해주는 방식으로 자료를 변형하는 경우, 기본 히스토그램에서는 $m = 100$ 개의 잡음을 추가하지만 최적 히스토그램에서는 $m = 20$ 혹은 $m = 50$ 개의 잡음만 추가하기 때문이다. 이는 차등정보보호를 적용하더라도 정보손실을 더 적게 발생시키는 알고리즘을 고안해 볼 가능성을 보여준다.

다만 이러한 알고리즘을 통해 얻는 히스토그램에 대해서는 최소최대 위험을 정보손실 측도로 사용할 때의 효율성을 논하기는 어렵다. 기존에 활용하던 거리 기반 정보손실 측도를 사용하여 통계량의 정보손실을 비교할 수 있다. 또한 최적 구조 알고리즘을 이용해 히스토그램을 생성하기 위해서는 다른 방식에 비하여 상당히 많은 계산 시간이 필요하다.

2. 차등정보보호 히스토그램: 교란 방식

(요약) 기본 히스토그램에 잡음을 추가하여 차등정보보호 히스토그램을 만들 수 있다. 이를 교란 히스토그램이라 하자. 정보손실을 줄이기 위해 보정된 교란 히스토그램, 양으로 보정된 교란 히스토그램, 동일 지지집합 교란 히스토그램을 만들 수 있다. 한편 교란 히스토그램에서 최적 구조 히스토그램을 구하여 교란 후 최적 구조 히스토그램을 생성할 수 있다. 총 다섯 가지 알고리즘을 설명한다.

이제 차등정보보호를 만족하는 히스토그램을 생성하는 알고리즘을 소개하자. 앞에서 원자료를 이용해 히스토그램을 만들 때 동일한 평활모수를 사용하여 동일한 너비의 막대를 가지는 히스토그램을 기본 히스토그램이라 하였다. 기본 히스토그램에 잡음을 추가하는 방식을 교란 히스토그램이라 한다. 교란 히스토그램을 생성하는 알고리즘을 조금씩 변형하여 정보손실을 축소시켜 새로운 차등정보보호 히스토그램 생성 알고리즘을 만들 수 있다.

이에 본 연구에서는 Wasserman과 Zhou(2010)에서 제안한 보정된 교란 히스토그램을 실험하였다. 또한 이를 조금씩 수정하여 양으로 보정된 히스토그램 및 동일 지지집합 교란 히스토그램을 제안하고 실험하였다. 또한 교란 히스토그램 및 최적 구조 알고리즘을 이용한 Xu et al.(2013)의 제안을 "교란 후 최적 구조 히스토그램"이라고 명명하고 실험하였다. 다음 각 소절에서 알고리즘을 차례로 정리한다.

가. 교란 히스토그램

최초의 α-차등정보보호 히스토그램은 Dwork(2006)의 교란(perturbed) 히스토그램이다. 교란 히스토그램은 원자료 히스토그램의 빈도를 잡음(noise)으로 교란(perturbation)하여 작성하는 히스토그램이다. Dwork(2006)은 라플라스(Laplace) 혹은 이중지수(double exponential) 분포에서 생성한 임의표본을 잡음(noise)으로 사용하였다. 평균이 0이고 산포모수[15]가 $b(>0)$인 이중지수분포의 밀도함수는 다음과 같다.

$$f(x) = (2|b|)^{-1} e^{-|x|/|b|},\ x \in R.$$

〈표 4〉 교란 히스토그램(Dwork, 2006)

(1) 원자료를 사용하여 기본 히스토그램을 작성한다.

$$\hat{f}_m^{original}(x) = \frac{1}{n}\sum_{j=1}^{m}\frac{C_j}{V_j} I(x \in B_j),\ x \in R^d.$$

(2) 정보보호 수준 $\alpha > 0$를 결정한다.

(3) 산포모수가 $b = 2/\alpha$인 이중지수 분포에서 임의표본 $W_j,\ j \le m$을 생성한다.

(4) 교란빈도 $D_j = C_j + W_j,\ j \le m$을 사용하여 히스토그램을 작성한다.

$$\hat{f}_m^{perturbed}(x) = \frac{1}{D}\sum_{j=1}^{m}\frac{D_j}{V_j} I(x \in B_j),\ x \in R^d,\ D = \sum_{j=1}^{m} D_j.$$

Dwork(2006)은 임의의 $\alpha > 0$에 대하여 교란빈도 $D_j,\ j \le m$는 원자료의 α-차등정보보호 자료임을 증명하였다. Dwork(2006)의 교란 히스토그램은 구현이 쉽고 정보보호 과정을 쉽게 이해할 수 있다는 장점을 가진다.

15) 산포모수가 b인 경우 분산은 $2b^2$, 표준편차는 $\sqrt{2}b$이다.

나. 보정된 교란 히스토그램

교란 히스토그램은 자료의 수 n이 너무 작거나 정보보호 수준 α를 너무 작게 잡으면 교란빈도가 음의 값을 가지는 경우가 발생할 수도 있다. 이러한 경우 교란 히스토그램은 밀도함수가 될 수 없으므로 Wasserman과 Zhou(2010)에 소개된 방법으로 교란 빈도를 보정할 수 있다.

〈표 5〉 보정된 교란 히스토그램(Wasserman과 Zhou, 2010)

(1) 원자료를 사용하여 기본 히스토그램을 작성한다.

$$\hat{f}_m^{original}(x) = \frac{1}{n}\sum_{j=1}^{m}\frac{C_j}{V_j}I(x \in B_j),\ x \in R^d.$$

(2) 정보보호 수준 $\alpha > 0$를 결정한다.

(3) 산포모수가 $b = 2/\alpha$인 이중지수 분포[16]에서 임의표본 $W_j,\ j \le m$을 생성한다.

(4) 보정된 교란빈도 $M_j = max\{C_j + W_j,\ 0\},\ j \le m$을 사용하여 히스토그램을 작성한다.

$$\hat{f}_m^{modified}(x) = \frac{1}{M}\sum_{j=1}^{m}\frac{M_j}{V_j}I(x \in B_j),\ x \in R^d,\ M = \sum_{j=1}^{m}M_j.$$

Wasserman과 Zhou(2010)은 임의의 $\alpha > 0$에 대하여 보정빈도 $M_j,\ j \le m$은 원자료의 α-차등정보보호 자료임을 증명하였다. 보정된 교란 히스토그램 작성법은 음의 교란빈도를 강제로 0으로 대체한다. 따라서 원자료의 히스토그램에서 빈도가 0이 아닌 막대가 보정된 교란 히스토그램에서 빈도가 0인 막대로 바뀔 수 있으며 반대의 경우도 발생한다. 이는 정보보호의 측면에서 도움이 될 수도 있으나 히스토그램의 모양이 급격히 바뀌어 유용성의 측면에서 도움이 되지 않을 수도 있다.

16) 합이 n인 가중치 $w_i,\ i \le n$이 주어진 경우, $b = w_*/\alpha$를 사용한다. w_*는 가장 큰 두 가중치의 합이다.

다. 양으로 보정된 교란 히스토그램

Dwork(2006)의 교란 히스토그램과 Wasserman과 Zhou(2010)의 보정된 교란 히스토그램에서는 자료가 존재하는 원자료 히스토그램의 막대가 자료가 존재하지 않는 막대로 바뀔 수도 있다. 이와 같은 결과를 보정하려면 다음과 같이 이중지수 분포에서 생성한 난수에 절대값을 취하여 교란 빈도가 항상 양수가 되도록 할 수도 있다.

〈표 6〉 양(positive)으로 보정된 교란 히스토그램

(1) 원자료를 사용하여 기본 히스토그램을 작성한다.

$$\hat{f}_m^{original}(x) = \frac{1}{n}\sum_{j=1}^{m}\frac{C_j}{V_j}I(x \in B_j),\ x \in R^d.$$

(2) 정보보호 수준 $\alpha > 0$를 결정한다.

(3) 산포모수가 $b = 2/\alpha$인 이중지수 분포[17]에서 임의표본 $W_j,\ j \le m$을 생성한다.

(4) 보정된 교란빈도 $D_j^+ = C_j + |W_j|,\ j \le m$을 사용하여 히스토그램을 작성한다.

$$\hat{f}_m^{positive}(x) = \frac{1}{D^+}\sum_{j=1}^{m}\frac{D_j^+}{V_j}I(x \in B_j),\ x \in R^d,\ D^+ = \sum_{j=1}^{m}D_j^+.$$

평균이 0인 이중지수 분포의 밀도함수는 $f_{|W|}(w) = 2f_W(w),\ w > 0$을 만족한다. 따라서 밀도함수의 비율 $f_{|W|}(w)/f_{|W|}(w') = f_W(w)/f_W(w'),\ w > 0,\ w' > 0$이 유지되어 양으로 보정된 교란빈도와 보정된 교란빈도는 유사한 성질을 가질 것으로 기대된다. 또한 원자료의 빈도가 0이 아닌 막대의 경우 교란된 히스토그램에서도 빈도가 0이 되지 않도록 강제하는 장점을 가지므로 유용성의 측면에서 좀 더 도움이 될 수 있다. 양으로 보정된 교란빈도가 원자료의 α-차등정보보호 자료임을 확인하는 엄밀한 이론적 연구가 필요하겠지만 본 연구에서는 단순히 비교의 측면에서 양으로 보정된 교란히스토그램을 실험 대상으로 고려하였다.

17) 합이 n인 가중치 $w_i,\ i \le n$이 주어진 경우, $b = w_*/\alpha$를 사용한다. w_*는 가장 큰 두 가중치의 합이다.

라. 동일 지지집합(support) 교란 히스토그램

양으로 교란된 히스토그램과 비슷한 원리로, 원자료의 히스토그램과 완전히 동일한 지지집합(support)를 가지는 교란 히스토그램을 작성할 수도 있다. 여기서 지지집합은 빈도가 0이 아닌 막대의 모임을 말한다.

〈표 7〉 동일 지지집합(support) 교란 히스토그램

(1) 원자료를 사용하여 기본 히스토그램을 작성한다.

$$\hat{f}_m^{original}(x) = \frac{1}{n}\sum_{j=1}^{m}\frac{C_j}{V_j}I(x \in B_j),\ x \in R^d.$$

(2) 정보보호 수준 $\alpha > 0$ 를 결정한다.

(3) 산포모수가 $b = 2/\alpha$ 인 이중지수 분포[18]에서 임의표본 $W_j,\ j \le m$ 을 생성한다.

(4) 양의 교란빈도 $D_j^+ = C_j + |W_j|,\ j \le m$ 을 계산한다.

(5) 동일 지지집합을 가지도록 보정된 교란빈도 $M_j^+ = D_j^+ I(C_j \neq 0),\ j \le m$ 을 사용하여 히스토그램을 작성한다.

$$\hat{f}_m^{support}(x) = \frac{1}{M^+}\sum_{j=1}^{m}\frac{M_j^+}{V_j}I(x \in B_j),\ x \in R^d,\ M^+ = \sum_{j=1}^{m}M_j^+.$$

보정된 양의 교란빈도는 빈도가 0인 막대를 그대로 유지하므로, 교란 히스토그램은 원자료의 히스토그램과 지지집합이 동일하다는 특징을 가진다. 따라서 동일한 정보보호 수준에서 시각적으로 좀 더 원자료의 히스토그램과 유사한 히스토그램을 작성할 수 있다는 장점이 있다. 다만 원래 빈도의 값이 0인 막대를 그대로 보존하므로 노출제어 측면에서 논란의 여지가 있을 수 있다. 따라서 동일 지지집합을 가지도록 보정된 교란빈도가 원자료의 α-차등정보보호 자료가 되도록 알고리즘을 보완하는 이론적 연구가 추가적으로 필요하며 본 연구에서는 단순히 비교의 측면에서 동일 지지집합 히스토그램을 실험 대상으로 고려하였다.

---

18) 합이 n 인 가중치 $w_i,\ i \le n$이 주어진 경우, $b = w_*/\alpha$ 를 사용한다. w_* 는 가장 큰 두 가중치의 합이다.

마. 교란 후 최적 구조 히스토그램

차등정보보호가 가지는 강점은 수준 α에서 주어진 쿼리에 대하여 범용적으로 정보보호를 한다는 것이다. 또한 차등정보보호를 만족하는 자료에 다른 함수를 취하면, 여전히 차등정보보호를 만족한다. 때문에 수준 α에서 정보보호를 하였다면, 적절한 알고리즘을 적용해서 정보손실을 줄이기 위한 노력을 하면 된다. 적절한 함수를 적용해도 여전히 동일한 수준의 차등정보보호를 만족하므로 노출위험 제어에 변화가 없기 때문이다. 이러한 맥락에서 정보손실을 줄이도록 Xu et al.(2013)에서 제안한 알고리즘을 살펴보자.

이용자가 요청한 m개의 막대를 가지는 히스토그램 구조에 대한 정보를 S_m이라 하고 원자료 히스토그램 H_m을 $H(S_m)$으로, 교란 히스토그램은 $\widetilde{H}(S_m)$으로 표현하자. 주어진 k에 대해 원자료에 대한 최적 구조 S_k^*는 히스토그램 $H(S_m)$으로부터 얻을 수 있다.

Xu et al.(2013)의 아이디어는 히스토그램 $H(S_m)$의 m개 빈도값에 잡음을 추가하는 것보다 k개 막대를 가지는 히스토그램에 k개의 잡음만 추가하는 것이 전체적으로 정보손실을 줄일 수도 있다는 것이다. 이에 잡음을 추가한 $\widetilde{H}(S_m)$에서 정보손실을 최소화시키는 최적 막대의 개수 k^*를 구하는 알고리즘을 제안하였다. 최적 k^*가 얻어지는 구조를 S_{k^*}라고 하면 차등정보보호 히스토그램은 $\widetilde{H}(S_{k^*})$로 표현할 수 있다. 이를 교란 후 최적 구조 히스토그램이라 부르도록 한다. 한편, 같은 문헌에서 히스토그램 구조를 먼저 교란하는 알고리즘도 제시하였으나 본 실험 연구의 범위에 포함시키지는 않았다.

〈표 8〉 교란 후 최적 구조 히스토그램

(1) 정보보호 수준 $\alpha > 0$에 대해 교란 히스토그램을 작성한다.

$$\hat{f}_m^{perturbed}(x) = \frac{1}{D}\sum_{j=1}^{m}\frac{D_j}{V_j}I(x \in B_j),\ x \in R^d,\ D = \sum_{j=1}^{m} D_j.$$

(2) Xu et al.(2013)이 제안한 알고리즘을 이용해 최적 막대의 개수 k^* 및 최적 구조 S_{k^*}를 얻어 히스토그램을 생성한다.

$$\hat{f}_{k^*}^{pert.optimal}(x) = \frac{1}{E}\sum_{j=1}^{k^*}\frac{E_j}{V_j}I(x \in B_j),\ x \in R^d,\ E = \sum_{j=1}^{k^*} E_j.$$

3. 차등정보보호 히스토그램: 재현 방식

(요약) 차등정보보호를 만족시키는 밀도함수 추정량에서 임의추출한 자료를 (차등정보보호) 재현자료라 하며, 이 재현자료로 히스토그램을 생성할 수 있다. 또한 적절한 알고리즘으로 차등정보보호를 만족시키는 최소최대 추정량을 구할 수 있다.

앞 소절에서 설명한 교란기법들은 히스토그램 작성에 필요한 빈도만 생성하기 때문에 히스토그램 작성에 사용된 재현자료는 존재하지 않는다. 하지만 Wasserman과 Zhou (2010)의 연구 결과에 따르면 α-차등정보보호를 만족하는 자료(microdata)를 직접 생성하여 α-차등정보보호 히스토그램을 작성할 수도 있다.

가. 재현된 교란 히스토그램

Wasserman과 Zhou(2010)은 교란된 빈도로 히스토그램을 작성한 후 이 히스토그램으로부터 재현자료를 임의추출하고 추출된 재현자료를 사용하여 히스토그램을 다시 작성하는 방법을 제안하였다.

〈표 9〉 재현된 교란 히스토그램

(1) 보정된 교란 히스토그램을 작성한다.

$$\hat{f}_m^{modified}(x) = \frac{1}{M}\sum_{j=1}^{m}\frac{M_j}{V_j}I(x \in B_j),\ x \in R^d,\ M = \sum_{j=1}^{m} M_j.$$

(2) 재현자료의 개수 k를 결정한다.

(3) 보정된 교란 히스토그램에서 재현자료 Z_i, $i \le k$를 임의생성한다.

(4) 재현자료를 사용하여 히스토그램을 작성한다.

$$\hat{f}_m^{syn.perturbed}(x) = \frac{1}{R}\sum_{j=1}^{m}\frac{R_j}{V_j}I(x \in B_j),$$

$$x \in R^d,\ R_j = \sum_{i=1}^{k} I(Z_i \in B_j),\ R = \sum_{j=1}^{m} R_j.$$

Wasserman과 Zhou(2010)은 임의의 $\alpha > 0$ 와 $k > 0$ 에 대하여 재현자료 Z_i, $i \le k$가 원자료의 α-차등정보보호 자료임을 증명하였다. 따라서 재현된 교란 히스토그램 작성법은 α-차등정보보호 히스토그램과 α-차등정보보호 재현자료 구성을 동시에 진행하고자 하는 경우 유용한 방법이다. 특히 재현자료를 원자료보다 더 많이 생성할 수도 있으며 (4)번 과정에 필요한 막대의 개수를 k의 크기에 따라 적절히 조정하는 것도 가능하다. 또한 보정된 교란 히스토그램이 아닌 다른 α-차등정보보호 히스토그램을 사용해도 무방하다.

나. 재현된 평활 히스토그램

다음으로 히스토그램의 빈도를 직접 교란하지 않고 α-차등정보보호 히스토그램을 작성하는 방법을 소개한다. 이 방법은 평활(smoothed) 히스토그램을 사용하여 α-차등정보보호 히스토그램을 작성하는 방법이다(Wasserman과 Zhou, 2010).

〈표 10〉 재현된 평활 히스토그램

(1) 원자료를 사용하여 기본 히스토그램을 작성한다.

$$\hat{f}_m^{original}(x) = \frac{1}{n}\sum_{j=1}^{m}\frac{C_j}{V_j} I(x \in B_j),\ x \in R^d.$$

(2) 평활수준 $\delta > 0$ 을 결정한다.

(3) 평활 히스토그램에서 재현자료 Z_i, $i \le k$를 임의생성한다.

$$\hat{f}_m^{\delta}(x) = (1-\delta)\hat{f}_m^{original}(x) + \delta / V,\ x \in R^d,\ V = \sum_{j=1}^{m} V_j.$$

(4) 재현자료를 사용하여 히스토그램을 작성한다.

$$\hat{f}_m^{syn.smoothed}(x) = \frac{1}{R}\sum_{j=1}^{m}\frac{R_j}{V_j} I(x \in B_j),$$

$$x \in R^d,\ R_j = \sum_{i=1}^{k} I(Z_i \in B_j),\ R = \sum_{j=1}^{m} R_j.$$

재현된 평활 히스토그램 작성 방법에는 정보보호 수준 $\alpha > 0$가 포함되지 않는 것처럼 보인다. 하지만 평활수준 δ와, 재현자료의 수 k, 막대의 개수 m 등은 정보보호 수준

과 이론적인 관계를 가진다. Wasserman과 Zhou(2010)은 아래 부등식이 성립하는 경우 평활 히스토그램에서 임의생성한 재현자료가 원자료의 α-차등정보보호 자료임을 증명하였다.

$$\alpha \geq klog\left(1+\frac{(1-\delta)m}{\delta n}\right). \qquad (0.1)$$

근사적으로 $\delta \to 0$, $m/n\delta \to 0$ 이면 $mk \leq n\alpha\delta$가 위 조건의 충분조건이 된다. 따라서 정보보호수준이 강할수록($\alpha \to 0$), 평활 수준이 작을수록($\delta \to 0$), 막대의 개수가 많을수록($m \to \infty$) 재현자료의 개수($k \to 0$)가 더 작아진다는 특징을 가진다.

제4장 차등정보보호 히스토그램 구현 및 평가

본 절에서는 앞에서 소개한 α-차등정보보호 히스토그램에 대하여 모의실험을 진행하고 그 결과를 서술한다. 먼저 실험을 통해 각각의 α-차등정보보호 히스토그램 작성법이 유한 표본에서 어떠한 장단점을 가지는지 파악한다. 나아가 실험 결과를 바탕으로 가계금융복지조사 자료의 소득 변수에 α-차등정보보호 히스토그램을 적용하고, 그 결과를 정리한다.

1. 차등정보보호 히스토그램: 모의실험

가계금융복지조사 자료의 소득 변수는 일부 고소득자의 영향으로 매우 치우친 분포를 가진다. 이러한 상황을 반영하기 위하여 두 개의 정규분포에 대한 혼합 정규분포(normal mixture distribution)를 사용하여 모의실험을 진행하였다.

모의실험에서는 3절에서 소개한 α-차등정보보호 히스토그램을 작성한 후, 혼합 정규분포 밀도함수와 히스토그램 사이의 거리를 계산하여 α-차등정보보호 히스토그램을 평가하였다. 동일한 정보보호 수준 α에 대하여 이 거리가 작을수록 α-차등정보보호 히스토그램이 원자료의 밀도함수와 비슷하다는 것을 의미하므로 더 좋은 방법이라고 평가할 수 있다. 한편 실제자료에서는 원자료의 밀도함수를 알 수 없으므로, 원자료의 히스토그램과 차등정보보호 히스토그램 사이의 거리도 참고로 제시하였다.

차등정보보호 히스토그램은 앞에서 설명한 교란 방식 히스토그램과 재현 방식 히스토그램을 고려하였다. 재현 히스토그램의 재현자료는 원자료의 수와 동일하게 생성하였으나, 평활 히스토그램을 사용하는 경우 제약식 (0.1)에서 등호를 만족하는 자료의 수가

100을 넘지 않으면 제외하였다. 실험에 사용된 밀도함수와 거리, 자료의 수와 정보보호 수준 등은 다음과 같다.

- 원자료 밀도함수 (단, ϕ는 정규분포의 밀도함수이다.)

  $f_X = (\phi(\mu_{1,}\sigma_1^2) + \phi(\mu_{2,}\sigma_2^2))/2\,,\ \mu_1 = -0.5,\ \mu_2 = 0.5,\ \sigma_1^2 = 0.09,\ \sigma_2^2 = 0.04$

- 정보보호 수준 : $\alpha = 10^{-k},\ k \in \{-2, -1, 0\}$
- 자료의 수 : $n = 100 \times 2^k,\ k \in \{0, 1, 2, 3, 4, 5\}$
- 막대의 수 : $m \in \{10, 20\}$
- 평활히스토그램의 평활상수 : $\delta \in \{0.1, 0.2\}$
- 두 밀도함수 f와 g사이의 거리 : $d(f, g) = \int_R |f(x) - g(x)|\,dx$

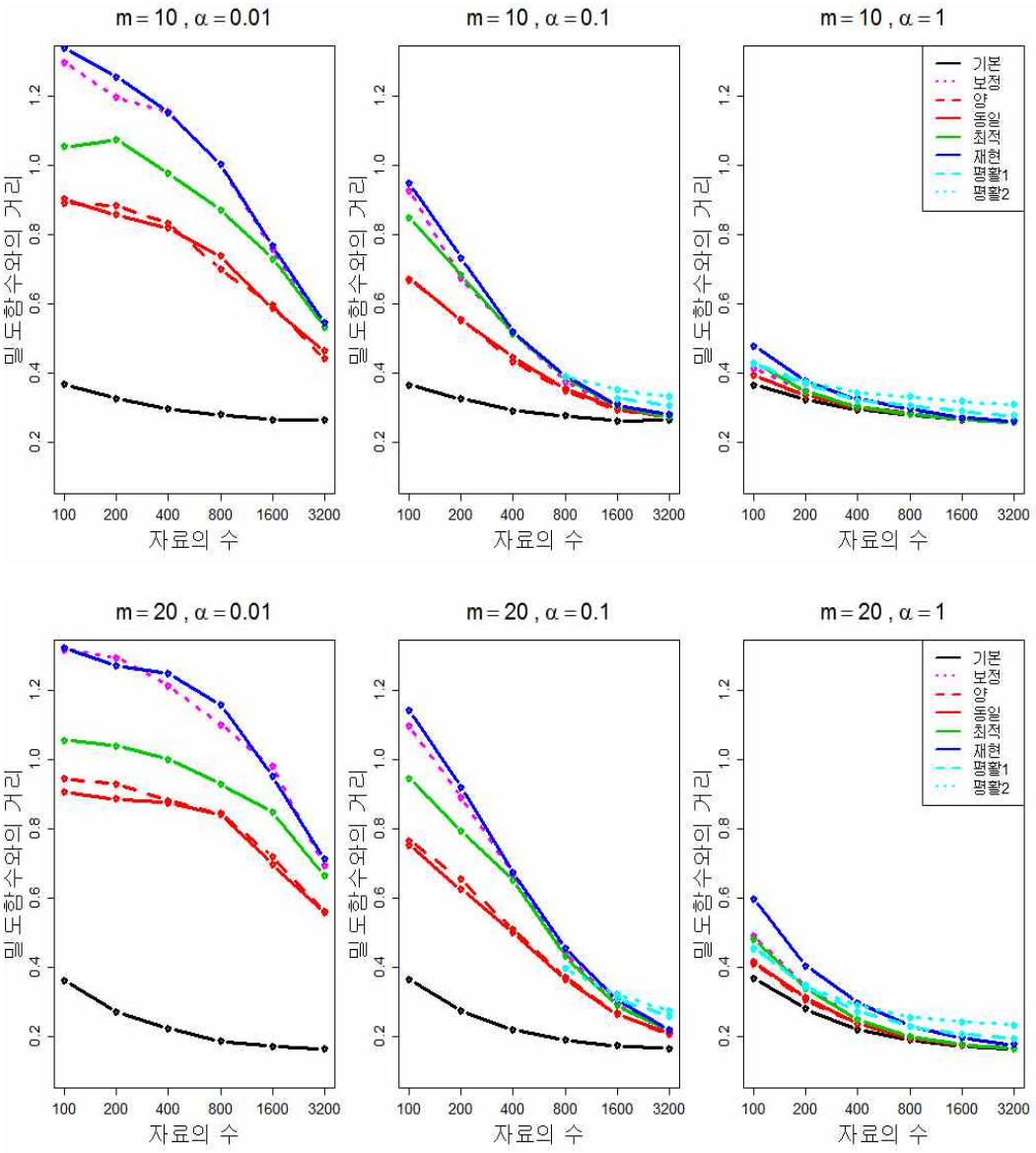
고정된 실험 조건에서 원자료 밀도함수로부터 n개의 임의표본을 생성하고, 이 임의표본을 사용하여 α-차등정보보호 히스토그램을 작성하고, 원자료 밀도함수와 기본 히스토그램에 대한 거리를 각각 계산하였다. 각각의 실험은 100회씩 반복하였으며 계산된 거리의 평균을 표와 그림으로 정리하였다. 교란 히스토그램은 밀도함수 추정량이 될 수 없는 경우가 발생하므로 실험 대상에서 제외하였다.

〈표 11〉 차등정보보호 히스토그램 생성 알고리즘

교란 방식	(보정) <표 5> 보정된 교란 히스토그램
	(양) <표 6> 양으로 보정된 교란 히스토그램
	(동일) <표 7> 동일 지지집합 교란 히스토그램
	(최적) <표 8> 교란 후 최적 구조 히스토그램
재현 방식	(재현) <표 9> 재현된 교란 히스토그램
	(평활) <표 10> 재현된 평활 히스토그램

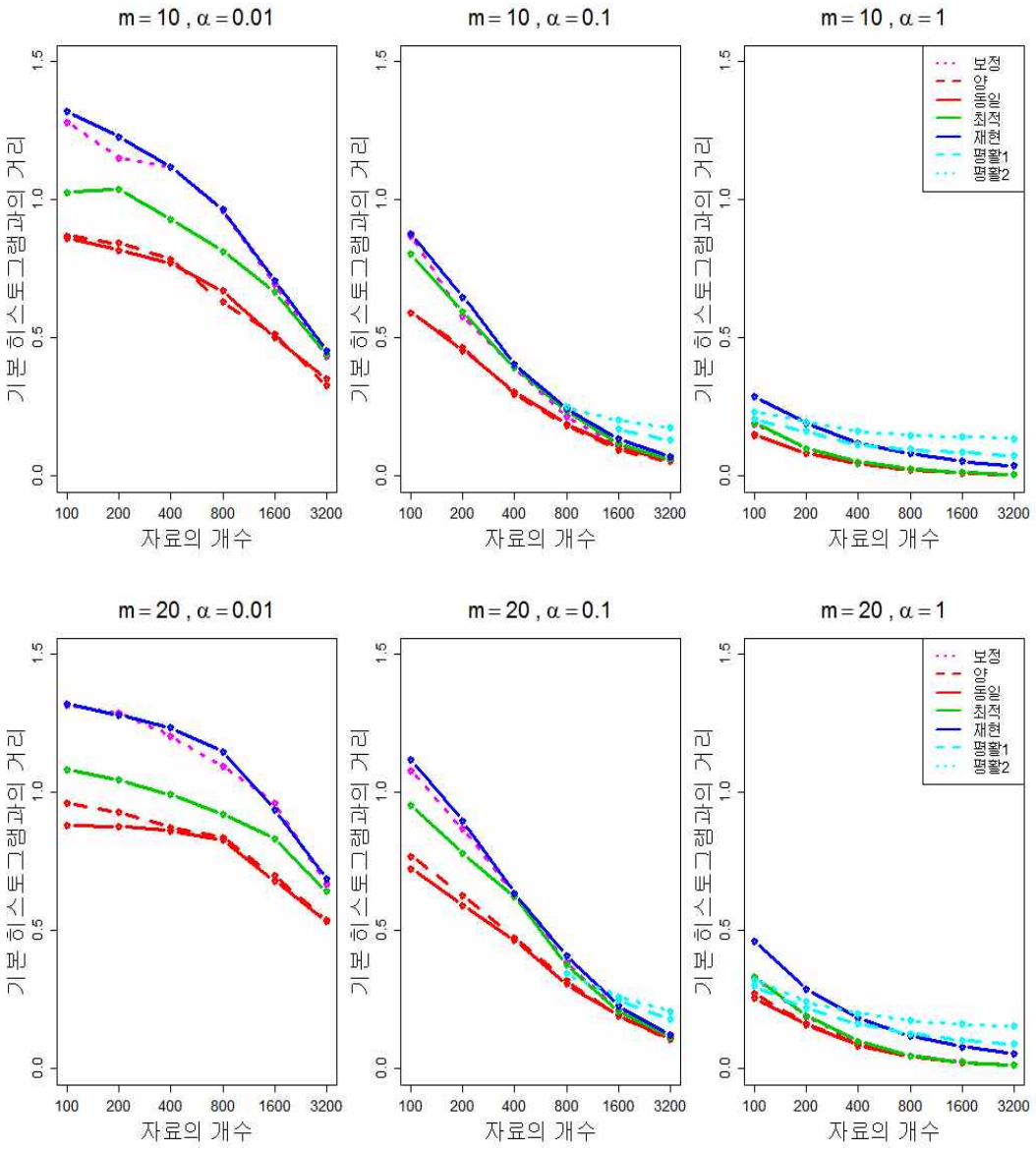
[그림 15]는 원자료 밀도함수와 차등정보보호 히스토그램 사이의 거리를 나타낸다. 정보보호 수준이 강할수록 (α 값이 작을수록) 정보손실이 크지만, 자료의 수가 많으면 정보손실이 작아지면서 각 방법별로 정보손실 차이가 줄어든다.

[그림 15]에서 보정된 교란 히스토그램의 정보손실이 가장 크며, 교란 후 최적 구조 히스토그램의 정보손실이 좀 더 적다. 그러나 양으로 보정된 교란 히스토그램 및 동일 지지집합 히스토그램이 이들보다 더 적은 정보손실을 가진다. 이로써 보정된 교란 히스토그램의 알고리즘을 수정하여 정보손실을 줄일 수 있다는 것을 확인하였다. 또한 자료의 크기가 충분히 크면 각 방법이 가지는 정보손실은 큰 차이를 보이지 않는다.



[그림 15] 원자료 밀도함수와 차등정보보호 히스토그램 사이의 거리

[그림 15]에서 재현 방식 히스토그램은 교란 방식 히스토그램보다 정보손실이 크다. 그러나 자료의 크기가 크면 그 차이가 급격히 줄어든다. 자료의 크기가 크면 재현 방식 중 재현된 교란 히스토그램이 재현된 평활 히스토그램보다 정보손실이 작다. 재현된 교란 히스토그램이 최소최대 추정량이었음을 주지할 때, 이는 자명한 결과일 수도 있다. [그림 15]와 관련된 표는 <부록>의 <부표 1> 및 <부표 2>에 수록되어 있다.



[그림 16] 기본(원자료) 히스토그램과 차등정보보호 히스토그램 사이의 거리

실제자료에서는 원자료의 밀도함수를 알 수 없으며 따라서 정보손실 역시 계산할 수 없다. 이러한 경우, 기본 히스토그램을 원자료의 적합한 추정량으로 간주하여 원자료 밀도함수 대신 사용할 수 있다. [그림 16]은 기본 히스토그램을 기준으로 차등정보보호 히스토그램이 가지는 정보손실을 계산한 결과이며 [그림 15]와 유사한 결과를 나타낸다. [그림 15]와 [그림 16]과 관련된 표는 <부록>의 <부표 3> 및 <부표 4>에 수록되어 있다.

앞에서 언급하였듯이 적절한 막대의 개수를 선택하는 것은 최적의 히스토그램을 작성하기 위한 매우 중요한 요소 중 하나이다. 따라서 동일한 실험 조건에서 막대의 수에 따른 정보손실의 변화를 파악하기 위하여 다음과 같은 조건에서 실험을 추가로 진행하였다.

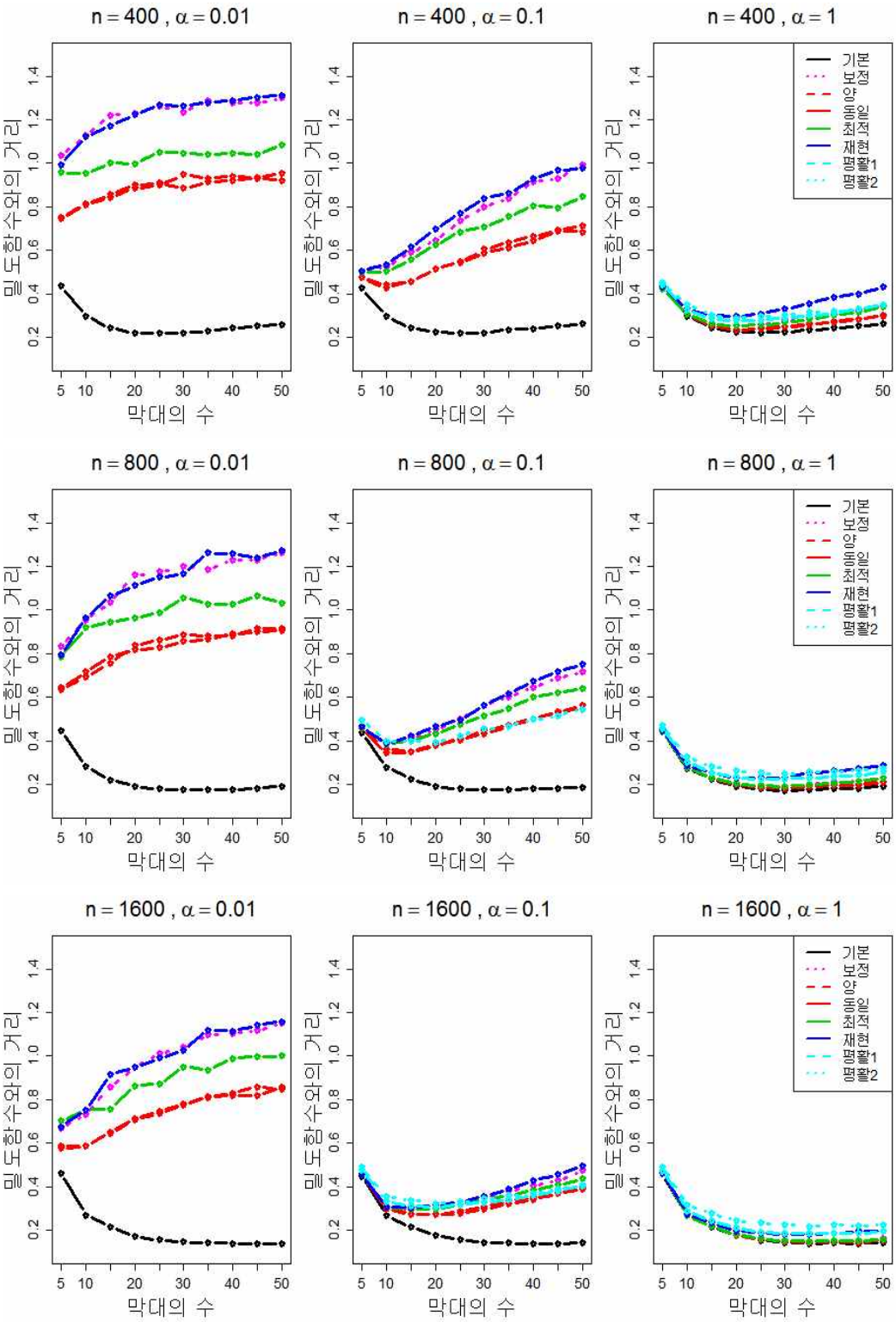
- 막대의 수 : $m = 5k,\ k \in \{1, 2, ..., 10\}$

[그림 17] 및 [그림 18]은 막대의 개수에 따른 정보손실을 나타낸다. 전체적으로 막대의 개수가 최적 개수에 가까울 때 정보손실이 줄어드는 것을 볼 수 있다. 또한 정보보호 수준이 강할수록 (α 값이 작을수록) 정보손실이 크지만, 같은 정보보호 수준에서는 자료의 수가 많을수록 정보손실 차이가 약간 완화되는 것을 알 수 있다. 실험의 결과를 간단히 정리하면 다음과 같다.

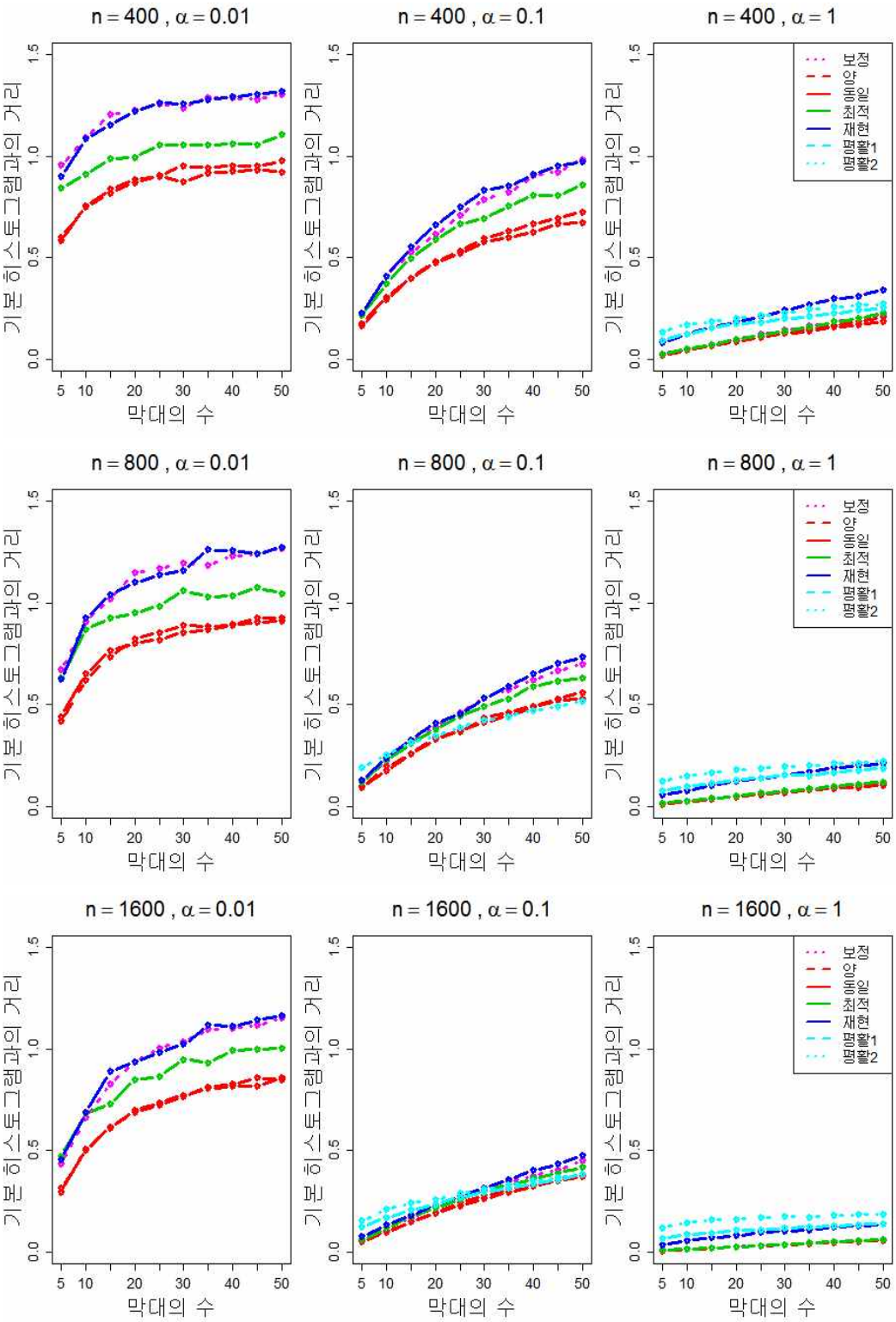
- $\alpha \leq 0.1$ 인 경우, n 이 작으면 막대의 개수가 많을수록 정보손실이 증가하는 경향을 보인다. 하지만 n 이 커지면 증가의 속도가 둔화되어 최적의 막대 개수가 발생하는 것을 확인할 수 있다.
- $\alpha \geq 1$ 인 경우, n 이 작으면 최적의 막대 개수가 발생하지만 n 이 커지면서 막대 개수가 증가함에 따라 정보손실이 계속 감소하는 것을 확인할 수 있다.

원자료의 히스토그램과 차등정보보호 히스토그램 사이의 정보손실은 [그림 18]에 나타냈다.

이상의 결과로부터 교란 방식이 가장 적은 정보손실을 가지며 재현 방식에서 정보손실이 가장 크게 발생하는 경향이 있음을 확인할 수 있다. 하지만 원자료의 수 n 과 정보보호 수준 α 가 큰(약한 정보보호) 경우 어느 방법을 사용해도 무관한 것으로 보인다. 재현 방식의 경우 다른 기법에 비하여 정보손실이 어느 정도 큰 편이지만 히스토그램과 함께 재현 자료를 얻을 수 있다는 장점이 있으므로 상황에 맞게 적당한 방법을 고려할 수 있을 것으로 판단된다. 참고로 재현된 평활 히스토그램의 경우, 평활 상수 δ를 더 작게 잡으면 정보손실을 나타내는 거리가 더 작아질 수 있다. 따라서 주어진 정보보호 수준 α와 막대의 개수 m



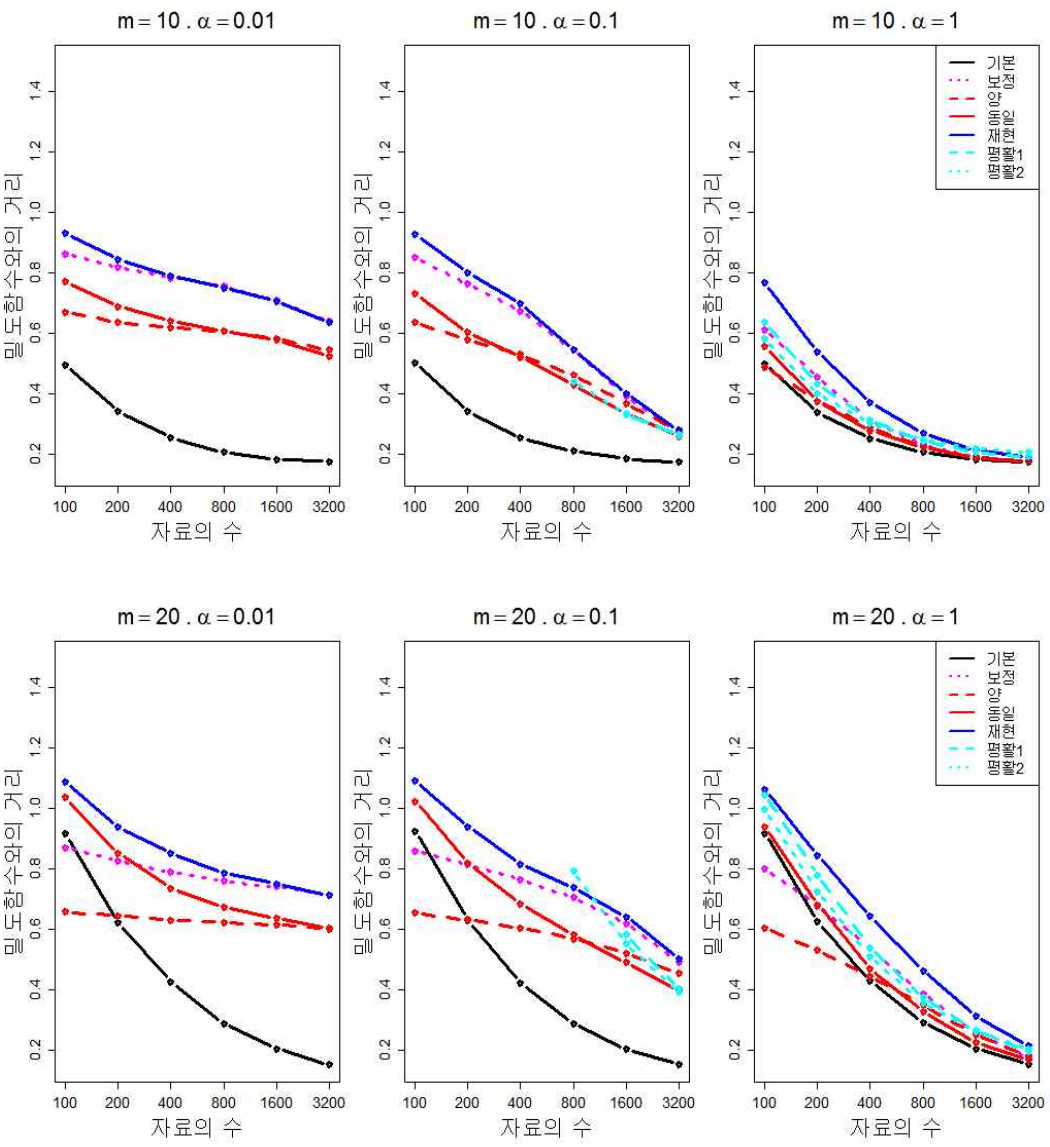
[그림 17] 막대 개수에 따른 원자료 밀도함수와 차등정보보호 히스토그램 사이의 거리



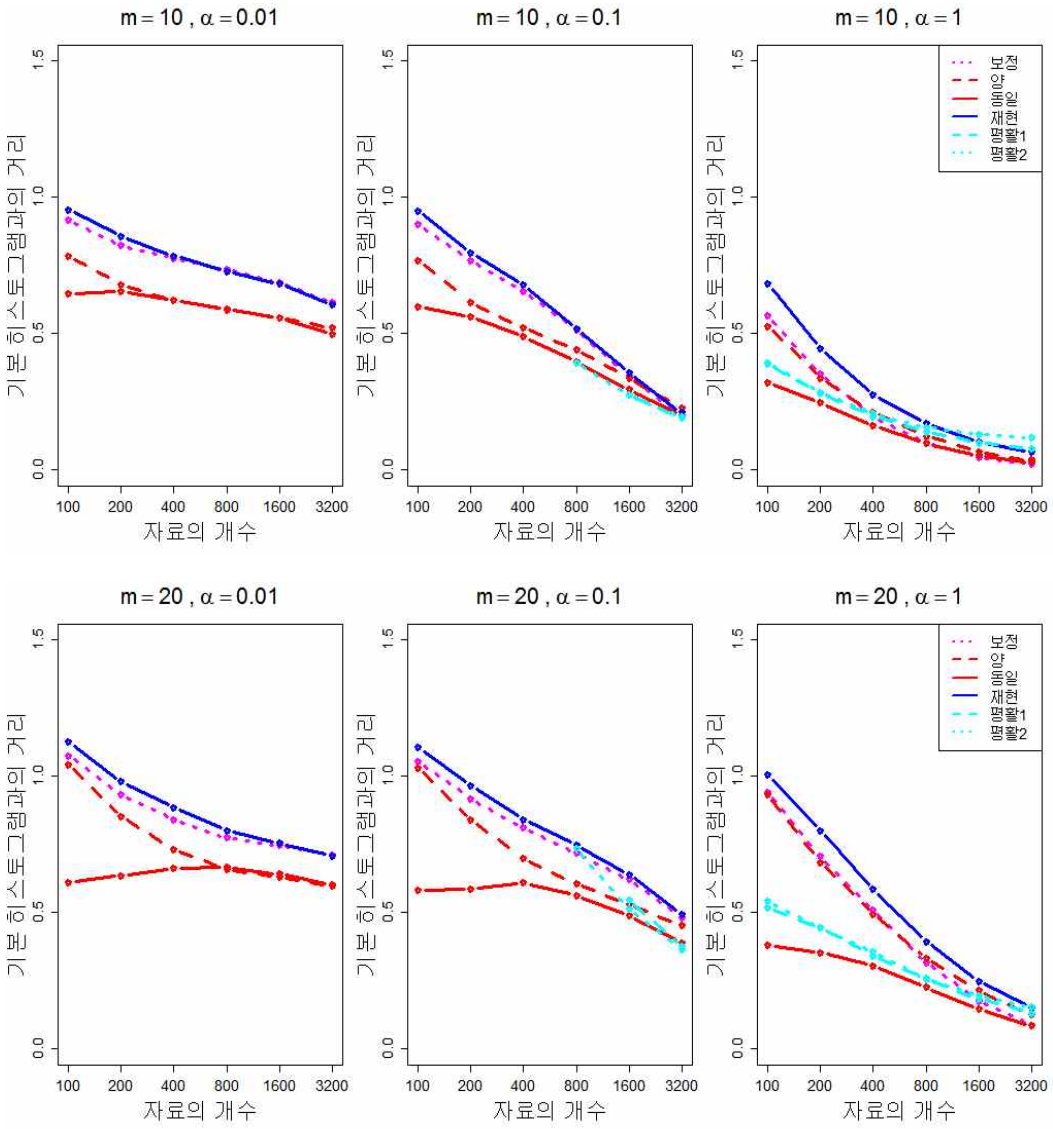
[그림 18] 막대 개수에 따른 기본(원자료) 히스토그램과 차등정보보호 히스토그램 사이의 거리

에 대하여, 제약식 (0.1)를 만족하면서 δ의 값을 최소로 하는 k의 값을 사용하여 비교하는 것이 좀 더 합당한 비교인 것으로 판단되지만 이번 연구에서는 고려하지 않았다.

다음으로 $d=2$인 경우에 대한 모의실험을 동일한 방법으로 진행하였으며 그 결과를 [그림 19] ~ [그림 22]에 도시하였다. 관련된 표는 <부록>에 수록한다. 실험의 결과는 $d=1$인 경우와 매우 유사하며 따라서 자세한 기술은 생략한다. 다만 재현 방식에



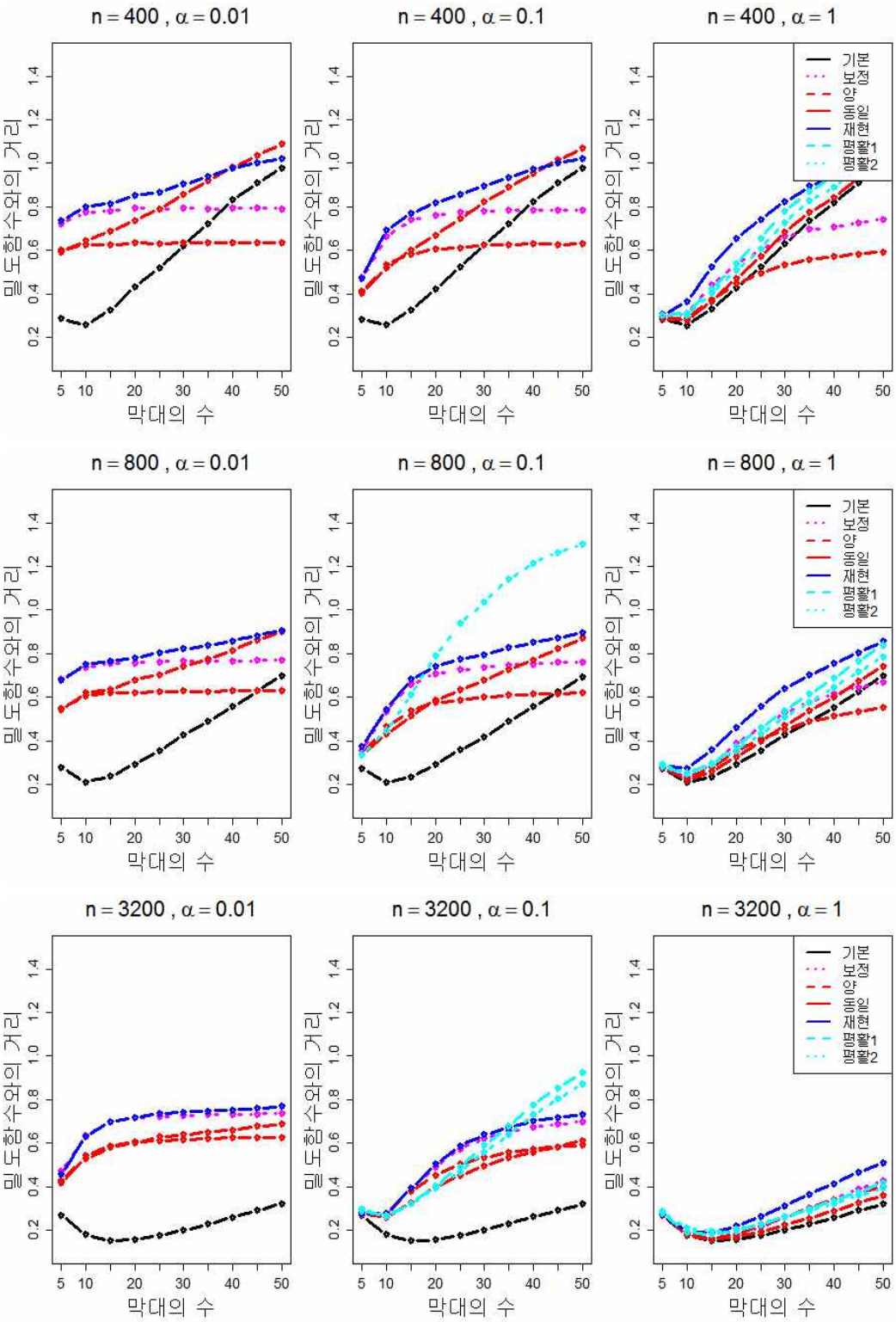
[그림 19] 원자료 밀도함수와 차등정보보호 히스토그램 사이의 거리($d=2$인 경우)



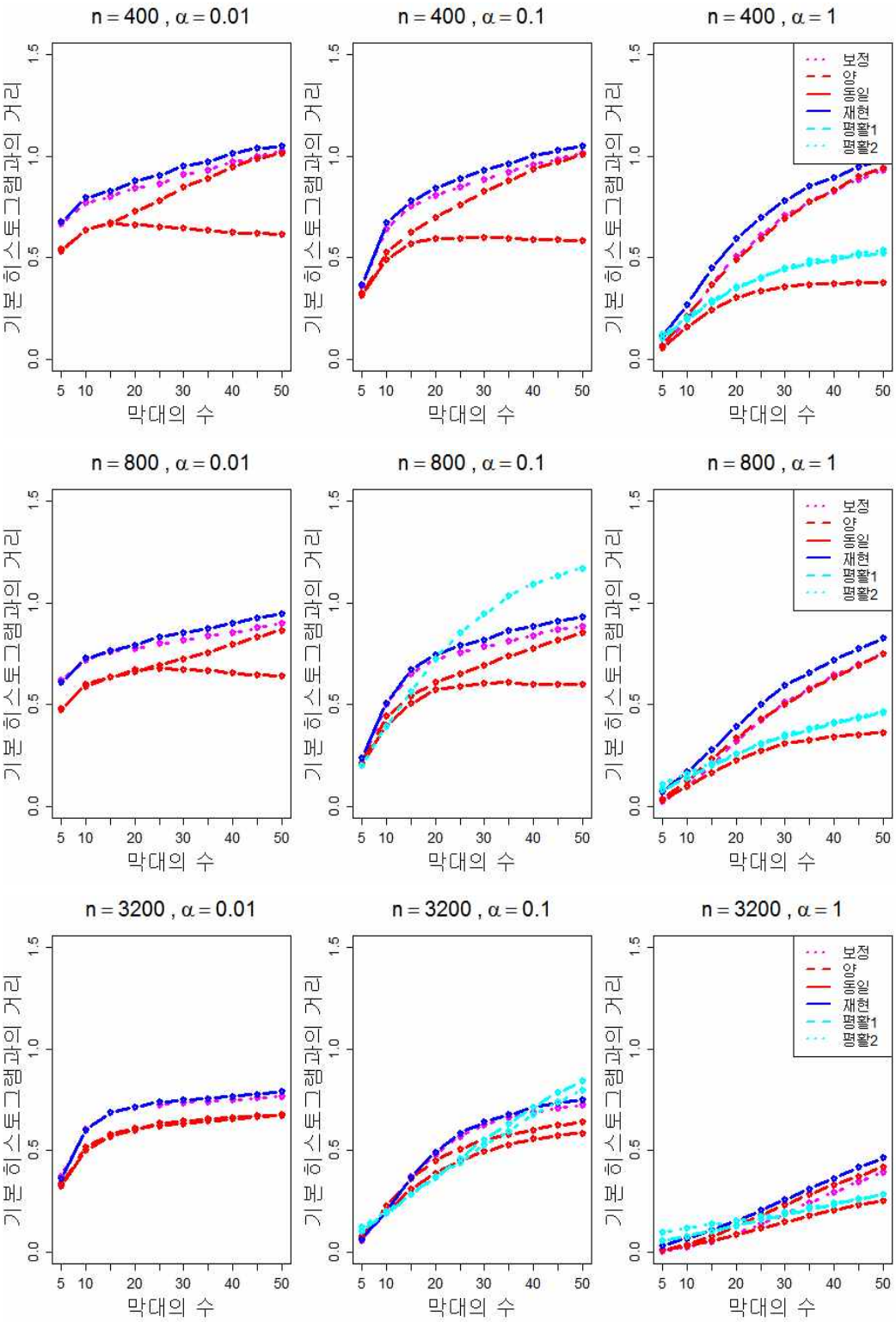
[그림 20] 기본(원자료) 히스토그램과 차등정보보호 히스토그램 사이의 거리($d = 2$인 경우)

서 재현된 교란 히스토그램보다 재현된 평활 히스토그램에서 정보손실이 더 적게 나타나는 것에 대해서는 그 이유를 더 생각해볼 필요가 있다. 교란 후 최적 구조 히스토그램은 기술적인 이유로 실험에서 제외하였으며 실험에 사용한 원자료 밀도함수는 다음과 같다.

$$f_X = (\phi(\mu_1, \Sigma_1) + \phi(\mu_2, \Sigma_2))/2,\ \mu_1 = -\mu_2 = (0.5, 0.5),\ \Sigma_1 = 0.09 \times I,\ \Sigma_2 = 0.04 \times I$$



[그림 21] 막대 개수에 따른 원자료 밀도함수와 차등정보보호 히스토그램 사이의 거리 ($d = 2$인 경우)



[그림 22] 막대 개수에 따른 기본(원자료) 히스토그램과 차등정보보호 히스토그램 사이의 거리 ($d = 2$ 인 경우)

2. 차등정보보호 히스토그램: 실제자료 적용 실험

통계청 보도자료 중 가계금융 복지조사 결과(2018) IV장 '가구소득 및 노후생활' 부분에서 가구소득과 관련된 조사 결과에 대하여 차등정보보호 적용 실험을 진행하였다. 모의실험 결과를 바탕으로 원자료를 사용한 경우와 정보보호 수준이 $\alpha = 1$과 $\alpha = 0.1$인 두 가지 경우를 비교한다. 독자의 이해를 돕기 위하여 부제목을 원 보도자료와 동일하게 설정하였으며 2017년에 조사된 소득에 대한 내용만 취사하여 정리하였다. 소득자료에 대한 조사결과는 대부분 가중표본평균(weighted sample mean)과 가중표본비율(weighted sample ratio)로 구성되어 있으며, 가중표본평균에 대해서는 Dwork(2006)의 교란기법을, 가중표본비율에 대해서는 모의실험에서 진행한 히스토그램에 가중치를 적용하여 작성하였다.[19] 독자의 편의를 위하여 필요한 경우 작성기법을 자세히 기술하였다.

가. 가구주 연령대별 가구소득

〈표 12〉 보도자료(표 4-2) 가구특성 및 가구소득 구간별 가구분포[1)]

(단위: %)

구 분		전체	1천만 미만	1~3천만 미만	3~5천만 미만	5~7천만 미만	7~10천만 미만	10천만 이상
2016년 전 체		100.0	10.0	25.1	22.0	16.2	14.0	12.7
2017년 전 체		100.0	9.3	24.5	22.0	15.9	14.4	13.9
가구주 연령대별	30세 미만	100.0	7.0	41.3	31.4	11.9	6.8	1.6*
	30~39세	100.0	2.5	18.0	29.4	23.0	16.1	11.1
	40~49세	100.0	2.5	16.0	22.1	19.7	20.1	19.6
	50~59세	100.0	3.5	18.3	20.3	17.2	18.3	22.4
	60세 이상	100.0	20.8	36.0	19.6	9.9	7.8	5.9

주: 1) 소득구간은 명목소득으로 작성되어 전년비교는 한계가 있음

* 상대표준오차(RSE)가 크거나(25% 이상) 표본규모가 과소하여 이용 시 유의

19) 차등정보보호 적용을 실험하고 결과를 비교하기 위하여 두 가지 시나리오가 가능하다. 첫째는 실제로 공표한 통계가 어떻게 달라지는지 살펴보는 것이다. 이를 통해 차등정보보호 적용으로 인하여 감내해야 하는 변화를 알 수 있다. 본 연구에서는 가중치를 반영하여 차등정보보호 자료를 만들어 실제로 보도자료에 제공된 통계와 얼마나 달라지는지 확인했다. 두 번째는 가계금융복지조사의 조사 대상을 어느 국가의 전체 인구라고 가정하고 차등정보보호 적용 전후의 변화를 살피는 것이다. 이 경우 가중치를 고려할 필요가 없으므로 계산이 훨씬 단순하나, 원자료로부터 생산한 통계가 보도자료의 <표 4-2>와 다르게 된다. 본 연구에서는 두 번째 비교는 생략하였다.

보도자료 <표 4-2>에 포함된 가구주 연령대별 가구소득 분포는 각 연령대별 가구소득에 대한 히스토그램이다. 가구소득을 기준으로 1천만 미만, 1~3천만 미만, 3~5천만 미만, 5~7천만 미만, 10천만 이상으로 구간을 정하였다. 각 구간에 대해 빈도를 계산한 후, 각 빈도를 합이 100이 되도록 조정한 히스토그램이 보도자료 <표 4-2>이다. 따라서 α-차등정보보호 히스토그램 역시 밀도 대신 빈도를 사용하여 합이 100이 되도록 조정하였으며, 보정된 교란 히스토그램, 동일 지지집합 히스토그램, 재현된 교란 히스토그램을 생성하였다. 다만 재현된 교란 히스토그램의 경우 정의(Wasserman과 Zhou, 2010)에 맞게 작성하기 위하여 빈도가 아닌 밀도를 사용하여 원자료와 동일한 개수의 재현자료를 임의로 추출하였다. 그 결과를 [그림 23] ~ [그림 28] 및 <표 16>에서 정리하였다.

[그림 23]에서는 전체 가구주를 대상으로 4종류의 가구소득 히스토그램을 나열하였다. 차등정보보호를 적용하지 않은 기본 히스토그램이나 차등정보보호를 적용한 다른 3종류(보정된 교란 방식, 동일 지지집합 방식, 재현된 교란 방식)가 정보 전달에 있어 큰 차이를 보이지 않았다. 정보보호 수준이 강력하여도($\alpha = 0.1$) 차등정보보호 히스토그램이 기본 히스토그램과 큰 차이를 보이지 않았다. 이는 전체 집단의 경우, <표 13>에서와 같이 가구소득 각 구간의 자료수가 모두 2,000 이상으로 충분히 크기 때문이다. 따라서 $\alpha = 0.1$인 경우에도 기본 히스토그램과 α-차등정보보호 히스토그램의 밀도값은 거의 차이가 나지 않는다.

〈표 13〉 전체 가구주에 대한 가구소득 빈도표

가구소득	1천만 미만	1~3천만 미만	3~5천만 미만	5~7천만 미만	7~10천만 미만	10천만 이상	합
빈도	2,144	5,017	4,005	2,750	2,428	2,296	18,640

[그림 24]는 30세 미만 가구주 그룹에 대하여 가구소득 히스토그램을 나타낸다. 상대적으로 약한 정보보호 수준($\alpha = 1$)에서는 기본 히스토그램과 α-차등정보보호 히스토그램의 밀도값에 어느 정도 차이가 나타나지만, 그 차이가 정보를 왜곡시키는 수준은 아니라고 판단된다. 그러나 강한 정보보호 수준($\alpha = 0.1$)에서는 이 차이는 더욱 뚜렷해지며, 고소득의 구간에서 변형이 많이 일어난 것을 확인할 수 있다. 이러한 현상이 발생하는 이유는 각 연령대별 가구소득의 경우 자료의 수가 전체 가구소득의 자료수보다 작기 때문이다. <표 14>의 연령별 빈도표를 참고하면 전체 자료에 비하여 특히 30세 미만 가구주의 수가 상당히 적은 것을 확인할 수 있다.

〈표 14〉 연령별 빈도표

연령대	30세 미만	30~39세	40~49세	50~59세	60세 이상	합
빈도	437	2,424	3,807	4,388	7,584	18,640

[그림 24]에서 30세 미만 가구주의 경우 $\alpha = 0.1$인 경우 보정된 교란 히스트그램과 재현된 교란 히스토그램에서 밀도값이 0이 되는 히스토그램 막대가 생성되었다. 이러한 결과는 30대 미만 가구주의 수가 <표 14>에서와 같이 437로 매우 적고, 정보보호 수준이 $\alpha = 0.1$로 강력한 경우 이중지수분포의 표준편차[20]가 대략 28.24이기 때문에 얼마든지 발생 가능한 것이다. 즉, 정보보호 수준이 강력하면, 작은 빈도값을 가지는 막대에 큰 잡음이 더해지면서, 해당 막대의 빈도값이 음수에서 0으로 조정되고, 그 효과로 다른 막대의 비율이 증가하여 정보의 왜곡이 발생할 수 있다.

〈표 15〉 30세 미만 가구주에 대한 가구소득 빈도표

가구소득	1천만 미만	1~3천만 미만	3~5천만 미만	5~7천만 미만	7~10천만 미만	10천만 이상	합
빈도	34	178	129	58	31	7	437

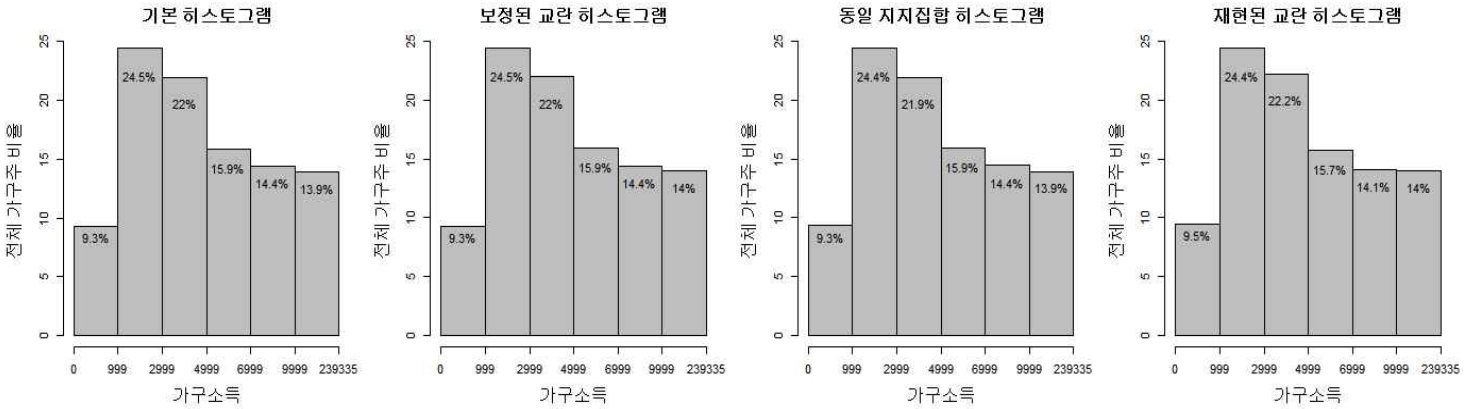
설명한 대로 α-차등정보보호 히스토그램에서 때때로 0에 아주 가까운 밀도값을 가지는 막대가 존재할 수 있다. 재현된 교란 히스토그램의 경우 이 구간에서 재현자료가 추출되지 않는 현상이 발생한다. 특히 10천만 이상 구간에 포함되는 가구소득의 최대값은 약 239천만이며 따라서 구간의 길이가 약 229천만이므로 밀도함수는 거의 0의 값을 가진다. 실험을 반복하면 [그림 24]에서 마지막 막대의 밀도값이 0에 가까운 경우가 더 많이 발생한다. 이러한 경우를 방지하려면 밀도값 대신 빈도값을 사용하여 임의 추출을 하는 것이 좀 더 현실적인 히스토그램을 제공할 것이라고 판단된다.

[그림 25] ~ [그림 28]의 다른 연령대에서는 [그림 24]와 달리 전체 연령의 히스토그램과 같이 기본 히스토그램과 유사한 차등정보보호 히스토그램이 생성되는 것을 확인할 수 있다. <표 14>의 연령별 빈도표를 볼 때, 자료의 크기가 충분해서 유용한 빈도표가 생성되는 것으로 여겨진다. 모의실험에서 확인한 것처럼 히스토그램 막대의 개수에 비하여 자료의 개수가 적으면 차등정보보호를 적용할 때 정보손실이 커져서 자료를 사용하기 어렵다는 것을 실제 자료 분석을 통해서도 알 수 있다.

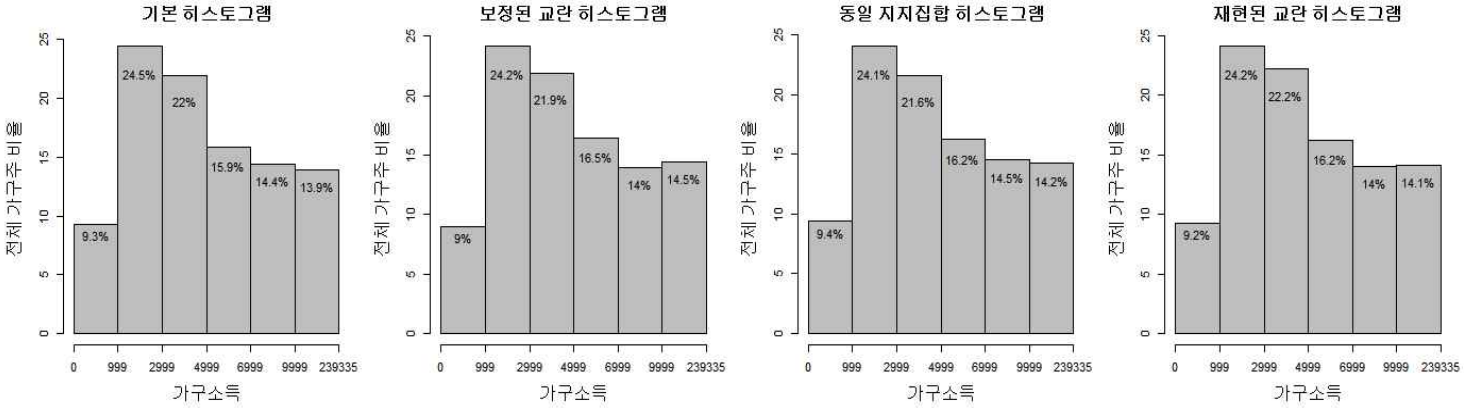
---

20) 정확한 표준편차는 $\sqrt{2}b = 2\sqrt{2}/\alpha = 20\sqrt{2}$이다.

$\alpha = 1$ 인 경우

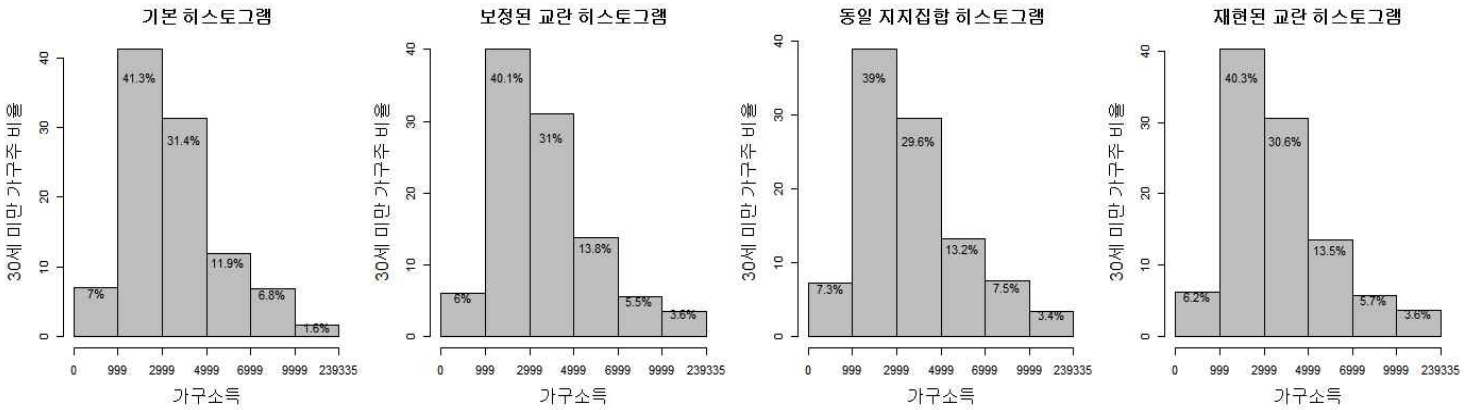


$\alpha = 0.1$ 인 경우

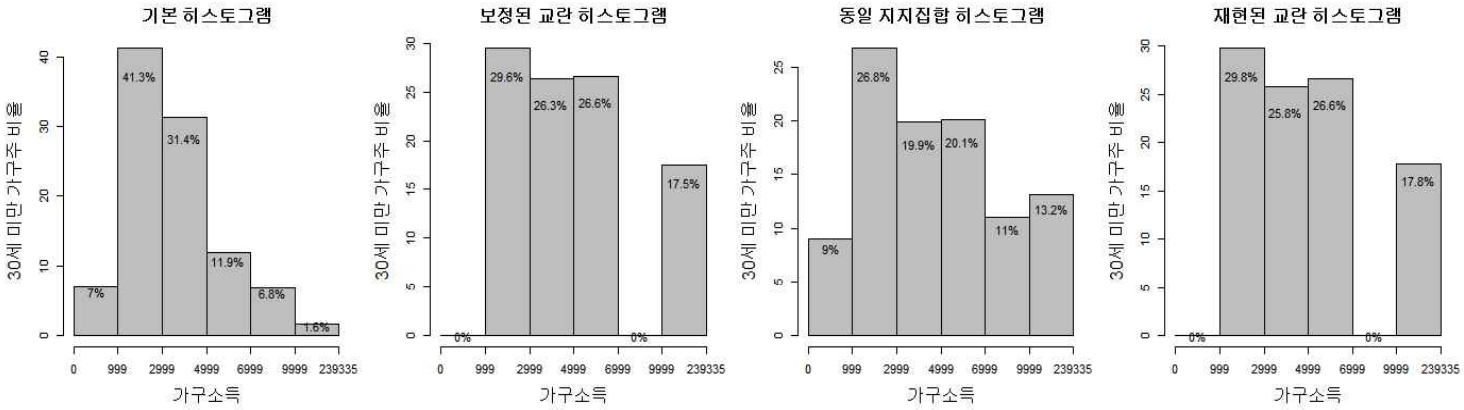


[그림 23] 가구소득 분포(전체 가구주)

$\alpha = 1$ 인 경우

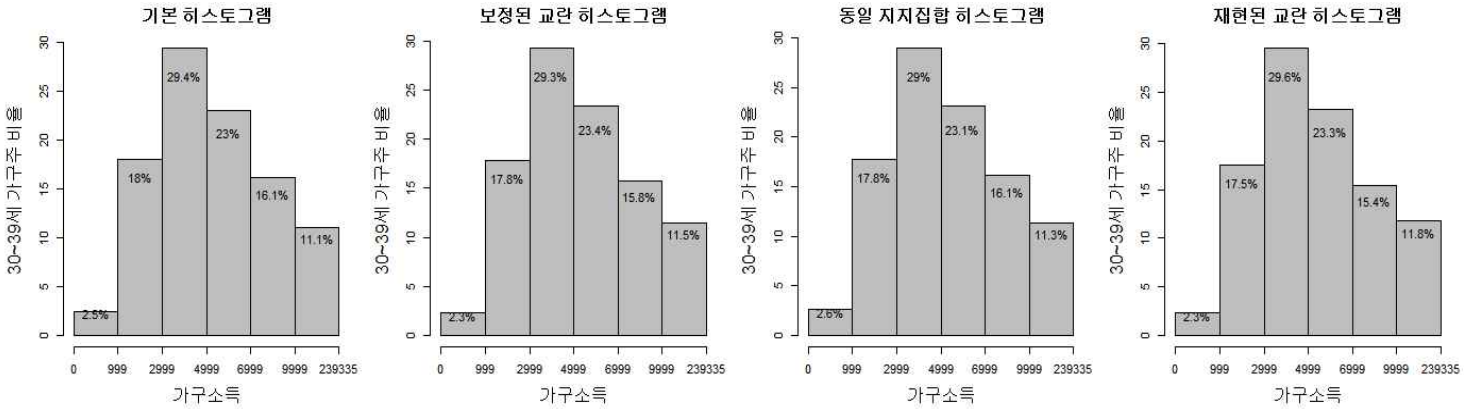


$\alpha = 0.1$ 인 경우

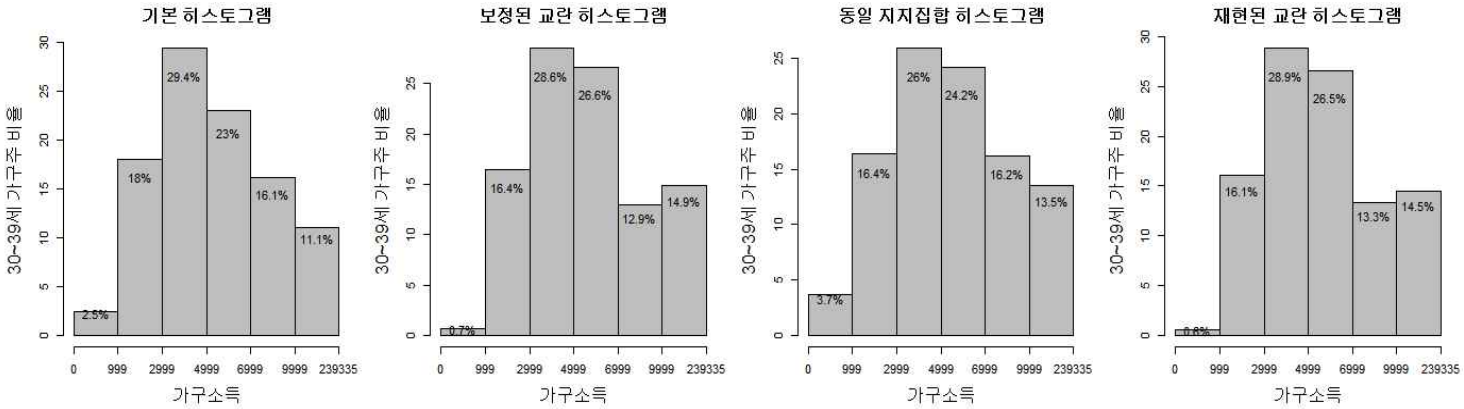


[그림 24] 가구소득 분포(30세 미만 가구주)

$\alpha = 1$ 인 경우

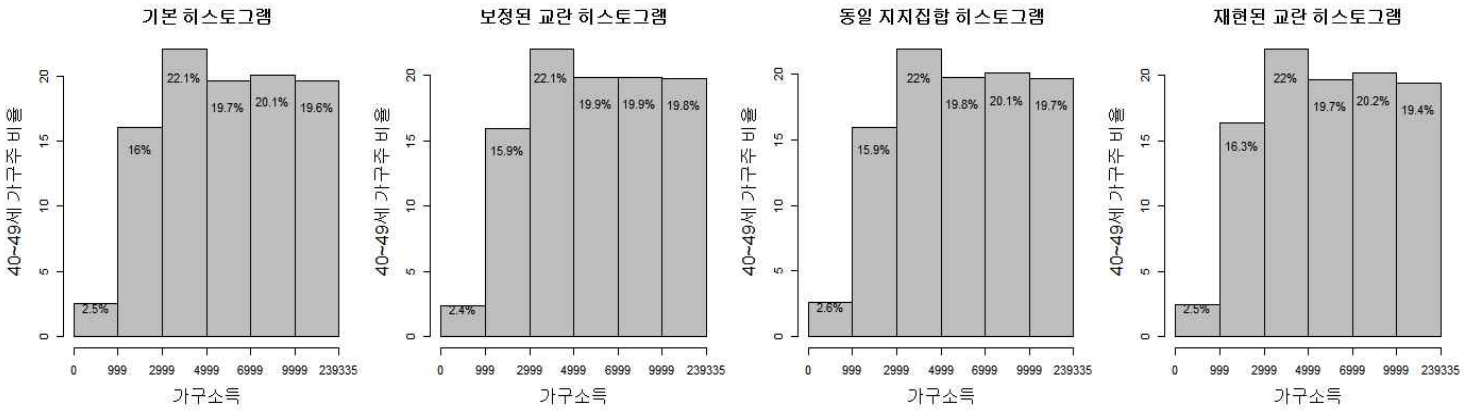


$\alpha = 0.1$ 인 경우

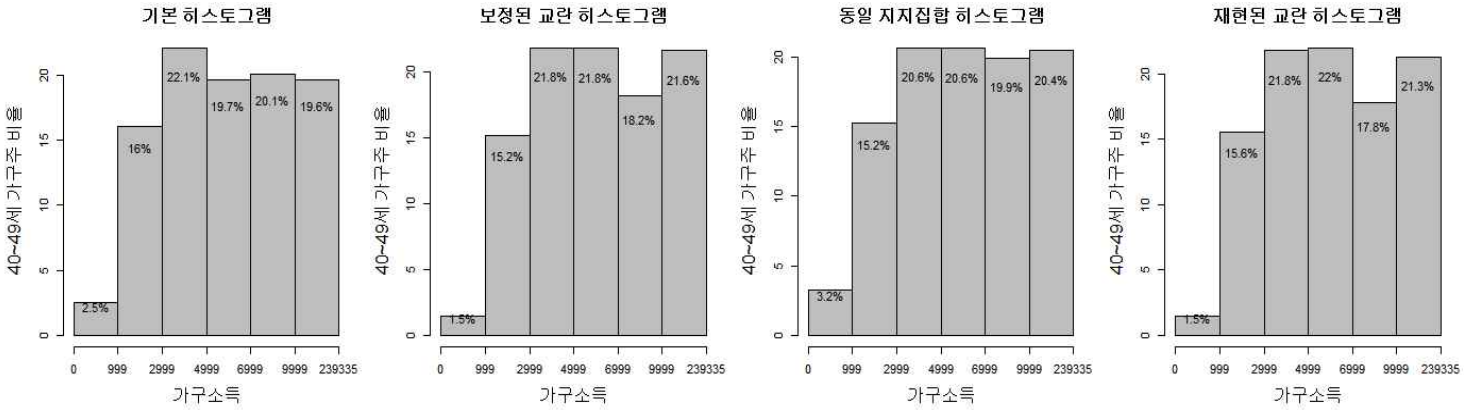


[그림 25] 가구소득 분포(30~39세 가구주)

$\alpha = 1$ 인 경우

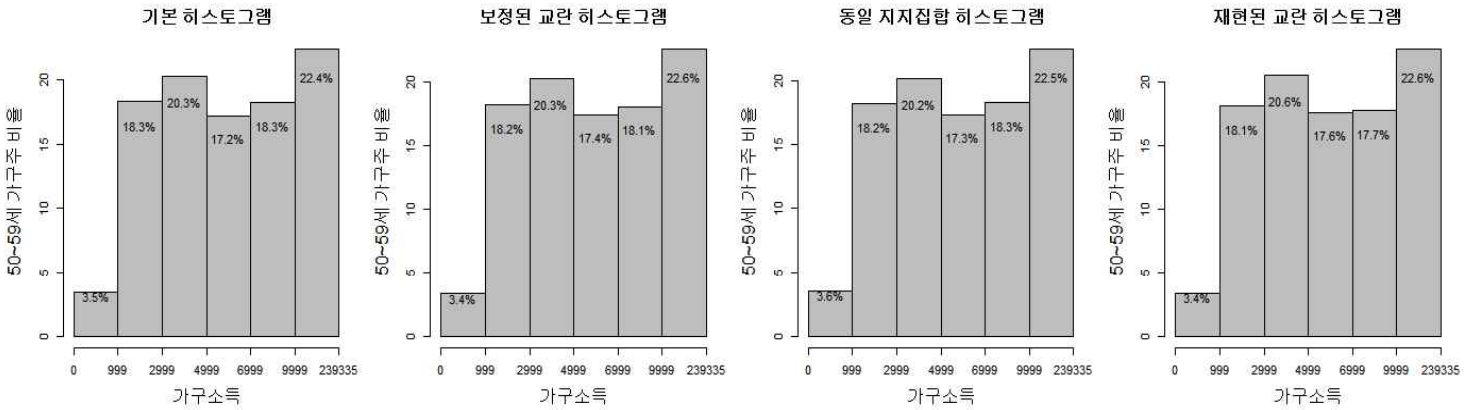


$\alpha = 0.1$ 인 경우

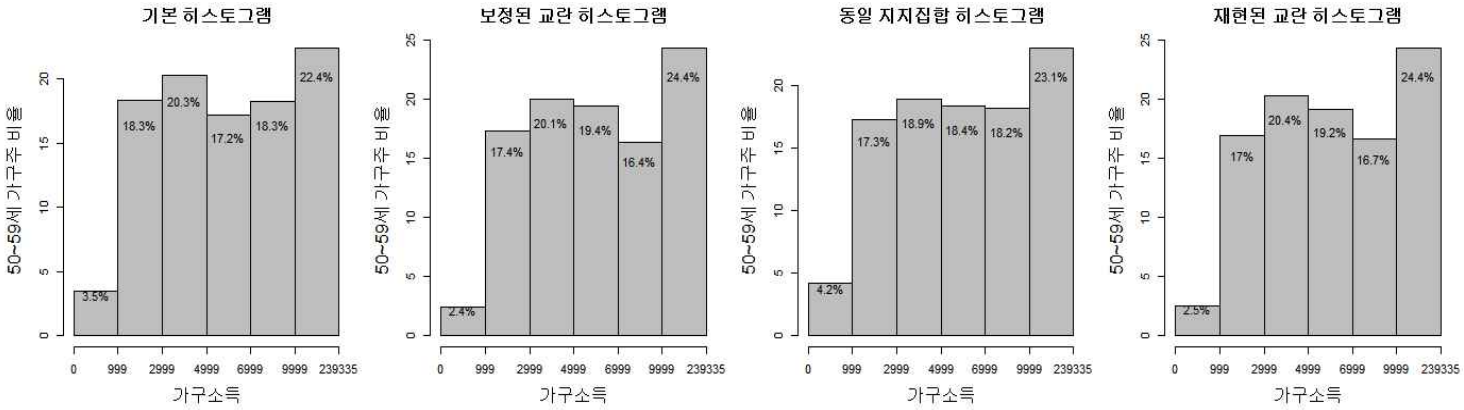


[그림 26] 가구소득 분포(40~49세 가구주)

$\alpha = 1$ 인 경우

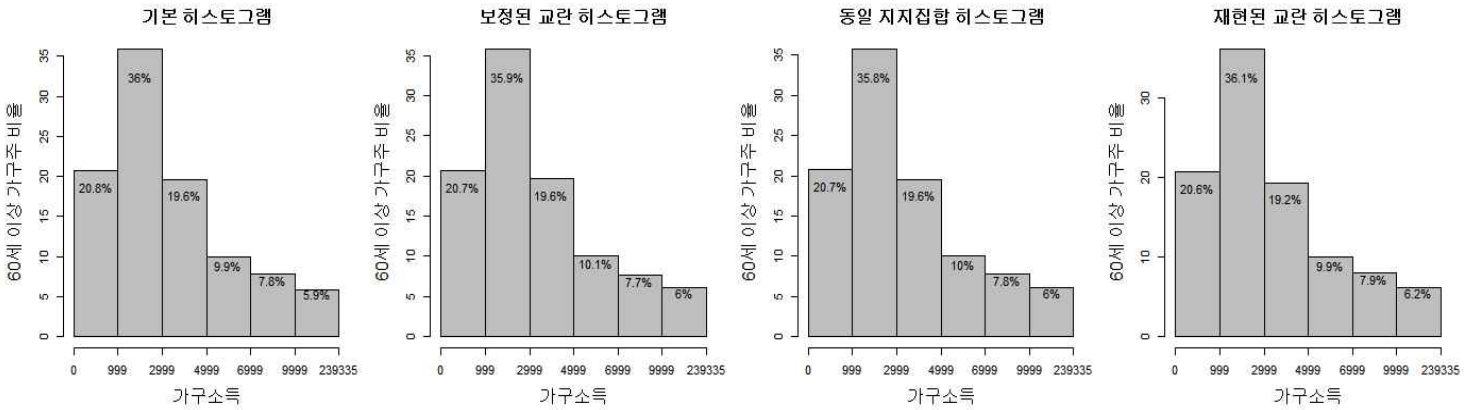


$\alpha = 0.1$ 인 경우

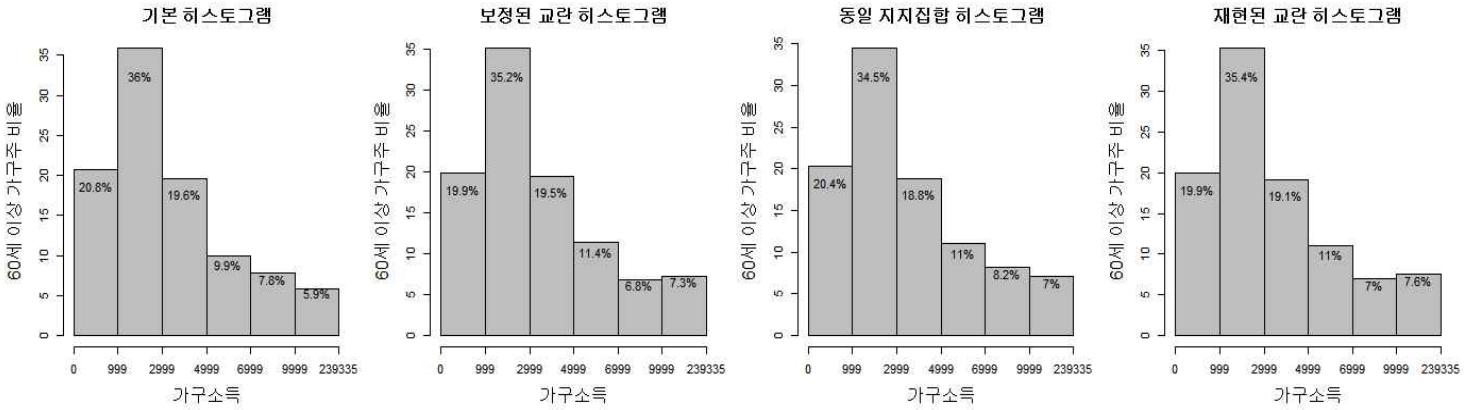


[그림 27] 가구소득 분포(50~59세 가구주)

$\alpha = 1$ 인 경우



$\alpha = 0.1$ 인 경우



[그림 28] 가구소득 분포(60세 이상 가구주)

〈표 16〉 가구주 특성별 가구소득 분포 비교

가구주	방식	소득 구간						정보보호 수준
		~1	1~3	3~5	5~7	7~10	10~	
전체	기본	9.312	24.485	21.984	15.875	14.440	13.903	$\alpha = 1$
	보정	9.280	24.457	21.976	15.934	14.393	13.959	
	동일	9.318	24.444	21.943	15.911	14.446	13.938	
	재현	9.501	24.447	22.237	15.714	14.083	14.018	
전체	기본	9.312	24.485	21.984	15.875	14.44	13.903	$\alpha = 0.1$
	보정	8.993	24.204	21.904	16.463	13.974	14.462	
	동일	9.369	24.084	21.581	16.22	14.498	14.249	
	재현	9.238	24.168	22.226	16.239	14.024	14.104	

가구주	방식	소득 구간						정보보호 수준
		~1	1~3	3~5	5~7	7~10	10~	
~30세	기본	6.978	41.324	31.397	11.897	6.838	1.567	$\alpha = 1$
	보정	6.035	40.108	30.956	13.799	5.550	3.551	
	동일	7.297	39.042	29.582	13.186	7.499	3.393	
	재현	6.250	40.349	30.601	13.476	5.692	3.632	
~30세	기본	6.978	41.324	31.397	11.897	6.838	1.567	$\alpha = 0.1$
	보정	0	29.562	26.348	26.633	0	17.457	
	동일	9.008	26.832	19.872	20.087	11.036	13.166	
	재현	0	29.844	25.762	26.572	0	17.822	

가구주	방식	소득 구간						정보보호 수준
		~1	1~3	3~5	5~7	7~10	10~	
30~39세	기본	2.473	17.981	29.404	23.004	16.085	11.054	$\alpha = 1$
	보정	2.284	17.817	29.315	23.376	15.757	11.451	
	동일	2.613	17.799	29.016	23.137	16.101	11.334	
	재현	2.323	17.527	29.624	23.305	15.445	11.776	
30~39세	기본	2.473	17.981	29.404	23.004	16.085	11.054	$\alpha = 0.1$
	보정	0.658	16.403	28.552	26.578	12.936	14.873	
	동일	3.706	16.374	25.978	24.182	16.228	13.532	
	재현	0.622	16.127	28.879	26.513	13.326	14.533	

〈표 16〉 가구주 특성별 가구소득 분포 비교(계속)

가구주	방식	소득 구간						정보보호 수준
		~1	1~3	3~5	5~7	7~10	10~	
40~49세	기본	2.492	16.030	22.109	19.658	20.086	19.624	$\alpha = 1$
	보정	2.387	15.944	22.077	19.874	19.893	19.824	
	동일	2.571	15.944	21.951	19.761	20.063	19.711	
	재현	2.452	16.314	22.033	19.667	20.172	19.362	
40~49세	기본	2.492	16.03	22.109	19.658	20.086	19.624	$\alpha = 0.1$
	보정	1.461	15.182	21.804	21.772	18.199	21.583	
	동일	3.22	15.23	20.642	20.611	19.864	20.432	
	재현	1.481	15.569	21.781	22.017	17.843	21.309	

가구주	방식	소득 구간						정보보호 수준
		~1	1~3	3~5	5~7	7~10	10~	
50~59세	기본	3.486	18.296	20.319	17.174	18.282	22.444	$\alpha = 1$
	보정	3.374	18.200	20.291	17.404	18.087	22.643	
	동일	3.558	18.187	20.171	17.301	18.273	22.509	
	재현	3.433	18.069	20.558	17.597	17.741	22.602	
50~59세	기본	3.486	18.296	20.319	17.174	18.282	22.444	$\alpha = 0.1$
	보정	2.395	17.362	20.05	19.419	16.379	24.394	
	동일	4.157	17.292	18.949	18.353	18.195	23.054	
	재현	2.468	16.958	20.359	19.158	16.668	24.388	

가구주	방식	소득 구간						정보보호 수준
		~1	1~3	3~5	5~7	7~10	10~	
60세~	기본	20.785	35.966	19.635	9.936	7.788	5.890	$\alpha = 1$
	보정	20.694	35.884	19.620	10.080	7.691	6.031	
	동일	20.741	35.816	19.553	10.045	7.835	6.010	
	재현	20.617	36.132	19.217	9.941	7.929	6.164	
60세~	기본	20.785	35.966	19.635	9.936	7.788	5.89	$\alpha = 0.1$
	보정	19.888	35.156	19.488	11.359	6.83	7.279	
	동일	20.368	34.528	18.846	10.985	8.234	7.039	
	재현	19.909	35.386	19.115	11.041	6.958	7.591	

마지막으로 차등정보보호를 적용하여 보도자료를 작성한다고 가정해보자. 여기서는 교란 후 보정($\alpha = 1$) 방식으로 생성한 히스토그램을 선택했다. <표 17>에서 위의 표는 기존의 보도자료에 사용된 히스토그램이다. 아래 표는 차등정보보호를 적용한 히스토그램이다. 대부분의 값이 유사하지만 빈도수가 작았던 30세 미만 고소득자의 경우 정보가 상당히 왜곡된 것을 확인할 수 있다. 결론적으로 차등정보보호를 실제로 적용한다면 자료의 크기를 고려하는 것이 매우 중요하며, 과다한 정보손실을 방지하기 위해서는 차등정보보호 적용 이전에 많은 실험과 점검이 요구된다. 한편 차등정보보호가 적용된 통계를 기존 통계를 대체하여 사용할 수 있을지, 그 범위나 형식의 구체적인 결정에 관하여는 별도의 논의가 필요하다.

〈표 17〉 원 보도자료 vs. 차등정보보호 적용 보도자료

원 히스토그램

구 분		전체	1천만 미만	1~3천만 미만	3~5천만 미만	5~7천만 미만	7~10천만 미만	10천만 이상
2017년 전 체		100.0	9.3	24.5	22.0	15.9	14.4	13.9
가구주 연령대별	30세 미만	100.0	7.0	41.3	31.4	11.9	6.8	1.6*
	30~39세	100.0	2.5	18.0	29.4	23.0	16.1	11.1
	40~49세	100.0	2.5	16.0	22.1	19.7	20.1	19.6
	50~59세	100.0	3.5	18.3	20.3	17.2	18.3	22.4
	60세 이상	100.0	20.8	36.0	19.6	9.9	7.8	5.9

차등정보보호 히스토그램 : 교란(보정, $\alpha = 1$) 방식

구 분		전체	1천만 미만	1~3천만 미만	3~5천만 미만	5~7천만 미만	7~10천만 미만	10천만 이상
2017년 전 체		100.0	9.3	24.5	22.0	15.9	14.4	14.0
가구주 연령대별	30세 미만	100.0	6.0	40.1	31.0	13.8	5.6	3.6
	30~39세	100.0	2.3	17.8	29.3	23.4	15.8	11.5
	40~49세	100.0	2.4	15.9	22.1	19.9	19.9	19.8
	50~59세	100.0	3.4	18.2	20.3	17.4	18.1	22.6
	60세 이상	100.0	20.7	35.9	19.6	10.1	7.7	6.0

나. 소득 5분위별 가구소득

〈표 18〉 보도자료(표 4-4) 소득 5분위 및 소득원천별 가구소득 평균 및 구성비(2017년)

(단위: 만원, %)

구 분		가구소득	근로소득	사업소득	재산소득	공적이전 소득	사적이전 소득
2017년	전 체	5,705	3,639	1,244	380	351	91
평균	1분위	1,057	328	98	86	392	153
	2분위	2,655	1,443	526	168	388	130
	3분위	4,464	2,718	1,054	232	383	76
	4분위	6,825	4,633	1,506	314	310	62
	5분위	13,521	9,072	3,035	1,098	280	36
소득 원천별 구성비	1분위	100.0	31.1	9.2	8.1	37.1	14.5
	2분위	100.0	54.3	19.8	6.3	14.6	4.9
	3분위	100.0	60.9	23.6	5.2	8.6	1.7
	4분위	100.0	67.9	22.1	4.6	4.5	0.9
	5분위	100.0	67.1	22.4	8.1	2.1	0.3

가구소득은 근로, 사업, 재산, 공적이전, 사적이전 소득으로 나누어진다. 차등정보보호 적용 결과를 소개하기 전에 먼저 각각의 소득 자료에 대하여 α-차등정보보호 가중표본평균(weighted sample mean)을 작성하는 방법에 대하여 설명한다. 각 소득 자료 X_i, $i \le n$에 대하여 자료의 가중치가 w_i, $i \le n$로 주어지는 경우 가중표본평균은 가중치가 적용된 자료에 대한 합을 이용해 다음과 같이 구성되어 있다.

$$\overline{X}_w^{original} = \sum_{i=1}^{n} Y_i, \;\; Y_i = \frac{w_i}{w} X_i, \; i \le n, \; w = \sum_{i=1}^{n} w_i .$$

Dwork(2006)이 제안한 교란 방식으로 작성된 α-차등정보보호 가중표본평균은 다음과 같다.

$$\overline{X}_w^{perturbed} = \sum_{i=1}^{n} Y_i + W, \;\; W \sim DE(\Delta_Y/\alpha)$$ [21].

단, $Y_{max} = max_{i \le n} Y_i$와 $Y_{min} = min_{i \le n} Y_i$은 가중치가 반영된 자료의 최대값과 최소값이며 $\Delta = |Y_{max} - Y_{min}|$이다.

21) $DE(b)$는 산포모수가 b인 이중지수 분포

한편 가구소득의 경우 다른 하위 소득의 합으로 구성된다. 따라서 가구소득과 하위 소득에 대하여 각각 정보보호 처리를 진행하는 경우, 가구소득과 하위 소득의 합이 불일치하는 현상이 발생한다. 예를 들어 자료가 $Z_i = X_i + Y_i,\ i \le n$로 구성되는 경우 세 자료 $Z = \sum_{i=1}^{n} Z_i,\ X = \sum_{i=1}^{n} X_i,\ Y = \sum_{i=1}^{n} Y_i$에 대한 α-차등정보보호 자료는 각각 다음과 같다.

$$S_Z = Z + W_Z, \quad S_X = X + W_X, \quad S_Y = Y + W_Y,$$

단, $W_Z \sim DE(\Delta_Z/\alpha)$, $W_X \sim DE(\Delta_X/\alpha)$, $W_Y \sim DE(\Delta_Y/\alpha)$이며 W_Z, W_X 및 W_Y는 독립이다.

따라서 α-차등정보보호 자료는 일반적으로 $S_Z \neq S_X + S_Y$ 즉, 합이 불일치하는 관계를 가진다. 하지만 Z에 대하여 차등정보보호를 적용하지 않고 다음을 Z의 α-차등정보보호 자료로 사용하면 합을 일치시킬 수 있다.[22]

$$S_X + S_Y = (X + W_X) + (Y + W_Y).$$

이제 차등정보보호 적용 전후 평균의 변화를 살펴보자. 평균에 차등정보보호를 적용하는 과정은 교란 과정과 동일하여 단순하다. <표 19>는 이러한 정보보호 처리 과정으로 작성된 소득 5분위 및 소득원천별 가구소득 평균을 나타낸다. 정보보호 수준이 지나치게 강력하지 않다면 원래의 평균을 비교적 잘 보여주는 것을 확인할 수 있다. 그러나 정보보호 수준이 커지면서 소득 4분위의 사적이전소득 항목에서 정보손실이 크게 생성된 것으로 나타난다. 이 역시 자료의 크기의 영향을 반영하는 것이라 판단된다. 참고로 <표 20>은 원래의 평균과 차등정보보호 평균의 차이를 나타내며, 정보보호 수준이 강력하면 정보손실이 커지는 경향을 확인시켜 준다.

---

22) 다음은 $S_X + S_Y$가 Z의 α-차등정보보호 자료임을 확인하는 간단한 과정이다.

$$\frac{Pr(X+Y+W_X+W_Y=z \mid X+Y=t)}{Pr(X+Y+W_X+W_Y=z \mid X+Y=t')} = \frac{Pr(W_X+W_Y=z-t)}{Pr(W_X+W_Y=z-t')}$$

$$= \frac{\int_R Pr(W_X=\tau)\,Pr(W_Y=z-t-\tau)\,d\tau}{\int_R Pr(W_X=\tau)\,Pr(W_Y=z-t'-\tau)\,d\tau}$$

$$\le \frac{\int_R Pr(W_X=\tau)\,Pr(W_Y=z-t'-\tau)\times exp(\alpha|t-t'|)\,d\tau}{\int_R Pr(W_X=\tau)\,Pr(W_Y=z-t'-\tau)\,d\tau} = exp(\alpha|t-t'|), \quad \forall\, t \neq t'$$

〈표 19〉 소득 5분위 및 소득원천별 가구소득 평균 결과 비교

	가구	근로	사업	재산	공적	사적	정보보호 수준
전체	5,705	3,639	1,244	380	351	91	원자료
1분위	1,057	328	98	86	392	153	
2분위	2,655	1,443	526	168	388	130	
3분위	4,464	2,718	1,054	232	383	76	
4분위	6,825	4,633	1,506	314	310	62	
5분위	13,521	9,072	3,035	1,098	280	36	
전체	5700	3636	1241	380	352	91	$\alpha = 1$
1분위	1059	330	99	86	392	151	
2분위	2652	1441	523	169	387	131	
3분위	4478	2718	1056	244	383	77	
4분위	6827	4646	1500	315	305	60	
5분위	13312	9066	2834	1092	284	35	
전체	5726	3638	1268	379	344	98	$\alpha = 0.1$
1분위	1061	334	104	75	396	151	
2분위	2709	1472	535	181	390	131	
3분위	4413	2756	945	231	398	84	
4분위	6817	4630	1589	309	285	4	
5분위	12901	8691	2781	1106	293	31	

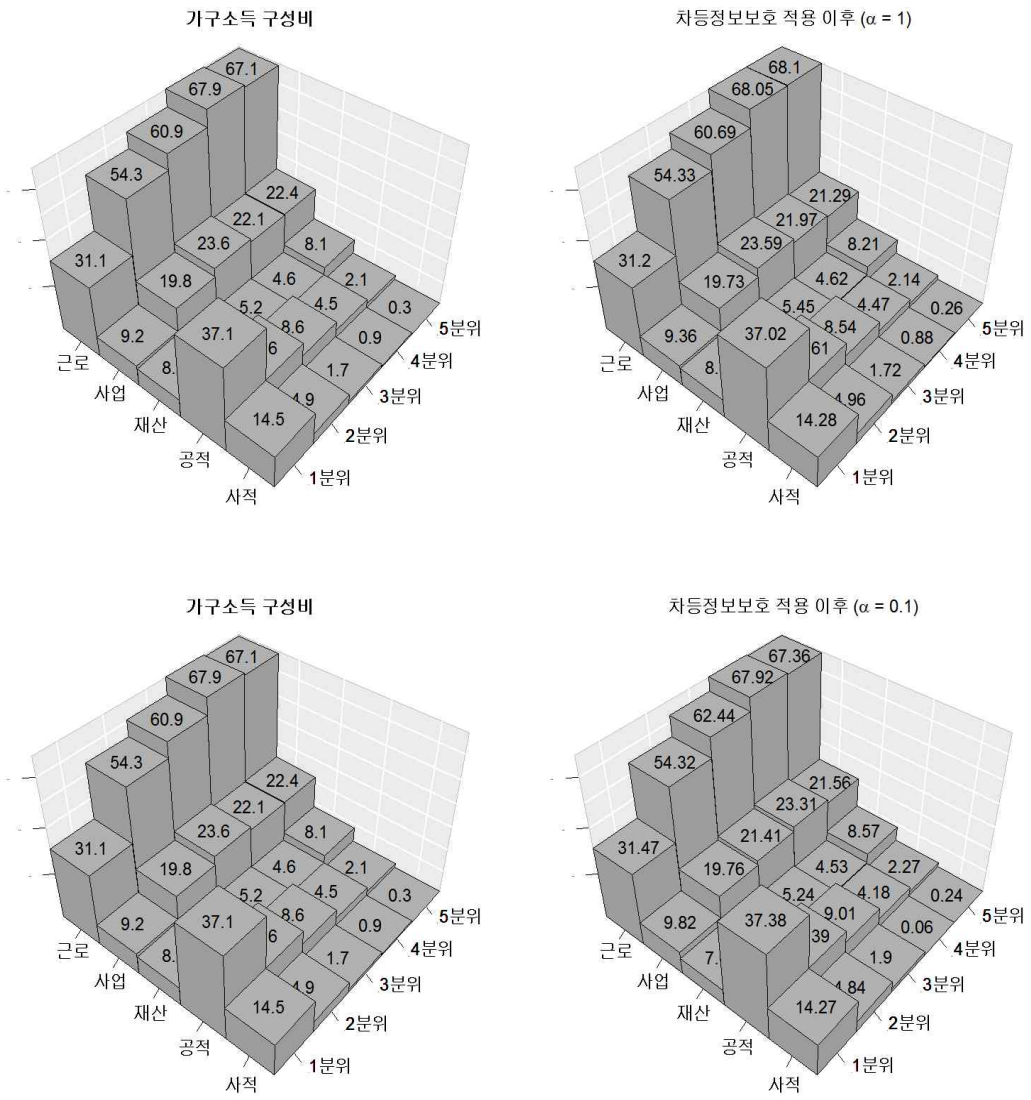
〈표 20〉 소득 5분위 및 소득원천별 가구소득 평균 : 정보손실(차이의 절대값) 비교

원자료 평균	가구	근로	사업	재산	공적	사적
전체	5,705	3,639	1,244	380	351	91
1분위	1,057	328	98	86	392	153
2분위	2,655	1,443	526	168	388	130
3분위	4,464	2,718	1,054	232	383	76
4분위	6,825	4,633	1,506	314	310	62
5분위	13,521	9,072	3,035	1,098	280	36

$\alpha = 1$	가구	근로	사업	재산	공적	사적
전체	5	3	3	0	1	0
1분위	2	2	1	0	0	2
2분위	3	2	3	1	1	1
3분위	14	0	2	12	0	1
4분위	2	13	6	1	5	2
5분위	209	6	201	6	4	1

$\alpha = 0.1$	가구	근로	사업	재산	공적	사적
전체	21	1	24	1	7	7
1분위	4	6	6	11	4	2
2분위	54	29	9	13	2	1
3분위	51	38	109	1	15	8
4분위	8	3	83	5	25	58
5분위	620	381	254	8	13	5

앞에서 설명한 소득 5분위 및 소득원천별 가구소득 평균을 구성비로 바꾸어 살펴보면 다음과 같다. [그림 29]는 소득원천별 가구소득 구성비를 그림으로 나타내었다. 평균과 마찬가지로 소득 4분위의 사적이전소득 항목에서 정보손실이 큰 것을 확인할 수 있다. <표 21>은 소득원천별 가구소득을 구성비로 나타낸 보도자료에 대해 차등정보보호 적용 결과를 보여준다. <표 22>는 원래의 구성비과 차등정보보호 구성비의 차이를 나타내며 참고로 제시하였다. 정보보호를 위해 차등정보보호를 적용한 결과를 수용할 수 있을 것인가에 관하여는 별도의 논의가 필요하다.



[그림 29] 소득원천별 가구소득 구성비의 변화

〈표 21〉 소득 5분위 및 소득원천별 가구소득 구성비 결과 비교

	가구	근로	사업	재산	공적	사적	정보보호 수준
전체	100	63.79	21.81	6.66	6.15	1.60	원자료
1분위	100	31.1	9.2	8.1	37.1	14.5	
2분위	100	54.3	19.8	6.3	14.6	4.9	
3분위	100	60.9	23.6	5.2	8.6	1.7	
4분위	100	67.9	22.1	4.6	4.5	0.9	
5분위	100	67.1	22.4	8.1	2.1	0.3	
전체	100	63.8	21.76	6.67	6.17	1.59	$\alpha = 1$
1분위	100	31.2	9.36	8.14	37.02	14.28	
2분위	100	54.33	19.73	6.37	14.61	4.96	
3분위	100	60.69	23.59	5.45	8.54	1.72	
4분위	100	68.05	21.97	4.62	4.47	0.88	
5분위	100	68.1	21.29	8.21	2.14	0.26	
전체	100	63.52	22.15	6.62	6	1.71	$\alpha = 0.1$
1분위	100	31.47	9.82	7.07	37.38	14.27	
2분위	100	54.32	19.76	6.69	14.39	4.84	
3분위	100	62.44	21.41	5.24	9.01	1.9	
4분위	100	67.92	23.31	4.53	4.18	0.06	
5분위	100	67.36	21.56	8.57	2.27	0.24	

〈표 22〉 소득 5분위 및 소득원천별 가구소득 구성비 : 정보손실(차이의 절대값) 비교

원자료 구성비(%)	근로	사업	재산	공적	사적
전체	63.79	21.81	6.66	6.15	1.60
1분위	31.1	9.2	8.1	37.1	14.5
2분위	54.3	19.8	6.3	14.6	4.9
3분위	60.9	23.6	5.2	8.6	1.7
4분위	67.9	22.1	4.6	4.5	0.9
5분위	67.1	22.4	8.1	2.1	0.3

$\alpha = 1$	근로	사업	재산	공적	사적
전체	0.01	0.05	0.01	0.02	0.01
1분위	0.1	0.16	0.04	0.08	0.22
2분위	0.03	0.07	0.07	0.01	0.06
3분위	0.21	0.01	0.25	0.06	0.02
4분위	0.15	0.13	0.02	0.03	0.02
5분위	1	1.11	0.11	0.04	0.04

$\alpha = 0.1$	근로	사업	재산	공적	사적
전체	0.27	0.34	0.04	0.15	0.11
1분위	0.37	0.62	1.03	0.28	0.23
2분위	0.02	0.04	0.39	0.21	0.06
3분위	1.54	2.19	0.04	0.41	0.2
4분위	0.02	1.21	0.07	0.32	0.84
5분위	0.26	0.84	0.47	0.17	0.06

다. 가구소득 및 순자산 분포

〈표 22〉 보도자료(표 4-7) 소득분위별 순자산 분포[23)]

2017년					
소득 / 순자산	1분위	2분위	3분위	4분위	5분위
전체	100.0	100.0	100.0	100.0	100.0
1분위	48.1	28.2	14.4	7.1	2.1
2분위	25.6	27.2	25.7	16.4	5.1
3분위	15.0	20.9	24.7	26.7	12.7
4분위	7.3	14.6	20.8	28.4	28.8
5분위	4.0	9.0	14.3	21.3	51.4

보도자료에 포함된 가구소득 및 순자산 분포는 가구소득과 순자산을 각각 5분위 구간으로 나누어 총 25개의 구간 조합을 구성한 후 각 구간에 속하는 가구의 비율을 추정한 분포이다. 이 분포는 각 구간에 포함된 가구 수에 대한 2차원 히스토그램이며, 가구소득 분위에 대하여 합이 100이 되도록 조정하여 정리되어 있다. α-차등정보보호 2차원 히스토그램을 동일한 방식으로 작성하여 <표 25> 및 <표 26>에 정리하였다. <표 26>는 각 2차원 히스토그램에 차등정보보호를 적용할 때 정보손실을 나타낸다. 정보보호 수준이 $\alpha = 0.1$로 강력한 경우에는 더해지는 잡음이 많으므로 정보손실도 많이 발생하는 것을 알 수 있다. 참고로 <표 24>는 자산분위 및 소득분위의 빈도표이며, 차등정보보호 적용 결과를 이해하는데 도움을 주기 위하여 제시하였다.

〈표 24〉 소득분위별 순자산 분포 : 빈도표

소득 / 순자산	1분위	2분위	3분위	4분위	5분위
1분위	2141	1062	503	236	64
2분위	1180	1083	954	573	167
3분위	629	846	885	940	410
4분위	321	557	729	951	923
5분위	160	346	510	733	1737

23) 통계청 보도자료 가계금융 복지조사 결과(2018) IV장 5. 가구소득 및 순자산 분포

〈표 25〉 소득분위별 순자산 분포 결과 비교($\alpha = 1$)

기본

소득 / 순자산	1분위	2분위	3분위	4분위	5분위
전체	100.0	100.0	100.0	100.0	100.0
1분위	48.1	28.2	14.4	7.1	2.1
2분위	25.6	27.2	25.7	16.4	5.1
3분위	15.0	20.9	24.7	26.7	12.7
4분위	7.3	14.6	20.8	28.4	28.8
5분위	4.0	9.0	14.3	21.3	51.4

보정

소득 / 순자산	1분위	2분위	3분위	4분위	5분위
전체	100.0	100.0	100.0	100.0	100.0
1분위	48.3	27.8	14.6	7.3	1.9
2분위	25.8	27.6	25.5	16.5	5.2
3분위	15	20.7	24.9	26.5	12.7
4분위	6.6	15	20.5	28.5	28.7
5분위	4.3	8.8	14.5	21.3	51.5

동일

소득 / 순자산	1분위	2분위	3분위	4분위	5분위
전체	100.0	100.0	100.0	100.0	100.0
1분위	47.6	28	14.4	7.4	2.1
2분위	25.7	27.8	25.8	16.3	5.2
3분위	14.9	20.6	24.8	26.6	12.8
4분위	7.6	14.5	20.8	28.4	28.7
5분위	4.2	9.2	14.3	21.3	51.1

재현

소득 / 순자산	1분위	2분위	3분위	4분위	5분위
전체	100.0	100.0	100.0	100.0	100.0
1분위	47.8	29.2	13.9	7	2
2분위	25.4	26.4	26.3	17	5.4
3분위	15.4	21.6	24.9	27	12
4분위	7.4	13.8	20.1	28.4	28.8
5분위	4.1	8.9	14.9	20.6	51.7

〈표 26〉 소득분위별 순자산 분포 결과 비교($\alpha = 0.1$)

기본

순자산 \ 소득	1분위	2분위	3분위	4분위	5분위
전체	100.0	100.0	100.0	100.0	100.0
1분위	48.1	28.2	14.4	7.1	2.1
2분위	25.6	27.2	25.7	16.4	5.1
3분위	15.0	20.9	24.7	26.7	12.7
4분위	7.3	14.6	20.8	28.4	28.8
5분위	4.0	9.0	14.3	21.3	51.4

보정

순자산 \ 소득	1분위	2분위	3분위	4분위	5분위
전체	100.0	100.0	100.0	100.0	100.0
1분위	47	28.5	14.2	7.5	9
2분위	28.1	26.3	29.5	15.8	3.4
3분위	15.9	23.1	23.6	26.7	14.1
4분위	7.1	13.7	22.5	29.2	16.8
5분위	1.9	8.3	10.2	20.7	56.7

동일

순자산 \ 소득	1분위	2분위	3분위	4분위	5분위
전체	100.0	100.0	100.0	100.0	100.0
1분위	45.7	28.5	13.7	9.6	2.9
2분위	25.7	25.6	24.3	16.4	5.9
3분위	14.2	19.1	24.3	28.3	12.8
4분위	10	16.4	23	25.3	29.4
5분위	4.5	10.3	14.6	20.4	49.1

재현

순자산 \ 소득	1분위	2분위	3분위	4분위	5분위
전체	100.0	100.0	100.0	100.0	100.0
1분위	46.9	31.7	12.8	6.9	4
2분위	24.3	30.6	28.6	16.1	3.2
3분위	13.7	21.8	24.7	26.3	12.1
4분위	10.2	7	20	29.9	27.3
5분위	4.9	8.9	14	20.8	53.4

〈표 27〉 소득분위별 순자산 분포 결과 : 정보손실(차이의 절대값) 비교

보정 ($\alpha = 1$)

소득 / 순자산	1분위	2분위	3분위	4분위	5분위
1분위	0.2	0.4	0.2	0.2	0.2
2분위	0.2	0.4	0.2	0.1	0.1
3분위	0	0.2	0.2	0.2	0
4분위	0.7	0.4	0.3	0.1	0.1
5분위	0.3	0.2	0.2	0	0.1

보정 ($\alpha = 0.1$)

소득 / 순자산	1분위	2분위	3분위	4분위	5분위
1분위	1.1	0.3	0.2	0.4	6.9
2분위	2.5	0.9	3.8	0.6	1.7
3분위	0.9	2.2	1.1	0	1.4
4분위	0.2	0.9	1.7	0.8	12
5분위	2.1	0.7	4.1	0.6	5.3

동일 ($\alpha = 1$)

소득 / 순자산	1분위	2분위	3분위	4분위	5분위
1분위	0.5	0.2	0	0.3	0
2분위	0.1	0.6	0.1	0.1	0.1
3분위	0.1	0.3	0.1	0.1	0.1
4분위	0.3	0.1	0	0	0.1
5분위	0.2	0.2	0	0	0.3

동일 ($\alpha = 0.1$)

소득 / 순자산	1분위	2분위	3분위	4분위	5분위
1분위	2.4	0.3	0.7	2.5	0.8
2분위	0.1	1.6	1.4	0	0.8
3분위	0.8	1.8	0.4	1.6	0.1
4분위	2.7	1.8	2.2	3.1	0.6
5분위	0.5	1.3	0.3	0.9	2.3

재현 ($\alpha = 1$)

소득 / 순자산	1분위	2분위	3분위	4분위	5분위
1분위	0.3	1	0.5	0.1	0.1
2분위	0.2	0.8	0.6	0.6	0.3
3분위	0.4	0.7	0.2	0.3	0.7
4분위	0.1	0.8	0.7	0	0
5분위	0.1	0.1	0.6	0.7	0.3

재현 ($\alpha = 0.1$)

소득 / 순자산	1분위	2분위	3분위	4분위	5분위
1분위	1.2	3.5	1.6	0.2	1.9
2분위	1.3	3.4	2.9	0.3	1.9
3분위	1.3	0.9	0	0.4	0.6
4분위	2.9	7.6	0.8	1.5	1.5
5분위	0.9	0.1	0.3	0.5	2

3. 재현자료로 활용 가능성

재현 방식으로 차등정보보호 히스토그램을 생성할 때 중간과정에서 재현자료가 얻어진다. 이 재현자료를 베이지안 재현자료처럼 사용할 수 있는지 간단히 살펴보자.

먼저 각 항목에 대하여 원자료와 동일한 크기의 재현자료를 생성한다. 재현자료는 각 항목마다 보정된 교란 히스토그램에서 얻으므로 변수 사이의 관계는 전혀 고려되지 않는다. 원래 마이크로데이터가 <표 28>과 같이 나타난다면, 재현자료는 <표 29>와 같이 나타난다. 가중치를 고려하지 않은 평균은 각 표의 맨 아랫줄에 수록하였다. 이 차등정보보호 재현자료는 <표 12>에서와 같이 6개의 막대를 가지는 히스토그램으로부터 생성되었다.

〈표 28〉 원자료의 가구소득 세부항목 마이크로데이터

	근로	사업	재산	공적	사적
1	140	1	1	442	140
2	1293	28000	326	1	1
3	15159	3000	21	1	1
4	1	1	22	240	180
5	7233	1	329	1	1
6	7197	1369	2301	1	1
…	…	…	…	…	…
평균	3288.2	1177.6	407.8	391.1	101.4

〈표 29〉 차등정보보호 자료의 가구소득 세부항목 마이크로데이터

	근로	사업	재산	공적	사적
1	18083.8	3834.0	4600.7	504.5	567.0
2	18439.1	6724.2	4999.4	200.6	399.8
3	18326.0	9510.3	6786.0	480.9	112.0
4	3506.1	2988.2	2804.1	126.1	670.0
5	8106.6	4664.1	6576.5	452.5	572.2
6	1842.8	6555.2	8359.1	585.0	189.4
…	…	…	…	…	…
평균	9675.1	5186.4	4757.7	603.2	381.3

실험 결과 원래 마이크로데이터를 사용하여 얻은 평균에 비하여 차등정보보호 재현자료를 사용하면 평균의 정확도가 크게 떨어진다. 또한 실험할 때마다 그 차이도 다양하였다. 평균을 직접 교란시켜 차등정보보호 평균을 얻을 때 자료의 크기가 충분히 크면 정보손실이 크지 않았던 반면에, 차등정보보호 재현자료로 평균을 구하는 것은 정보손실이 커서 바람직하지 않음을 알 수 있다. 다만 차등정보보호 재현자료는 평균이 아니라 6개 막대를 가지는 히스토그램을 생성하는 과정에서 얻은 것임을 주의할 필요가 있다. 차등정보보호는 쿼리마다 매커니즘이 달라지므로 평균이라는 쿼리에 더욱 적절한 밀도함수 추정량을 고려할 수 있으며, 나아가 보도자료에 맞추어 6개 막대를 가지는 히스토그램을 밀도함수 추정량으로 사용하였으므로 정보손실이 큰 것은 충분히 예상되는 결과이기 때문이다.

같은 재현자료에 대하여 상관계수 행렬을 구해보면 <표 30> 및 <표 31>과 같이 나타난다. 각 변수별로 차등정보보호를 적용하여 생성한 재현자료에서 상관관계가 거의 사라지는 것을 알 수 있다. 처음부터 상관관계를 고려하지 않았으므로 이는 당연한 결과이다.

〈표 30〉 원자료의 가구소득 세부항목 상관계수 행렬

	근로	사업	재산	공적	사적
근로	1.0000	-0.0973	0.1931	-0.1914	-0.1541
사업	-0.0973	1.0000	0.0897	-0.0709	-0.0734
재산	0.1931	0.0897	1.0000	0.0741	-0.0167
공적	-0.1914	-0.0709	0.0741	1.0000	0.0443
사적	-0.1541	-0.0734	-0.0167	0.0443	1.0000

〈표 31〉 차등정보보호 자료의 가구소득 세부항목 상관계수 행렬

	근로	사업	재산	공적	사적
근로	1.0000	0.0023	-0.0007	0.0026	-0.0193
사업	0.0023	1.0000	0.0086	-0.0050	0.0000
재산	-0.0007	0.0086	1.0000	0.0002	0.0021
공적	0.0026	-0.0050	0.0002	1.0000	0.0007
사적	-0.0193	0.0000	0.0021	0.0007	1.0000

지금까지 설명한 차등정보보호 재현자료는 히스토그램 산출을 목적으로 생성하였기 때문에 히스토그램은 비교적 정확하게 얻어지지만, 평균이나 상관계수를 계산할 때는

정확성이 떨어지는 것을 확인할 수 있다. 이는 차등정보보호가 쿼리마다 다르게 적용됨을 감안할 때 자연스러운 현상이다.

반면, Wasserman과 Zhou(2010)가 논의한 대로 최소최대 위험 관점에서 재현자료가 원자료와 동일한 정보를 가진다는 것이 증명되어 있음을 고려한다면 이러한 실험결과가 의아할 수 있다. 최소최대 위험은 점근적으로(asymptotically) 효율성이 증명된 정보손실 측도이지만, 실험은 유한표본으로 진행된다는 것을 주의할 필요가 있다. 또한 최소최대 위험은 자료를 생성하는 매커니즘에 대한 정보손실 측도이지만, 전통적인 거리기반 정보손실은 데이터마다 측정되는 정보손실 측도이다. 향후 전통적인 정보손실 측도를 어떻게 활용할지 더 생각해 볼 필요가 있다.

다음으로 2개 항목을 선택하여 6×6 개의 막대를 가지는 이차원 히스토그램을 작성하면서 동일한 실험을 수행했다. 단변량 히스토그램과 마찬가지로 평균과 상관계수를 구해보면 재현자료를 생성할 때마다 그 결과가 달라지는 것을 확인할 수 있다. 근로소득과 재산소득에서 원자료의 평균은 각각 3288 및 408이나, 재현자료를 이용하여 얻은 평균은 실험할 때마다 크게 달라진다. 또한 <표 32>의 결과를 보면 정보손실이 크다는 것을 알 수 있다. 상관계수에 관하여는 <표 33>에서 재현자료의 상관계수가 자료를 생성할 때마다 달라지는 것을 보여준다.

〈표 32〉 가구소득 세부항목 평균

	근로	재산
원자료	3288.2	407.8

	근로	재산
재현자료	83154.4	48129.1

〈표 33〉 가구소득 세부항목 중 근로소득과 재산소득의 상관계수(원자료: 0.1931)

반복 회차	1	2	3	4	5	…
상관계수	0.2032	0.0968	0.0762	-0.1335	-0.1514	…

차등정보보호로 노출위험 수준이 보장되는 재현자료를 이용해 베이지안 재현자료를 대신할 수 있다면 매우 바람직하다. 베이지안 재현자료는 자료마다 모형 구축 작업이 쉽지 않기 때문이다. 또한 노출위험 제어도 정량적으로 보장되지 않는다. 이에 비하여 차등정보보호 재현자료는 노출위험 제어가 보장되지만 전통적 의미에서 정보손실이 크다는 단점이 있다. 이를 보완하기 위하여 다음과 같은 간단한 알고리즘 아이디어를 제안한다. 이는 최소최대 추정량 중에서 고전적인 정보손실을 작게 가지는 재현자료를 선택하는 과정이다.

단계0. 검정력 혹은 검정오차를 감안하여 정보보호 수준 α를 결정한다.

단계1. 보정된 교란 히스토그램, 평활 히스토그램, 기저함수를 이용한 분포함수 등으로부터 밀도함수를 추정한다.

단계2. 추정한 밀도함수에 α-차등정보보호를 적용한다. (또는 주어진 쿼리에 대하여 차등정보보호를 만족하는 일반화 이중지수분포를 고안한다.)

단계3. α-차등정보보호를 만족하는 밀도함수 추정량에서 재현자료를 생성한다. 다만 차등정보보호를 만족시키는 조건을 고려하면서 밀도함수 추정량의 정확성을 높이기 위한 작업을 수행한다. 해당 작업으로 히스토그램 막대 개수 조정, 적절한 기저함수 선택 등을 생각할 수 있다. 다음은 Wasserman(2012)이 제안한 과정이다.

$$X_1, \ldots, X_n \xrightarrow[estimation]{density} \hat{p} \xrightarrow{privatize} \hat{\boldsymbol{p}}^{*} \xrightarrow{sample} Z_1, \ldots, Z_k$$

단계4. 단계 3을 R회 반복한다. 요청받은 쿼리에 대하여 원자료 및 R개 재현자료 세트에서 얻은 결과 사이의 정보손실을 계산한다. 정보손실이 가장 적은 재현자료 세트를 제공한다.

참고로 단계4를 수행하면 동일한 쿼리에 여러 번 응답하는 것이므로 차등정보보호 수준 α가 나뉘며, 이를 차등정보보호 예산이 소모된다고도 표현한다. 따라서 처음에 규정된 수준으로 차등정보보호가 보장된다고 말하기 어려우며 또다시 정보손실-노출위험 상충관계가 발생한다. 향후 다양한 시도와 활발한 논의가 지속되기를 기대한다.

제5장 결론

본 연구에서는 차등정보보호의 실무 적용 가능성 여부를 검토하기 위해 히스토그램을 위주로 차등정보보호 적용을 실험하여 보았다. 실험 결과를 논하기 위해 먼저 노출위험-정보손실 관점에서 기존의 매스킹, 베이지안 재현자료, 차등정보보호 자료를 비교하였다. 이는 차등정보보호의 장점이나 비교우위를 확인하기 위한 작업이다. 다음으로 차등정보보호를 통계학적 관점에서 살펴보았다. 차등정보보호 모수를 가설검정의 검정력 혹은 검정오차로 이해하여 노출위험을 해석하고, 정보손실은 통계적 추론의 관점에서

이해할 수 있음을 살펴보았다. 또한 차등정보보호 자료를 생성하는 방식으로 교란 방식과 재현 방식이 있음을 소개하면서 필요한 이론적 설명을 제시하였다.

이어 차등정보보호 적용에 관한 모의실험을 실시하고, 실제자료를 이용해 차등정보보호 적용실험을 수행한 결과를 수록하였다. 정보손실을 일정 수준 이하로 유지하기 위해서는 자료의 크기가 중요하며, 정보보호 수준에 따라서 자료의 크기가 더 클 필요가 있다는 것을 확인하였다. 예를 들어 히스토그램을 생성할 때 각 구간마다 적절한 크기의 자료가 필요하다. 실제자료에 차등정보보호를 적용하여 제공할 수 있으나, 제공 전에 다양한 측면의 분석을 통해 정보손실이 지나치게 많이 발생하는 것은 아닌지 점검할 필요가 있다. 차등정보보호 적용은 자료 특성에 따라 주의가 요구될 수 있기 때문이다. 예를 들어 계층 구조가 있을 때 등은 더 세심한 검토가 필요하다. 정리하면 최소최대 위험을 정보손실 측도로 활용할 때는 고전적인 정보손실 측도를 적용한 결과를 함께 검토할 필요가 있다. 또한 차등정보보호 자료의 이용자 수용성 여부는 별도의 논의가 필요하다 (Ruggles et al., 2018).

마지막으로 차등정보보호 재현자료의 활용 가능성을 검토하였다. 데이터센터 방문 이전 사전분석용으로 재현자료를 활용하는 것은 효율적인 대안으로 판단되며, 캐나다에서 이미 시행하고 있기도 하다. 게다가 차등정보보호 재현자료는 베이지안 재현자료보다 생성이 쉬우면서도 노출위험 제어가 보장되는 장점이 있다. 이는 실무적으로 큰 이점이다. 실제자료를 이용해 간단히 실험한 결과, 밀도함수 추정량으로 6개 막대를 가지는 히스토그램을 사용하고 차등정보보호를 만족하도록 생성한 재현자료는 히스토그램 이외의 통계에 대해서는 정보손실의 편차가 매우 컸다. 이는 차등정보보호를 적용하여 노출위험 수준은 보장하나, 최소최대 위험 이외의 정보손실 측도에 대해서는 그 수준을 보장하지 못한다는 것을 의미한다. 다만 향후 알고리즘 개선으로 정보손실을 좀 더 축소하여 재현자료를 생성할 가능성은 존재한다.

본 연구에서는 데이터센터 혹은 데이터 거래소에서 데이터 정보보호에 활용할 수 있는 다음과 같은 재현자료 활용 프로세스를 제안한다. 데이터 정보보호 분야의 국내 연구가 미비한 상황이기 때문에, 향후 구체적인 논의를 시작할 수 있는 지점을 제공하기 위하여 아이디어 차원의 프로세스를 간단히 제안하는 바이다.

단계0. 각 데이터베이스마다 민감 정도를 검토한다. 개인정보 노출 가설에 대한 검정력 혹은 검정오차를 검토한다. 자료의 민감성 및 검정력(검정오차) 등을 감안하여 데이터베이스마다 정보보호 수준 α를 결정한다.

단계1. 이용 신청이 발생한 데이터베이스 D에 대하여 각 변수별로 차등정보보호를

만족하는 재현자료를 생성하여 이용자 접근이 용이한 방식으로 제공한다. 자료 분석 목적이 명확한 경우 요청받은 쿼리에 적합한 차등정보보호 재현자료를 제공할 수 있다. 단계1은 탐색용 자료를 제공하는 절차에 해당된다.

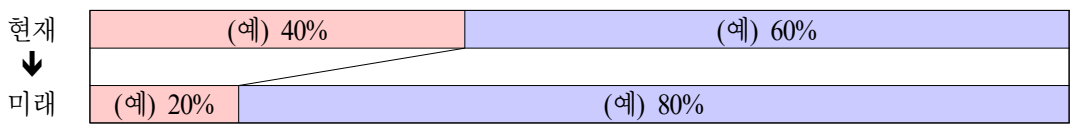
단계2. 분석계획 수립을 마치고 본격적인 분석을 요구하는 이용자에게 선택한 변수 범위에 대하여 베이지안 재현자료를 생성하여 원격 분석용 등으로 제공한다.

단계3. 이용자가 작성한 코드를 제출받아 원자료 분석을 수행하고 분석결과에 대하여 통계적 노출제어 가이드라인[24]을 적용한다. 가이드라인 준수 여부 검토를 마친 분석결과물을 제공한다. 만약 쿼리가 자료연계인 경우에는 단계2로 돌아가거나 인가파일 제공 등의 물리적 노출제어 과정을 수행한다.

참고로 통계적 노출제어[25]와 물리적 노출제어는 다음 <표 34>와 같이 나누어 생각할 수 있다. 효율적 자료제공을 위하여 통계적 노출제어만으로 데이터를 공유할 수 있는 플랫폼이 많아지는 것이 바람직하다고 판단된다.

〈표 34〉 데이터 정보보호 방안 분류

<table>
<tr><td>이용자</td><td colspan="2">심층 이용자</td><td colspan="3">불특정 다수</td></tr>
<tr><td>종류</td><td colspan="2">물리적 노출제어</td><td colspan="3">통계적 노출제어(SDC)</td></tr>
<tr><td rowspan="2">배포
형식</td><td>microdata</td><td>macrodata
(analysis output)</td><td>microdata</td><td colspan="2">macrodata
(analysis output)</td></tr>
<tr><td>인가파일</td><td>가공 통계표
원격 접속(RAS)
데이터센터(RDC)</td><td>공공이용파일</td><td colspan="2">온라인 분석 시스템
(OAS)</td></tr>
<tr><td>전략/
기법</td><td colspan="2">접근제어 및 결과통제
➢ 결과물에 SDC 처리 필요</td><td>매스킹
기법</td><td>재현자료
(SD)</td><td>플랫폼
맞춤형</td><td>차등정보보호
(DP)</td></tr>
</table>



24) 통계개발원에서는 이용자가 자료 분석을 마치고 반출을 요청하는 분석결과물에 대한 통계적 노출제어 가이드라인 연구 과제를 수행한 바 있으며, 향후 업데이트 예정이다. 이 가이드라인은 이용자가 자료 분석을 요청하는 시점부터 이용자가 숙지하도록 제공된다.

25) [용어] Masking → Statistical Disclosure Control → Statistical Data Confidentiality → Data Privacy
매스킹 → 통계적 노출제어 → 통계적 비밀보호 → 데이터 정보보호
[의미] 통계 및 데이터를 제공할 때 발생할 수 있는 개인정보 노출위험을 제어하고,
노출위험을 제어할 때 발생할 수 있는 데이터 유용성 감소를 완화하기 위한 통계적 방안 모색

한편 데이터 공유 및 개인정보보호라는 큰 목표를 이루기 위하여 각 주제별로 세부 과제가 정교하게 선정되고 수행될 필요가 있다. 세부 주제의 예를 제시하면 다음과 같다.

- 밀도함수 추정 및 차등정보보호 적용[26)]
- 이용 빈도가 높은 자료에 대한 베이지안 재현자료 생성
- 이용 빈도가 높은 쿼리에 대하여 이용자 편의성 확보를 위한 플랫폼 개발
- 각계의 관심을 유도할 수 있는 다양한 연구 네트워크 구성 및 운영
- 데이터 정보보호(data privacy)에 관하여 알기 쉬운 글(article) 기고 활동

4차 산업혁명 인프라 구축에 데이터 공유는 필수불가결한 항목이다. 더불어 개인정보보호 조치가 있어야 데이터 공유가 가능하다. 차등정보보호, 베이지안 재현자료 등 신기술의 적절한 활용으로 향후 개인정보보호와 데이터 공유가 모두 가능해지기를 기대한다.

26) 데이터베이스별 차등정보보호 적용 범위, 조건부 분포에 관한 차등정보보호, 최소최대 추정량 연구 및 차등정보보호 적용 알고리즘 개선, 유용성의 측도를 유한표본에서 계산하는 적절한 방법, 새로운 노출위험 측도에 관한 연구 등이 2019년 12월 연구협의체에서 제안되었다.

참고문헌

박민정(2014), “가계금융·복지조사 마이크로데이터 매스킹 방안 평가”, 통계개발원.
박민정(2016). “통계적 노출제어 최신동향 검토 - 차등정보보호를 중심으로-”, 통계개발원.
박민정, 김정연(2017). “재현자료 작성 방법론 검토”, 통계개발원.
박민정, 이용희, 권성훈(2018), “차등정보보호에 관한 연구”, 통계개발원.

Duchi, J. C. Jordan, M. I. and Wainwright, M. J.(2018), Minimax optimal procedures for locally private estimation, Journal of the American Statistical Association, 113(521), 182-215.

Dwork, C.(2006), Differential privacy, In 33rd International Colloquium on Automata, Languages and Programming, Part II (ICALP 2006), Springer, Venice, Italy, 1-12.

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, TCC 2006, Lecture Notes in Computer Science, vol 3876, Springer, Berlin, Heidelberg.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. and de Wolf, P. P.(2012), Statistical Disclosure Control, John Wiley & Sons Ltd.

Jagadish, H. V., Koudas, N., Muthukrishnan, S., Poosala, V., Sevcik, K.C., and Suel, T.(1998), Optimal histograms with quality guarantees, VLDB, 275-286.

Reiter, J. P.(2019), Differential privacy and federal data release, Annual Review of Statistics and Its Application, 6:13.1-13.17.

Templ, M. and Meindl, B.(2008), Robustification of microdata masking methods and the comparison with existing method. Privacy in Statistical Database, Lecture Notes in Computer Science, Springer, 5262: 177-189.

Xu, J., Zhang, Z., Xiao, X., Yang, Y., Yu, G. and Winslett, M.(2013), Differential private histogram publication, The VLDB Journal, 22(6), 797-822.

Wasserman, L. and Zhou., S.(2010), A statistical framework for the differential privacy, Journal of the American Statistical Association, 105(489), 375-389.

Wasserman, L.(2012), Minimaxity, Statistical Thinking and Differential Privacy, Journal of Privacy and Confidentiality, 4(1), 51-63.

Ruggles, S.(2018), Implication of differential privacy for Census Bureau data and scientific research, Working paper No. 2018-6, Task force on differential privacy for Census data, Institute for Social Research and Data Innovation, Minnesota Population Center at the University of Minnesota.

<부록>

1. 최적 구조 히스토그램 생성 알고리즘

히스토그램을 제공할 때, m개의 막대로 구성된 히스토그램 $H_m (= C_m)$으로부터 $k(\leq m)$개의 막대로 구성된 히스토그램 H_k를 새로 그릴 수 있다. 새로운 구조를 가지는 히스토그램 H_k가 원래의 분포를 더 잘 표현한다는 것은 자료의 참 분포를 H_0라 할 때 $Error(H_m, H_0) \geq Error(H_k, H_0)$라는 것을 의미한다. 히스토그램 H_k를 만들 때 막대를 병합하는 많은 경우의 수 중에서 가장 적절한 구조를 가지는 알고리즘은 다음과 같다.

병합으로 인해 각 막대마다 다양한 길이를 가지는 히스토그램 H_k를 다음과 같이 표현하자.

$$H_k = \{(l_1, r_1, h_1),\ ...,\ (l_k, r_k, h_k)\} = H(C_m, B_k).$$

단, $h = average(\{c_l,\ ...,\ c_r\}) = \bar{c}(l, r)$ 또는 $h = median(\{c_l,\ ...,\ c_r\})$이며, $B_k = \{l_1, r_1,\ ...,\ r_k\} \subset B_m$, $l_1 = b_0$, $l_{i+1} = r_i$, $r_k = b_m$이다.

이상적으로는 $Error(H_k, H_0)$를 최소화하는 H_k를 찾는 것이 바람직하겠지만, 참 분포 H_0를 알 수 없는 경우가 대부분이다. 따라서 다음과 같은 $SSE(H_k, C_m)$를 최소화하는 히스토그램을 최적 히스토그램으로 정의한다. 다음 최소화시키는 알고리즘은 m개의 막대를 k의 막대로 병합하는 많은 방법 중에서 최적의 해를 구하는 방법이 된다.

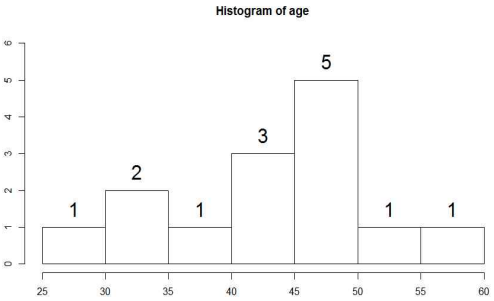
$$SSE(H_k, C_m) = \sum_{j=1}^{k} \sum_{i=l_j}^{i=r_j} (h_j - c_i)^2.$$

Jagadish et al.(1998)은 재귀적으로 위 SSE 최소값을 계산하는 알고리즘(dynamic programming algorithm)을 제안하였다. 최소 SSE를 $T(m, k)$라고 표현하면 최적해는 다음 식에서 구할 수 있다. 또한 알고리즘이 진행되는 과정에서 얻어지는 j값들로 B_k를 구할 수 있다.

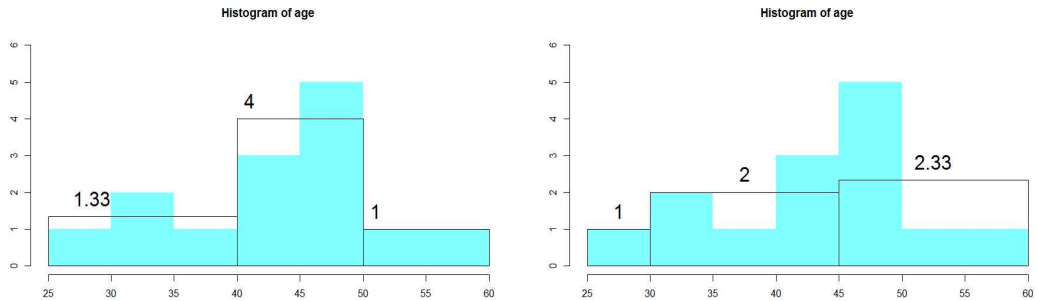
$$T(i, k) = min_{k-1 \leq j \leq i-1} \{T(j, k-1) + SSE(C_m, j+1, i)\}.$$

단, $SSE(C_m, l, r) = \sum_{i=l}^{r}(c_i - \bar{c}(l,r))^2$ 이다.

예를 들어 다음과 같은 히스토그램이 있다고 하자(Xu et al., 2013). 이 히스토그램은 $m = 7$개의 막대로 구성되어 있으며, 각 막대의 너비는 균일하다.



이제 이 히스토그램에 대하여, $k = 3$ 인 히스토그램 중에서 최적 구조를 가지는 히스토그램을 그리면 아래 그림의 왼쪽과 같다. 반면 오른쪽 그림은 최적 구조가 아닌 사례를 보여준다. 참고로 두 그림에서 H_m은 경계선 없이 하늘색으로 나타나 있다.



일반적으로 k 값이 작아질수록 즉, 더 적은 개수의 막대를 사용해서 히스토그램을 만들수록 $SSE(H_k, C_m)$는 커지는 경향이 있다. 이는 기준이 C_m이기 때문에 자연스러운 현상이다. 물론 기준을 H_0로 한다면 H_k가 C_m보다 더 적절할 수도 있다.

2. 모의실험 결과($d = 1$ 인 경우)

〈부표 1〉 원자료 밀도함수와 생성된 히스토그램 사이의 거리(막대 개수가 10)

			교란 방식				재현 방식		
α	자료수	기본	보정	양	동일	잡음	재현	평활1	평활2
0.01	100	0.3674	1.2986	0.8918	0.9050	1.0551	1.3407	NA	NA
0.01	200	0.3274	1.1975	0.8837	0.8578	1.0751	1.2559	NA	NA
0.01	400	0.2967	1.1554	0.8338	0.8192	0.9771	1.1544	NA	NA
0.01	800	0.2800	1.0035	0.7002	0.7387	0.8724	1.0050	NA	NA
0.01	1600	0.2669	0.7589	0.5992	0.5869	0.7315	0.7691	NA	NA
0.01	3200	0.2639	0.5321	0.4429	0.4645	0.5339	0.5472	NA	NA
0.1	100	0.3663	0.9268	0.6692	0.6728	0.8500	0.9483	NA	NA
0.1	200	0.3267	0.6747	0.5556	0.5537	0.6839	0.7332	NA	NA
0.1	400	0.2921	0.5163	0.4346	0.4464	0.5153	0.5209	NA	NA
0.1	800	0.2777	0.3766	0.3501	0.3556	0.3880	0.3925	NA	0.3927
0.1	1600	0.2634	0.2991	0.2931	0.2983	0.3061	0.3091	0.3280	0.3535
0.1	3200	0.2640	0.2750	0.2758	0.2760	0.2757	0.2808	0.3067	0.3328
1	100	0.3666	0.4155	0.3933	0.3953	0.4316	0.4786	0.4292	0.4259
1	200	0.3245	0.3496	0.3361	0.3357	0.3473	0.3786	0.3677	0.3764
1	400	0.2956	0.3026	0.3014	0.3023	0.3035	0.3252	0.3215	0.3443
1	800	0.2810	0.2825	0.2839	0.2845	0.2829	0.2990	0.3076	0.3329
1	1600	0.2670	0.2677	0.2683	0.2682	0.2677	0.2732	0.2913	0.3191
1	3200	0.2582	0.2586	0.2586	0.2587	0.2584	0.2621	0.2777	0.3097
10	100	0.3679	0.3683	0.3656	0.3668	0.3687	0.4368	0.3650	0.3710
10	200	0.3151	0.3157	0.3149	0.3150	0.3153	0.3508	0.3240	0.3388
10	400	0.2927	0.2927	0.2928	0.2929	0.2929	0.3175	0.3045	0.3283
10	800	0.2764	0.2764	0.2766	0.2765	0.2764	0.2937	0.2932	0.3214
10	1600	0.2685	0.2685	0.2686	0.2686	0.2685	0.2760	0.2868	0.3172
10	3200	0.2581	0.2581	0.2582	0.2582	0.2581	0.2620	0.2756	0.3073

〈부표 2〉 원자료 밀도함수와 생성된 히스토그램 사이의 거리(막대 개수가 20)

			교란 방식				재현 방식		
α	자료수	기본	보정	양	동일	잡음	재현	평활1	평활2
0.01	100	0.3638	1.3186	0.9448	0.9073	1.0571	1.3238	NA	NA
0.01	200	0.2722	1.2963	0.9300	0.8868	1.0410	1.2732	NA	NA
0.01	400	0.2235	1.2135	0.8823	0.8765	1.0026	1.2483	NA	NA
0.01	800	0.1876	1.1023	0.8467	0.8422	0.9309	1.1578	NA	NA
0.01	1600	0.1724	0.9807	0.7197	0.6993	0.8490	0.9535	NA	NA
0.01	3200	0.1662	0.6958	0.5637	0.5601	0.6667	0.7125	NA	NA
0.1	100	0.3671	1.0995	0.7670	0.7543	0.9464	1.1443	NA	NA
0.1	200	0.2745	0.8912	0.6547	0.6254	0.7946	0.9209	NA	NA
0.1	400	0.2203	0.6697	0.5110	0.5024	0.6533	0.6745	NA	NA
0.1	800	0.1900	0.4440	0.3722	0.3657	0.4319	0.4563	NA	0.3973
0.1	1600	0.1736	0.2930	0.2673	0.2658	0.2900	0.3080	0.3118	0.3232
0.1	3200	0.1662	0.2149	0.2059	0.2106	0.2198	0.2216	0.2576	0.2753
1	100	0.3694	0.4912	0.4108	0.4164	0.4814	0.5993	0.4563	0.4533
1	200	0.2802	0.3466	0.3072	0.3137	0.3408	0.4042	0.3435	0.3477
1	400	0.2212	0.2481	0.2368	0.2375	0.2488	0.2991	0.2722	0.2925
1	800	0.1897	0.1986	0.1960	0.1972	0.2008	0.2307	0.2311	0.2571
1	1600	0.1735	0.1775	0.1760	0.1764	0.1778	0.1966	0.2078	0.2441
1	3200	0.1651	0.1664	0.1665	0.1662	0.1668	0.1770	0.1960	0.2343
10	100	0.3643	0.3666	0.3567	0.3606	0.3662	0.4758	0.3556	0.3507
10	200	0.2759	0.2767	0.2739	0.2739	0.2763	0.3542	0.2810	0.2866
10	400	0.2212	0.2215	0.2206	0.2205	0.2217	0.2745	0.2311	0.2570
10	800	0.1909	0.1910	0.1908	0.1909	0.1909	0.2251	0.2075	0.2410
10	1600	0.1736	0.1736	0.1737	0.1737	0.1736	0.1936	0.1952	0.2344
10	3200	0.1660	0.1660	0.1660	0.1661	0.1661	0.1772	0.1881	0.2282

〈부표 3〉 기본 히스토그램과 차등정보보호 히스토그램 사이의 거리(막대 개수가 10)

		교란 방식				재현 방식		
α	자료수	보정	양	동일	잡음	재현	평활1	평활2
0.01	100	1.2814	0.8699	0.8614	1.0260	1.3197	NA	NA
0.01	200	1.1525	0.8426	0.8170	1.0395	1.2291	NA	NA
0.01	400	1.1173	0.7846	0.7716	0.9305	1.1175	NA	NA
0.01	800	0.9612	0.6284	0.6693	0.8137	0.9669	NA	NA
0.01	1600	0.6963	0.5149	0.4998	0.6666	0.7054	NA	NA
0.01	3200	0.4328	0.3269	0.3515	0.4423	0.4515	NA	NA
0.1	100	0.8685	0.5912	0.5915	0.8026	0.8754	NA	NA
0.1	200	0.5779	0.4669	0.4544	0.5949	0.6481	NA	NA
0.1	400	0.3925	0.2944	0.3037	0.3911	0.4047	NA	NA
0.1	800	0.2156	0.1841	0.1883	0.2346	0.2434	NA	0.2504
0.1	1600	0.1117	0.0955	0.1059	0.1184	0.1353	0.1689	0.2027
0.1	3200	0.0531	0.0545	0.0544	0.0606	0.0708	0.1286	0.1739
1	100	0.1883	0.1487	0.1482	0.1920	0.2867	0.2051	0.2330
1	200	0.0984	0.0823	0.0828	0.0982	0.1920	0.1613	0.1963
1	400	0.0495	0.0436	0.0454	0.0511	0.1184	0.1133	0.1611
1	800	0.0253	0.0223	0.0240	0.0249	0.0811	0.0966	0.1479
1	1600	0.0123	0.0112	0.0108	0.0121	0.0523	0.0857	0.1432
1	3200	0.0059	0.0058	0.0057	0.0063	0.0361	0.0742	0.1357
10	100	0.0199	0.0190	0.0175	0.0196	0.2170	0.0941	0.1493
10	200	0.0101	0.0087	0.0093	0.0096	0.1452	0.0811	0.1350
10	400	0.0048	0.0046	0.0047	0.0053	0.1074	0.0726	0.1337
10	800	0.0025	0.0023	0.0022	0.0025	0.0748	0.0687	0.1332
10	1600	0.0012	0.0012	0.0011	0.0012	0.0502	0.0688	0.1327
10	3200	0.0006	0.0006	0.0006	0.0006	0.0364	0.0671	0.1318

〈부표 4〉 기본 히스토그램과 차등정보보호 히스토그램 사이의 거리(막대 개수가 20)

		교란 방식				재현 방식		
α	자료수	보정	양	동일	잡음	재현	평활1	평활2
0.01	100	1.3145	0.9621	0.8799	1.0825	1.3200	NA	NA
0.01	200	1.2880	0.9282	0.8778	1.0440	1.2794	NA	NA
0.01	400	1.2035	0.8732	0.8618	0.9916	1.2348	NA	NA
0.01	800	1.0949	0.8379	0.8264	0.9208	1.1482	NA	NA
0.01	1600	0.9623	0.6984	0.6797	0.8331	0.9369	NA	NA
0.01	3200	0.6682	0.5381	0.5331	0.6413	0.6864	NA	NA
0.1	100	1.0788	0.7685	0.7232	0.9541	1.1185	NA	NA
0.1	200	0.8664	0.6262	0.5915	0.7805	0.8971	NA	NA
0.1	400	0.6295	0.4729	0.4634	0.6208	0.6345	NA	NA
0.1	800	0.3860	0.3188	0.3057	0.3774	0.4085	NA	0.3426
0.1	1600	0.2126	0.1917	0.1932	0.2076	0.2286	0.2453	0.2610
0.1	3200	0.1069	0.1040	0.1087	0.1145	0.1220	0.1793	0.2080
1	100	0.3256	0.2709	0.2536	0.3318	0.4622	0.2981	0.3211
1	200	0.1882	0.1630	0.1590	0.1922	0.2884	0.2189	0.2412
1	400	0.0976	0.0881	0.0834	0.0982	0.1840	0.1605	0.2003
1	800	0.0471	0.0464	0.0453	0.0468	0.1183	0.1271	0.1735
1	1600	0.0247	0.0232	0.0237	0.0243	0.0795	0.1034	0.1611
1	3200	0.0125	0.0120	0.0116	0.0121	0.0532	0.0891	0.1534
10	100	0.0371	0.0367	0.0335	0.0389	0.3081	0.1197	0.1694
10	200	0.0194	0.0188	0.0182	0.0196	0.2194	0.1013	0.1566
10	400	0.0103	0.0097	0.0096	0.0098	0.1512	0.0859	0.1492
10	800	0.0050	0.0046	0.0047	0.0049	0.1093	0.0793	0.1471
10	1600	0.0024	0.0024	0.0024	0.0024	0.0735	0.0761	0.1455
10	3200	0.0012	0.0012	0.0012	0.0012	0.0511	0.0737	0.1438

3. 모의실험 결과($d = 2$ 인 경우)

〈부표 5〉 원자료 밀도함수와 생성된 히스토그램 사이의 거리(막대 개수가 10)

			교란 방식			재현 방식		
α	자료수	기본	보정	양	동일	재현	평활1	평활2
0.01	100	0.498594	0.871168	0.668024	0.782479	0.943973	NA	NA
0.01	200	0.348274	0.818122	0.636288	0.690447	0.842576	NA	NA
0.01	400	0.255767	0.772952	0.626551	0.643799	0.800696	NA	NA
0.01	800	0.209655	0.738596	0.605603	0.618842	0.750164	NA	NA
0.01	1600	0.187062	0.708332	0.587023	0.57688	0.703997	NA	NA
0.01	3200	0.177943	0.633799	0.542703	0.527941	0.628609	NA	NA
0.1	100	0.499426	0.856155	0.636162	0.733122	0.944525	NA	NA
0.1	200	0.3481	0.752485	0.593622	0.604562	0.803215	NA	NA
0.1	400	0.258707	0.662059	0.533881	0.516415	0.693117	NA	NA
0.1	800	0.209811	0.535256	0.465687	0.429525	0.542808	NA	0.446403
0.1	1600	0.186633	0.390756	0.36742	0.340231	0.4025	0.337296	0.33182
0.1	3200	0.178882	0.271672	0.277793	0.255142	0.275972	0.264216	0.267804
1	100	0.495444	0.618454	0.486188	0.541934	0.746983	0.611192	0.580901
1	200	0.346683	0.456088	0.378772	0.379582	0.532342	0.43015	0.416916
1	400	0.253947	0.309464	0.285974	0.278109	0.365655	0.31116	0.302907
1	800	0.209099	0.231968	0.224329	0.219336	0.274208	0.247231	0.250646
1	1600	0.186042	0.194122	0.192739	0.189882	0.214856	0.210265	0.221548
1	3200	0.176493	0.17875	0.178938	0.177989	0.190988	0.19178	0.206971
10	100	0.502365	0.479017	0.441215	0.494431	0.653436	0.473258	0.436892
10	200	0.343426	0.339369	0.323776	0.337903	0.455272	0.330478	0.315337
10	400	0.255162	0.254271	0.249598	0.25252	0.326401	0.251722	0.251121
10	800	0.209002	0.208983	0.207514	0.208113	0.251064	0.210404	0.219843
10	1600	0.185192	0.185255	0.1849	0.184956	0.209166	0.18973	0.204775
10	3200	0.176171	0.176215	0.176121	0.176114	0.187958	0.182423	0.199865

〈부표 6〉 원자료 밀도함수와 생성된 히스토그램 사이의 거리(막대 개수가 20)

			교란 방식			재현 방식		
α	자료수	기본	보정	양	동일	재현	평활1	평활2
0.01	100	0.925103	0.869978	0.663647	1.041768	1.098193	NA	NA
0.01	200	0.629342	0.816447	0.642216	0.851669	0.938369	NA	NA
0.01	400	0.430766	0.793277	0.633266	0.737797	0.851354	NA	NA
0.01	800	0.292306	0.756758	0.62095	0.680017	0.778932	NA	NA
0.01	1600	0.207342	0.732639	0.614533	0.639458	0.747097	NA	NA
0.01	3200	0.156246	0.7178	0.605731	0.599875	0.715445	NA	NA
0.1	100	0.925277	0.869064	0.66314	1.016382	1.094237	NA	NA
0.1	200	0.635306	0.813333	0.62969	0.820856	0.923273	NA	NA
0.1	400	0.421225	0.761869	0.604832	0.67051	0.818581	NA	NA
0.1	800	0.291012	0.709288	0.57431	0.585402	0.742406	NA	0.790549
0.1	1600	0.205701	0.628266	0.520946	0.491524	0.639357	0.577665	0.560686
0.1	3200	0.15406	0.490542	0.451704	0.39691	0.501958	0.401744	0.394005
1	100	0.932118	0.795713	0.606885	0.952533	1.071675	1.057344	1.018162
1	200	0.626295	0.67976	0.532938	0.678563	0.845267	0.769981	0.72741
1	400	0.426869	0.536132	0.444013	0.471173	0.652821	0.537255	0.509933
1	800	0.291995	0.387215	0.349893	0.325316	0.459863	0.374908	0.35552
1	1600	0.208154	0.254577	0.25475	0.228709	0.317174	0.263211	0.261353
1	3200	0.154811	0.176656	0.180725	0.166699	0.217472	0.19494	0.203533
10	100	0.935076	0.736747	0.617186	0.932787	1.059904	0.85722	0.779604
10	200	0.628595	0.560496	0.493375	0.624864	0.777097	0.589916	0.537206
10	400	0.421972	0.403211	0.370993	0.413122	0.558278	0.403975	0.373217
10	800	0.29167	0.286776	0.273882	0.285415	0.391641	0.281453	0.271089
10	1600	0.205743	0.204978	0.200114	0.202718	0.274684	0.204804	0.208068
10	3200	0.154569	0.154579	0.153124	0.153662	0.199477	0.16024	0.17513

〈부표 7〉 기본 히스토그램과 차등정보보호 히스토그램 사이의 거리(막대 개수가 10)

		교란 방식			재현 방식		
α	자료수	보정	양	동일	재현	평활1	평활2
0.01	100	0.915814	0.781672	0.646626	0.954124	NA	NA
0.01	200	0.823137	0.679052	0.655936	0.855732	NA	NA
0.01	400	0.776184	0.620861	0.623746	0.785162	NA	NA
0.01	800	0.735153	0.591389	0.587951	0.72801	NA	NA
0.01	1600	0.685847	0.55926	0.556644	0.682713	NA	NA
0.01	3200	0.612755	0.521211	0.496756	0.606812	NA	NA
0.1	100	0.899354	0.765506	0.59676	0.950713	NA	NA
0.1	200	0.765834	0.614499	0.56347	0.796868	NA	NA
0.1	400	0.652555	0.5206	0.488882	0.677055	NA	NA
0.1	800	0.513593	0.440819	0.396193	0.517166	NA	0.393198
0.1	1600	0.342744	0.334352	0.293847	0.354678	0.275723	0.274253
0.1	3200	0.19778	0.225571	0.199067	0.208833	0.189838	0.194883
1	100	0.564181	0.52563	0.321268	0.683636	0.389259	0.392027
1	200	0.350654	0.337438	0.245781	0.446368	0.279577	0.283772
1	400	0.193555	0.21144	0.161876	0.276926	0.196318	0.205622
1	800	0.09868	0.123297	0.096155	0.17095	0.137756	0.156171
1	1600	0.04562	0.066181	0.051644	0.101139	0.099685	0.131201
1	3200	0.021711	0.035567	0.027857	0.065891	0.077379	0.116382
10	100	0.099923	0.131021	0.058776	0.390432	0.145148	0.181847
10	200	0.04826	0.062555	0.035993	0.27365	0.106697	0.143243
10	400	0.022495	0.030121	0.020709	0.196519	0.078573	0.116832
10	800	0.010176	0.01508	0.010767	0.131647	0.064189	0.108447
10	1600	0.00466	0.007498	0.005708	0.091729	0.058034	0.10619
10	3200	0.002115	0.003797	0.002887	0.06155	0.05552	0.106635

〈부표 8〉 기본 히스토그램과 차등정보보호 히스토그램 사이의 거리(막대 개수가 20)

		교란 방식			재현 방식		
α	자료수	보정	양	동일	재현	평활1	평활2
0.01	100	1.075155	1.043252	0.610712	1.126823	NA	NA
0.01	200	0.9321	0.851476	0.635458	0.980548	NA	NA
0.01	400	0.839855	0.731962	0.662041	0.886008	NA	NA
0.01	800	0.773513	0.659955	0.665595	0.799976	NA	NA
0.01	1600	0.745826	0.629417	0.641395	0.753911	NA	NA
0.01	3200	0.709303	0.599135	0.602782	0.708908	NA	NA
0.1	100	1.053165	1.03041	0.580204	1.10515	NA	NA
0.1	200	0.918513	0.839474	0.587175	0.966022	NA	NA
0.1	400	0.809948	0.698571	0.608277	0.841627	NA	NA
0.1	800	0.718265	0.604699	0.563525	0.746983	NA	0.734606
0.1	1600	0.621403	0.530729	0.488584	0.639865	0.543698	0.513361
0.1	3200	0.482307	0.452155	0.387092	0.49277	0.376968	0.365957
1	100	0.940044	0.934082	0.380901	1.004129	0.518923	0.541468
1	200	0.70553	0.68387	0.353882	0.80031	0.44517	0.443315
1	400	0.508838	0.495054	0.304613	0.584979	0.339733	0.354931
1	800	0.316112	0.332566	0.225069	0.393609	0.256194	0.257428
1	1600	0.1749	0.213492	0.145564	0.248594	0.183928	0.195357
1	3200	0.08765	0.125381	0.084146	0.150482	0.130172	0.153353
10	100	0.330171	0.470498	0.082136	0.615392	0.205352	0.264075
10	200	0.173288	0.246051	0.069704	0.466856	0.166712	0.210557
10	400	0.08495	0.121941	0.051676	0.354455	0.130437	0.169482
10	800	0.040817	0.060471	0.032184	0.260245	0.098426	0.139629
10	1600	0.019346	0.030704	0.01795	0.179428	0.077426	0.121387
10	3200	0.009034	0.015443	0.009704	0.124839	0.06608	0.1157

집필진

- 박민정(통계청 통계개발원 통계방법연구실 사무관)
- 권성훈(건국대학교 교수)
- 정재호(통계청 통계개발원 통계방법연구실 주무관)

연구보고서 2019-20

차등정보보호 적용 실험 연구

인 쇄	2020년 5월 18일
발 행	2020년 5월 19일
발 행 인	통계개발원장 전영일
발 행 처	통계청 통계개발원
	35220 대전광역시 서구 한밭대로 713
	TEL.(042)366-7100 FAX.(042)366-7123
홈페이지	http://sri.kostat.go.kr
ISSN(Online)	

