통계개발원 연구보고서

개인정보보호와 데이터 활용을 위한 통계적 방법론(Ⅱ)

– 데이터 정보보호를 위한 최신 기법의 활용 –

[1부] 매크로데이터를 위한 통계적 비밀보호 방안

- 매크로데이터의 비밀보호방법에 대한 고찰 (2008) 정동욱 · 김경미
- 그리드별 통계서비스의 비밀보호 방안 (2018) 박민정 · 신우람 · 박천영 · 서영오

[2부] 재현자료 생성 기법 및 활용

- 재현자료 작성 방법론 검토 (2017) 박민정 · 김정연

[3부] 차등정보보호 이론과 적용

- 통계적 노출제어 최신동향 검토 – 차등정보보호를 중심으로 (2016) 박민정
- 차등정보보호에 관한 연구 (2018) 박민정 · 이용희 · 권성훈
- 차등정보보호 적용 실험 연구 (2019) 박민정 · 권성훈 · 정재호

개인정보보호와 데이터 활용을 촉진시키는 연구서의 발간에 붙이며~

데이터는 영원한 생명체와 같다. 사람은 죽지만 데이터는 남는다. 기업은 소멸해도 데이터는 존재한다. 데이터는 4차 산업혁명에 생명을 불어넣어 데이터 세계경제를 탄생시켰으며, 21세기 미래성장의 원천이 되었다. 그러나 실제로 이의 실현적 측면에서는 「데이터 활용」의 전제 조건인 데이터의 공유에서부터 현실의 벽에 부딪히고 있다. 「개인정보보호」라는 양날의 검이 앞을 가로막고 있기 때문이다.

최근 데이터 3법이 법제화되면서 길이 열렸다. 개인정보보호법·정보통신망법·신용정보법이 모두 국회를 통과하면서 데이터 이용 규제가 대폭 완화되고, 빅데이터, 공공데이터, 국가통계데이터를 적절히 연계하여 활용할 수 있는 계기가 마련되었기 때문이다. 따라서 바이오, 금융, 인공지능 산업을 망라하는 새로운 데이터경제 성장의 지평이 열린 것이다.

그러나 다른 한편으로는 인권과 연관된 개인정보에 대해 법안이 허술하게 만들어졌다는 시민사회의 비판에 직면하고 있다. 이는 데이터의 연계가 정확성을 가질수록 개인정보 노출위험이 증폭된다는 것을 의미한다. 이렇듯 데이터의 활용과 개인정보 보호라는 두 측면의 상호충돌 관계가 현재 우리가 당면한 딜레마이다.

각 나라의 통계기관은 오랫동안 이 문제를 다루어왔다. 개별 정보의 노출위험을 제어하면서도 유용한 데이터를 제공하는 것은 국가통계기관이 수행해야 할 가장 중요한 업무 중 하나이기 때문이다. 한국 통계청은 통계개발원(Statistics Research Institute)을 중심으로 안전하고도 유용하며, 사용하기 편리한 데이터를 제공하기 위한 통계적 비밀보호 방안을 꾸준히 연구해오고 있다.

이러한 연구는 어디에서 출발하여, 어떠한 발전들이 있었을까?

통계적 비밀보호 기법의 진화

통계적 비밀보호 분야의 발전 초창기에는 데이터를 적절히 변형하여 제공하는 다양한 방법이 연구되었으며, 매스킹(masking) 기법이 그 시작이었다. 다양한 매스킹 방법들의 안정성과 유용성 정도를 판단하기 위해서 개인정보 노출위험과 자료의

정보손실을 측정하고 비교할 필요성이 대두되었으며, 이에 대한 연구가 발전하기 시작했다.

매스킹 기법을 사용할 때 노출위험과 정보손실 관점에서 합리적인 결론을 도출하기 어렵다는 것이 알려졌고, 동시에 매스킹 기법을 대체하기 위한 새로운 시도들이 이루어졌다. 그 결과 베이지안 재현자료(Bayesian synthetic data)와 차등정보보호(differential privacy)가 출현하였다. 최근에는 미국 통계청을 중심으로 차등정보보호가 국가통계에서 실용화되고 있으며, 애플, 구글 등의 데이터 관련 기업에서도 이미 이러한 방법들을 활용하고 있다.

방법론의 빠른 발전과 함께 용어 사용에도 변화가 일어났다. 초창기에는 매스킹(masking)이라는 용어가 주로 사용되지만, 노출위험과 정보손실 측도 연구를 포괄하여 통계적 노출제어(statistical disclosure control/limitation) 혹은 통계적 비밀보호(statistical data confidentiality)라는 용어가 널리 사용되었다. 차등정보보호(differential privacy)의 출현 이후 최근에는 데이터 정보보호(data privacy)라는 용어가 보편적으로 사용되고 있다.

통계개발원은 국내에서 아직 생소한 통계적 비밀보호, 즉 「데이터 정보보호」의 이해와 활용도를 높이기 위하여 그간의 주요 연구를 정리하여 두 권의 책자로 제작하게 되었다. 제1권은 「마이크로데이터에 적용한 비밀보호기법: 이론과 활용」을, 제2권은 「데이터 정보보호를 위한 최신 기법의 활용」을 다루고 있다.

제1권은 매스킹에 관한 국내·외 연구를 종합하여 정리하였다. 매스킹 기법의 소개부터 실무 활용을 위한 안내 그리고 가이드라인 프로그램의 작성을 제시하였다. 하지만 2006년 이래 초창기에 연구되었던 매스킹 기법으로 데이터 정보보호를 수행하기에는 한계가 많았다. 데이터 제공에 있어 특히 안전성과 유용성을 정량적으로 보장할 수 없었으며, 이는 데이터 제공을 위한 과학적인 비밀보호 방안으로서 매스킹 기법에 한계가 있다는 것을 의미한다.

이에 따라 제2권은 매스킹 기법의 대안으로 제시된 베이지안 재현자료 및 차등정보보호에 대한 연구결과를 중심으로 구성되었다. 즉 통계개발원이 2016년부터 발전시켜 오고 있는 데이터 정보보호를 위한 최신 기법을 활용한 대용량의 빈도표 제공, 재현자료, 차등정보보호의 이론과 적용을 다루고 있다.

연구서 1권과 2권을 각각 더 자세히 살펴보면 다음과 같다.

마이크로데이터에 적용한 비밀보호 기법: 이론과 활용 (제1권)

제1권 제1부는 데이터 제공 시 개인정보보호의 문제를 인식하고, 이를 해결하기 위한 첫 발걸음을 담고 있다. 이는 2006년 통계개발원의 개원와 함께 도전한 혁신적인 연구과제이다. 제1장에서는 통계 작성을 위하여 사용하는 상세 정보인 마이크로데이터를 통계와 함께 제공할 때 국제적으로 적용하는 제도들과 통계 기법들을 정리했다. 제2장에서는 당시 국내에서 활용도가 낮았던 마이크로데이터를 국민들이 이용할 수 있도록 제공하기 위한 노력을 담고 있다. 이를 위해 전통적인 비밀보호 방법인 매스킹 기법을 중심으로 그 개념과 상세 알고리즘 등 전반적인 내용을 국내에 처음으로 소개하고 있다.

제1권 제2부는 매스킹 기법의 활용을 다루는 네 개의 장으로 구성되어 있다. 제3장은 인구주택 총조사의 마이크로데이터를 안전하게 제공하기 위하여 매스킹 기법을 어떻게 적용할지 연구한 결과이다. 제4장은 이 연구결과를 바탕으로 민감한 개인정보를 포함하고 있는 가구조사 마이크로데이터의 비밀보호를 위한 매스킹 기법의 적용 방안을 제시하였다. 제5장에는 당시 최신 기법인 승법잡음모형을 가계조사 마이크로데이터에 적용한 연구 결과가 수록되어 있다. 한편, 2012년부터 본격적으로 실시되기 시작한 가계금융복지조사의 마이크로데이터 제공을 위한 매스킹 방안이 2013년에 심층연구 되었고, 그 연구 결과는 제6장에 제시되어 있다.

특정 마이크로데이터의 비밀보호를 위하여 수십 가지의 매스킹 방안이 만들어질 수 있다. 다양한 매스킹 기법이 적용된 결과들 중에서 원래 마이크로데이터를 대신하여 공표할 결과를 어떻게 선택하고 결정할 것인가? 안전성과 유용성은 그 결정을 위한 적절한 기준이다. 통계개발원은 이를 측정하기 위하여 개발된 다양한 노출위험 측도와 정보손실 측도를 정리하였다. 제1권 제3부의 제7장은 이러한 측도에 관한 정리와 더불어 제6장의 매스킹 방안들을 평가한 결과를 보여준다. 즉 제6장을 통해서는 매스킹에 관한 세부 기법이 전체적으로 정리되었다면, 제7장에서는 각종 측도를 활용한 평가 방안이 제시되었다고 할 수 있다.

제1권 제4부에서는 기존의 연구 결과를 실무에 도입하여 활용할 수 있도록 돕는 작업을 다루고 있다. 먼저 제8장에는 통계개발원의 매스킹에 관한 연구 결과를 정리하고 R패키지를 이용하여 실제 매스킹 작업을 수행할 수 있도록 작성한 안내서를 결과로 수록하였다. 이 내용은 문서로 만들어진 매스킹 실무 활용 가이드라인에

해당한다. 그러나 자료제공 업무수행 시 통계 관련 전공자가 아닌 경우에는 이 문서를 숙지해서 R프로그램을 코딩하는 것이 어렵기 때문에 보다 손쉽게 활용할 수 있는 가이드라인이 필요했다. 이에 가이드라인으로서 문서를 대신하여 엑셀로 구동되는 프로그램을 작성하였고, 그러한 가이드라인 프로그램을 작성하는 방법을 제9장에서 제시하였다. 이를 통해 업무자들이 손쉽게 제9장의 가이드라인 프로그램을 활용하여 기초적인 수준의 노출위험 측정을 수행할 수 있도록 하였다.

주지하는 바와 같이 제1권에 소개된 매스킹 기법은 데이터 정보보호를 수행하기에는 여러 한계가 노출되었다. 그러나 이러한 한계를 오히려 기회로 발전시킬 수 있지 않을까라는 사고의 전환을 통해 이 연구는 다음 단계로 나아가게 되었다. 바로 그 응답이 제2권에 제시되어 있다.

데이터 정보보호를 위한 최신 기법의 활용 (제2권)

제2권 제1부는 마이크로데이터의 집계 결과물인 매크로데이터에 대한 비밀보호 방안을 다루고 있다. 다양한 유형의 매크로데이터에 관한 매스킹 기법은 2008년에 이미 충실히 다루어졌으며, 제1장에 그 결과가 수록되어 있다. 다만 위계구조를 가지는 대용량 빈도표의 제공을 위한 비밀보호 문제의 해법은 제시되어 있지 않았다. 통계청은 통계지리정보서비스(SGIS)를 통하여 위계구조를 가지는 빈도표를 제공하고 있고, 이를 해결하기 위한 연구가 수행되었다. 제2장은 이에 대한 새로운 알고리즘을 제시하면서 노출위험과 정보손실에 관한 논의도 함께 다루었다.

제2권 제2부는 재현자료(synthetic data)를 다루고 있다. 재현자료는 베이지안 다중대체 기법을 활용하여 원래의 마이크로데이터와 유사한 분포를 가지도록 생성한 자료를 말한다. 제3장은 재현자료 생성과 관련된 이론을 정리하고, 최근 해외에서 제안된 R패키지를 활용하여 재현자료를 생성할 수 있음을 보여주었다. 재현자료에 관한 연구들에서 노출위험을 제어한 상태로 정보손실을 줄일 수 있는 다중대체 모형의 구축에 대해 다루고 있다. 그러나 재현자료의 노출위험에 대해 논란이 있으며, 모형 구축 자체가 쉽지 않은 문제가 있다. 그럼에도 자료제공의 문제를 단계적으로 해결해 나가는데 재현자료의 활용 가치는 높다고 할 수 있으며, 지속적인 연구가 필요하다.

제2권 제3부에서는 최근 가장 주목을 받고 있는 차등정보보호를 다루고 있다. 제4장은 차등정보보호를 다른 방법들과 비교하여 소개하고 있으며, 제5장에서는

차등정보보호에 관한 개념, 이론 및 활용사례를 종합하여 정리하였다. 제6장에는 매스킹 기법, 재현자료, 차등정보보호에 관하여 노출위험과 정보손실을 중심으로 이론적 체계를 제시하였다. 더불어 실제 통계청 보도자료에 차등정보보호를 적용하면 어떠한 결과를 얻는지에 대한 실험 결과와 차등정보보호 의의 및 향후 연구방향을 제시하였다.

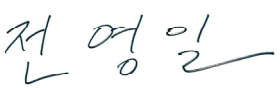
개인정보보호와 데이터 활용을 위한 핸드북

본 책자는 통계적 비밀보호, 데이터 정보보호 분야에 있어 지난 15년 가까운 기간 동안 통계개발원의 주요 연구를 집대성한 것이다. 통계개발원은 매스킹 기법, 노출위험 측도, 정보손실 측도, 대용량 빈도표의 비밀보호 등 정보보호와 관련된 실용적인 주제를 상세히 다루어 왔다. 또한 재현자료, 차등정보보호 등 혁신적인 통계적 대안을 탐구하였다. 각 연구보고서는 국가통계기관의 자료제공 관점에서 데이터 정보보호의 이론과 활용법을 제시하였다.

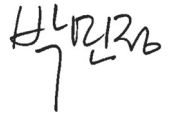
안타깝게도 국내에서 데이터 정보보호 분야의 기술적인 연구 사례를 찾기가 어려운 상황을 고려할 때 본 책자는 현 시점에서 이에 관한 거의 유일한 국문 자료일 것이다. 21세기 사회발전의 원천이 되는 「데이터 활용」과 「개인정보보호」 문제의 실타래를 풀기 위하여 '데이터 정보보호' 분야에 진입하거나 혁신적인 길을 모색하고자 하는 연구자들에게 본 책자가 도움이 되기를 바란다.

가까이 놓고 언제든지 활용할 수 있는 살아있는 핸드북이 되기를 희망한다면 지나친 기대일까? 생명체인 데이터를 다루는 모든 사람에게 유익하기를!

SRI 통계개발원장 전영일



SRI 데이터정보보호센터장 박민정



목 차

[1부] 매크로데이터를 위한 통계적 비밀보호 방안

제1장 매크로데이터의 비밀보호방법에 대한 고찰 ····· 3

제1절 서 론 ····· 3

제2절 집계표 데이터의 비밀보호방법 ····· 4

1. 집계표 자료형태 ····· 4
2. 민감한 셀 찾는 방법 ····· 6
3. 비밀보호 방법 ····· 10

제3절 산업구조통계의 비밀보호 현황 ····· 25

1. 광업 및 제조업 통계조사 작성 개요 ····· 26
2. 비밀보호 방법 ····· 28
3. 현행 SODC 방법 적용에 따른 한계점 ····· 37

제4절 실증분석 ····· 39

1. 분석 대상 ····· 39
2. 비밀보호 현황 및 한계점 ····· 40
3. 보조셀감추기(CCS)방법의 적용 ····· 41
4. 타당성 검증 ····· 47
5. 실증분석 결과 요약 ····· 50

제5절 결론 ····· 50

참고문헌 ····· 53

〈부 록〉 ····· 54

제2장 그리드별 통계서비스의 비밀보호 방안 ····· 89

제1절 서 론 ····· 90

제2절 위계 구조를 가지는 대량의 빈도표 제공 ····· 91

1. 빈도표의 위계 구조 ····· 91
2. SGIS 시스템의 자료 구조 ····· 97
3. 이용자가 생성하는 빈도표 ····· 98
4. 대용량 빈도표 제공 시스템의 고려 사항 ····· 99

제3절 빈도표를 위한 매스킹 기법 ····· 100

1. 빈도표에서 개별정보 노출 ····· 100

2. 빈도표를 위한 매스킹 기법 ····· 105
3. 작은 셀 조정에서 노출위험과 정보손실 ····· 112
제4절 개선된 작은 셀 조정(BSCA) 알고리즘 제안 ····· 114
1. 정보손실과 노출위험 관계에 대한 새로운 해석 ····· 114
2. 정보손실 축소 방안 ····· 117
3. 노출위험 제어 방안 ····· 120
4. BSCA 알고리즘 ····· 122
5. 적용 결과 ····· 124
6. BSCA R패키지 개발 ····· 128
제5절 맺음말 ····· 129

참고문헌 ····· 131
〈부 록〉 ····· 132

[2부] 재현자료 생성 기법 및 활용

제3장 재현자료 작성 방법론 검토 ····· 143
제1절 서 론 ····· 144
제2절 베이지안 통계의 이해 ····· 146
1. 통계적 추론 ····· 146
2. 베이지안 통계 ····· 148
3. 베이지안 붓스트랩 ····· 150
제3절 다중대체 ····· 152
1. 결측값과 대체 ····· 152
2. 단순대체 ····· 153
3. 다중대체 ····· 154
4. 순차회귀 다중대체 ····· 156
제4절 완전 재현자료 ····· 158
1. 아이디어 ····· 158
2. 완전 재현자료 생성 ····· 159
3. 결합규칙 ····· 165
제5절 부분 재현자료 ····· 167
1. 아이디어 ····· 167
2. 부분 재현자료 생성 ····· 168

3. 결합규칙 ····· 170
4. 부분 재현자료 생성: 결측값이 존재하는 경우 ····· 171
제6절 R패키지 활용 ····· 172
1. 재현자료 생성 방법론의 발전 ····· 173
2. R패키지 synthpop 소개 ····· 173
제7절 맺음말 ····· 177

참고문헌 ····· 179
〈부 록〉 ····· 181

[3부] 차등정보보호 이론과 적용

제4장 통계적 노출제어 최신동향 검토 - 차등정보보호를 중심으로 - ····· 213
제1절 서론 ····· 213
제2절 통계적 노출제어 기법 ····· 215
1. 노출위험과 정보손실 ····· 216
2. 매스킹 기법 ····· 219
3. 온라인 분석 서버와 매스킹 ····· 220
4. 재현자료 활용 ····· 222
제3절 차등정보보호 ····· 224
1. 정보보호의 신개념 ····· 227
2. 차등정보보호 구현 ····· 230
제4절 차등정보보호 적용 사례 ····· 234
1. 뉴욕시 택시 운전자 데이터베이스 : 승객의 개별정보보호 ····· 234
2. 뉴욕시 택시 운전자 데이터베이스 : 운전자의 개별정보보호 ····· 237
제5절 맺음말 ····· 240

참고문헌 ····· 242
〈부 록〉 ····· 243

제5장 차등정보보호에 관한 연구 ····· 252
제1절 서 론 ····· 253
제2절 개념적 이해 ····· 260
1. 빅데이터 시대와 개인정보보호 ····· 260

2. 차등정보보호의 기준과 성질 ····· 265
3. 차등정보보호의 확률적 정의 ····· 287
제3절 이론적 이해 ····· 291
1. 차등정보보호 ····· 292
2. 국소차등정보보호 ····· 304
3. 차등정보보호의 방법의 평가 ····· 310
제4절 활용 사례 ····· 315
1. 직접적인 잡음 첨가에 의한 차등정보보호 자료 생성 ····· 315
2. 통계적 모형을 이용한 차등정보보호 자료 생성 ····· 319
3. 차등정보보호를 만족하는 재현자료 생성 ····· 323
제5절 맺음말 ····· 329

참고문헌 ····· 331

제6장 차등정보보호 적용 실험 연구 ····· 334
제1절 서 론 ····· 335
제2절 차등정보보호 이해 ····· 338
1. 노출위험과 정보손실 ····· 338
2. 차등정보보호와 통계학 ····· 346
제3절 차등정보보호 히스토그램 생성 알고리즘 ····· 358
1. 원자료의 히스토그램 ····· 358
2. 차등정보보호 히스토그램: 교란 방식 ····· 365
3. 차등정보보호 히스토그램: 재현 방식 ····· 371
제4절 차등정보보호 히스토그램 구현 및 평가 ····· 373
1. 차등정보보호 히스토그램: 모의실험 ····· 373
2. 차등정보보호 히스토그램: 실제자료 적용 실험 ····· 384
3. 재현자료로 활용 가능성 ····· 407
제5절 결론 ····· 410

참고문헌 ····· 414
〈부 록〉 ····· 415

표 목차

〈표 1-1〉 집계표 예시(한국) ····· 5
〈표 1-2-1〉 집계표 예시(미국) ····· 5
〈표 1-2-2〉 집계표 예시(미국) ····· 6
〈표 1-3〉 빈도데이터의 민감한 셀 찾기(예제) ····· 6
〈표 1-4〉 (3,80) - dominance rule(예제) ····· 7
〈표 1-5〉 20-percent rule(예제) ····· 8
〈표 1-6〉 20/50-ambiguity rule(예제) ····· 9
〈표 1-7〉 자료형태에 따른 전통적인 비밀보호 기법(Cox,2008) ····· 10
〈표 1-8〉 전통적 반올림(예제) ····· 11
〈표 1-9〉 (4×5) Original Data ····· 12
〈표 1-10〉 (5×6) Original Data ····· 16
〈표 1-11〉 HCCS방법 적용 결과 ····· 17
〈표 1-12〉 (3×3)Original data ····· 19
〈표 1-13〉 (3×3) CCS방법 적용 결과 ····· 19
〈표 1-14〉 CCS방법에 따른 정보제한총량 ····· 20
〈표 1-15〉 Network에 의한 최소경로 탐색(a_{33}) ····· 22
〈표 1-16〉 Network에 의한 최소경로 탐색(a_{23}) ····· 22
〈표 1-17〉 원자료 집계 결과 ····· 24
〈표 1-18〉 보조셀감추기(CCS)방법 적용 결과 ····· 24
〈표 1-19〉 CTA방법 적용 결과 ····· 25
〈표 1-20〉 집계표 예시 ····· 26
〈표 1-21〉 광업 및 제조업 통계조사 항목 ····· 27
〈표 1-22〉 코크스,석유정제품 및 핵연료 제조업(D23) ····· 29
〈표 1-23〉 셀값찾기 Step1 ····· 33
〈표 1-24〉 셀값찾기 Step2 ····· 33
〈표 1-25〉 셀값찾기 Step3 ····· 34
〈표 1-26〉 셀값찾기 Step4 ····· 34
〈표 1-27〉 셀값찾기 Step5 ····· 34
〈표 1-28〉 선형관련식에 의한 감추어진 셀「X」값 찾기 최종 결과 ····· 35

〈표 1-29〉 감추어진 셀「X」값 찾기 (D2713) ····· 38
〈표 1-30〉 분석대상의 주요지표 ····· 39
〈표 1-31〉 소분류별 사업체수(D272) ····· 41
〈표 1-32〉 세분류별 사업체수(D2721) ····· 42
〈표 1-33〉 보조셀감추기(CCS) 결과 (소분류) ····· 43
〈표 1-34〉 보조셀감추기(CCS) 결과 (세분류) ····· 43
〈표 1-35-1〉 보조감추기셀 찾기($a_{81}=2$) ····· 45
〈표 1-35-2〉 보조감추기셀 찾기(a_{71}, a_{53}, a_{63}) ····· 45
〈표 1-36〉 보조셀감추기(CCS) 결과(세세분류) ····· 46
〈표 1-37〉 CCS와 SODC방법 적용결과 비교 (감추어진 셀의 수) ····· 48
〈표 1-38〉 CCS와 SODC방법 적용결과 비교 (감추어진 정보량) ····· 48
〈표 1-39〉 보조셀감추기(CCS) 결과 (세분류) ····· 49

〈부표 1-1〉 종사자규모별 주요지표의 보조셀감추기 방법 적용 결과 ····· 54
〈부표 1-2〉 출하액규모별 주요지표의 보조셀감추기 방법 적용 결과 ····· 72

〈표 2-1〉 대전 행정 지역 : 지역 수준별 규모 ····· 93
〈표 2-2〉 그리드 '다바' 지역 : 그리드 수준별 규모 ····· 93
〈표 2-3〉 행정지역 변수를 가지는 '인구' 자료 마이크로데이터 형식 ····· 94
〈표 2-4〉 행정지역별 '인구' 자료 빈도표 형식 (변수 4개, 각 지역단위마다 작성) ····· 94
〈표 2-5〉 행정지역별 '인구' 빈도표 형식 (변수 3개, 각 지역단위마다 작성) ····· 95
〈표 2-6〉 '인구' 자료에 대한 빈도표 유형 ····· 95
〈표 2-7〉 빈도표 유형별로 생성되는 빈도의 개수 ····· 96
〈표 2-8〉 그리드별 '가구' 자료 마이크로데이터 형식 (변수 1개) ····· 97
〈표 2-9〉 '가구' 자료에 대한 빈도표 유형 ····· 97
〈표 2-10〉 이용자가 생성하는 빈도표 : 상위 지역 수준으로 집계하는 경우 ····· 98
〈표 2-11〉 이용자가 생성하는 빈도표 : 변수의 수를 줄여 집계 ····· 99
〈표 2-12〉 빈도표에서 발생하는 노출의 종류 ····· 101
〈표 2-13〉 빈도표를 통한 개체 식별 예제 ····· 101
〈표 2-14〉 빈도표를 통한 속성 및 그룹 속성 노출 예제 ····· 102
〈표 2-15〉 빈도표를 통한 그룹 속성 노출 예제 ····· 102

〈표 2-16〉 지역 기준 변경으로 발생하는 차이에 의한 노출 예제 ····· 103
〈표 2-17〉 변수 범주 변화로 발생하는 차이에 의한 노출 예제 ····· 103
〈표 2-18〉 안전하게 제공된 빈도표 1 ····· 104
〈표 2-19〉 안전하게 제공된 빈도표 2 ····· 104
〈표 2-20〉 안전하게 제공된 빈도표 1과 빈도표 2를 연계했을 때 발생하는 노출 ····· 104
〈표 2-21〉 $B=5$를 기준으로 한 라운딩 예제 ····· 107
〈표 2-22〉 랜덤 라운딩의 원리 ($B=3$일 때) ····· 107
〈표 2-23〉 $B=5$를 기준으로 한 랜덤 라운딩 예제 ····· 108
〈표 2-24〉 $B=5$를 기준으로 한 제어된 라운딩 예제 ····· 108
〈표 2-25〉 SCA1에서 상위수준 빈도표의 집계방식 ····· 110
〈표 2-26〉 행정지역별 '인구' 자료에서 기저 빈도표 작은 셀에 SCA를 적용한 결과 ····· 110
〈표 2-27〉 각 지역 수준별 기저 빈도표들의 빈도수 분포 ····· 111
〈표 2-28〉 SCA2에서 상위수준 빈도표의 집계방식 ····· 111
〈표 2-29〉 SCA가 의도하는 정보손실 ····· 112
〈표 2-30〉 빈도표 유형별 정보손실 비율 ····· 112
〈표 2-31〉 빈도표 유형별 정보손실 범주(중간 규모 셀 대상) ····· 113
〈표 2-32〉 정보손실 절대값이 셀 크기의 50를 초과하는 비율(중간 규모 셀 대상) ····· 113
〈표 2-33〉 SCA2를 적용할 때 작은 셀 노출 예제 ····· 114
〈표 2-34〉 상위 수준 빈도값을 구성하는 기저 빈도표의 빈도값 ····· 119
〈표 2-35〉 SCA와 빈도표 위계 구조로부터 알아낼 수 있는 빈도 참값에 대한 정보 ····· 120
〈표 2-36〉 상위 수준 빈도값을 결정하기 위한 고려 사항 ····· 120
〈표 2-37〉 BSCA를 통한 상위 수준 빈도표에서의 빈도값 결정 예제 ····· 122
〈표 2-38〉 '인구' 자료 행정지역별 빈도표 정보손실 비율 ····· 124
〈표 2-39〉 '가구' 자료 그리드별 빈도표에서 정보손실 비율 ····· 125
〈표 2-40〉 BSCA 적용 사례 : '가구' 자료 7개 범주에 대한 빈도값 변화 ····· 126
〈표 2-41〉 BSCA를 적용해서 전체 빈도표를 제공하는 명령어 예제 ····· 128
〈표 2-42〉 일부 빈도표에 BSCA를 적용하는 R패키지 명령어 예제 ····· 129

〈부표 2-1〉 통계지리정보서비스 연혁 ····· 132
〈부표 2-2〉 노출제어 처리 기법 ····· 135
〈부표 2-3〉 노출제어 처리 기법 평가 기준 ····· 135

〈부표 2-4〉 빈도표의 분포에 따른 엔트로피와 노출위험 ····· 137
〈부표 2-5〉 일반적인 빈도표 ····· 138

〈표 3-1〉 이용자 유형에 따른 노출제어(비밀보호) 전략 ····· 144

〈부표 3-1〉 $\hat{\beta}_0$의 추정량 ····· 205
〈부표 3-2〉 $\hat{\beta}_2$의 추정량 ····· 206
〈부표 3-3〉 $\hat{\beta}_3$의 추정량 ····· 207
〈부표 3-4〉 $\hat{\beta}_4$의 추정량 ····· 208
〈부표 3-5〉 $\hat{\beta}_5$의 추정량 ····· 209

〈표 4-1〉 동적 부차표본추출 방식 ····· 222
〈표 4-2〉 네 종류의 통계적 노출제어 방식 ····· 226
〈표 4-3〉 절대적 정보보호 vs. 차등정보보호 ····· 228
〈표 4-4〉 차등정보보호 정의 및 의미 ····· 230

〈부표 4-1〉 공표 범위(시도 변수 V1 포함 여부)에 따른 유일한 개체수와 노출위험 ····· 244
〈부표 4-2〉 노출위험 0.1% 이하 유지를 위한 모집단의 최소크기(Bethlehem 외, 1990) ····· 246
〈부표 4-3〉 Reiter(2015)의 CPS 자료 설명, 표적의 유형, 노출위험 측정 결과 ····· 247

〈표 5-1〉 간단한 데이터 예제에서 노출위험 측정 ····· 254
〈표 5-2〉 국소 감추기를 통한 매스킹 처리 예제 ····· 255
〈표 5-3〉 용어 구분 ····· 315
〈표 5-4〉 빈도표에 대한 차등정보보호 자료 생성 사례 ····· 316
〈표 5-5〉 평균에 대한 차등정보보호 자료 생성 사례 ····· 317
〈표 5-6〉 빈도표에 대한 반복적인 차등정보보호 자료 생성 사례 ····· 318
〈표 5-7〉 전국 거주지별 목적지 로 출퇴근하는 인구수 ····· 324
〈표 5-8〉 목적지 에 대해, 거주지별 출퇴근하는 인구수 ····· 325

〈표 6-1〉 미국의 센서스 자료에 대한 개인정보보호 정책 변화 ····· 336

〈표 6-2〉 용어 정리 ····· 348
〈표 6-3〉 기본 히스토그램 ····· 362
〈표 6-4〉 교란 히스토그램(Dwork, 2006) ····· 366
〈표 6-5〉 보정된 교란 히스토그램(Wasserman과 Zhou, 2010) ····· 367
〈표 6-6〉 양(positive)으로 보정된 교란 히스토그램 ····· 368
〈표 6-7〉 동일 지지집합(support) 교란 히스토그램 ····· 369
〈표 6-8〉 교란 후 최적 구조 히스토그램 ····· 370
〈표 6-9〉 재현된 교란 히스토그램 ····· 371
〈표 6-10〉 재현된 평활 히스토그램 ····· 372
〈표 6-11〉 차등정보보호 히스토그램 생성 알고리즘 ····· 374
〈표 6-12〉 보도자료(표 4-2) 가구특성 및 가구소득 구간별 가구분포1) ····· 384
〈표 6-13〉 전체 가구주에 대한 가구소득 빈도표 ····· 385
〈표 6-14〉 연령별 빈도표 ····· 386
〈표 6-15〉 30세 미만 가구주에 대한 가구소득 빈도표 ····· 386
〈표 6-16〉 가구주 특성별 가구소득 분포 비교 ····· 393
〈표 6-17〉 원 보도자료 vs. 차등정보보호 적용 보도자료 ····· 395
〈표 6-18〉 보도자료(표 4-4) 소득 5분위 및 소득원천별 가구소득 평균 및 구성비(2017년) ··· 396
〈표 6-19〉 소득 5분위 및 소득원천별 가구소득 평균 결과 비교 ····· 398
〈표 6-20〉 소득 5분위 및 소득원천별 가구소득 평균 : 정보손실(차이의 절대값) 비교 ····· 399
〈표 6-21〉 소득 5분위 및 소득원천별 가구소득 구성비 결과 비교 ····· 401
〈표 6-22〉 소득 5분위 및 소득원천별 가구소득 구성비 : 정보손실(차이의 절대값) 비교 ····· 402
〈표 6-23〉 보도자료(표 4-7) 소득분위별 순자산 분포 ····· 403
〈표 6-24〉 소득분위별 순자산 분포 : 빈도표 ····· 403
〈표 6-25〉 소득분위별 순자산 분포 결과 비교($\alpha = 1$) ····· 404
〈표 6-26〉 소득분위별 순자산 분포 결과 비교($\alpha = 0.1$) ····· 405
〈표 6-27〉 소득분위별 순자산 분포 결과 : 정보손실(차이의 절대값) 비교 ····· 406
〈표 6-28〉 원자료의 가구소득 세부항목 마이크로데이터 ····· 407
〈표 6-29〉 차등정보보호 자료의 가구소득 세부항목 마이크로데이터 ····· 407
〈표 6-30〉 원자료의 가구소득 세부항목 상관계수 행렬 ····· 408
〈표 6-31〉 차등정보보호 자료의 가구소득 세부항목 상관계수 행렬 ····· 408
〈표 6-32〉 가구소득 세부항목 평균 ····· 409

〈표 6-33〉 가구소득 세부항목 중 근로소득과 재산소득의 상관계수(원자료: 0.1931) …… 409
〈표 6-34〉 데이터 정보보호 방안 분류 …… 412

〈부표 6-1〉 원자료 밀도함수와 생성된 히스토그램 사이의 거리(막대 개수가 10) …… 417
〈부표 6-2〉 원자료 밀도함수와 생성된 히스토그램 사이의 거리(막대 개수가 20) …… 418
〈부표 6-3〉 기본 히스토그램과 차등정보보호 히스토그램 사이의 거리(막대 개수가 10) …… 419
〈부표 6-4〉 기본 히스토그램과 차등정보보호 히스토그램 사이의 거리(막대 개수가 20) …… 420
〈부표 6-5〉 원자료 밀도함수와 생성된 히스토그램 사이의 거리(막대 개수가 10) …… 421
〈부표 6-6〉 원자료 밀도함수와 생성된 히스토그램 사이의 거리(막대 개수가 20) …… 422
〈부표 6-7〉 기본 히스토그램과 차등정보보호 히스토그램 사이의 거리(막대 개수가 10) …… 423
〈부표 6-8〉 기본 히스토그램과 차등정보보호 히스토그램 사이의 거리(막대 개수가 20) …… 424

그림 목차

[그림 1-1] 〈표 1-12〉에 대한 network table ····· 21
[그림 1-2] 코크스,석유정제품 및 핵연료 제조업 분류체계 ····· 32
[그림 1-3] 제1차 금속산업 분류체계(중분류 27) ····· 40

[그림 2-1] 국가지점번호 부여 방법 ····· 92
[그림 2-2] 이상적인 노출위험-정보손실 관계 ····· 116
[그림 2-3] SCA2를 적용하는 경우 노출위험-정보손실 관계 ····· 117
[그림 2-4] $B=3$일 때, 고정 그리드 규칙(Fixed Grid Rule) ····· 118
[그림 2-5] $B=5$일 때 고정 그리드 규칙(Fixed Grid Rule) ····· 118
[그림 2-6] 노출제어를 위한 빈도값 이동 ····· 121
[그림 2-7] BSCA를 적용하는 경우 노출위험-정보손실 관계 ····· 128

[그림 3-1] 결측값과 다중대체(Rubin, 1987) ····· 155
[그림 3-2] 자료구조 ····· 159
[그림 3-3] 완전 재현자료 생성을 이해하기 위한 자료구조 ····· 160
[그림 3-4] 다변량 정규분포를 이용한 재현자료 생성 ····· 163
[그림 3-5] 원 자료와 재현자료를 이용해 얻은 통계량 비교 실험 ····· 164
[그림 3-6] 베이지안 붓스트랩을 기반으로 한 재현자료 생성 ····· 166
[그림 3-7] 부분 재현자료 생성을 이해하기 위한 자료구조 ····· 169
[그림 3-8] 부분 재현자료 생성 - 결측값이 존재하는 경우 ····· 171

[부그림 3-1] 베타분포 의 확률밀도함수 ····· 187

[그림 4-1] 노출위험-정보손실 지도(박민정, 2014) ····· 218
[그림 4-2] 온라인 분석 시스템의 일반적인 구조(Krenzke 외, 2013) ····· 221
[그림 4-3] 비상호적 vs. 상호적 자료제공 방식 ····· 225
[그림 4-4] 분석도구가 총합일 때 차등정보보호 구현(ϵ=0.1, 1회 실험) ····· 232
[그림 4-5] 분석도구가 총합일 때 차등정보보호 구현(ϵ=0.1, 반복 실험) ····· 233
[그림 4-6] 분석도구가 총합일 때 차등정보보호 구현(ϵ=2.0, 반복 실험) ····· 233
[그림 4-7] A씨의 이동 경로 ····· 235
[그림 4-8] 차등정보보호를 적용하여 제공되는 A씨의 이동 경로 관련 정보(ϵ=0.5) ····· 236

[그림 4-9] 차등정보보호를 적용하여 제공되는 A씨의 이동 경로 관련 정보(ϵ=4.02) ····· 236
[그림 4-10] 차등정보보호를 적용하여 제공되는 택시 운전자 1인의 평균 속도(ϵ=0.5) ····· 237
[그림 4-11] 차등정보보호를 적용하여 제공되는 전체 택시 운전자들의 평균 속도(ϵ=0.5) ····· 238
[그림 4-12] 차등정보보호를 적용하여 제공되는 택시 운전자들의 평균 속도(ϵ=4.28) ····· 238
[그림 4-13] 차등정보보호를 적용하여 제공되는 택시 운전자들의 평균 속도(ϵ=0.5) ····· 239
[그림 4-14] 차등정보보호를 적용하여 제공되는 택시 운전자들의 평균 속도(ϵ=4.08) ····· 240

[그림 5-1] 위험-유용성 지도 위에 표현된 다양한 매스킹 시나리오 ····· 256
[그림 5-2] 최근 제안된 위험-유용성 지도 사례 ····· 257
[그림 5-3] 한 개인의 참여여부에 따른 두 개의 시나리오와 차등정보보호 ····· 269
[그림 5-4] 차등정보보호에 의한 노출 위험 제어 ····· 280
[그림 5-5] 국소적 차등정보보호와 광역적/전역적 차등정보보호 ····· 285
[그림 5-6] 이중지수분포의 확률밀도함수 ····· 290
[그림 5-7] 정보보호 수준에 따른 검정 오차 ····· 295
[그림 5-8] 이중지수(Double exponential or Laplace)분포의 밀도함수 ····· 297
[그림 5-9] 라플라스 잡음을 추가해 차등정보보호 자료를 생성하는 과정 ····· 298
[그림 5-10] 국소차등정보보호와 차등정보보호의 비교 ····· 304
[그림 5-11] Warner(1965) 방법을 따른 비율 추정 예제 ····· 306
[그림 5-12] 국소차등정보보호 자료 생성 : 순차적 교호 방법 vs. 순차적 비교호 방법 ····· 307
[그림 5-13] 차등정보보호를 만족하는 자료제공 사례 ····· 320
[그림 5-14] 평균 출퇴근 거리에 대한 원자료(네모)와 차등정보보호 자료(별) ····· 327
[그림 5-15] OnTheMap의 통계적 비밀보호 모형 ····· 328

[그림 6-1] 위험-유용성 지도(노출위험-자료유용성) ····· 339
[그림 6-2] 위험-유용성 지도(정확성-노출위험) ····· 340
[그림 6-3] 위험-유용성 지도(노출위험-정보손실) ····· 340
[그림 6-4] 매스킹 방안에 대한 위험-유용성 지도 예제 ····· 342
[그림 6-5] 이상적인 완전 재현자료의 위험-유용성 지도 ····· 343
[그림 6-6] 재현자료의 위험-유용성 지도 ····· 344
[그림 6-7] 차등정보보호 자료의 위험-유용성 지도 ····· 345
[그림 6-8] 통계학을 바라보는 관점의 차이 ····· 347
[그림 6-9] 차등정보보호 수준에 따른 검정 오차 ····· 351
[그림 6-10] 차등정보보호 자료의 위험-유용성 지도 ····· 354

[그림 6-11] 빈도수를 이용해 작성한 히스토그램: 막대 개수에 따른 변화 ····· 359
[그림 6-12] 비율을 이용해 작성한 히스토그램: 막대 개수에 따른 변화 ····· 360
[그림 6-13] 막대의 개수에 따른 히스토그램의 정보손실 ····· 361
[그림 6-14] 기본 히스토그램 및 최적 구조 히스토그램 ····· 364
[그림 6-15] 원자료 밀도함수와 차등정보보호 히스토그램 사이의 거리 ····· 375
[그림 6-16] 기본(원자료) 히스토그램과 차등정보보호 히스토그램 사이의 거리 ····· 376
[그림 6-17] 막대 개수에 따른 원자료 밀도함수와 차등정보보호 히스토그램 사이의 거리 ····· 378
[그림 6-18] 막대 개수에 따른 기본(원자료) 히스토그램과 차등정보보호 히스토그램 사이의 거리 ··· 379
[그림 6-19] 원자료 밀도함수와 차등정보보호 히스토그램 사이의 거리($d = 2$인 경우) ····· 380
[그림 6-20] 기본(원자료) 히스토그램과 차등정보보호 히스토그램 사이의 거리($d = 2$인 경우) ·· 381
[그림 6-21] 막대 개수에 따른 원자료 밀도함수와 차등정보보호 히스토그램 사이의 거리
($d = 2$인 경우) ····· 382
[그림 6-22] 막대 개수에 따른 기본(원자료) 히스토그램과 차등정보보호 히스토그램 사이의 거리
($d = 2$인 경우) ····· 383
[그림 6-23] 가구소득 분포(전체 가구주) ····· 387
[그림 6-24] 가구소득 분포(30세 미만 가구주) ····· 388
[그림 6-25] 가구소득 분포(30~39세 가구주) ····· 389
[그림 6-26] 가구소득 분포(40~49세 가구주) ····· 390
[그림 6-27] 가구소득 분포(50~59세 가구주) ····· 391
[그림 6-28] 가구소득 분포(60세 이상 가구주) ····· 392
[그림 6-29] 소득원천별 가구소득 구성비의 변화 ····· 400

1부

매크로데이터를 위한 통계적 비밀보호 방안

1. 매크로데이터의 비밀보호방법에 대한 고찰
정동욱 · 김경미

2. 그리드별 통계서비스의 비밀보호 방안
박민정 · 신우람 · 박천영 · 서영오

제1장

매크로데이터의 비밀보호방법에 대한 고찰

정동욱 · 김경미

제1절 서 론

우리나라 산업구조는 대내외 환경변화에 따라 전반적으로 다양한 측면에서의 구조적인 변화기에 진입하고 있다. 통계수요 측면에서 보면, 종전의 개별 업종 중심의 접근(Sectoral Approach)에서 경제주체 간의 협력과 보완관계를 중시하는 클러스터 접근(Cluster Approach)이 요구되는 구조적인 전환기를 맞고 있으며, 소득수준 향상으로 교육, 의료, 지식기반서비스업 등이 국민경제의 새로운 동력으로 떠오르고 서비스업에 대한 관심이 커지기 시작하면서 새로운 통계 수요를 창출하고 있다.

자료제공 측면에서 보면, 통계법 개정(2007.10.)과 아울러 통계이용자들의 자료요구에 대응하기 위해 응답자의 개별정보 노출을 제한하면서 활용성을 증대시킬 방안모색의 필요성이 대두되고 있다. 그러나 자료수집 시 개인 및 사업체의 절대적 비밀보호라는 서약을 따르고 있지만, 각 종 정보의 DB화로 통계적 자료결합이나 분석과정에서 개인 또는 사업체의 비밀보호 자료가 공개될 위험은 더욱 커져가고 있다. 따라서 세분화된 정보까지 요구하는 이용자와 응답자의 개별정보 비밀보호라는 제약을 동시에 만족하기 위한 방법에 대한 연구의 필요성이 대두되고 있는 시기이다. 이와 같은 현안을 해결하기 위해 각국은 여러 방법을 모색하고 있는데, 미국, 캐나다, 호주 등에서는 노출위험이 있는 자료에 대해 다양한 방법론을 적용하여 비밀보호된 자료를 제공하고 있으며, UNECE/Eurostat에서는 “통계자료비밀보호 국제공동회의(Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality)”를 개최하여 각국 의 연구결과 및 적용사례에 대한 토의를 활발하게 진행하고 있다.

✍ 이 연구는 2008년 하반기 통계개발원 연구보고서에 수록된 자료입니다.

통계청 통계개발원에서도 통계자료의 비밀보호에 대한 문제를 인식하고 국제공동연구사업의 일환으로 2006년부터 3개년간 「통계적 비밀보호기법(Masking)」의 연구를 추진하고 있는데, '06～'07년 기간 중에는 마이크로데이터(이산형데이터 및 연속형데이터)의 비밀보호 기법연구를 수행하였다. 이어 실시되는 본 연구는 미국 국립보건통계센터(National Center for Health Statistics :NCHS)에서 동 분야에 대한 선행연구를 수행한 Dr. Lawerence Henry COX의 연구를 토대로 매크로데이터(빈도데이터 및 크기데이터)에 관한 비밀보호 방법에 관한 연구이다.

본 연구에서 다루는 내용은 자료제공 형태 중 집계표(매크로데이터) 자료에 대한 비밀보호방법 연구로, 공개범위 제한 및 측정방법과 집계표에서 그 결과의 공개위험 정도를 평가하는 방법 등 노출위험이 있는 민감한 셀의 정보를 보호하여 응답기관 또는 응답자의 개별정보 노출을 제한하는 방법을 연구하는 것이다. 제2절에서는 집계표를 구성하는 자료형태인 빈도데이터(Count Data)와 크기데이터(Magnitude Data)에 따른 전통적인 비밀보호 기법에 대한 방법론과 예제를 소개하였다. 이는 집계표 데이터의 비밀보호방법론에 관해 전반적으로 고찰하였을 뿐만 아니라 통계청에서 적용 가능한 방법을 찾기 위한 기초자료로 활용될 것이다. 제3절에서는 「광업 및 제조업통계조사」의 집계표를 자세히 살펴봄으로써 산업구조통계의 비밀보호 현황과 한계점을 살펴본다. 특히, 노출위험이 있는 셀만 감추는 현행 비밀보호 기법(Suppressing Only the Disclosure Cells: SODC)의 한계점에 대해 선형관련식을 통해 비밀보호의 정도를 재조명해 본다. 이를 통해 연구된 한계점은 실증분석을 통한 새로운 기법 적용의 필요성과 결과를 도출하기 위한 근거로 활용한다. 제4절에서는 광업 및 제조업통계조사 23개 중분류 중 매출액, 자산비중 등을 고려하여 적합한 중분류를 선정한 후, 기존의 SODC방식과 새로 적용해본 기법인 보조셀감추기(Complementary Cell Suppression: CCS)방법의 적용 결과를 비교하여 현 상황에서 적합한 비밀보호 기법을 제시하고자 한다.

제2절 집계표 데이터의 비밀보호방법

1. 집계표 자료형태

집계표를 구성하고 있는 자료는 크게 ①빈도데이터(count data) ②크기데이터(magnitude data)로 구분할 수 있다. 「빈도데이터」는 사업체수와 같이 개수 자체에 의미가 있는 자료형태이며, 「크기데이터」는 매출액, 종사자수 등과 같이 규모에 의미가 있는 형태이다. 자료형태에 따라 노출위험이 있는 셀을 결정하는 방법에 차이가 있다. 따라서

적용하는 비밀보호 방법 또한 상이하기 때문에 집계표를 어떻게 설계하는가는 중요한 과정이다.(Cox, 2008)

한국과 미국은 집계표 구성에서 다소 차이가 있는데 제조업통계조사[1]를 통해 아래와 같이 비교해 보고자 한다. 한국의 집계표 형태는 <표 1-1>과 같이 산업분류, 지역(시도), 종사자규모, 매출액규모 등을 조합하여 표측(Row)항목을 구성하고, 조사항목중 사업체수, 종사자수, 출하액, 유형자산 등 주요항목을 표두(Column)에 배치하고 있다. 즉, 빈도데이터와 크기데이터가 혼합되어 있으며, 비밀보호를 위한 기준변수(Key variable)로서 사업체수가 이용되고 있다.

〈표 1-1〉 집계표 예시(한국)

산업분류	종사자규모	사업체수	월평균 종사자수	급여액	출하액	…	유형자산 연말잔액
D19	5～9명	1,181	7,212	84,543	565,117	…	110,002
가죽, 가방	10～19명	435	5,696	90,273	619,753	…	124,855
및	:	:	:	:	:	:	:
신발 제조업	300～499명	2	X	X	X	X	X
	500명이상	3	1,964	56,025	324,478	…	227,235

반면 미국의 제조업통계조사의 경우는 빈도데이터와 크기데이터의 집계표가 구분되어 각각 제공되고 있으며, 노출위험 여부를 결정하는데 상호 영향을 미치지 않도록 설계되어 있다. 이와 같은 집계표 구성형태 의 차이로 인해 비밀보호 방법을 적용함에 있어서도 각각에 적합한 방법이 다를 것이다. 이는 제3절에서 다시 언급하기로 한다.

〈표 1-2-1〉 집계표 예시(미국)

Table 1. **Statistics for All Manufacturing by State: 2005**

[Data based on the 2005 Annual Survey of Manufactures (ASM). For information on confidentiality protection, sampling error, nonsampling error, and explanation of terms, see Note at end of table. For meaning of abbreviations and symbols, see introductory text. For explanation of terms, see Appendix A]

Geographic area	All employees		Production workers			Value added ($1,000)	Total cost of materials[2] ($1,000)	Total value of shipments[2] ($1,000)	Relative standard error of estimate[3] (percent) for column—	
	Number[1]	Payroll ($1,000)	Number[1]	Hours (1,000)	Wages ($1,000)					
	A	B	C	D	E	F	G	H	A	F
United States	**13 168 822**	**579 890 961**	**9 230 151**	**19 069 641**	**337 490 264**	**2 204 094 956**	**2 555 491 508**	**4 735 383 666**	**1**	**1**
Alabama	264 159	10 016 239	205 496	424 471	6 879 496	38 372 840	49 984 530	87 840 548	1	1
Alaska	10 810	383 553	9 079	19 895	285 130	1 805 427	4 834 957	6 570 763	1	2
Arizona	158 995	7 604 132	99 610	206 893	3 742 100	23 736 129	20 091 994	43 233 559	2	1
Arkansas	185 669	6 325 492	149 268	304 055	4 504 044	26 981 983	31 429 409	58 188 229	1	1
California	1 397 881	65 922 850	881 594	1 809 756	30 786 986	217 546 032	218 653 094	434 238 357	1	1
Colorado	128 468	6 031 574	85 813	180 366	3 192 860	19 462 751	18 075 862	37 419 999	2	2

1) 제조업통계조사 : 한국-광업 및 제조업통계조사, 미국-Annual Survey of Manufacture.

〈표 1-2-2〉 집계표 예시(미국)

Table 2. **Statistics for Industry Groups and Industries: 2005 and Earlier Years**

[Data based on the 2005 Annual Survey of Manufactures (ASM). For information on confidentiality protection, sampling error, nonsampling error, and explanation of terms, see Note 2 at end of table. For meaning of abbreviations and symbols, see introductory text. For explanation of terms, see Appendix A]

NAICS and NAICS-based code[1]	Industry group and industry		All employees		Production workers			Value added ($1,000)	Total cost of materials ($1,000)	Total value of shipments[3] ($1,000)	Relative standard error of estimate[4] (percent) for column—	
			Number[2]	Payroll ($1,000)	Number[2]	Hours (1,000)	Wages ($1,000)					
			A	B	C	D	E	F	G	H	A	F
31-33	**Manufacturing**	2005	13 168 822	579 890 961	9 230 151	19 069 641	337 490 264	2 204 094 956	2 555 491 508	4 735 383 666	1	1
		2004	13 394 079	569 703 575	9 365 130	19 283 817	332 873 474	2 041 433 991	2 283 143 958	4 308 970 620	1	1
		2003	13 872 958	567 602 408	9 796 581	19 853 892	330 480 113	1 923 414 910	2 095 279 074	4 015 387 243	1	1
		2002	14 664 385	575 165 127	10 319 528	20 431 721	336 540 063	1 889 290 940	2 022 158 348	3 914 719 163	X	X

2. 민감한 셀 찾는 방법

민감한 셀을 찾는 방법은 자료형태에 따라 다소 차이가 있는데, 다음에서는 빈도데이터와 크기데이터로 구분하여 민감한 셀을 찾는 방법을 살펴보고자 한다.

가. 빈도데이터(Count Data)

빈도데이터(Count Data)는 기준(Base) B값을 정의함으로써 노출위험이 있는 셀을 결정한다. 즉, 임의의 셀 값이 B 이하인 경우 해당 셀을 PDC(Primary Disclosure Cell)로 정의하고 해당 셀의 자료제공을 제한한다. <표 1-3>은 기준 B가 2인 경우 노출위험이 있는 셀을 표시한 것이다.

〈표 1-3〉 빈도데이터의 민감한 셀 찾기(예제)

산업분류 / 종사자규모	D27211	D27212	D27213	D27219
5 ~ 9	9	40	9	14
⋮	⋮	⋮	⋮	⋮
300 ~ 499	0	1	1	0
500명 이상	1	0	1	0

나. 크기데이터(Magnitude Data)

크기데이터(Magnitude data)는 ① (n,k) - dominance rule ② p-percent rule ③ p/q ambiguity rule 등을 적용하여 해당 셀의 노출위험여부를 결정한다. 각각의 방법에 대해 자세히 살펴보면 다음과 같다.

1) (n,k) - dominance rule

이 방법은 임의의 셀에 대한 노출위험 여부를 결정할 때, 해당 셀을 구성하고 있는 각각(사업체별)의 값이 셀 합계의 K% 이상을 차지하는 경우 노출위험이 있는 민감한 셀로 정의하는 방법이다.

민감한 셀을 탐색하는 과정을 보여주기 위해, 우선 임의의 셀을 구성하는 L개의 사업체별 자료를 식(1.1)과 같이 내림차순으로 순차적으로 정렬하고, 셀 합계와 상위 n개 자료의 합을 아래 식(1.2), (1.3)과 같이 정의하자.

$$x_1 \geq x_2 \geq \cdots \geq x_n \geq x_{n+1} \geq \cdots \geq x_l \tag{1.1}$$

$$V(x) = \sum_{i=1}^{l} x_i \tag{1.2}$$

$$D(x_n) = \sum_{j=1}^{n} x_j \tag{1.3}$$

이 때 식(1.4)를 만족하면 임의의 셀은 노출위험이 있는 셀로 정의되어, 해당 셀에 대한 자료제공을 제한한다. 이와 관련한 선형민감도를 표현하면 식(1.5)와 같다.

$$D(x_n) > (\frac{k}{100})\, V(x) \tag{1.4}$$

$$S_{n,k}(X) = \sum_{i=1}^{n} X_i - (X_{(n+1)^+} \cdot (\frac{k}{100-k})) > 0 \tag{1.5}$$

〈표 1-4〉 (3,80) - dominance rule(예제)

x_1	x_2	x_3	x_4	x_5	$V(x)$
100	50	40	30	20	240

<표 1-4>의 예제를 통해 자세히 살펴보자. $n=3,\ k=80$으로 정의하면, 위의 (1.2) (1.3)식에 의해 $D(x_3)=190,\ V(x)=240$이다. 민감한 셀인지 여부를 판단하기 위해 (1.4)식을 이용해 계산해보면, 아래와 같이 $D(x_3)$의 값이 $V(x)$의 k%값 보다 작으므로 이 셀은 노출위험이 있는 셀이 아님을 알 수 있다.

$$D(x_3) = 190 < (\frac{k}{100})\, V(x) = (\frac{80}{100}) \times 240 = 192$$

2) p-percent rule

이 방법은 자료에서 두 번째로 큰 값에 해당되는 데이터를 이용해 제일 큰 값에 해당하는 데이터를 얼마나 추정할 수 있는가 하는 정도로 노출위험을 판단하는 방법이다.

민감한 셀을 탐색하는 과정을 살펴보자. 임의의 셀을 구성하는 L개의 사업체별 자료를 내림차순으로 순차적으로 정렬한 결과를 아래의 식(1.6)과 같이 정의하자.

$$x_1 \geq x_2 \geq \ldots \geq x_n \geq x_{n+1} \geq \ldots \geq x_l \tag{1.6}$$

이때 x_2값을 이용하여 x_1값을 p% 이내에서 추정할 수 있다면 이 셀을 노출위험이 있는 민감한 셀로 판단한다. 즉, 아래의 식(1.7)을 만족하면 이 셀을 노출위험이 있는 셀로 정의하며, 선형민감도는 식 (1.8)과 같이 표현할 수 있다.

$$V(x) - x_2 < (100+p)/100 \times x_1 \tag{1.7}$$

$$S_{p\%}(X) = x_1 - (\frac{100}{p}) \times x_{3+} > 0 \tag{1.8}$$

〈표 1-5〉 20-percent rule(예제)

x_1	x_2	x_3	x_4	x_5	$V(x)$
100	50	40	30	20	240

<표 1-5>의 예제를 통해 $p=20$인 20-percent rule을 적용해 보자. 노출위험 여부를 알아보기 위해, 식(1.7)을 이용해 계산해 보면 아래와 같이 (1.7)의 부등식을 만족하지 않기 때문에 이 경우 노출위험이 발생하지 않는다고 할 수 있다.

$$240 - 50 > ((100+20)/100) \times 100$$

3) p/q-ambiguity rule

이 방법은 p-percent rule을 보완한 방법이다. 먼저 앞서 언급한 p% rule의 수식(1.7)을 아래와 같이 다시 정리해 보자.

$$V(x) - x_2 < (1 + \frac{p}{100}) \times x_1$$

$$\Leftrightarrow V(x) - x_2 - x_1 < (\frac{p}{100}) \times x_1$$
$$\Leftrightarrow x_{3+} < (\frac{p}{100}) \times x_1$$

$$\Leftrightarrow (\frac{p}{100}) \times x_1 - x_{3+} > 0$$

$$\Leftrightarrow x_1 - (\frac{100}{p}) \times x_{3+} > 0 \quad (1.8)$$

만약, 두 번째로 큰 값(x_2)를 알고 있다면 식(1.8)은 다음의 식과 같이 유도될 수 있다.

$$x_{3+} > L = (\frac{q}{100}) \times x_{3+} \quad (1.9)$$

(1.9)의 식을 (1.8)의 식에 대입하면, 식(1.10)을 도출할 수 있다.

$$x_1 - (\frac{q}{p}) \times x_{3+} > 0 \quad (1.10)$$

즉, p/q-ambiguity rule은 p% rule을 좀 더 강화한 것으로서 비밀보호된 자료 중 원시자료의 일부 개별정보를 알 수 있는 경우 p/q-ambiguity rule을 이용하여 노출위험이 있는 셀을 찾는다.

이 방법에서 선형민감도는 식(1.11)과 같이 표현할 수 있으며 q=100인 경우는 p% rule의 선형민감도와 동일하게 된다.

$$S_{pq}(X) = x_1 - (\frac{q}{p}) \times x_{3+} > 0 \quad (1.11)$$

〈표 1-6〉 20/50-ambiguity rule(예제)

x_1	x_2	x_3	x_4	x_5	$V(x)$
100	50	40	30	20	240

<표 1-6>의 예제를 통해 살펴보면, $V(x) = 240,\ \ p = 20,\ \ q = 50$ 조건에서 식(1.10)을 이용해 계산하면 아래와 같이 (1.10)식과 같은 부등식을 만족하지 않아 노출위험이 발생하지 않음을 알 수 있다.

$$100 - (\frac{50}{20}) \times 90 \ = \ -125 < 0$$

이와 같이 민감한 셀을 찾는 방법에 대해 자료형태별로 알아보았다. 민감한 셀이란 노출위험이 있는 셀로서, 집계된 자료를 통해 개인 또는 사업체의 개별정보를 재계산할 수 있는 경우이다. 이렇게 선정된 민감한 셀을 비밀보호하는 방법 역시 빈도데이터(Count data)인지 크기데이터(Magnitude data)인지에 따라 전통적으로 적용하는 방법이 있는데, 다음 장에서는 데이터 형태에 따라 적용될 수 있는 방법과 각각의 방법에 대한 자세한 내용을 소개코자 한다.

3. 비밀보호 방법

비밀보호 방법은 자료유형에 따라 적용 가능한 기법이 서로 다르다. 빈도데이터(Count data)의 경우, 작은 값을 민감한 셀로 정의하여 비밀보호가 이루어지기 때문에 ①반올림(Rounding) ②셀감추기(Cell suppression) ③자료변조(data perturbation) 또는 ④분류구조변경(data collapsing) 등과 같은 방법이 주로 적용된다.

반면 크기데이터(Magnitude data)의 경우, 셀 값을 결정하는 데 영향력이 큰 값을 갖고 있는 개별정보의 제공 여부에 따라 민감한 셀이 결정되기 때문에 주로 셀감추기(Cell suppression)방법이 적용된다.

〈표 1-7〉 자료형태에 따른 전통적인 비밀보호 기법(Cox,2008)

빈도데이터	크기데이터
반올림(Rounding) 셀 감추기(Cell suppression) 자료변조(data perturbation) 분류구조변경(data collapsing) 기타(swapping, switching 등)	셀 감추기(Cell suppression)

본 장에서는 비밀보호기법 중 주로 적용되고 있는 반올림, 셀감추기, CTA방법을 중심으로 적용방법을 상세히 소개하고자 한다.

가. 반올림(Rounding)방법

1) 전통적 반올림방법(Conventional Rounding)

〈표 1-8〉 전통적 반올림(예제)

2	6	3	7	6	7	31
6	7	6	5	7	1	32
3	6	5	7	6	7	34
6	7	6	6	7	6	38
1	6	8	2	6	5	28
18	32	28	27	32	26	163

⇒

0	5	5	5	5	5	25
5	5	5	5	5	0	25
5	5	5	5	5	5	30
5	5	5	5	5	5	30
0	5	5	0	5	5	20
15	25	25	20	25	20	130

기준(Base) = 5

① if [i, j] ≥ 5 then [i, j] = 5

② if [i, j] ≤ 2 then [i, j] = 0

③ if [i, j] ≥ 3 and [i, j] < 5 then [i, j] = 5

전통적인 반올림방법은 흔히 알고 있는 일반적인 반올림방법을 의미한다. 즉, 기준(Base: B)을 10으로 할 경우 0 이상 10 미만은 0으로 변환하고 10 이상은 10으로 변환하는 방법이다. B를 5로 할 경우는 위 <표 1-8>의 예제와 같이 변환될 수 있다.

일반적인 수식으로 표현하면 아래와 같이 나타낼 수 있다.

$$r = \begin{cases} B & , r \geq \frac{B+1}{2} \ (B : odd) \text{ or } r \geq \frac{B}{2} \ (B : even) \\ 0 & , otherwise \end{cases}$$

2) 제어된 반올림방법(Controlled Rounding)

제어된 반올림방법은 모든 셀이 기준(Base: B)의 배수가 되도록 더하기(+), 빼기(−)를 반복하는 방법이다. 즉, B가 5인 경우 모든 셀 [i, j]가 5의 배수가 되도록 더하고 빼기를 반복한다. 이 방법은 특정한 고리를 가지고 같은 수를 더하고 뺌으로써 주변합을 바꾸지 않으면서 자료를 변환하는 방법이다.

〈표 1-9〉 (4×5) Original Data

	Col1	Col2	Col3	Col4	Col5	Row Sums
Row1	37	3	30	6	4	80
Row2	1	16	23	5	15	60
Row3	30	15	8	27	10	90
Row4	7	1	4	7	21	40
Col Sums	75	35	65	45	50	270

위 <표 1-9>의 예제를 가지고 기준이 5인 제어된 반올림방법을 적용하는 과정을 자세히 살펴보자. 먼저, 표에서 5의 배수가 아닌 셀 중 하나를 고른다. 본 예제에서는 첫 번째 열, 두 번째 행의 셀, 즉 셀(1,2)를 선택하기로 한다. 그 다음으로, 첫 번째 선택한 셀에서 시작하여 행과 열이 이어지도록 6개의 셀을 골라 교체고리(alternating cycle)를 만든다.이 때 선택되는 셀은 마찬가지로 5의 배수가 아닌 셀 중에 고른다. 본 예제에서는 셀(1,2) 셀(1,4) 셀(3,4) 셀(3,3) 셀(4,3) 셀(4,2)로 정하였다. 이는 아래 [제어된 반올림방법의 변환과정]의 표 R1과 같다.

이제 정해진 교체고리에 어떤 수 d를 더하고 빼는데, 두 가지 경우를 생각할 수 있다. 첫 번째 셀을 더하기로 시작하여 더하기 빼기를 반복하는 경우와 빼기로 시작하여 이를 반복하는 경우이다. 첫 번째 셀을 더하기로 시작하는 경우 셀(1,2)에는 d를 더하고, 셀(1,4)에는 d를 빼고, 셀(3,4)에는 d를 더하고, 셀(3,3)에는 d를 빼고, 셀(4,3)에 d를 더하고, 셀(4,2)에 d를 뺀다. 이를 양의 교체고리라고 하자. 반대로 음의 교체고리 즉, 빼기로 시작하는 경우는 셀(1,2) 셀(3,4) 셀(4,3)이 빼는 셀이 되고 셀(1,4) 셀 (3,3) 셀 (4,2)가 더하는 셀이 된다. 아래의 표에서는 더하는 셀은 격자무늬, 빼는 셀은 빗금무늬로 구분하였다. 교체고리를 만든 후 결정해야 할 것은 더하거나 빼질 수 d이다. 양의 교체고리에서 더하거나 빼질 수 d+ 는 더하는 셀인 셀(1,2) 셀(3,4) 셀(4,3)과 빼는 셀인 셀(1,4) 셀(3,3) 셀 (4,2)을 5의 배수로 만들기까지 최대로 더하거나 뺄 수 있는 수이다. 음의 교체고리에서의 d - 는 더하는 셀인 셀(1,4) 셀(3,3) 셀 (4,2)과 빼는 셀인 셀(1,2) 셀(3,4) 셀(4,3)을 5의 배수로 만들기까지 최대로 더하거나 뺄 수 있는 수이다. 즉, 예제에서 d+는 1이고 d - 는 2가 된다.

d를 정한 후 마지막으로 결정해야할 사항은 양의 교체고리를 이용할 것인지 음의 교체고리를 이용할 것인지 하는 것이다. 즉, 더하기로 시작할 것인지 빼기로 시작할

것인지를 정하는 것이다. 이는 p+ 와 p - 를 계산하여 결정한다. 양의 교체고리의 확률을 p+ 로, 음의 교체고리의 확률을 p - 로 난수를 발생하여 랜덤하게 교체고리를 선택하는데 p+ 와 p - 는 아래와 같이 계산된다. 본 예제는 음의 교체고리를 사용하는 방법을 선택하였다.

$$P_+ = \frac{d_-}{d_+ + d_-} \quad , \quad P_- = \frac{d_+}{d_+ + d_-}$$

[제어된 반올림방법의 변환 과정]

Original	Col1	Col2	Col3	Col4	Col5	Row Sums
Row1	37	3	30	6	4	80
Row2	1	16	23	5	15	60
Row3	30	15	8	27	10	90
Row4	7	1	4	7	21	40
Col Sums	75	35	65	45	50	270

↳

R1	Col1	Col2	Col3	Col4	Col5	Row Sums
Row1	37	1	30	8	4	80
Row2	1	16	23	5	15	60
Row3	30	15	10	25	10	90
Row4	7	3	2	7	21	40
Col Sums	75	35	65	45	50	270

↳

R2	Col1	Col2	Col3	Col4	Col5	Row Sums
Row1	37	2	30	8	3	80
Row2	1	15	24	5	15	60
Row3	30	15	10	25	10	90
Row4	7	3	1	7	22	40
Col Sums	75	35	65	45	50	270

↳ ***R3***

	Col1	Col2	Col3	Col4	Col5	Row Sums
Row1	38	2	30	8	2	80
Row2	0	15	25	5	15	60
Row3	30	15	10	25	10	90
Row4	7	3	0	7	23	40
Col Sums	75	35	65	45	50	270

↳ ***R4***

	Col1	Col2	Col3	Col4	Col5	Row Sums
Row1	35	5	30	5	5	80
Row2	0	15	25	5	15	60
Row3	30	15	10	25	10	90
Row4	10	0	0	10	20	40
Col Sums	75	35	65	45	50	270

↳ ***Final***

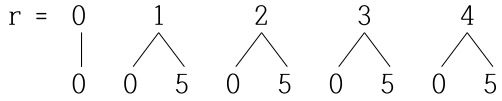
	Col1	Col2	Col3	Col4	Col5	Row Sums
Row1	35	5	30	5	5	80
Row2	0	15	25	5	15	60
Row3	30	15	10	25	10	90
Row4	10	0	0	10	20	40
Col Sums	75	35	65	45	50	270

모든 셀의 값이 B의 배수가 되도록, R1에서 R4와 같이 교체고리를 만들어 더하고 빼는 과정을 반복하면 위와 같이 최종적으로 변환된다. 이 방법은 반올림 방법 중 주변합을 바꾸지 않으면서 변환된다는 점에서 큰 의미를 가진다.

3) Zero restriction 50/50 Rounding

이 방법은 기존의 반올림방법과 유사한 방법으로서 확률적인 개념을 적용한 최근의 기법이다. 기존 방법이 반올림된 원래의 값을 추측할 수 있는 반면, 이 방법은 확률적 접근으로 의하여 랜덤하게 반올림한다. 이로 인해 원래의 값을 추측하기가 어렵다. 이 때 반올림할 확률은 50/50을 기준으로 하며, 자세한 방법은 아래와 같다.

예제) 기준(Base :B) = 5



. $P(R(r)=B) = \frac{B-1}{2B}$

. $P(R(r)=0) = \frac{B+1}{2B}$

예제의 경우 각각의 데이터가 0이 되거나 B가 될 경우를 고려해 볼 때, 0이 될 확률과 B가 될 확률은 위와 같이 계산되며 평균과 분산은 다음과 같이 계산할 수 있다.

· $E(r) = \sum r \cdot P(r)$

$= \frac{1}{B}(0 + \ ... \ + B-1)$

$= \frac{1}{B} \cdot \frac{B(B-1)}{2} = \frac{B(B-1)}{2B} = \frac{B-1}{2}$

$\because$ assume r ~ U(0, B—1), $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$

· $V(r) = E(r^2) - [E(r)]^2$

$= \frac{1}{B}(0^2+1^2+...+(B-1)^2) - (\frac{B-1}{2})^2$

$= \frac{1}{B} \cdot \frac{(B-1)B(2B-1)}{6} - (\frac{B^2-2B+1}{4})$

$= \frac{1}{12}(B^2-1)$

$\because \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$

나. 셀감추기 방법(Cell Suppression)

1) 개 요

셀감추기(CS)방법은 집계표 자료 중 노출위험이 있는 셀을 보호하기 위한 비밀보호 기법으로 일반적으로 이용되고 있다. 이 방법은 사업체수와 같은 빈도데이터(count data)와 종사자수, 매출액 등과 같은 크기데이터(magnitude data) 모두에 적용 가능하기 때문에 그 활용도가 높다. CCS방법을 적용한 미국의 사례를 보면, 1970년대 이후 미센서스국(Census Bureau)을 중심으로 집계표 자료에 대한 비밀보호 기법으로 이 방법이 적용되었는데 특히, 산업구조통계에서는 Economic Census(1994)와 ASM (Annual Survey of Manufacture) 등에 적용되고 있다.

2) 셀감추기(Cell Suppression :CS) 방법론

먼저 일반적인 셀감추기(Cell Suppression :CS) 방법에 대해 살펴보자. 이 방법은 응답자의 비밀과 관련하여 노출위험이 있다고 생각되는 즉, 민감하다고 판단되는 셀을 집계표 공표 시 감추어 보여주지 않는 방법을 말한다. 단지 민감한 셀만을 감춘다고 하여 SODC(Suppressing Only the Disclosure Cells)라고 부르기도 한다.

노출위험이 발생한다고 생각되는 경계점은 각 나라마다 다른데, 사업체수 등의 빈도데이터(count data)를 예로 들면 미국의 센서스국은 5, 미국의 국세청(IRS)는 3, 뉴질랜드는 3, 한국은 2를 그 기준으로 하고 있다. 다시 말해, 한국 통계청의 경우에는 사업체수가 2개 이하에 해당되는 셀은 감추어 보여주지 않고 집계표를 제공하고 있음을 말한다. 이 방법의 단점은 단순히 민감한 셀만을 감춤으로 인해, 집계표 내의 주변합이나 관련된 다른 집계표를 이용하여 감추어진 셀이 계산 가능한 경우가 많이 발생한다는 것이다. 이는 다음의 예제를 통해 쉽게 알 수 있다.

〈표 1-10〉 (5×6) Original Data

	Col1	Col2	Col3	Col4	Col5	Col6	
Row1	2	6	3	7	6	7	31
Row2	6	7	6	5	7	1	32
Row3	3	6	5	7	6	7	34
Row4	6	7	6	6	7	6	38
Row5	1	6	8	2	6	5	28
	18	32	28	27	32	26	163

<표 1-10>과 같은 예제에서 노출위험의 경계점을 3으로 하여 민감한 셀에 대해서만 셀감추기 한 경우, 그 주변합만을 이용한 아래의 선형관계식을 통해 그 값을 쉽게 알 수 있다.

- Col 6 : $7 + D_{26} + 7 + 6 + 5 = 26 \Rightarrow D_{26} = 1$
- Col 4 : $7 + 5 + 7 + 6 + D_{54} = 27 \Rightarrow D_{54} = 2$
- Col 3 : $D_{13} + 6 + 5 + 6 + 8 = 28 \Rightarrow D_{13} = 3$
- Row 1 : $D_{11} + 6 + 3 + 7 + 6 + 7 = 31 \Rightarrow D_{11} = 2$
- Row 3 : $D_{31} + 6 + 5 + 7 + 6 + 7 = 34 \Rightarrow D_{31} = 3$
- Row 5 : $D_{51} + 6 + 8 + 2 + 6 + 5 = 28 \Rightarrow D_{51} = 1$

셀 감추기 방법의 이와 같은 점을 보완한 비밀보호방법이 보조 감추기 즉, CCS(Complementary Cell Suppression)방법이다.

3) 보조셀 감추기(CCS) 방법론

CCS 방법은 먼저 일반적인 셀감추기 방법인 CS방법과 동일하게 집계표의 각 셀 중 응답자 비밀과 관련하여 직접적인 노출위험이 있는 PDC(Primary Disclosure Cell)을 정의한다. 그리고 정의된 셀의 값이 관련 집계표의 다른 정보들을 이용하여 계산되지 않도록 하는 CSC(Complementary Suppression Cell)을 찾는다. 즉, PDC가 다른 셀과의 선형관련성으로 재계산되지 않도록 하는 CSC를 찾는다. 이 과정에서 정의 되는 PDC와 CSC의 값 모두를 공표하지 않는 방법이 CCS방법이다.

CCS 방법에서의 주요 관건은 PDC(Primary Disclosure Cell)가 노출되지 않도록 하는 CSC(Complementary Suppression Cell)를 찾는 과정이다. 간단히 생각하면 PDC가 존재하는 행과 열에 두 개 이상의 셀이 감추어지도록 하면 될 것이다. 이러한 방법을 직관적인 보조셀감추기(Heuristic Complementary Cell Suppression :HCCS)이라고 한다. <표 1-11>의 예제에서 이 방법을 이용하여, 감추어진 PDC가 존재하는 행과 열에 두 개 이상의 셀이 감추어지도록 하면 다음과 같다.

〈표 1-11〉 HCCS방법 적용 결과

	Col1	Col2	Col3	Col4	Col5	Col6	
Row1	2	6	3	7	6	7	31
Row2	6	7	6	5	7	1	32
Row3	3	6	5	7	6	7	34
Row4	6	7	6	6	7	6	38
Row5	1	6	8	2	6	5	28
	18	32	28	27	32	26	163

<표 1-11>과 같이 감추기를 한 경우, 감추어진 셀을 계산해 낼 수 없는 것처럼 보이지만 아래의 식을 이용하면 그 값을 찾을 수 있게 된다.

- Equation (1)

$$\text{Row2} + \text{Row5} = (6+7+6+D_{24}+7+D_{26}) + (D_{51}+6+8+D_{54}+6+D_{56})$$

- Equation (2)

$$\text{Col4} + \text{Col6} = (7+D_{24}+7+6+D_{26}) + (7+D_{26}+7+6+D_{56})$$

- Equation (1) − Equation (2) = $D_{51} = 1$

따라서 CSC(Complementary Suppression Cell)를 찾는 과정에 있어서 위와 같이 행과 열에 두 개의 셀이 감추어지도록 하는 것이 아니라, 셀들의 선형관련성으로 계산되지 않도록 하는 수학적인 접근방법이 필요함을 알 수 있다. 아울러 감추어지는 셀의 수와 정보량을 최소화하는 CSC를 찾아야 함은 물론이다.

이러한 논리적인 접근을 위해 결합기법(combinational technique), 선형기법(linear programming), 수학적 네트워크기법(mathematical network)과 같은 3가지 접근법이 주로 적용된다. 3가지 방법 모두 수학적인 측면에서는 상이하나, 자료제공이 제한되는 셀의 수와 정보량을 최소화하는 목적함수(Objective function)를 정의하고, 이를 이용하여 그 해(解)를 찾아나간다는 점에서는 그 맥락이 같다.

CCS방법의 절차는 「감추어진 셀의 수」와 「감추어진 정보량」중 어떤 것을 최소화 할 것인가에 따라 접근방법이 다소 차이가 있으나, 기본적인 절차는 다음과 같다.

(1) 기준(Base :B)을 정한다. 즉 PDC(Primary Disclosure Cell)을 구분할 수 있는 경계점을 지정하는 것으로서, 사업체수의 경우 n개 이하 셀 정보를 제한한다면 B=n이 된다.
(2) CSC(Complementary Suppression Cell)을 선형모형식 또는 수학적 네트워크 등을 이용하여 찾는다.
(3) PDC와 CSC 모두를 감추기 한 후, 자료를 제공한다.

가) 일반적인 CCS방법

다음의 예제는 「감추어진 정보량」을 최소화 할 수 있도록 CSC(Complementary Suppression Cell)를 찾는 방법이다. <표 1-12>의 (3×3)정방행렬의 데이터를 가지고 CCS 방법을 적용하는 경우를 생각해 보면 아래와 같은 과정으로 설명할 수 있다.

〈표 1-12〉 (3×3)Original data

	Col1	Col2	Col3	합 계
Row1	10	5	10	25
Row2	15	10	1	26
Row3	20	15	1	36
합 계	45	30	12	87

(Step1) 노출위험의 경계점 Base를 b=2로 가정하면, PDC(Primary Disclosure Cell)는 $X_{23}=1, X_{33}=1$이 된다.

(Step2) $X_{23}=1, X_{33}=1$ 정보만 제한한다면, 집계표 내 다른 셀과의 선형관련성으로 인해 비밀보호가 이루어지지 않는다.

$$X_{21}+X_{22}+X_{23}=26 \Leftrightarrow 15+10+X_{23}=26$$

$$\therefore X_{23}=1$$

그렇기 때문에 가능한 CSC(Complementary Suppression Cell)의 쌍(pair)을 선형모형식으로 찾으면 $(X_{21}, X_{31})(X_{22}, X_{32})$인데, 공개가 제한되는 정보량이 최소가 되도록 하기 위해서는 (X_{22}, X_{32})가 최적의 CSC로 정해진다.

(Step3) PDC : (X_{23}, X_{33})
CSC : (X_{22}, X_{32})

〈표 1-13〉 (3×3) CCS방법 적용 결과

	Col1	Col2	Col3	합 계
Row1	10	5	10	25
Row2	15	10	1	26
Row3	20	15	1	36
합 계	45	30	12	87

노출위험이 있는 셀에 대한 CSC의 도출 결과는 <표 1-13>과 같으며, 동 결과의 효율성(efficiency)은 <표 1-14>와 같이 측정할 수 있다.

〈표 1-14〉 CCS방법에 따른 정보제한총량

(단위 : 개, %)

	PDC[2]		CSC[3]		합 계	
		구성비		구성비		구성비
- 감추어진 셀의 수	2	22.2	2	22.2	4	44.4
- 감추어진 정보량	2	2.3	25	28.7	27	31.0

즉, PDC(Primary Disclosure Cell) (X_{23}, X_{33})에 대한 노출위험을 보호하기 위해 보조적으로 2개의 CSC(Complementary Suppression Cell) (X_{22}, X_{32})을 추가함으로써, 3×3 table의 노출위험이 있는 셀 값을 보호하였다. 이 과정에서 「감추어진 셀의 수」는 2개(PDC) → 4개(PDC 2개+CSC 2개)로 증가하였고, 노출위험이 있는 2.3% (PDC)의 정보를 보호하기 위해 28.7% (CSC)의 정보를 추가적으로 감추기 함으로써 「감추어진 정보량」은 31.0% (PSC+CSC)로 증가하였다.

<표 1-13> CCS결과에 대한 정도 검증을 위하여, 선형관련성을 이용해 확인해 보면 다음과 같다.

Equations) $X_{22} + X_{23} = 11 \ (eq1)$ $X_{32} + X_{33} = 16 \ (eq2)$

$X_{22} + X_{32} = 25 \ (eq3)$ $X_{23} + X_{33} = 2 \ (eq4)$

$0 \le X_{33} \le 2 \quad (eq5)$ $0 \le X_{23} \le 2 \quad (eq6)$

☞ X_{ij} :"or"$iginal\ value \quad i : \#\ of\ rows \quad j : \#\ of\ columns$

solution) $9 \le X_{22} \le 11 \ (eq7)$ $14 \le X_{32} \le 16 \ (eq8)$

eq7, eq8, eq3으로 가능한 $(X_{22}, X_{32}, X_{23}, X_{33})$의 조합은 아래와 같다.

- Feasible solution 1 : (9, 16, 2, 0)
- Feasible solution 2 : (10, 15, 1, 1)
- Feasible solution 3 : (11, 14, 0, 2)

즉, CCS방법을 이용하여 위와 같이 4개의 셀(PDC: 2개, CSC: 2개)을 감추기 한 경우, 선형관련식을 이용하여 감추어진 셀에 대한 unique한 solution을 찾을 수 없기 때문에, <표 1-13>은 성공적으로 비밀보호가 수행된 것으로 평가할 수 있다.

---

2) PDC(Primary Disclosure Cell)

3) CSC(Complementary Suppression Cell)

나) 네트워크(network)이론을 통한 CCS방법

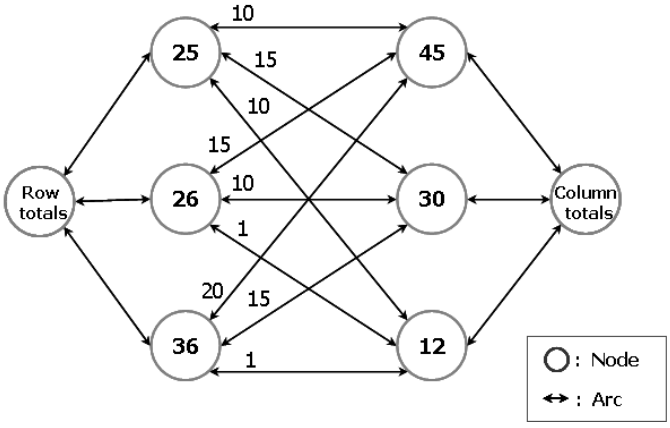
CCS방법을 네트워크(network) 이론으로 접근하면 다음과 같이 일반화 할 수 있다. 목적함수(1)는 「감추어진 정보량」을 최소화하는 것이며, 목적함수(2)는 「감추어진 셀의 수」를 최소화하는 방법이다.

목적함수(1) : $V(a) = \min \sum a_{ij} z_{ij}$

목적함수(2) : $V(a) = \min \sum z_{ij}$

☞ a_{ij} : 셀값, $z_{ij} = 1$ (감춘 셀), $z_{ij} = 0$ (기타 셀)

<표 1-12>에서 「감추어진 정보량」을 최소화하도록 네트워크테이블(network table)을 작성하면 [그림 1-1]과 같다.



[그림 1-1] 〈표 1-12〉에 대한 network table

위 [그림 1-1]은 <표 1-12>에 대한 네트워크 테이블(network table)을 나타낸 것이다. 네트워크 테이블을 작성하는 방법을 살펴보면, [그림 1-1]에서 보는바와 같이 노드(Node)의 두 번째 열에 데이터의 행(Row)의 합계를 표시하고 노드의 세 번째 열에는 데이터의 열(Column)의 합계를 표시한다. 또한, 행에서 열으로 보내어지는 값을 아크(Arc ;↔)위에 나타내어 도식화 한다. 이때 행과 열간에는 서로 왕복할 수 있는 경로를 갖기 때문에 양쪽이 열린 화살표(↔)로 표시한다. 참고로 3차원 이상의 집계표를 network table로 도식화하는 경우, 평면으로 도식화 가능하도록 2차원 집계표로 분리・축소하여 각각 나타내어야 한다.

우선 <표 1-12>데이터의 기준(Base :B)은 B=2로 가정했기 때문에 [그림 1-1]에서 PDC(Primary Disclosure Cell)은 a_{23}, a_{33}임을 알 수 있다. PDC(a_{23}, a_{33}) 중 a_{33}에서 출발하여 경로를 찾는 방법을 살펴보자. 1개의 PDC 자료를 보호하기 위해서는 최소 3개의 CSC(Complementary Suppression Cell)가 필요한데, 이는 집계표 내 다른 셀과의 선형관련성으로 인해 PDC가 재계산되어 비밀보호가 이루어지지 않을 수 있는 경우를 막기 위해서이다. 이와 같이 4step 내에 a_{33}로 도착할 수 있는 모든 경로를 고려하여 이에 대한 $V(a) = \sum a_{ij} Z_{ij}$ 값의 합을 계산한 후, 목적함수(1)를 만족하는 최소의 경로를 찾는다.

〈표 1-15〉 Network에 의한 최소경로 탐색(a_{33})

가능한 경로	감추어질 셀 값	비 고
(3,3) → (2,3) → (2,2) → (3,2) → (3,3)	27	최소경로
(3,3) → (1,3) → (1,2) → (3,2) → (3,3)	31	
(3,3) → (1,3) → (1,1) → (3,1) → (3,3)	41	
(3,3) → (2,3) → (2,1) → (3,1) → (3,3)	37	

<표 1-15>의 결과를 보면, PDC a_{33}에 대한 CSC는 a_{23}, a_{22}, a_{32}로 정해진다. 이 경로는 「감추어진 정보량」와 「감추어진 셀의 수」를 모두 최소화하는 결과이다.

동일한 방법으로 a_{23}에 대한 최소경로를 찾으면 <표 1-16>과 같다.

〈표 1-16〉 Network에 의한 최소경로 탐색(a_{23})

가능한 경로	감추어질 셀 값	비 고
(2,3) → (1,3) → (1,2) → (2,2) → (2,3)	26	최소경로
(2,3) → (3,3) → (3,2) → (2,2) → (2,3)	27	
(2,3) → (1,3) → (1,1) → (2,1) → (2,3)	41	
(2,3) → (3,3) → (3,1) → (2,1) → (2,3)	37	

<표 1-16>의 결과를 보면, PDC a_{23}에 대한 CSC는 a_{13}, a_{12}, a_{22}로 정해진다. 그러나 네트워크(network)에 의한 최소경로 탐색결과는 과대감추기(over suppression)를 발생할 수 있다. 이 때문에 네트워크 탐색 후, CSC의 적정성을 선형관련식으로 판단하여 과대감추기된 셀을 제거할 필요가 있다. 이를 통해 최종적으로 「감추어진 정보량」을 최소화하는 CSC를 찾게 된다. 위 예제에 대한 선형관련식은 다음과 같다.

Equations) $a_{12} + a_{13} = 15 \ (eq1)$ $a_{22} + a_{23} = 11 \ (eq2)$
$a_{32} + a_{33} = 16 \ (eq3)$
$0 \leq a_{33} \leq 2 \ (eq4)$ $0 \leq a_{23} \leq 2 \ (eq5)$

☞ X_{ij} :"or"$iginal\ value$ $i : \#\ of\ rows$ $j : \#\ of\ columns$

위 선형식에서 PDC (a_{23}, a_{33})에 대한 CSC (a_{22}, a_{32}) 만으로도 감추어진 셀의 유일한 해를 찾을 수 없기 때문에, (a_{12}, a_{13})을 CSC로 추가하는 것은 과대감추기(over suppression)가 됨을 알 수 있다. 이와 같은 경우 과대감추기한 셀은 감추기에서 제외한다. 따라서 최종 CCS 결과는 PDC(Primary Disclosure cell) : (a_{23}, a_{33}), CSC(Complementary Suppression Cell) : (a_{22}, a_{32}) 이다.

3셀감추기(CS) 방법은 결합기법(combinational technique), 선형기법(linear programming), 수학적 네트워크기법(mathematical network)과 같은 3가지 접근법이 주로 적용된다. 빈도데이터와 크기데이터 모두에 적용가능한 장점이 있으며, 자료제공이 제한되는 셀의 수와 정보량을 최소화하는 목적함수(Objective function)를 만족시키는 해(解)를 찾아나간다는 점에서는 그 맥락이 같다. 또한, 직관적인 셀감추기(Heuristic Complementary Cell Suppression :HCCS)와 SODC(Suppression Only the Disclosure Cell)방법은 비밀보호된 셀 값이 재계산될 가능성이 높기 때문에 보조셀감추기(Complementary Cell Suppression) 방법을 적용하는 것이 효과적임을 살펴보았다.

다. CTA방법(Controlled Tabular Adjustment)

1) 개 요

CTA방법은 집계표 형태의 데이터에 대한 비밀보호 기법 중의 하나로서, 주로 크기데이터(Magnitude data)에 적용된다. 이 방법은 노출위험이 있는 자료를 감추거나 제거하지 않고 셀 값을 조정하는 방법이다. 보조셀감추기(CCS)방법이 노출위험이 있는 셀 값을 제한(감추기, suppression)한다는 것과 다르게 CTA방법은 모든 셀 값을 제공한다는 데에 큰 장점이 있다. 이에 따라 미국에서는 보조셀감추기(CCS)방법의 대안으로 대두되고 있다. 그러나 이 방법은 주변합(marginal total)이 원자료의 집계결과와 달라지기 때문에 우리나라의 경우 이용 상 혼선을 초래할 수 있어 적용 시 사전검토가 필요하다.

2) CTA 방법론

CTA방법이 적용되는 과정을 자세히 살펴보자. 우선 앞에서 소개한「민감한 셀 찾기」방법 중 해당 자료형태에 맞는 방법을 이용하여 노출위험이 있는 셀을 정의한다. 그리고 정의된 민감한 셀이 노출위험이 없는 셀이 되기 위한 값(safe cell value)이 무엇인지

찾는다. 이때 노출위험이 없는 셀이 되기 위한 값은 원자료에 임의의 값을 더하거나 빼면서 탐색하며, 선형모형 프로그램을 이용하여 리밸런싱(rebalancing) 과정을 수행한 후 셀 값을 조정한다. 이러한 과정을 모형으로 표현하면 아래와 같다.

$$\min \sum_{i=1}^{n} (y_i^+ + y_i^-)$$

$$\text{s.t } T(y) = 0 \ ,$$
$$y_i^- \geq p_i(1 - I_i) \ ,$$
$$y_i^+ \geq p_i I_i \ ,$$
$$0 \leq y_i^- \ ,$$
$$y_i^+ \leq e_i \ ,$$
$$I_i : binary \implies 0 \rightarrow (-),\ 1 \rightarrow (+)$$

where $y = y^+ - y^-$,
p_i : 민감한 셀을 보호하기 위한 상하한값,
a : 원자료,
$a + y^+ - y^-$: 조정된 자료

〈표 1-17〉 원자료 집계 결과

167	317	1,284	587	4,490	3,981	2,442	1,150	70(21)	**14,488**
57(1)	1487	172	667	1,006	327	1,683	1,138	46(7)	**6,583**
616	202	1,899	1,098	2,172	3,825	4,372	300(40)	787	**15,271**
0	36(10)	0	16(4)	0	0	65	0	140(40)	**257**
840	**2,042**	**3,355**	**2,368**	**7,668**	**8,133**	**8,562**	**2,588**	**1,043**	**36,599**

노출위험이 있는 셀(7개) : a_{21}, a_{42}, a_{44}, a_{38}, a_{51}, a_{52}, a_{54}
괄호() 안의 숫자 : safe cell value

〈표 1-18〉 보조셀감추기(CCS)방법 적용 결과

CSC(X)	317	1,284	CSC(X)	4,490	3,981	2,442	1,150	PDC(X)	**14,488**
PDC(X)	1,487	172	667	1,006	327	1,683	CSC(X)	PDC(X)	**6,583**
616	CSC(X)	1,899	1,098	2,172	3,825	4,372	PDC(X)	787	**15,271**
0	PDC(X)	0	PDC(X)	0	0	65	0	PDC(X)	**257**
840	**2,042**	**3,355**	**2,368**	**7,668**	**8,133**	**8,562**	**2,588**	**1,043**	**36,599**

감추어진 셀 수 : 11개(30%)
감추어진 정보량 : 2,759개(7.5%)

〈표 1-19〉 CTA방법 적용 결과

167	317	1,276(-8)	587	4,490	3,981	2,442	1,150	91(+21)	**14,501**
56(-1)	1,487	172	667	1,006	327	1679(-4)	1,138	39(-7)	**6,571**
617(+1)	196(-6)	1899	1,095(-3)	2172	3,825	4,371(-1)	260(-40)	797(+10)	**15,232**
0	26(-10)	0	12(-4)	0	0	70(+5)	0	180(+40)	**288**
840	**2,026**	**3,347**	**2,361**	**7,668**	**8,133**	**8,562**	**2,548**	**1,107**	**36,592**

<표 1-17>의 예제를 가지고 보조셀감추기(CCS) 방법과 CTA방법을 적용해 보기로 하자. 예제의 민감한 셀은 p-percent rule(p=20%)을 적용하여 선정하였으며, 괄호 안의 수를 더하거나 뺐을 때 노출위험이 없는 값이 된다. <표 1-18>는 보조셀감추기 방법을 적용한 결과이며 <표 1-19>은 CTA방법을 적용한 결과이다. CTA방법은 위에 나타낸 모형을 바탕으로 행과 열의 균형을 유지하도록 리밸런싱(rebalancing)하면서 노출위험이 있는 셀 값에 일정한 값을 가감하여 민감하지 않은 셀이 되도록 조정하였다.

보조셀감추기 방법을 적용한 경우 <표 1-18>와 같이 감추어진 셀의 수가 11개이며 CTA방법을 적용한 경우 <표 1-19>과 같이 바꾸어진 셀의 수가 13개이다. 보조셀감추기 방법에서 감추어진 셀의 수에 비해 CTA방법에서 바꾸어지는 셀의 수가 더 많으나, CTA 방법의 경우 감추어지는 셀 없이 셀 값 모두를 제공한다는 차이점이 있다. 즉, 이 방법의 경우 제한되는 정보량은 없는 것이다. 그러나 원자료 집계결과와 주변 합이 다르기 때문에 우리나라에 이 같은 기법을 적용하기 위해서는 충분한 검토가 필요할 것으로 생각된다.

제3절 산업구조통계의 비밀보호 현황

매크로 데이터란 개별자료의 집계표 형태를 의미하는데, 통계조사별로 발간되는 보고서에 수록되는 집계표가 이에 속한다. 이러한 집계표를 구성하고 있는 자료는 크게 ①빈도데이터(count data)와 ②크기데이터(magnitude data)로 구분할 수 있다. 빈도데이터는 사업체수와 같이 개수 자체에 의미가 있는 자료형태이며, 크기데이터는 매출액, 종사자수 등과 같이 그 규모에 의미가 있는 형태이다. 두 데이터는 그 형태에 따라 비밀보호 방법에 있어 적용상 차이가 있다. 앞서 언급한 바와 같이 일반적으로「빈도데이터」에는 셀감추기(Cell suppression), 반올림(Rounding), 자료변조(Data perturbation) 등을 「크기데이터」는 셀감추기(Cell suppression)를 적용하고 있다.(Cox, 2008) 만약 두 데이터 형태가 혼합되어 제공되고 있는 경우는 셀감추기 방법만이 적용 가능하다. 이 밖의 방법으로는 분류구조를 변경하는 방법으로, 표 크기를 축소하거나 특성을 재분류하여 노출위험이 있는 셀이 나타나지 않도록 조정하는 방법이 있다.

통계청에서 제공되고 있는 산업관련 매크로 데이터의 형태를 보면, 빈도데이터와 크기데이터가 혼합되어 하나의 집계표에 제공된다. 즉, 사업체수를 중심으로 주요 지표들을 산업분류, 종사자규모, 매출액규모 등으로 그룹화하여 주요 항목에 관한 정보를 제공하는데, 그 형태는 <표 1-20>과 같다.

〈표 1-20〉 집계표 예시

산업분류	종사자규모	사업체수	월평균 종사자수	급여액	출하액	…	유형자산 연말잔액
D19 가죽, 가방 및 신발 제조업	5~9명	1,181	7,212	84,543	565,117	…	110,002
	10~19명	435	5,696	90,273	619,753	…	124,855
	:	:	:	:	:	:	:
	300~499명	2	X	X	X	X	X
	500명이상	3	1,964	56,025	324,478	…	227,235

현재 통계청에서 작성하고 있는 산업구조통계는 10종에 달한다. 이들에 대한 집계표 제공 시 적용하고 있는 비밀보호방법은 제2절에서 언급한 방법 중 노출위험이 있는 셀만 감추는 SODC(Suppressing Only the Disclosure Cells) 방법이다. 사업체 수를 기준변수로 셀의 노출위험 여부를 결정하며, 그 기준(base)은 2로 한다. 즉, 해당 셀의 사업체 수가 2 이하인 경우는 노출위험이 있는 셀로 결정하고, 노출위험이 있는 셀로 결정되면 동일한 행의 모든 자료는「X」표시로 제공하고 본래의 정보는 감추어진다. 다시 말해, 노출위험이 있는 셀만 감추는 SODC(Suppressing Only the Disclosure Cells) 방법으로서 기준을 2로 적용하고 있다.

본 절에서는 산업구조통계 중 「광업 및 제조업 통계조사」를 대상으로 노출위험이 있는 셀에 대한 비밀보호 현황과 그 문제점을 살펴보고자 한다.

1. 광업 및 제조업 통계조사 작성 개요

「광업 및 제조업 통계조사」는 한국표준산업분류상 「C. 광업」과「D. 제조업」을 영위하는 사업체 중 종사자수 5인 이상의 사업체를 대상으로 매년 실시되고 있다. 조사결과가 공표되는 형태는 ①산업분류별 ②산업분류별×종사자규모별 ③산업분류별×매출액규모별로 「전국편과 지역편」「산업편과 품목편」등으로 구분해 공표된다. 주요지표 도출을 위한 조사항목은 14개 항목으로서 <표 1-21>와 같다.

〈표 1-21〉 광업 및 제조업 통계조사 항목

구 분	조 사 항 목
일반사항	1)사업체명 및 소재지 2)본사소속 및 타공장 유무 3)본사명 및 소재지 4)경영조직 5)자산과 자본
고용사항	6)종사자수 및 연간급여액
수입사항	7)연간출하액 및 수입액
비용사항	8)연간제조원가 및 판매비와 관리비 9)내국소비세
자산사항	10)유형자산 11)무형자산 및 컴퓨터소프트웨어
기타사항	12)연간제품출하액 및 재고액 내역 13)연간 임가공 수입액의 품목별 내역 14)재고액

<표 1-21>의 항목 중 주요지표를 중심으로 「광업 및 제조업 통계조사 보고서」 및 국가통계포털(KOSIS)을 통해 제공되고 있는 주요 집계표의 종류는 아래와 같다.

1-1) 산업세세분류×종사자규모별 주요지표
1-2) 시도×산업중분류×종사자규모별 주요지표
2-1) 산업세세분류×출하액규모별 주요지표
2-2) 시도×산업중분류×출하액규모별 주요지표
3-1) 산업세세분류별 사업체수 및 종사상지위 및 성별 종사자수
3-2) 시도×산업중분류별 사업체수 및 종사상지위 및 성별 종사자수
4-1) 산업세세분류별 사업체수 및 출하액 및 주요생산비
4-2) 시도×산업중분류별 사업체수 및 출하액 및 주요생산비
5-1) 산업세세분류별 사업체수 및 재고액
5-2) 시도×산업중분류별 사업체수 및 재고액
6-1) 산업세세분류별 유형고정자산
6-2) 시도×산업중분류별 유형고정자산
7-1) 산업세세분류별 부지면적 및 임차료
7-2) 시도×산업중분류별 부지면적 및 임차료
7-3) 시군구×산업대분류별 부지면적 및 임차료

2. 비밀보호 방법

「광업 및 제조업 통계조사」는 <표 1-21>와 같이 빈도데이터(count data)와 크기데이터(magnitude data)가 하나의 집계표에 표두 항목으로 제공되고 있다. 제2절에서 소개한 미국의 ASM(Annual Survey of Manufacture)이 빈도데이터와 크기데이터를 서로 다른 집계표로서 제공되고 있는 점과는 차이가 있는데, 이런 점에서 우리나라의 비밀보호 방법은 미국과 달리 변형된 방법을 적용해야 함을 알 수 있다.

현행「광업 및 제조업 통계조사」는 노출위험이 있다고 판단되는 셀을 사업체수 2개를 기준으로 정하고 있다. 즉, 사업체수가 2개 이하인(base=24)) 경우 자료제공을 제한하는 방법이다. 예를 들어 산업세세분류별×종사자규모별 집계표에서 해당 셀의 사업체수가 1개이거나 2개인 경우, 사업체수를 제외한 해당 분류 내 셀과 관련된 모든 정보의 자료제공을 제한하고「X」표시로 나타낸다. 이는 노출위험이 있는 셀만 감추는 전통적인 SODC(Suppressing Only the Disclosure Cells) 방법을 적용한 경우이다. 그러나 이 방법은 제2절에서 언급한 바와 같이, 집계표내의 행과 열의 선형 관련식으로 계산이 가능하기 때문에 쉽게 「X」표시된 값을 계산하거나 추측할 수 있다.

이에 본 절에서는「2006년 기준 광업 및 제조업통계조사」결과를 통해 통계청의 산업구조 통계관련 비밀보호기법 적용 현황과 그 한계점을 살펴보고자 한다.

「2006년 기준 광업 및 제조업통계조사」결과 중 한국표준산업분류코드 D23에 해당하는「코크스 석유정제품 및 핵연료제조업」의 산업 및 종사자규모별 집계표의 예를 살펴보자. <표 1-22>는 이에 해당하는 데이터로서 현재 「2006년 기준 광업 및 제조업 통계조사 보고서」와 국가통계포털(KOSIS)에서 제공하고 있는 형태로, 분류 내에 사업체수가 1개이거나 2개인 경우 동일한 행의 모든 정보를 「X」표시로 감추기하고 있다. 이 과정에서 우선 눈에 띄는 점은 사업체 분포도가 상대적으로 적은 D2321 분류의 경우, 8개 범주의 종사자규모[5] 중 「500명 이상」의 정보만이 제공되고 있다는 것을 알 수 있다. 이는 공표를 위한 세부 분류단계를 정함에 있어, 산업을 특성을 고려하지 않고 모든 산업에 그 기준을 동일하게 적용한 데 기인한 것으로 판단된다. 그러므로 세부분류 단계의 공표 기준을 정함에 있어, 산업의 특성을 반영하여 정보량과 실효성을 감안한 기준을 정하는 것에 관한 재검토가 필요할 것이라 생각된다.

---

4) Base=2 : 사업체수가 1개 또는 2개인 경우에 한함(단, 0인 경우는 제외)

5) 종사자규모 : ① 5~9명 ② 10~19명 ③ 20~49명 ④ 50~99명 ⑤ 100~199명 ⑥ 200~299 ⑦ 300~499 ⑧ 500명이상

〈표 1-22〉 코크스,석유정제품 및 핵연료 제조업(D23)

산업 및 종사자규모		사업체수	월평균 종사자수	급여액	출하액
23	계	118	9,791	674,424	69,840,546
	5 ~ 9	38	257	5,967	217,730
	10 ~ 19	35	469	11,144	211,887
	20 ~ 49	32	966	28,308	630,663
	50 ~ 99	6	406	17,052	392,182
	100 ~ 199	1	X	X	X
	500명 이상	6	7,514	603,092	67,817,139
232	계	117	9,168	642,376	69,730,433
	5 ~ 9	38	257	5,967	217,730
	10 ~ 19	35	469	11,144	211,887
	20 ~ 49	32	966	28,308	630,663
	50 ~ 99	6	406	17,052	392,182
	100 ~ 199	1	X	X	X
	500명 이상	5	6,891	571,044	67,707,026
2321	계	8	7,176	583,927	68,408,790
	5 ~ 9	0	0	0	0
	10 ~ 19	0	0	0	0
	20 ~ 49	1	X	X	X
	50 ~ 99	1	X	X	X
	100 ~ 199	1	X	X	X
	500명 이상	5	6,891	571,044	67,707,026
23210	계	8	7,176	583,927	68,408,790
	5 ~ 9	0	0	0	0
	10 ~ 19	0	0	0	0
	20 ~ 49	1	X	X	X
	50 ~ 99	1	X	X	X
	100 ~ 199	1	X	X	X
	500명 이상	5	6,891	571,044	67,707,026

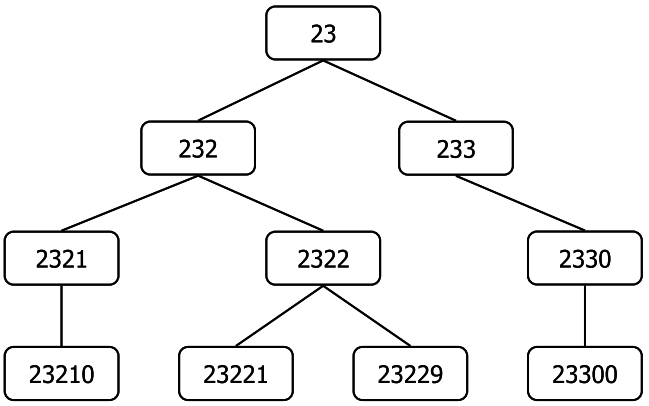
〈표 1-22〉 코크스,석유정제품 및 핵연료 제조업(D23)(계속)

산업 및 종사자규모		사업체수	월평균 종사자수	급여액	출하액
232	계	117	9,168	642,376	69,730,433
	5 ~ 9	38	257	5,967	217,730
	10 ~ 19	35	469	11,144	211,887
	20 ~ 49	32	966	28,308	630,663
	50 ~ 99	6	406	17,052	392,182
	100 ~ 199	1	X	X	X
	500명 이상	5	6,891	571,044	67,707,026
2322	계	109	1,992	58,449	1,321,643
	5 ~ 9	38	257	5,967	217,730
	10 ~ 19	35	469	11,144	211,887
	20 ~ 49	31	922	27,076	568,481
	50 ~ 99	5	344	14,262	323,545
	100 ~ 199	0	0	0	0
	500명 이상	0	0	0	0
23221	계	66	1,210	37,211	1,099,890
	5 ~ 9	22	150	3,231	170,564
	10 ~ 19	23	315	7,300	177,910
	20 ~ 49	18	552	17,452	469,511
	50 ~ 99	3	193	9,228	281,905
	100 ~ 199	0	0	0	0
	500명 이상	0	0	0	0
23229	계	43	782	21,238	221,753
	5 ~ 9	16	107	2,736	47,166
	10 ~ 19	12	154	3,844	33,977
	20 ~ 49	13	370	9,624	98,970
	50 ~ 99	2	X	X	X
	100 ~ 199	0	0	0	0
	500명 이상	0	0	0	0

〈표 1-22〉 코크스,석유정제품 및 핵연료 제조업(D23)(계속)

산업 및 종사자규모		사업체수	월평균 종사자수	급여액	출하액
233	계	1	X	X	X
	5 ~ 9	0	0	0	0
	10 ~ 19	0	0	0	0
	20 ~ 49	0	0	0	0
	50 ~ 99	0	0	0	0
	100 ~ 199	0	0	0	0
	500명 이상	1	X	X	X
2330	계	1	X	X	X
	5 ~ 9	0	0	0	0
	10 ~ 19	0	0	0	0
	20 ~ 49	0	0	0	0
	50 ~ 99	0	0	0	0
	100 ~ 199	0	0	0	0
	500명 이상	1	X	X	X
23300	계	1	X	X	X
	5 ~ 9	0	0	0	0
	10 ~ 19	0	0	0	0
	20 ~ 49	0	0	0	0
	50 ~ 99	0	0	0	0
	100 ~ 199	0	0	0	0
	500명 이상	1	X	X	X

한편, 비밀보호 방법론적인 관점에서 살펴보면, <표 1-22>는 비밀보호를 위해 SODC(Suppressing Only the Disclosure Cells) 방법이 적용되었는데, 이 방법의 보완점을 파악하기 위해 우선 집계표내의 선형관련식을 통해 감추기한 셀「X」의 값을 계산할 수 있는지 알아보고자 한다. 집계표 내의 선형관련식을 만들기 위해 중분류 D23의 분류체계를 보면 [그림 1-2]와 같다.



D23 코크스, 석유정제품 및 핵연료	D232 석유정제품 제조업	D2321 원유 정제처리업	D23210 원유 정제처리업
		D2322 석유정제물 재처리업	D23221 윤활유 및 그리스 제조업 D23229 기타 석유정제물 재처리업
	D233 핵연료 가공업	D2330 핵연료 가공업	D23300 핵연료 가공업

[그림 1-2] 코크스,석유정제품 및 핵연료 제조업 분류체계[6)]

[그림 1-2]의 분류체계를 이용해 선형관련식을 만들면 다음과 같다.

$$D23_i = D232_i + D233_i$$

$$D232_i = D2321_i + D2322_i \qquad D2321_i = D23210_i$$
$$D2322_i = D23221_i + D23229_i$$
$$D233_i = D2330_i \qquad D2330_i = D23300_i$$

$$i = 1,2,...,8(\text{종사자규모})$$

선형관련식을 이용하여 감추기한 셀「X」의 값을 하향식(Top down) 접근방식으로 단계별로 계산하면 다음과 같다.

6) D231(코크스 및 관련제품 제조업)은 조사대상 사업체가 없기 때문에 제외

〈표 1-23〉 셀값찾기 Step1

산업 및 종사자규모		사업체수	월평균 종사자수	급여액	출하액
D23	계	118	9,791	674,424	69,840,546
	5 ~ 9명	38	257	5,967	217,730
	10 ~ 19명	35	469	11,144	211,887
	20 ~ 49명	32	966	28,308	630,663
	50 ~ 99명	6	406	17,052	392,182
	100 ~ 199명	1	179	8,861	570,945
	500명 이상	6	7,514	603,092	67,817,139

노출위험이 있는 「100~199명」 분류에서 「X」로 표시되어 감추어진 월평균종사자수, 급여액 및 출하액 등의 값은 관련식 (1.12)에 의해서 <표 1-23>과 같이 모든 값을 계산할 수 있다.

$$X_{5j} = X_{+j} - \sum_{i \neq 5} X_{ij} \quad j = 1, 2, \ldots \tag{1.12}$$

〈표 1-24〉 셀값찾기 Step2

산업 및 종사자규모		사업체수	월평균 종사자수	급여액	출하액
D23	100 ~ 199명	1	179	8,861	570,945
D232	100 ~ 199명	1	179	8,861	570,945
D233	100 ~ 199명	0	0	0	0

소분류 D232의 「100~199명」 분류에 감추어진 셀의 값은 <표 1-23>의 step 1에서 선형관련식으로 계산된 값과 관련식 (1.13)를 이용하여 <표 1-24>와 같이 계산된다.

$$D23_i = D232_i + D233_i \tag{1.13}$$

이와 같은 계산단계를 선형관련식을 이용하여 순차적으로 진행하면, 다음과 같이 중분류 D23에 감추어진 모든 「X」의 값이 <표 1-28>과 같이 계산된다. 즉, 감추어진 모든 셀의 값을 찾아낼 수 있다.

〈표 1-25〉 셀값찾기 Step3

산업 및 종사자규모		사업체수	월평균 종사자수	급여액	출하액
D232	20 ~ 49명	32	966	28,308	630,663
D2321	20 ~ 49명	1	44	1,232	62,182
D2322	20 ~ 49명	31	922	27,076	568,481

선형관련식 : $D232_i = D2321_i + D2322_i$

〈표 1-26〉 셀값찾기 Step4

산업 및 종사자규모		사업체수	월평균 종사자수	급여액	출하액
D232	50 ~ 99명	6	406	17,052	392,182
D2321	50 ~ 99명	1	62	2,790	68,637
D2322	50 ~ 99명	5	344	14,262	323,545
D232	100 ~ 199명	1	179	8,861	570,945
D2321	100 ~ 199명	1	179	8,861	570,945
D2322	100 ~ 199명	0	0	0	0

선형관련식 : $D232_i = D2321_i + D2322_i$

〈표 1-27〉 셀값찾기 Step5

산업 및 종사자규모		사업체수	월평균 종사자수	급여액	출하액
합 계		118	9,791	674,424	69,840,546
D232	소 계	117	9,168	642,376	69,730,433
D233	소 계	1	623	32,048	110,113
	5 ~ 9명	0	0	0	0
	10 ~ 19명	0	0	0	0
	20 ~ 49명	0	0	0	0
	50 ~ 99명	0	0	0	0
	100 ~ 199명	0	0	0	0
	500명 이상	1	623	32,048	110,113

선형관련식 : $D23_i = D232_i + D233_i \qquad D233 = \sum_{i=1} D233_i$

〈표 1-28〉 선형관련식에 의한 감추어진 셀「X」값 찾기 최종 결과

산업 및 종사자규모		사업체수	월평균 종사자수	급여액	출하액
D23	계	118	9,791	674,424	69,840,546
	5 ~ 9	38	257	5,967	217,730
	10 ~ 19	35	469	11,144	211,887
	20 ~ 49	32	966	28,308	630,663
	50 ~ 99	6	406	17,052	392,182
	100 ~ 199	1	179	8,861	570,945
	500명 이상	6	7,514	603,092	67,817,139
D232	계	117	9,168	642,376	69,730,433
	5 ~ 9	38	257	5,967	217,730
	10 ~ 19	35	469	11,144	211,887
	20 ~ 49	32	966	28,308	630,663
	50 ~ 99	6	406	17,052	392,182
	100 ~ 199	1	179	8,861	570,945
	500명 이상	5	6,891	571,044	67,707,026
D2321	계	8	7,176	583,927	68,408,790
	5 ~ 9	0	0	0	0
	10 ~ 19	0	0	0	0
	20 ~ 49	1	44	1,232	62,182
	50 ~ 99	1	62	2,790	68,637
	100 ~ 199	1	179	8,861	570,945
	500명 이상	5	6,891	571,044	67,707,026
D23210	계	8	7,176	583,927	68,408,790
	5 ~ 9	0	0	0	0
	10 ~ 19	0	0	0	0
	20 ~ 49	1	44	1,232	62,182
	50 ~ 99	1	62	2,790	68,637
	100 ~ 199	1	179	8,861	570,945
	500명 이상	5	6,891	571,044	67,707,026
D2322	계	109	1,992	58,449	1,321,643
	5 ~ 9	38	257	5,967	217,730
	10 ~ 19	35	469	11,144	211,887
	20 ~ 49	31	922	27,076	568,481
	50 ~ 99	5	344	14,262	323,545
	100 ~ 199	0	0	0	0
	500명 이상	0	0	0	0

〈표 1-28〉 선형관련식에 의한 감추어진 셀「X」값 찾기 최종 결과(계속)

산업 및 종사자규모		사업체수	월평균 종사자수	급여액	출하액
D23221	계	66	1,210	37,211	1,099,890
	5 ~ 9	22	150	3,231	170,564
	10 ~ 19	23	315	7,300	177,910
	20 ~ 49	18	552	17,452	469,511
	50 ~ 99	3	193	9,228	281,905
	100 ~ 199	0	0	0	0
	500명 이상	0	0	0	0
D23229	계	43	782	21,238	221,753
	5 ~ 9	16	107	2,736	47,166
	10 ~ 19	12	154	3,844	33,977
	20 ~ 49	13	370	9,624	98,970
	50 ~ 99	2	151	5,034	41,640
	100 ~ 199	0	0	0	0
	500명 이상	0	0	0	0
D233	계	1	623	32,048	110,113
	5 ~ 9	0	0	0	0
	10 ~ 19	0	0	0	0
	20 ~ 49	0	0	0	0
	50 ~ 99	0	0	0	0
	100 ~ 199	0	0	0	0
	500명 이상	1	623	32,048	110,113
D2330	계	1	623	32,048	110,113
	5 ~ 9	0	0	0	0
	10 ~ 19	0	0	0	0
	20 ~ 49	0	0	0	0
	50 ~ 99	0	0	0	0
	100 ~ 199	0	0	0	0
	500명 이상	1	623	32,048	110,113
D23300	계	1	623	32,048	110,113
	5 ~ 9	0	0	0	0
	10 ~ 19	0	0	0	0
	20 ~ 49	0	0	0	0
	50 ~ 99	0	0	0	0
	100 ~ 199	0	0	0	0
	500명 이상	1	623	32,048	110,113

3. 현행 SODC[7] 방법 적용에 따른 한계점

통계청에서 실시하고 있는 산업구조통계는 모두 사업체수가 2개 이하인 분류와 관련된 행의 자료를 민감한 셀로 정의하고 보호하는 SODC방법을 적용하고 있다. 그러나 이 방법은 <표 1-23>~<표 1-27>에 제시된 바와 같이, 집계표내의 행과 열의 선형관련식을 이용하여 비밀보호 된 셀 값「X」를 찾을 수 있다.

즉, 중분류 D23의 경우 민감한 셀에 대한 비밀보호가 무의미하다고 볼 수 있으며, 해당 셀에「X」표시 대신 실제 집계자료를 제공해 준 것과 다르지 않다. 결과적으로 SODC방법 적용에 따라, 이용자는「X」값을 찾기 위해 선형관련식을 풀어야 하는 어려움만 있을 뿐, 그 값을 알아낼 수 있다는 데에 그 한계점을 갖는다.

한편, 중분류 D23과 같이 모든 분류에서 감추어진 셀「X」값이 모두 계산되는 것은 아니다. <표 1-29>의 중분류 D27에 속하는 세분류 D2713의 경우를 살펴보자. 먼저 세세분류 D27131의 종사자규모「200~299명」의 감추어진 셀은 그 합계가 주어져 있으므로 쉽게 계산된다. 다음으로 세분류 D2713의 종사자규모「300~499명」,「500명 이상」에서 감추어진 셀 값을 계산하기 위해 선형관련식을 만들면 (1.14)과 같다.

$$D2713_i = D271_i - (D2711_i + D2712_i + D2719_i) \tag{1.14}$$

$D2713_i$ 값을 계산하기 위해 필요한 값 중 $D2711_i$와 $D2719_i$ 값이 사업체수가 2개 이하로서 감추어진 셀이다. 따라서 하나의 식에 2개의 미지수가 발생하여 선형관련식 (1.14)을 이용하여 감추어진 셀「X」값을 계산할 수 없다. 그러나 이와 같은 경우는 우연히 같은 범주의 종사자 규모에 해당하는 사업체 수가 2개 이하로 감추어진 경우이며, 이와 같은 우연한 경우를 제외하고는 SODC방법은 노출위험이 있는 셀을 보호하는데 한계점을 가진다.

7) SODC : Suppressing Only the Disclosure Cells

〈표 1-29〉 감추어진 셀「X」값 찾기 (D2713)

산업 및 종사자규모		사업체수	월평균 종사자수	급여액	출하액
D2713	계	695	18,481	465,891	6,791,073
	5 ~ 9명	221	1,511	26,753	290,258
	10 ~ 19명	217	2,917	59,758	715,556
	20 ~ 49명	178	5,543	124,725	1,680,559
	50 ~ 99명	55	3,578	94,269	1,414,480
	100 ~ 199명	17	2,260	69,610	1,372,520
	200 ~ 299명	4	1,125	33,716	335,659
	300 ~ 499명	1	X	X	X
	500명 이상	2	X	X	X
D27131	계	33	1,503	38,322	280,166
	5 ~ 9명	6	44	705	6,969
	10 ~ 19명	9	115	2,398	27,605
	20 ~ 49명	10	314	7,361	47,650
	50 ~ 99명	4	347	8,163	53,909
	100 ~ 199명	3	405	9,403	44,351
	200 ~ 299명	1	278	10,292	99,682
	300 ~ 499명	0	0	0	0
	500명 이상	0	0	0	0
D27132	계	662	16,978	427,569	6,510,907
	5 ~ 9명	215	1,467	26,048	283,289
	10 ~ 19명	208	2,802	57,360	687,951
	20 ~ 49명	168	5,229	117,364	1,632,909
	50 ~ 99명	51	3,231	86,106	1,360,571
	100 ~ 199명	14	1,855	60,207	1,328,169
	200 ~ 299명	3	847	23,424	235,977
	300 ~ 499명	1	X	X	X
	500명 이상	2	X	X	X

이와 같이 현재 통계청에서 제공되고 있는 집계표 자료는 SODC(Suppressing Only the Disclosure Cells)방법 적용에 따른 한계점으로 인하여, 노출위험이 있는 자료에 대한 비밀보호가 이루어지지 않고 있다. 따라서 다음 절에서는 실증분석을 통해 SODC 방법을 보완한「분류구조변경(Cell Collapsing)방법」과 「보조셀감추기(Complementary Cell Suppression) 방법」을 중심으로 집계표 형태의 자료에 대한 비밀보호 방법을 적용해 보고자 한다.

제4절 실증분석

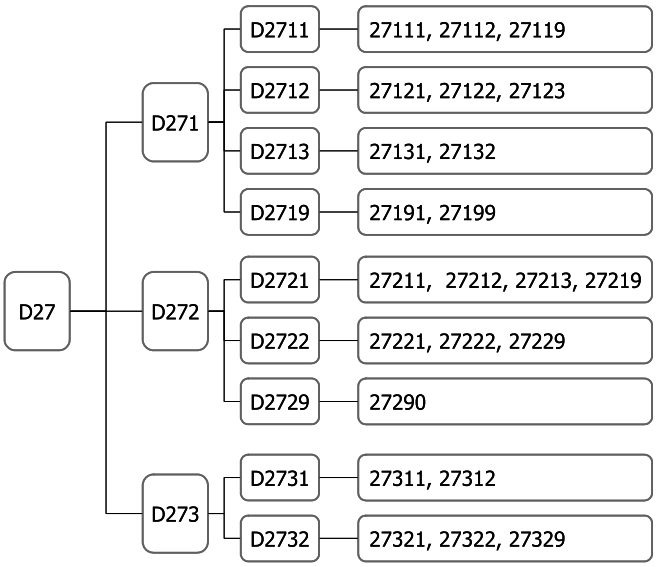
1. 분석 대상

「광업 및 제조업통계조사」는 C.광업 3개, D.제조업 23개 중분류로 구성되어 있다. 분석대상은 이 중 매출액, 유형자산 비중과 소분류 내세분류 및 세세분류별 분포를 기준으로 선정하였다. 분석대상으로 선정한 중분류는 「27. 제1차 금속산업」으로, 분류체계는 [그림 1-3]과 같이 소분류 3개, 세분류 9개, 세세분류 23개로 세분류별내 세세분류가 2개 이상으로 적절하게 분포되어 있다. 또한 <표 1-30>에서 볼 수 있듯이 출하액과 유형자산 규모는 제조업 전체의 약 10% 정도로 큰 비중을 차지한다. 이에 따라 이 분류를 통해 집계표 형태 자료에 관한 비밀보호 방법의 적용에 관한 실증분석을 해보고자 한다.

〈표 1-30〉 분석대상의 주요지표

(단위 : 개, 명, 10억 원)

	사업체수	월평균 종사자수	출하액	부가가치	유형자산 연말잔액
제조업 전체	119,181	2,910,935	909,067	326,844	310,176
중분류27	3,173	117,684	88,721	25,282	30,990
구성비	2.7%	4.0%	9.8%	7.7%	10.0%



[그림 1-3] 제1차 금속산업 분류체계(중분류 27)

2. 비밀보호 현황 및 한계점

다른 산업구조통계조사와 마찬가지로「광업 및 제조업통계조사」결과도 노출위험이 있는 셀만 감추는 SODC(Suppressing Only the Disclosure Cells) 방법을 적용하고 있다. 민감한 셀을 결정하는 기준(base)은 2이다. 즉, 사업체수가 1~2개인 경우 <표 1-29>의「300명~499명」셀과 같이 노출위험이 있는 셀로 정의한 후 해당 셀과 관련된 모든 자료의 제공을 제한한다.

그러나 「광업 및 제조업통계조사」는 물론 산업구조 통계조사에 일반적으로 적용하고 있는 SODC방법은 제2절에서 언급한 바와 같이 집계표내의 열과 행의 선형관련식으로 비밀보호된 셀 값을 계산할 수 있기 때문에 적절한 비밀보호가 이루어지지 않는다.

또한, 집계표 내에 빈도데이터[8]와 크기데이터[9]가 혼합되어 있기 때문에 비밀보호 여부는 전적으로 빈도데이터의 기준(base) 이하 여부에 따라 결정된다. 즉, 사업체수와 기타 크기데이터가 서로 상관(correlate)되어 있기 때문에 사업체수에 따라 노출위험 셀 여부가 결정된다. 이는 크기데이터의 민감한 셀을 결정하는 방법인 (n,k) - dominance rule, p-percent rule 등을 적용하는 데 한계점을 갖게 된다.

---

8) 빈도데이터 : Count data
9) 크기데이터 : Magnitude data

따라서 현재의 자료제공 형태를 유지하면서 효과적으로 자료의 비밀을 보호하는 방법은 셀감추기 방법 중 하나인 보조셀감추기(CCS)를 적용하는 것이 바람직하다. 본 절에서는 분석대상으로 선정된 중분류 27에 대한 보조셀감추기(CCS) 방법의 적용결과를 실증분석하였다. 중 분류 27에 속하는 소분류 272를 중심으로 보조셀감추기 방법을 적용한 비밀보호 과정을 자세히 설명하고 있으며, 중분류 27 전체의 적용 결과표는 부록에 수록하였다.

3. 보조셀감추기(CCS)방법의 적용

CCS방법은 제2절에서 언급한 바와 같이, 노출위험이 있는 셀의 비밀을 보호하기 위해 보조감추기셀(Complementary Supression Cells :CSC)을 찾아 집계표 내의 행과 열의 선형관련식으로 계산되지 않도록 하는 것이다. 이 방법은 열과 행의 총합 및 분포의 변형 등이 없기 때문에 일반적으로 적용되는 방법이다. 그러나 1개의 노출 위험이 있는 민감한 셀(Primary Disclosure Cell :PDC)을 보호하기 위해 일반적으로 3개의 보조감추기셀(CSC)을 추가해야 한다. 이에 따라 보조셀감추기(CCS) 방법은 SODC방법에 비해 노출위험이 있는 셀을 완전히 보호할 수 있는 반면, 감추어진 셀 수는 2~3배 이상 증가할 수 있기 때문에 비밀보호 측면과 정보공개 측면이 상충될 수 있다.

본 절에서는 보조셀감추기(CCS) 방법을 적용한 소분류 272에 대한 결과를 우선 제시하고, 「감추어진 셀 수」와 「감추어진 정보량」을 계산함으로써 적용방법의 효율성을 제고하고자 한다.

〈표 1-31〉 소분류별 사업체수(D272)

산업분류 / 종사자규모	D272	D2721	D2722	D2729
계	748	217	467	64
5 ~ 9	241	72	147	22
10 ~ 19	232	79	137	16
20 ~ 49	169	47	101	21
50 ~ 99	59	11	45	3
100 ~ 199	26	4	21	1
200 ~ 299	8	0	7	1
300 ~ 499	7	2	5	0
500명 이상	6	2	4	0

우선 SODC 방법에 의한 비밀보호 현황을 파악하기 위해 하향식(Top down) 접근방식으로 살펴보면, 기준(base) 2에서 노출위험이 있는 셀은 세분류 D2721과 D2729에서 각각 2개가 발생한다. 그러나 4개의 셀에 대한 관련정보를 「X」표시로 제한하더라도 다음과 같은 선형관련식으로 셀 값을 계산할 수 있다.

$$D2721_i = D272_i - (D2722_i + D2729_i) \tag{1.15}$$

$$D2729_i = D272_i - (D2721_i + D2722_i) \tag{1.16}$$

<표 1-32>에서 계산된 결과로 세분류 D2721의 비밀보호 정도를 보면, 감추어진 셀 8개 중 「50～99명」 행은 관련식(1.17)으로 계산한 후, (1.18)식에 계산된 값을 대입하면 감추어진 셀 값은 모두 계산된다.

$$D27219_i = \sum_{i=1}^{8} D27219_i - \sum_{j \neq i} D27219_j \tag{1.17}$$

$$D27213_i = D2721_i - (D27211_j + D27212_i + D27219_i) \tag{1.18}$$

〈표 1-32〉 세분류별 사업체수(D2721)

산업분류 / 종사자규모	D2721	D27211	D27212	D27213	D27219
계	217	22	128	32	35
5 ~ 9	72	9	40	9	14
10 ~ 19	79	5	53	9	12
20 ~ 49	47	5	25	10	7
50 ~ 99	11	0	7	(2)	(2)
100 ~ 199	4	X(2)	X(2)	0	0
200 ~ 299	0	0	0	0	0
300 ~ 499	(2)	0	X(1)	X(1)	0
500명 이상	(2)	X(1)	0	X(1)	0

나머지 6개의 감추어진 셀 값은 우연히 같은 범주의 셀이 민감한 셀로 선정되어 감추어짐에 따라 보조셀감추기(CCS)방법과 비슷한 형태의 조합이 이루어져 계산이 불가능하다. 기준(base) 2 이하의 SODC방법에 의한 결과, 노출위험이 있는 모든 셀에 대한 비밀보호가 이루어지지 않았다. 따라서 보조셀감추기 방법으로 추가적인 비밀보호가 필요하며, <표 1-31>, <표 1-32>에 대한 소분류 D272의 종사자규모별 집계표의 보조셀 감추기(CCS) 결과는 <표 1-33>과 같다.

[그림 1-3]의 중분류 D27 분류체계를 참조하여, 상위분류의 노출위험셀을 비밀보호하면서 하위분류를 감추기 하는 하향식(Top down) 접근방식으로 보조셀감추기를 수행한 결과, <표 1-33>과 같이 소분류 단계에서는 노출위험이 있는 셀이 없다. 즉, 소분류 단위로 자료를 제공할 경우, 모든 셀 값이 제공 가능함을 알 수 있다.

다음으로 세분류 단계를 살펴보자. 세분류에 관한 보조셀감추기를 위해서는 PDC(Primary disclosure cell)를 중심으로 경로를 구성하여 감추어지는 정보량이 최소인 것을 찾아 보조감추기 셀(CSC)로 정해야 한다. 그 결과는 <표 1-34>와 같으며 그 방법은 다음과 같다.

〈표 1-33〉 보조셀감추기(CCS) 결과 (소분류)

산업분류 / 종사자규모	D271	D272	D273
계	1,805	748	620
5 ~ 9	625	241	162
10 ~ 19	540	232	195
20 ~ 49	418	169	182
50 ~ 99	130	59	59
100 ~ 199	57	26	15
200 ~ 299	11	8	7
300 ~ 499	10	7	0
500명 이상	14	6	0

〈표 1-34〉 보조셀감추기(CCS) 결과 (세분류)

산업분류 / 종사자규모	D2721	D2722	D2729
계	217	467	64
5 ~ 9	72	147	22
10 ~ 19	79	137	16
20 ~ 49	47	101	21
50 ~ 99	CSC(11)	45	CSC(3)
100 ~ 199	CSC(4)	21	PDC(1)
200 ~ 299	0	CSC(7)	PDC(1)
300 ~ 499	PDC(2)	CSC(5)	0
500명 이상	PDC(2)	CSC(4)	0

보조감추기셀(Complementary Suppression Cell :CSC)을 찾는 과정은 셀들의 선형관련성으로 감춘 셀이 계산되지 않도록 하는 수학적인 접근방법이 필요하며, 아울러 감추어지는 셀의 수와 정보량을 최소화하는 보조감추기셀을 찾아야 한다.

목적함수(1) : $V(a) = \min \sum a_{ij} z_{ij}$

목적함수(2) : $V(a) = \min \sum z_{ij}$

☞ a_{ij} : 셀값, $z_{ij} = 1$ (감춘셀), $z_{ij} = 0$ (기타 셀)

이러한 논리적인 접근을 위해 결합기법(combinational technique), 선형기법(linear programming), 수학적 네트워크기법(mathematical network)과 같은 3가지 접근법이 주로 적용된다. 3가지 방법 모두 수학적인 측면에서는 상이하나, 자료제공이 제한되는 셀의 수와 정보량을 최소화하는 목적함수(Objective function)를 정의하고, 이를 이용하여 그 해(解)를 찾아나간다는 점에서는 그 맥락이 같다. 이 때 사용되는 목적함수는 위와 같이 감추어지는 정보량을 최소화하는 관점에서의 목적함수(1)과 감추어지는 셀의 수를 최소화하는 관점에서의 목적함수(2)로 나타낼 수 있다. 이러한 목적함수를 이용하여 보조감추기셀을 찾아간다.

위 내용의 세부과정을 살펴보면 다음과 같은 단계로 요약된다.

Step 1 : <표 1-35-1>과 같이 임의의 PDC(Primary disclosure cell) a_{81}을 출발점으로 가능한 모든 경로(Circuit)를 찾은 후, 각 경로에 해당하는 셀 값의 합을 구한다.

Step 2 : 경로 중 「0」 값을 포함하고 있는 경우, 이를 찾아 제외 시킨다. 왜냐하면 「0」값은 보고서를 통해 이용자에게 제공해야 하는 정보로서 보조감추기셀로 이용할 수 없기 때문이다.

Step 3 : Step1, Step2에서 남은 경로 중 감추어진 셀 값이 최소인 경로를 선택한다. <표 1-35-1>

Step 4 : 모든 PDC에 대한 보조감추기 경로를 찾아 총합을 최소화하는 셀을 선택한다.<표 1-34>

위와 같은 과정으로 a_{81}에 대한 보조감추기셀은 a_{82}, a_{72}, a_{71}된다.

〈표 1-35-1〉 보조감추기셀 찾기($a_{81}=2$)

PDC	Circuit			감추어진 셀값의 합	찾기
(8,1)=2	(8,2)	(7,2)	(7,1)	13	최소경로
	(8,2)	(6,2)	(6,1)	13	X
	(8,2)	(5,2)	(5,1)	31	
	(8,2)	(4,2)	(4,1)	62	
	(8,2)	(3,2)	(3,1)	154	
	(8,2)	(2,2)	(2,1)	222	
	(8,2)	(1,2)	(1,1)	225	
	(8,3)	(7,3)	(7,1)	4	X
	(8,3)	(6,3)	(6,1)	3	X
	(8,3)	(5,3)	(5,1)	8	X
	(8,3)	(4,3)	(4,1)	16	X
	(8,3)	(3,3)	(3,1)	70	X
	(8,3)	(2,3)	(2,1)	97	X
	(8,3)	(1,3)	(1,1)	96	X

동일한 방법으로 PDC a_{71}, a_{53}, a_{63}에 대한 최소경로 탐색결과는 아래와 같다.

〈표 1-35-2〉 보조감추기셀 찾기(a_{71}, a_{53}, a_{63})

PDC	Circuit			감추어진 셀값의 합	찾기
(7,1)=2	(7,2)	(8,2)	(8,1)	13	최소경로
(5,3)=1	(5,1)	(4,1)	(4,3)	19	최소경로
(6,3)=1	(6,2)	(5,2)	(5,3)	30	최소경로

위와 같은 최소경로 탐색결과를 바탕으로 보조감추기셀을 찾은 결과, a_{41}, a_{43}, a_{51}, a_{52}, a_{62}, a_{72}, a_{82}가 보조감추기셀(Complementary Suppression Cell : CSC)로 결정되었다. 그러나 이 경우 각각의 PDC에 대한 최소경로의 탐색에서 나아가 감추어지는 셀 전체의 관점에서 보면, 5번행에서 과대감추기(Over Suppression)가 발생되었음을 알 수 있다. 따라서 과대감추기셀인 a_{52}는 보조감추기셀에서 제외한다.

〈표 1-36〉 보조셀감추기(CCS) 결과(세세분류)

산업분류 / 종사자규모	D2721	D27211	D27212	D27213	D27219
계	217	22	128	32	35
5 ~ 9	72	9	40	9	14
10 ~ 19	79	5	53	9	12
20 ~ 49	47	CSC(5)	CSC(25)	CSC(10)	CSC(7)
50 ~ 99	CSC(11)	0	CSC(7)	PDC(2)	PDC(2)
100 ~ 199	4	PDC(2)	PDC(2)	0	0
200 ~ 299	0	0	0	0	0
300 ~ 499	PDC(2)	0	PDC(1)	PDC(1)	0
500명 이상	PDC(2)	PDC(1)	0	PDC(1)	0

〈표 1-36〉 보조셀 감추기 결과(세세분류)(계속)

산업분류 / 종사자규모	D2722	D27221	D27222	D27229
계	467	182	240	45
5 ~ 9	147	51	75	21
10 ~ 19	137	53	75	9
20 ~ 49	101	42	49	10
50 ~ 99	45	CSC(23)	20	PDC(2)
100 ~ 199	CSC(21)	CSC(7)	12	PDC(2)
200 ~ 299	CSC(7)	PDC(2)	CSC(4)	PDC(1)
300 ~ 499	CSC(5)	PDC(2)	CSC(3)	0
500명 이상	CSC(4)	PDC(2)	PDC(2)	0

〈표 1-36〉 보조셀 감추기 결과(세세분류)(계속)

산업분류 / 종사자규모	D2729	D27290
계	64	64
5 ~ 9	22	22
10 ~ 19	16	16
20 ~ 49	21	21
50 ~ 99	CSC(3)	CSC(3)
100 ~ 199	PDC(1)	PDC(1)
200 ~ 299	PDC(1)	PDC(1)
300 ~ 499	0	0
500명 이상	0	0

4. 타당성 검증

보조셀감추기(CCS) 방법을 적용해본 결과, 기존의 SODC[10]방법과 본 절에서 적용한 방법에 어떠한 차이가 있는지 비교하여 그 차이에 대해 살펴보고자 한다. 또한 보조셀 감추기 방법에 의해 적용된 위의 결과가 선형관련식에 의해 재계산되지 않는지를 확인하는 검증을 실시하고자 한다.

먼저, 소분류 D272의 종사자규모별 자료에 대해 보조셀감추기 방법을 적용한 결과와 기존 적용하고 있는 SODC방법의 결과에 대해 「감추어진 셀 수」와 「감추어진 정보량」 측면에서 각각 비교해 보았다. 본절에서 적용한 보조셀감추기 방법이 기존 방법보다 감추어진 셀 수」와「감추어진 정보량」은 각각 증가하는 것을 <표 1-37>을 통해 알 수 있다. 그러나 보조셀감추기 방법으로 노출 위험이 있는 민감한 셀(PDC)을 비밀보호하기 위해 추가적으로 제한된 정보량은 약 6.8%로, SODC방법이 민감한 셀(PDC)을 선형관련식으로 재계산해낼 수 있다는 한계점을 극복한다는 점에서 높은 수준은 아니다.

10) SODC : Suppressing Only the Disclosure Cells

〈표 1-37〉 CCS와 SODC방법 적용결과 비교 (감추어진 셀의 수)

산업분류		D272	D2721	D2722	D2729
원자료	셀수(A)	24	8	8	8
CCS방법	감추어진 셀수(B)	10	4	3	3
	C=(B/A)×100	41.7	50.0	37.5	37.5
SODC방법	감추어진 셀수(D)	4	2	0	2
	E=(B/A)×100	16.7	25.0	0.0	25.0
Difference	**C - E**	**25.0**	**25.0**	**37.5**	**12.5**

〈표 1-38〉 CCS와 SODC방법 적용결과 비교 (감추어진 정보량)

산업분류		D272	D2721	D2722	D2729
원자료	정보량(H)	748	217	467	64
CCS방법	감추어진 정보량(I)	40	19	16	5
	J=(I/H)×100	5.4	8.8	3.4	7.8
SODC방법	감추어진 정보량(K)	6	4	0	2
	L=(K/H)×100	0.8	1.8	0.0	3.1
Difference	**J - L**	**4.6**	**7.0**	**3.4**	**4.7**

본 절에서 제시한 실증분석 결과, 보조셀감추기(CCS) 방법은 집계표 내 선형관련식에 의해 감추어진 셀 값이 재계산될 수 있는 SODC방법의 한계점을 보완하였다. 그러나 추가적으로 보조셀을 찾아 감추기 해야 하는 등 제한되는 셀 수는 평균 2배 수준, 정보량은 구간에 따라 차이를 보이고 있으나 SODC방법의 결과보다는 증가한다. 이와 같은 감추어진 정보량의 증가는 보조셀감추기(CCS)방법의 문제점이 아니라, SODC방법에서 충분한 비밀보호를 위해 정보를 제한하지 않은 차이로 기인된다.

다음으로는 보조셀감추기 결과가 집계표 내 행과 열의 선형관련식에 의해 재계산되는지 <표 1-39>을 이용하여 디버깅(debugging) 과정을 살펴본다.

〈표 1-39〉 보조셀감추기(CCS) 결과 (세분류)

산업분류 / 종사자규모	D272	D2721	D2722	D2729
계	748	217	467	64
5 ~ 9	241	72	147	22
10 ~ 19	232	79	137	16
20 ~ 49	169	47	101	21
50 ~ 99	59	CSC(11)	45	CSC(3)
100 ~ 199	26	CSC(4)	21	PDC(1)
200 ~ 299	8	0	CSC(7)	PDC(1)
300 ~ 499	7	PDC(2)	CSC(5)	0
500명 이상	6	PDC(2)	CSC(4)	0

<표 1-39>에서 감추어진 셀은 10개의 셀a_{41}, a_{51}, a_{71}, a_{81}, a_{62}, a_{72}, a_{82}, a_{43}, a_{53}, a_{63} 이다. 선형관련식으로 표현하면 다음과 같은데, 이들 식을 통해 감추어진 셀 값에 대한 유일한 해(Unique solution)를 찾을 수 없으며, 보조감추기셀(CSC) 값의 합도 최소이기 때문에 보조셀감추기(CCS)가 최적(Optimal)으로 이루어진 것으로 평가한다.

$$
\begin{aligned}
a_{41} + a_{43} &= 14 \\
a_{51} + a_{53} &= 5 \\
a_{62} + a_{63} &= 8 \\
a_{71} + a_{72} &= 7 \\
a_{81} + a_{82} &= 6 \\
a_{41} + a_{51} + a_{71} + a_{81} &= 19 \\
a_{62} + a_{72} + a_{82} &= 16 \\
a_{43} + a_{53} + a_{63} &= 5 \\
a_{ij} &\geq 1
\end{aligned}
$$

<표 1-36>에 대한 결과도 위와 같은 동일한 방법을 적용하여 선형관련식으로 나타낼 수 있으며, 이 경우도 유일한 해(Unique solution)를 찾을 수 없다.

5. 실증분석 결과 요약

실증분석을 통해 얻은 결론을 요약하면 다음과 같다. 집계표 구성형태를 어떻게 할 것인가의 측면에서 살펴보면, 첫째로 기존 집계표 형태를 유지하는 방안과 둘째로 집계표 구성형태를 재조정하는 방안을 생각해 볼 수 있다.

먼저 기존 집계표 형태를 유지하기 위해서는 노출위험이 있는 셀만 감추는 기존방법(Suppression Only the Disclosure Cells: SODC)을 보완한 보조셀감추기(CCS)방법을 적용하여 추가적인 셀감추기를 실시해야 한다. 그러나 이 방법을 적용할 때, 모든 세부분류가 산업별로 동일한 기준으로 적용되고 있기 때문에, 산업별 특성에 따라 특정 셀에서 0 값 또는 기준(base) 이하의 노출위험이 있는 셀이 편중되어 나타나고 있다. 이와 같은 특징은 보조셀감추기 방법 적용 시 실행가능한 감추기경로(feasible circuit)를 찾지 못하는 경우가 나타날 수 있기 때문에 특정 집계표 전체를 감추기 해야 하는 문제점이 발생한다. 따라서 산업별 또는 세부분류별 분포를 미리 파악하여 「분류구조변경(Cell Collapsing)방법」과 「보조셀감추기(CCS)방법」을 병행하여 노출위험 셀을 감추기 하는 방법을 적용해야 한다.

두 번째로 집계표 구성형태를 재조정하는 방법을 살펴보자. 기존 집계표의 대부분은 사업체수(빈도데이터)와 기타 항목(크기데이터)이 혼합되어 있다. 그러나 노출위험 셀을 결정하고 비밀보호를 적용하는 기준변수(key variable)는 빈도데이터인 사업체수에 의존하고 있기 때문에 자료의 종류에 따라 다양한 비밀보호 기법을 적용하는 데 한계점을 갖게 된다. 따라서 집계표 구성형태를 재검토하여 빈도데이터와 크기데이터를 구분하여 집계표를 작성하는 것이 바람직하다고 판단된다.

끝으로 노출위험이 있는 셀을 결정하는 기준(Base)에 대해 살펴보면, 통계청에서는 그 기준을 2로 정의하고 있다. 그러나 이 경우 특정 업체의 정보를 알고 있을 경우 다른 업체의 정보가 추정가능하기 때문에 비밀보호가 효과적으로 이루어졌다고 보기 어렵다. 따라서 최소한 기준(Base)을 3이상으로 상향하는 것을 검토할 필요가 있다.

제5절 결론

본 연구는 집계표 데이터의 비밀보호방법에 관한 실증연구로서, 미국 국립보건통계센터(National Center for Health Statistics)에서 동 분야에 대한 선행연구를 수행한 Dr. Lawerence Henry cox의 연구를 바탕으로 하였다. 이 연구의 목적은 집계표 데이터의 비밀보호 기법에 관한 고찰을 통해 통계작성자가 이 문헌연구와 예시 등을 바탕으로,

추후 집계표 데이터의 비밀보호기법을 적용할 때 용이하게 접근할 수 있도록 방법론을 소개하고, 통계청 ×산업구조통계를 중심으로 집계표 데이터의 비밀보호현황분석과 한계점을 도출하는 데 있다. 또한, 한계점을 극복하기 위한 새로운 기법을 선정하여 실증분석을 통해 우리나라의 실정에 적합한 비밀보호 기법을 도출하는 데 기여하고자 한다.

이에 따라 제2절에서는 전통적으로 적용하고 있는 비밀보호기법 선행연구 결과를 참조하여 방법론을 소개하였다. 외국의 적용사례를 통한 우리나라에 적합한 방법론 선정을 위해, 먼저 우리나라와 미국의 제조업통계를 중심으로 집계표의 자료형태를 비교하여 차이점을 분석하였다. 이는 비밀보호기법이 집계표를 구성하고 있는 자료형태 즉, 빈도데이터(Count Data) 및 크기데이터(Magnitude Data) 여부에 따라 각각 다르게 적용되고 있기 때문이다. 자료형태에 따른 비밀보호 대상을 찾기 위해 노출위험이 있는 민감한 셀을 탐색하는 방법과 이 셀들을 보호할 수 있는 방법론도 본 절에서 살펴보았다.

제3절에서는 통계청에서 작성 중인 산업구조통계의 비밀보호 현황을 광업 및 제조업 통계조사를 중심으로 살펴보았다. 광업 및 제조업통계조사는 노출위험 셀을 결정하는 기준(Base)을 2로 정하고 노출위험이 있는 셀만 감추기 하는 SODC(Suppression Only the Disclosure Cells)방법을 적용하고 있다. 그러나 이 방법은 집계표 내의 선형관련식을 이용하여 대부분의 감추기한 셀 값을 재계산할 수 있기 때문에 민감한 셀의 비밀보호에 한계점이 있음을 발견하였다.

제4절에서는 제3절에서 도출된 한계점을 보완하기 위해, 광업 및 제조업 통계조사 23개 중분류 중 매출액, 유형자산 비중과 소분류 내 세분류 및 세세분류별 분포를 고려하여 실증분석대상 중분류를 선정하여 보완된 비밀보호 기법을 적용하고 그 결과를 분석하였다. 실증분석 결과, 기존 집계표 형태를 유지하기 위해서는 노출위험이 있는 셀만 감추는기존의 SODC방법을 보완한 보조셀감추기(CCS) 방법을 적용하여 추가적인 셀감추기를 실시해야 한다는 결론을 도출하였다. 그러나 모든 세부분류가 산업별로 동일한 기준으로 적용되고 있기 때문에, 산업별 특성에 따라 특정 셀에서 0 값 또는 기준(base) 이하의 노출위험이 있는 셀이 편중되어 나타나고 있다. 또한 보조셀감추기(CCS) 방법 적용 시 실행가능한 감추기경로(feasible circuit)를 찾지 못하는 경우가 나타날 수 있기 때문에 특정 집계표 전체를 감추기 해야 하는 문제점이 발생한다. 따라서 산업별 또는 세부 분류별 분포를 미리 파악하여 「분류구조변경(Cell Collapsing) 방법」과 「보조셀감추기(CCS)방법」을 병행하여 노출 위험 셀을 감추기 하는 방법을 적용해야 한다.

그러나 현재 제공되고 있는 산업구조통계의 집계표 구성형태가 빈도데이터(count

data)인 사업체수와 크기데이터(magnitude data)인 기타 항목이 혼합되어 있기 때문에, 노출위험 셀을 결정하는 기준변수(key variable)는 빈도데이터인 사업체수에 의존할 수밖에 없다. 이에 따라 현재의 집계표형태를 가지고는 적용할 수 있는 비밀보호 기법이 매우 한정적이다. 따라서 집계표 구성형태에 따라 다양하고 효과적인 비밀보호기법을 적용하기 위해서는 빈도데이터와 크기데이터를 구분하여 구성하는 형태로 집계표를 재조정하고, 이에 해당하는 다양한 비밀보호기법을 적용하는 것에 관련한 추가적인 후속연구가 필요할 것이다.

정보에 대한 이용자와 응답자는 창과 방패의 관계이다. 정보공개를 확대하는 것에 앞서 중요하게 생각해야 하는 것은 정보의 원천인 응답자의 민감한 정보를 보호하는 것이다. 고갈되어 가는 원천에서 좋은 정보의 생산을 기대하기는 어렵기 때문이다. 본 연구를 계기로 현재 적용 중인 비밀보호기법이 재검토되어 응답자의 개별정보 노출에 대한 불안감을 제거함은 물론 정보이용자와 정보제공자(응답자) 간의 중재기능으로 본 연구결과가 활용되기를 기대한다.

참고문헌

Cox, L.H. (2008), “Statistical Disclosure Limitation Methods For Tabular Data and Effects on Data Quality”.

(1980), “Suppression Methodology and Statistical Disclosure Control,” *Journal of the American statistical Association* **75**, 377-385.

(1981), “Linear Sensitivity Measures in Statistical Disclosure Control,” *Journal of Statistical Planning and Inference* **5**, 153-164.

(1994), “Matrix Masking Methods for Disclosure Limitation in Microdata,” *Survey Methdodology* **20**, 165-169.

(1995), “Network Models for Complementary Cell Suppression,” *Journal of the American statistical Association* **90**, 1453-1462.

(2000), “Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints,” *Journal of the American statistical Association* **95**, 916-928.

Gonzalez, J.F. and L.H. Cox (2005), “Software for Tabular Data Protection,” *Statistics in Medicine* **24(4)**, 659-669.

<부 록>

〈부표 1-1〉 종사자규모별 주요지표의 보조셀감추기 방법 적용 결과

산업별		종사자규모별	2006			
			사업체수	종사자수	급여액	출하액
제 1차 금속산업	D27	계	3,173	117,484	3,858	88,721
		5 ~ 9	1,028	6,954	128	1,709
		10 ~ 19	967	13,113	278	3,794
		20 ~ 49	769	23,538	567	10,955
		50 ~ 99	248	16,982	475	9,330
		100 ~ 199	98	13,624	437	9,348
		200 ~ 299	26	6,350	203	3,011
		300 ~ 499	17	6,303	220	7,157
		500명 이상	20	30,620	1,550	43,417
제 1차 철강산업	D271	계	1,805	74,240	2,708	60,595
		5 ~ 9	625	4,235	78	1,060
		10 ~ 19	540	7,255	154	2,066
		20 ~ 49	418	13,016	323	6,448
		50 ~ 99	130	8,729	248	5,427
		100 ~ 199	57	7,739	264	5,302
		200 ~ 299	11	2,682	92	1,543
		300 ~ 499	10	3,985	155	5,584
		500명 이상	14	26,599	1,393	33,165
제철, 제강 및 합금철 제조업	D2711	계	47	26,522	1,384	32,921
		5 ~ 9	12	79	1	23
		10 ~ 19	8	98	3	46
		20 ~ 49	7	216	5	132
		50 ~ 99	2	X	X	X
		100 ~ 199	4	532	23	338
		200 ~ 299	1	X	X	X
		300 ~ 499	2	X	X	X
		500명 이상	11	24,537	1,312	31,384

〈부표 1-1〉의 계속

산업별		종사자규모별	2006			
			사업체수	종사자수	급여액	출하액
제철 및 제강업	D27111	계	29	25,914	1,365	32,507
		5 ~ 9	7	47	0	8
		10 ~ 19	3	38	1	21
		20 ~ 49	2	…	…	…
		50 ~ 99	1	…	…	…
		100 ~ 199	2	…	…	…
		200 ~ 299	1	…	…	…
		300 ~ 499	2	…	…	…
		500명 이상	11	24,537	1,312	31,384
합금철 제조업	D27112	계	18	608	19	413
		5 ~ 9	5	32	1	16
		10 ~ 19	5	60	1	25
		20 ~ 49	5	162	4	124
		50 ~ 99	1	…	…	…
		100 ~ 199	2	…	…	…
		200 ~ 299	0	0	0	0
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
철강압연·압출 및 연신제품	D2712	계	359	12,989	404	9,194
		5 ~ 9	117	797	16	276
		10 ~ 19	90	1,205	25	385
		20 ~ 49	85	2,774	72	1,459
		50 ~ 99	38	2,672	85	1,529
		100 ~ 199	21	3,010	104	2,207
		200 ~ 299	4	854	30	643
		300 ~ 499	4	1,677	72	2,695
		500명 이상	0	0	0	0
열간 압연 및 압출제품 제조업	D27121	계	147	4,287	122	2,274
		5 ~ 9	43	293	5	60
		10 ~ 19	34	447	9	137
		20 ~ 49	47	1,534	38	559
		50 ~ 99	17	1,165	38	830
		100 ~ 199	5	X	X	X
		200 ~ 299	1	X	X	X
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0

보조셀 감추기 경로 없음! 산업분류 통합 필요!

〈부표 1-1〉의 계속

산업별		종사자규모별	2006			
			사업체수	종사자수	급여액	출하액
냉간 압연 및 압출제품 제조업	D27122	계	77	3,565	122	3,252
		5 ~ 9	19	129	3	69
		10 ~ 19	21	284	6	103
		20 ~ 49	21	703	19	656
		50 ~ 99	9	713	23	276
		100 ~ 199	4	X	X	X
		200 ~ 299	1	X	X	X
		300 ~ 499	2	X	X	X
		500명 이상	0	0	0	0
철강선 제조업	D27123	계	135	5,137	161	3,668
		5 ~ 9	55	375	8	147
		10 ~ 19	35	474	10	145
		20 ~ 49	17	537	15	245
		50 ~ 99	12	794	24	422
		100 ~ 199	12	1,709	61	1,046
		200 ~ 299	2	X	X	X
		300 ~ 499	2	X	X	X
		500명 이상	0	0	0	0
철강관 제조업	D2713	계	695	18,481	466	6,791
		5 ~ 9	221	1,511	27	290
		10 ~ 19	217	2,917	60	716
		20 ~ 49	178	5,543	125	1,681
		50 ~ 99	55	3,578	94	1,414
		100 ~ 199	17	2,260	70	1,373
		200 ~ 299	4	1,125	34	336
		300 ~ 499	1	X	X	X
		500명 이상	2	X	X	X
주철관 제조업	D27131	계	33	1,503	38	280
		5 ~ 9	6	44	1	7
		10 ~ 19	9	115	2	28
		20 ~ 49	10	314	7	48
		50 ~ 99	4	347	8	54
		100 ~ 199	3	405	9	44
		200 ~ 299	1	…	…	…
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0

보조셀 감추기 경로 없음! 산업분류 통합 필요!

〈부표 1-1〉의 계속

산업별		종사자규모별	2006			
			사업체수	종사자수	급여액	출하액
강관 제조업	D27132	계	662	16,978	428	6,511
보조셀 감추기 경로 없음! 산업분류 통합 필요!		5 ~ 9	215	1,467	26	283
		10 ~ 19	208	2,802	57	688
		20 ~ 49	168	5,229	117	1,633
		50 ~ 99	51	3,231	86	1,361
		100 ~ 199	14	1,855	60	1,328
		200 ~ 299	3	847	23	236
		300 ~ 499	1	…	…	…
		500명 이상	2	…	…	…
기타 철강산업	D2719	계	704	16,248	454	11,689
		5 ~ 9	275	1,848	35	470
		10 ~ 19	225	3,035	67	920
		20 ~ 49	148	4,483	121	3,175
		50 ~ 99	35	X	X	X
		100 ~ 199	15	1,937	68	1,385
		200 ~ 299	2	X	X	X
		300 ~ 499	3	X	X	X
		500명 이상	1	X	X	X
절단가공 및 표면처리강재 생산업	D27191	계	698	16,181	453	11,685
보조셀 감추기 경로 없음! 산업분류 통합 필요!		5 ~ 9	273	1,832	34	469
		10 ~ 19	221	2,984	66	918
		20 ~ 49	148	4,483	121	3,175
		50 ~ 99	35	2,324	64	2,376
		100 ~ 199	15	1,937	68	1,385
		200 ~ 299	2	…	…	…
		300 ~ 499	3	1,269	44	2,129
		500명 이상	1	…	…	…
그외 기타 철강산업	D27199	계	6	67	1	4
보조셀 감추기 경로 없음! 산업분류 통합 필요!		5 ~ 9	2	…	…	…
		10 ~ 19	4	51	1	2
		20 ~ 49	0	0	0	0
		50 ~ 99	0	0	0	0
		100 ~ 199	0	0	0	0
		200 ~ 299	0	0	0	0
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0

〈부표 1-1〉의 계속

산업별		종사자규모별	2006			
			사업체수	종사자수	급여액	출하액
제 1차 비철금속산업	D272	계	748	26,147	743	24,156
		5 ~ 9	241	1,641	31	504
		10 ~ 19	232	3,184	72	1,318
		20 ~ 49	169	5,044	120	3,529
		50 ~ 99	59	4,252	120	2,888
		100 ~ 199	26	3,741	118	3,008
		200 ~ 299	8	1,946	60	1,084
		300 ~ 499	7	2,318	65	1,573
		500명 이상	6	4,021	158	10,252
비철금속 제련, 정련 및 합금	D2721	계	217	6,315	191	11,164
		5 ~ 9	72	478	9	209
		10 ~ 19	79	1,116	27	683
		20 ~ 49	47	1,402	35	2,155
		50 ~ 99	11	X	X	X
		100 ~ 199	4	553	15	554
		200 ~ 299	0	0	0	0
		300 ~ 499	2	X	X	X
		500명 이상	2	X	X	X
동 제련, 정련 및 합금 제조업	D27211	계	22	1,180	48	5,098
		5 ~ 9	9	61	1	22
		10 ~ 19	5	73	2	37
		20 ~ 49	5	X	X	X
		50 ~ 99	0	0	0	0
		100 ~ 199	2	X	X	X
		200 ~ 299	0	0	0	0
		300 ~ 499	0	0	0	0
		500명 이상	1	X	X	X
알루미늄 제련, 정련 및 합금	D27212	계	128	2,853	75	2,196
		5 ~ 9	40	272	5	101
		10 ~ 19	53	754	18	489
		20 ~ 49	25	X	X	X
		50 ~ 99	7	X	X	X
		100 ~ 199	2	X	X	X
		200 ~ 299	0	0	0	0
		300 ~ 499	1	X	X	X
		500명 이상	0	0	0	0

〈부표 1-1〉의 계속

산업별		종사자규모별	2006			
			사업체수	종사자수	급여액	출하액
연 및 아연제련, 정련 및 합금	D27213	계	32	1,694	54	3,030
		5 ~ 9	9	56	1	19
		10 ~ 19	9	133	3	56
		20 ~ 49	10	321	6	98
		50 ~ 99	2	X	X	X
		100 ~ 199	0	0	0	0
		200 ~ 299	0	0	0	0
		300 ~ 499	1	X	X	X
		500명 이상	1	X	X	X
기타 비철금속제련,정련 및 합금	D27219	계	35	588	15	841
		5 ~ 9	14	89	2	67
		10 ~ 19	12	156	3	101
		20 ~ 49	7	X	X	X
		50 ~ 99	2	X	X	X
		100 ~ 199	0	0	0	0
		200 ~ 299	0	0	0	0
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
비철금속압연,압출 및 연신제품	D2722	계	467	18,357	516	12,429
		5 ~ 9	147	1,019	19	256
		10 ~ 19	137	1,854	41	577
		20 ~ 49	101	3,050	72	1,129
		50 ~ 99	45	3,252	90	1,948
		100 ~ 199	21	X	X	X
		200 ~ 299	7	X	X	X
		300 ~ 499	5	X	X	X
		500명 이상	4	X	X	X
동 압연, 압출 및 연신제품 제조업	D27221	계	182	7,891	216	7,316
		5 ~ 9	51	356	7	109
		10 ~ 19	53	750	18	281
		20 ~ 49	42	1,257	27	512
		50 ~ 99	23	X	X	X
		100 ~ 199	7	X	X	X
		200 ~ 299	2	X	X	X
		300 ~ 499	2	X	X	X
		500명 이상	2	X	X	X

〈부표 1-1〉의 계속

산업별		종사자규모별	2006			
			사업체수	종사자수	급여액	출하액
알루미늄 압연,압출 및 연신제품	D27222	계	240	9,214	262	4,433
		5 ~ 9	75	521	9	123
		10 ~ 19	75	988	21	267
		20 ~ 49	49	1,446	35	464
		50 ~ 99	20	1,400	38	503
		100 ~ 199	12	1,799	54	627
		200 ~ 299	4	X	X	X
		300 ~ 499	3	X	X	X
		500명 이상	2	X	X	X
기타 비철금속 압연,압출 및 연신제품	D27229	계	45	1,252	38	679
		5 ~ 9	21	142	3	24
		10 ~ 19	9	116	2	29
		20 ~ 49	10	347	10	153
		50 ~ 99	2	X	X	X
		100 ~ 199	2	X	X	X
		200 ~ 299	1	X	X	X
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
기타 제 1차 비철금속산업	D2729	계	64	1,475	37	563
		5 ~ 9	22	144	3	39
		10 ~ 19	16	214	5	58
		20 ~ 49	21	592	13	245
		50 ~ 99	3	X	X	X
		100 ~ 199	1	X	X	X
		200 ~ 299	1	X	X	X
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
기타 제 1차 비철금속산업	D27290	계	64	1,475	37	563
		5 ~ 9	22	144	3	39
		10 ~ 19	16	214	5	58
		20 ~ 49	21	592	13	245
		50 ~ 99	3	X	X	X
		100 ~ 199	1	X	X	X
		200 ~ 299	1	X	X	X
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0

〈부표 1-1〉의 계속

산업별		종사자규모별	2006			
			사업체수	종사자수	급여액	출하액
금속 주조업	D273	계	620	17,097	407	3,970
		5 ~ 9	162	1,078	19	145
		10 ~ 19	195	2,674	51	411
		20 ~ 49	182	5,478	124	978
		50 ~ 99	59	4,001	107	1,015
		100 ~ 199	15	2,144	55	1,037
		200 ~ 299	7	1,722	51	384
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
철강 주조업	D2731	계	392	12,320	303	2,318
		5 ~ 9	75	519	9	60
		10 ~ 19	115	1,595	30	231
		20 ~ 49	141	4,255	97	691
		50 ~ 99	43	2,858	80	643
		100 ~ 199	12	X	X	X
		200 ~ 299	6	X	X	X
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
선철주물 주조업	D27311	계	217	7,566	189	1,470
		5 ~ 9	36	265	5	28
		10 ~ 19	62	869	16	116
		20 ~ 49	83	2,546	56	401
		50 ~ 99	24	1,648	46	337
		100 ~ 199	7	X	X	X
		200 ~ 299	5	X	X	X
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
강주물 주조업	D27312	계	175	4,754	113	847
		5 ~ 9	39	254	4	33
		10 ~ 19	53	726	14	115
		20 ~ 49	58	1,709	41	291
		50 ~ 99	19	1,210	34	307
		100 ~ 199	5	X	X	X
		200 ~ 299	1	X	X	X
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0

〈부표 1-1〉의 계속

산업별		종사자규모별	2006			
			사업체수	종사자수	급여액	출하액
비철금속 주조업	D2732	계	228	4,777	104	1,652
		5 ~ 9	87	559	9	85
		10 ~ 19	80	1,079	21	179
		20 ~ 49	41	1,223	27	287
		50 ~ 99	16	1,143	28	372
		100 ~ 199	3	X	X	X
		200 ~ 299	1	X	X	X
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
알루미늄주물 주조업	D27321	계	138	2,985	64	700
		5 ~ 9	51	325	5	47
		10 ~ 19	48	653	12	115
		20 ~ 49	25	781	18	211
		50 ~ 99	12	862	20	246
		100 ~ 199	2	…	…	…
		200 ~ 299	0	0	0	0
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
동주물 주조업	D27322	계	39	849	20	211
		5 ~ 9	17	110	2	18
		10 ~ 19	10	124	3	23
		20 ~ 49	10	283	7	58
		50 ~ 99	1	…	…	…
		100 ~ 199	0	0	0	0
		200 ~ 299	1	…	…	…
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
기타 비철금속 주조업	D27329	계	51	943	21	742
		5 ~ 9	19	124	2	20
		10 ~ 19	22	302	6	41
		20 ~ 49	6	159	3	17
		50 ~ 99	3	189	5	37
		100 ~ 199	1	…	…	…
		200 ~ 299	0	0	0	0
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0

보조셀 감추기 경로 없음! 산업분류 통합 필요!

보조셀 감추기 경로 없음! 산업분류 통합 필요!

보조셀 감추기 경로 없음! 산업분류 통합 필요!

〈부표 1-1〉의 계속

산업별		종사자규모별	2006			
			사업체수	주요생산비	부가가치	유형자산
제철 및 제강업	D27111	계	29	20,909	12,000	15,501
		5 ~ 9	7	5	3	2
		10 ~ 19	3	17	4	3
		20 ~ 49	2	…	…	…
		50 ~ 99	1	…	…	…
		100 ~ 199	2	…	…	…
		200 ~ 299	1	…	…	…
		300 ~ 499	2	…	…	…
		500명 이상	11	20,103	11,686	15,124
합금철 제조업	D27112	계	18	321	103	148
		5 ~ 9	5	12	3	4
		10 ~ 19	5	15	15	6
		20 ~ 49	5	117	12	30
		50 ~ 99	1	…	…	…
		100 ~ 199	2	…	…	…
		200 ~ 299	0	0	0	0
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
철강압연·압출 및 연신제품	D2712	계	359	7,200	2,042	3,789
		5 ~ 9	117	188	86	79
		10 ~ 19	90	292	98	139
		20 ~ 49	85	1,128	333	324
		50 ~ 99	38	1,152	388	502
		100 ~ 199	21	1,676	500	650
		200 ~ 299	4	281	342	128
		300 ~ 499	4	2,483	295	1,968
		500명 이상	0	0	0	0
열간 압연 및 압출제품 제조업	D27121	계	147	1,748	532	621
		5 ~ 9	43	41	20	26
		10 ~ 19	34	100	39	39
		20 ~ 49	47	406	152	153
		50 ~ 99	17	667	174	251
		100 ~ 199	5	X	X	X
		200 ~ 299	1	X	X	X
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0

보조셀 감추기 경로 없음! 산업분류 통합 필요!

〈부표 1-1〉의 계속

산업별		종사자규모별	2006			
			사업체수	주요생산비	부가가치	유형자산
냉간 압연 및 압출제품 제조업	D27122	계	77	2,482	806	1,300
		5 ~ 9	19	32	34	14
		10 ~ 19	21	81	24	54
		20 ~ 49	21	539	119	105
		50 ~ 99	9	214	65	109
		100 ~ 199	4	X	X	X
		200 ~ 299	1	X	X	X
		300 ~ 499	2	X	X	X
		500명 이상	0	0	0	0
철강선 제조업	D27123	계	135	2,970	704	1,868
		5 ~ 9	55	115	32	39
		10 ~ 19	35	111	35	45
		20 ~ 49	17	183	62	66
		50 ~ 99	12	271	149	142
		100 ~ 199	12	784	249	352
		200 ~ 299	2	X	X	X
		300 ~ 499	2	X	X	X
		500명 이상	0	0	0	0
철강관 제조업	D2713	계	695	4,657	2,229	1,976
		5 ~ 9	221	189	103	85
		10 ~ 19	217	485	238	226
		20 ~ 49	178	1,193	499	548
		50 ~ 99	55	1,053	398	380
		100 ~ 199	17	792	594	265
		200 ~ 299	4	199	154	130
		300 ~ 499	1	X	X	X
		500명 이상	2	X	X	X
주철관 제조업	D27131	계	33	152	126	147
		5 ~ 9	6	5	2	4
		10 ~ 19	9	17	11	11
		20 ~ 49	10	29	19	14
		50 ~ 99	4	31	26	40
		100 ~ 199	3	28	16	17
		200 ~ 299	1	…	…	…
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0

보조셀 감추기 경로 없음! 산업분류 통합 필요!

〈부표 1-1〉의 계속

산업별		종사자규모별	2006			
			사업체수	주요생산비	부가가치	유형자산
강관 제조업	D27132	계	662	4,505	2,104	1,828
보조셀 감추기 경로 없음! 산업분류 통합 필요!		5 ~ 9	215	185	100	81
		10 ~ 19	208	469	226	215
		20 ~ 49	168	1,163	480	534
		50 ~ 99	51	1,022	372	340
		100 ~ 199	14	764	578	248
		200 ~ 299	3	156	102	67
		300 ~ 499	1	…	…	…
		500명 이상	2	…	…	…
기타 철강산업	D2719	계	704	8,951	2,783	3,719
		5 ~ 9	275	327	144	200
		10 ~ 19	225	640	278	297
		20 ~ 49	148	2,551	639	615
		50 ~ 99	35	X	X	X
		100 ~ 199	15	853	550	263
		200 ~ 299	2	X	X	X
		300 ~ 499	3	X	X	X
		500명 이상	1	X	X	X
절단가공 및 표면처리강재 생산업	D27191	계	698	8,950	2,781	3,715
보조셀 감추기 경로 없음! 산업분류 통합 필요!		5 ~ 9	273	326	143	199
		10 ~ 19	221	639	277	294
		20 ~ 49	148	2,551	639	615
		50 ~ 99	35	1,922	471	498
		100 ~ 199	15	853	550	263
		200 ~ 299	2	…	…	…
		300 ~ 499	3	1,666	480	1,208
		500명 이상	1	…	…	…
그외 기타 철강산업	D27199	계	6	1	2	4
보조셀 감추기 경로 없음! 산업분류 통합 필요!		5 ~ 9	2	…	…	…
		10 ~ 19	4	1	1	3
		20 ~ 49	0	0	0	0
		50 ~ 99	0	0	0	0
		100 ~ 199	0	0	0	0
		200 ~ 299	0	0	0	0
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0

〈부표 1-1〉의 계속

산업별		종사자규모별	2006			
			사업체수	주요생산비	부가가치	유형자산
제 1차 비철금속산업	D272	계	748	19,971	4,751	4,650
		5 ~ 9	241	376	132	105
		10 ~ 19	232	985	339	268
		20 ~ 49	169	2,501	954	487
		50 ~ 99	59	2,129	820	599
		100 ~ 199	26	2,458	608	712
		200 ~ 299	8	886	238	232
		300 ~ 499	7	1,323	322	395
		500명 이상	6	9,312	1,339	1,852
비철금속 제련, 정련 및 합금	D2721	계	217	9,487	1,979	1,479
		5 ~ 9	72	165	44	32
		10 ~ 19	79	522	163	111
		20 ~ 49	47	1,445	612	160
		50 ~ 99	11	X	X	X
		100 ~ 199	4	395	162	100
		200 ~ 299	0	0	0	0
		300 ~ 499	2	X	X	X
		500명 이상	2	X	X	X
동 제련, 정련 및 합금 제조업	D27211	계	22	4,466	842	477
		5 ~ 9	9	14	7	4
		10 ~ 19	5	30	7	12
		20 ~ 49	5	X	X	X
		50 ~ 99	0	0	0	0
		100 ~ 199	2	X	X	X
		200 ~ 299	0	0	0	0
		300 ~ 499	0	0	0	0
		500명 이상	1	X	X	X
알루미늄 제련, 정련 및 합금	D27212	계	128	1,804	404	319
		5 ~ 9	40	75	26	21
		10 ~ 19	53	384	107	73
		20 ~ 49	25	X	X	X
		50 ~ 99	7	X	X	X
		100 ~ 199	2	X	X	X
		200 ~ 299	0	0	0	0
		300 ~ 499	1	X	X	X
		500명 이상	0	0	0	0

〈부표 1-1〉의 계속

산업별		종사자규모별	2006			
			사업체수	주요생산비	부가가치	유형자산
제 1차 금속산업	D27	계	3,173	64,609	25,282	30,990
		5 ~ 9	1,028	1,190	523	533
		10 ~ 19	967	2,693	1,127	1,091
		20 ~ 49	769	8,095	2,820	2,304
		50 ~ 99	248	6,935	2,524	2,331
		100 ~ 199	98	6,838	2,566	2,371
		200 ~ 299	26	1,982	1,046	888
		300 ~ 499	17	6,006	1,329	3,732
		500명 이상	20	30,869	13,348	17,740
제 1차 철강산업	D271	계	1,805	42,038	19,157	25,132
		5 ~ 9	625	722	338	369
		10 ~ 19	540	1,449	633	671
		20 ~ 49	418	4,993	1,488	1,526
		50 ~ 99	130	4,196	1,295	1,406
		100 ~ 199	57	3,586	1,715	1,481
		200 ~ 299	11	851	672	455
		300 ~ 499	10	4,683	1,007	3,337
		500명 이상	14	21,557	12,008	15,888
제철, 제강 및 합금철 제조업	D2711	계	47	21,230	12,103	15,649
		5 ~ 9	12	18	5	6
		10 ~ 19	8	32	19	9
		20 ~ 49	7	120	17	39
		50 ~ 99	2	X	X	X
		100 ~ 199	4	265	72	302
		200 ~ 299	1	X	X	X
		300 ~ 499	2	X	X	X
		500명 이상	11	20,103	11,686	15,124

〈부표 1-1〉의 계속

산업별		종사자규모별	2006			
			사업체수	주요생산비	부가가치	유형자산
연 및 아연제련, 정련 및 합금	D27213	계	32	2,492	661	626
		5 ~ 9	9	14	5	2
		10 ~ 19	9	40	16	11
		20 ~ 49	10	74	26	13
		50 ~ 99	2	X	X	X
		100 ~ 199	0	0	0	0
		200 ~ 299	0	0	0	0
		300 ~ 499	1	X	X	X
		500명 이상	1	X	X	X
기타 비철금속제련,정련 및 합금	D27219	계	35	724	72	56
		5 ~ 9	14	61	6	5
		10 ~ 19	12	68	33	16
		20 ~ 49	7	X	X	X
		50 ~ 99	2	X	X	X
		100 ~ 199	0	0	0	0
		200 ~ 299	0	0	0	0
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
비철금속압연,압출 및 연신제품	D2722	계	467	10,103	2,581	3,013
		5 ~ 9	147	187	73	62
		10 ~ 19	137	426	155	143
		20 ~ 49	101	884	265	276
		50 ~ 99	45	1,297	692	486
		100 ~ 199	21	X	X	X
		200 ~ 299	7	X	X	X
		300 ~ 499	5	X	X	X
		500명 이상	4	X	X	X
동 압연, 압출 및 연신제품 제조업	D27221	계	182	5,963	1,524	1,500
		5 ~ 9	51	86	24	21
		10 ~ 19	53	211	74	58
		20 ~ 49	42	417	104	112
		50 ~ 99	23	X	X	X
		100 ~ 199	7	X	X	X
		200 ~ 299	2	X	X	X
		300 ~ 499	2	X	X	X
		500명 이상	2	X	X	X

〈부표 1-1〉의 계속

산업별		종사자규모별	2006			
			사업체수	주요생산비	부가가치	유형자산
알루미늄 압연,압출 및 연신제품	D27222	계	240	3,566	943	1,427
		5 ~ 9	75	86	40	34
		10 ~ 19	75	195	72	81
		20 ~ 49	49	345	129	139
		50 ~ 99	20	407	106	187
		100 ~ 199	12	495	144	185
		200 ~ 299	4	X	X	X
		300 ~ 499	3	X	X	X
		500명 이상	2	X	X	X
기타 비철금속 압연,압출 및 연신제품	D27229	계	45	574	114	86
		5 ~ 9	21	15	9	7
		10 ~ 19	9	20	9	4
		20 ~ 49	10	122	33	26
		50 ~ 99	2	X	X	X
		100 ~ 199	2	X	X	X
		200 ~ 299	1	X	X	X
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
기타 제 1차 비철금속산업	D2729	계	64	381	191	158
		5 ~ 9	22	24	15	11
		10 ~ 19	16	37	20	14
		20 ~ 49	21	172	77	50
		50 ~ 99	3	X	X	X
		100 ~ 199	1	X	X	X
		200 ~ 299	1	X	X	X
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
기타 제 1차 비철금속산업	D27290	계	64	381	191	158
		5 ~ 9	22	24	15	11
		10 ~ 19	16	37	20	14
		20 ~ 49	21	172	77	50
		50 ~ 99	3	X	X	X
		100 ~ 199	1	X	X	X
		200 ~ 299	1	X	X	X
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0

〈부표 1-1〉의 계속

산업별		종사자규모별	2006			
			사업체수	주요생산비	부가가치	유형자산
금속 주조업	D273	계	620	2,601	1,374	1,208
		5 ~ 9	162	92	53	59
		10 ~ 19	195	259	155	152
		20 ~ 49	182	600	378	292
		50 ~ 99	59	609	409	326
		100 ~ 199	15	794	243	178
		200 ~ 299	7	245	136	201
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
철강 주조업	D2731	계	392	1,355	962	832
		5 ~ 9	75	36	24	31
		10 ~ 19	115	139	92	73
		20 ~ 49	141	406	282	222
		50 ~ 99	43	364	279	198
		100 ~ 199	12	X	X	X
		200 ~ 299	6	X	X	X
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
선철주물 주조업	D27311	계	217	873	595	556
		5 ~ 9	36	16	11	18
		10 ~ 19	62	66	49	44
		20 ~ 49	83	229	169	131
		50 ~ 99	24	194	142	106
		100 ~ 199	7	X	X	X
		200 ~ 299	5	X	X	X
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
강주물 주조업	D27312	계	175	482	367	275
		5 ~ 9	39	20	13	13
		10 ~ 19	53	73	43	29
		20 ~ 49	58	177	113	91
		50 ~ 99	19	170	137	93
		100 ~ 199	5	X	X	X
		200 ~ 299	1	X	X	X
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0

〈부표 1-1〉의 계속

산업별		종사자규모별	2006			
			사업체수	주요생산비	부가가치	유형자산
비철금속 주조업	D2732	계	228	1,245	412	376
		5 ~ 9	87	56	29	28
		10 ~ 19	80	120	63	80
		20 ~ 49	41	194	96	70
		50 ~ 99	16	245	130	128
		100 ~ 199	3	X	X	X
		200 ~ 299	1	X	X	X
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
알루미늄주물 주조업	D27321	계	138	484	220	231
		5 ~ 9	51	31	17	18
		10 ~ 19	48	79	40	53
		20 ~ 49	25	146	67	50
		50 ~ 99	12	179	70	80
		100 ~ 199	2	…	…	…
		200 ~ 299	0	0	0	0
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
동주물 주조업	D27322	계	39	121	87	61
		5 ~ 9	17	12	6	4
		10 ~ 19	10	16	7	4
		20 ~ 49	10	39	20	11
		50 ~ 99	1	…	…	…
		100 ~ 199	0	0	0	0
		200 ~ 299	1	…	…	…
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0
기타 비철금속 주조업	D27329	계	51	640	105	84
		5 ~ 9	19	14	6	6
		10 ~ 19	22	26	16	22
		20 ~ 49	6	10	8	9
		50 ~ 99	3	22	16	26
		100 ~ 199	1	…	…	…
		200 ~ 299	0	0	0	0
		300 ~ 499	0	0	0	0
		500명 이상	0	0	0	0

보조셀 감추기 경로 없음! 산업분류 통합 필요!

보조셀 감추기 경로 없음! 산업분류 통합 필요!

보조셀 감추기 경로 없음! 산업분류 통합 필요!

〈부표 1-2〉 출하액규모별 주요지표의 보조셀감추기 방법 적용 결과

산업별		출하액규모별	2006			
			사업체수	종사자수	급여액	출하액
제 1차 금속산업	D27	계	3,173	117,484	3,858,282	88,720,949
		5억 미만	282	2,564	35,375	88,797
		5 ~ 10	435	4,002	67,594	322,549
		10 ~ 50	1,342	19,551	405,746	3,225,060
		50 ~ 100	408	11,118	277,122	2,876,762
		100 ~ 300	373	16,305	430,108	6,416,958
		300 ~ 400	74	5,206	148,359	2,558,808
		400 ~ 1000	162	14,746	444,863	9,749,943
		1000 ~ 5000	68	11,316	421,053	12,742,127
		5000억 이상	29	32,676	1,628,062	50,739,945
제 1차 철강산업	D271	계	1,805	74,240	2,708,160	60,594,683
		5억 미만	160	1,504	22,022	51,234
		5 ~ 10	264	2,432	40,528	196,217
		10 ~ 50	756	10,428	221,599	1,779,890
		50 ~ 100	209	5,427	141,308	1,482,073
		100 ~ 300	200	8,250	217,612	3,408,533
		300 ~ 400	49	2,636	74,466	1,713,896
		400 ~ 1000	102	8,061	248,113	6,023,474
		1000 ~ 5000	47	7,731	303,428	8,442,434
		5000억 이상	18	27,771	1,439,084	37,496,932
제철, 제강 및 합금철 제조업	D2711	계	47	26,522	1,383,708	32,920,911
		5억 미만	3	50	665	565
		5 ~ 10	4	34	378	3,418
		10 ~ 50	12	143	3,266	26,951
		50 ~ 100	4	185	6,998	29,240
		100 ~ 300	3	64	1,899	38,985
		300 ~ 400	0	0	0	0
		400 ~ 1000	5	337	11,059	282,120
		1000 ~ 5000	6	X	X	X
		5000억 이상	10	X	X	X

〈부표 1-2〉의 계속

산업별		출하액규모별	2006			
			사업체수	종사자수	급여액	출하액
제철 및 제강업	D27111	계	29	25,914	1,364,880	32,507,442
		5억 미만	1	X	X	X
		5 ~ 10	3	X	X	X
		10 ~ 50	7	94	2,122	17,184
		50 ~ 100	1	X	X	X
		100 ~ 300	1	X	X	X
		300 ~ 400		…	…	…
		400 ~ 1000	1	X	X	X
		1000 ~ 5000	5	X	X	X
		5000억 이상	10	23,970	1,285,630	30,924,030
합금철 제조업	D27112	계	18	608	18,828	413,469
		5억 미만	2	X	X	X
		5 ~ 10	1	X	X	X
		10 ~ 50	5	49	1,144	9,767
		50 ~ 100	3	X	X	X
		100 ~ 300	2	X	X	X
		300 ~ 400	0			
		400 ~ 1000	4	X	X	X
		1000 ~ 5000	1	X	X	X
		5000억 이상	0	0	0	0
철강압연·압출 및 연신제품	D2712	계	359	12,989	404,379	9,194,093
		5억 미만	17	159	1,879	5,124
		5 ~ 10	40	356	5,973	29,578
		10 ~ 50	141	1,769	37,150	349,310
		50 ~ 100	46	1,059	28,392	320,689
		100 ~ 300	45	1,741	49,127	802,092
		300 ~ 400	20	1,050	29,432	694,147
		400 ~ 1000	31	3,151	104,276	1,830,643
		1000 ~ 5000	16	2,434	90,334	2,602,555
		5000억 이상	3	1,270	57,816	2,559,955
열간 압연 및 압출제품 제조업	D27121	계	147	4,287	122,257	2,274,232
		5억 미만	8	94	820	2,296
		5 ~ 10	24	248	4,264	17,490
		10 ~ 50	51	748	15,773	117,041
		50 ~ 100	18	419	10,039	114,081
		100 ~ 300	22	827	23,486	413,405
		300 ~ 400	9	522	12,752	305,792
		400 ~ 1000	9	618	20,395	510,564
		1000 ~ 5000	6	811	34,728	793,563
		5000억 이상	0	0	0	0

〈부표 1-2〉의 계속

산업별		출하액규모별	2006			
			사업체수	종사자수	급여액	출하액
냉간 압연 및 압출제품 제조업	D27122	계	77	3,565	121,547	3,251,709
		5억 미만	3	X	X	X
		5 ~ 10	5	36	618	3,304
		10 ~ 50	30	428	9,632	85,588
		50 ~ 100	9	294	9,998	68,252
		100 ~ 300	10	350	9,214	168,402
		300 ~ 400	6	298	10,229	217,982
		400 ~ 1000	6	500	13,901	338,027
		1000 ~ 5000	6	767	24,825	1,192,779
		5000억 이상	2	X	X	X
철강선 제조업	D27123	계	135	5,137	160,575	3,668,152
		5억 미만	6	X	X	X
		5 ~ 10	11	72	1,091	8,784
		10 ~ 50	60	593	11,745	146,681
		50 ~ 100	19	346	8,355	138,356
		100 ~ 300	13	564	16,427	220,285
		300 ~ 400	5	230	6,451	170,373
		400 ~ 1000	16	2,033	69,980	982,052
		1000 ~ 5000	4	856	30,781	616,213
		5000억 이상	1	X	X	X
철강관 제조업	D2713	계	695	18,481	465,891	6,791,073
		5억 미만	72	666	8,381	23,991
		5 ~ 10	91	838	13,746	67,839
		10 ~ 50	311	4,724	98,236	715,722
		50 ~ 100	89	2,452	56,597	631,001
		100 ~ 300	88	4,055	98,622	1,419,595
		300 ~ 400	8	588	15,933	280,552
		400 ~ 1000	27	2,781	82,800	1,536,408
		1000 ~ 5000	7	X	X	X
		5000억 이상	2	X	X	X
주철관 제조업	D27131	계	33	1,503	38,322	280,166
		5억 미만	0	0	0	0
		5 ~ 10	6	49	842	4,205
		10 ~ 50	14	333	7,442	37,245
		50 ~ 100	7	271	6,133	52,778
		100 ~ 300	5	572	13,613	86,256
		300 ~ 400	0	0	0	0
		400 ~ 1000	1	…	…	…
		1000 ~ 5000	0	0	0	0
		5000억 이상	0	0	0	0

보조셀 감추기 경로 없음! 산업분류 통합 필요!

〈부표 1-2〉의 계속

산업별		출하액규모별	2006			
			사업체수	종사자수	급여액	출하액
강관 제조업	D27132	계	662	16,978	427,569	6,510,907
		5억 미만	72	666	8,381	23,991
		5 ~ 10	85	789	12,904	63,634
		10 ~ 50	297	4,391	90,794	678,477
		50 ~ 100	82	2,181	50,464	578,223
		100 ~ 300	83	3,483	85,009	1,333,339
		300 ~ 400	8	588	15,933	280,552
		400 ~ 1000	26	2,503	72,508	1,436,726
		1000 ~ 5000	7	1,690	62,870	962,441
		5000억 이상	2	…	…	…
기타 철강산업	D2719	계	704	16,248	454,182	11,688,606
		5억 미만	68	629	11,097	21,554
		5 ~ 10	129	1,204	20,431	95,382
		10 ~ 50	292	3,792	82,947	687,907
		50 ~ 100	70	1,731	49,321	501,143
		100 ~ 300	64	2,390	67,964	1,147,861
		300 ~ 400	21	998	29,101	739,197
		400 ~ 1000	39	1,792	49,978	2,374,303
		1000 ~ 5000	18	1,868	76,411	3,261,836
		5000억 이상	3	1,844	66,932	2,859,423
절단가공 및 표면처리강재 생산업	D27191	계	698	16,181	453,217	11,684,950
		5억 미만	65	591	10,635	20,384
		5 ~ 10	127	X	X	X
		10 ~ 50	291	X	X	X
		50 ~ 100	70	1,731	49,321	501,143
		100 ~ 300	64	2,390	67,964	1,147,861
		300 ~ 400	21	998	29,101	739,197
		400 ~ 1000	39	1,792	49,978	2,374,303
		1000 ~ 5000	18	1,868	76,411	3,261,836
		5000억 이상	3	1,844	66,932	2,859,423
그외 기타 철강산업	D27199	계	6	67	965	3,656
		5억 미만	3	38	462	1,170
		5 ~ 10	2	X	X	X
		10 ~ 50	1	X	X	X
		50 ~ 100	0	0	0	0
		100 ~ 300	0	0	0	0
		300 ~ 400	0	0	0	0
		400 ~ 1000	0	0	0	0
		1000 ~ 5000	0	0	0	0
		5000억 이상	0	0	0	0

보조셀 감추기 경로 없음! 산업분류 통합 필요!

〈부표 1-2〉의 계속

산업별		출하액규모별	2006			
			사업체수	종사자수	급여액	출하액
제 1차 비철금속산업	D272	계	748	26,147	743,172	24,156,073
		5억 미만	52	515	6,193	16,043
		5 ~ 10	77	745	12,399	56,870
		10 ~ 50	295	3,967	77,995	755,995
		50 ~ 100	108	2,148	52,219	750,232
		100 ~ 300	118	3,881	104,319	2,091,985
		300 ~ 400	20	X	X	X
		400 ~ 1000	47	4,695	135,116	2,893,741
		1000 ~ 5000	21	3,585	117,625	4,299,693
		5000억 이상	10	X	X	X
비철금속 제련, 정련 및 합금	D2721	계	217	6,315	190,652	11,164,291
		5억 미만	11	X	X	X
		5 ~ 10	18	246	3,506	12,636
		10 ~ 50	79	902	18,345	214,678
		50 ~ 100	39	587	13,467	264,743
		100 ~ 300	38	736	19,742	669,194
		300 ~ 400	8	X	X	X
		400 ~ 1000	13	678	21,248	882,112
		1000 ~ 5000	6	584	18,794	975,065
		5000억 이상	5	X	X	X
동 제련, 정련 및 합금 제조업	D27211	계	22	1,180	47,553	5,097,829
		5억 미만	2	X	X	X
		5 ~ 10	1	X	X	X
		10 ~ 50	7	54	1,283	17,280
		50 ~ 100	4	52	1,214	24,969
		100 ~ 300	1	X	X	X
		300 ~ 400	1	X	X	X
		400 ~ 1000	2	X	X	X
		1000 ~ 5000	2	X	X	X
		5000억 이상	2	X	X	X
알루미늄 제련, 정련 및 합금	D27212	계	128	2,853	74,784	2,195,555
		5억 미만	3	X	X	X
		5 ~ 10	8	72	1,023	5,982
		10 ~ 50	43	438	8,512	117,646
		50 ~ 100	27	418	9,813	185,484
		100 ~ 300	31	570	14,971	545,814
		300 ~ 400	5	527	15,659	162,845
		400 ~ 1000	7	388	11,855	540,248
		1000 ~ 5000	4	X	X	X
		5000억 이상	0	0	0	0

〈부표 1-2〉의 계속

산업별		출하액규모별	2006			
			사업체수	종사자수	급여액	출하액
연 및 아연제련, 정련 및 합금	D27213	계	32	1,694	53,589	3,029,562
		5억 미만	2	X	X	X
		5 ~ 10	5	X	X	X
		10 ~ 50	13	183	3,138	32,369
		50 ~ 100	4	63	1,405	26,613
		100 ~ 300	4	X	X	X
		300 ~ 400	0	0	0	0
		400 ~ 1000	2	X	X	X
		1000 ~ 5000	0	0	0	0
		5000억 이상	2	X	X	X
기타 비철금속제련,정련 및 합금	D27219	계	35	588	14,726	841,345
		5억 미만	4	43	736	1,557
		5 ~ 10	4	45	775	2,580
		10 ~ 50	16	227	5,412	47,383
		50 ~ 100	4	54	1,035	27,677
		100 ~ 300	2	X	X	X
		300 ~ 400	2	X	X	X
		400 ~ 1000	2	X	X	X
		1000 ~ 5000	0	0	0	0
		5000억 이상	1	X	X	X
비철금속압연,압출 및 연신제품	D2722	계	467	18,357	515,907	12,428,609
		5억 미만	38	296	3,959	12,281
		5 ~ 10	45	368	6,646	33,657
		10 ~ 50	190	2,636	50,896	475,729
		50 ~ 100	64	1,482	36,487	455,909
		100 ~ 300	69	2,787	75,342	1,234,673
		300 ~ 400	11	X	X	X
		400 ~ 1000	30	3,594	100,713	1,775,749
		1000 ~ 5000	15	3,001	98,831	3,324,628
		5000억 이상	5	X	X	X
동 압연, 압출 및 연신제품 제조업	D27221	계	182	7,891	216,454	7,316,333
		5억 미만	16	149	2,109	4,691
		5 ~ 10	10	82	1,515	7,193
		10 ~ 50	67	935	18,065	179,701
		50 ~ 100	25	461	11,270	178,631
		100 ~ 300	33	1,117	27,465	560,330
		300 ~ 400	3	X	X	X
		400 ~ 1000	17	1,769	48,697	965,443
		1000 ~ 5000	8	X	X	X
		5000억 이상	3	X	X	X

〈부표 1-2〉의 계속

산업별		출하액규모별	2006			
			사업체수	종사자수	급여액	출하액
알루미늄 압연,압출 및 연신제품	D27222	계	240	9,214	261,854	4,433,370
		5억 미만	16	110	1,338	5,043
		5 ~ 10	29	242	4,302	21,850
		10 ~ 50	105	1,507	28,791	259,375
		50 ~ 100	35	919	22,597	246,057
		100 ~ 300	28	1,306	36,893	508,431
		300 ~ 400	7	X	X	X
		400 ~ 1000	13	1,825	52,016	810,306
		1000 ~ 5000	5	1,406	44,893	815,492
		5000억 이상	2	X	X	X
기타 비철금속 압연,압출 및 연신제품	D27229	계	45	1,252	37,599	678,906
		5억 미만	6	37	512	2,547
		5 ~ 10	6	44	829	4,614
		10 ~ 50	18	194	4,040	36,653
		50 ~ 100	4	102	2,620	31,221
		100 ~ 300	8	364	10,984	165,912
		300 ~ 400	1	X	X	X
		400 ~ 1000	0	0	0	0
		1000 ~ 5000	2	X	X	X
		5000억 이상	0	0	0	0
기타 제 1차 비철금속산업	D2729	계	64	1,475	36,613	563,173
		5억 미만	3	X	X	X
		5 ~ 10	14	131	2,247	10,577
		10 ~ 50	26	429	8,754	65,588
		50 ~ 100	5	79	2,265	29,580
		100 ~ 300	11	358	9,235	188,118
		300 ~ 400	1	X	X	X
		400 ~ 1000	4	423	13,155	235,880
		1000 ~ 5000	0	0	0	0
		5000억 이상	0	0	0	0
기타 제 1차 비철금속산업	D27290	계	64	1,475	36,613	563,173
		5억 미만	3	X	X	X
		5 ~ 10	14	131	2,247	10,577
		10 ~ 50	26	429	8,754	65,588
		50 ~ 100	5	79	2,265	29,580
		100 ~ 300	11	358	9,235	188,118
		300 ~ 400	1	X	X	X
		400 ~ 1000	4	423	13,155	235,880
		1000 ~ 5000	0	0	0	0
		5000억 이상	0	0	0	0

〈부표 1-2〉의 계속

산업별		출하액규모별	2006			
			사업체수	종사자수	급여액	출하액
금속 주조업	D273	계	620	17,097	406,950	3,970,193
		5억 미만	70	545	7,160	21,520
		5 ~ 10	94	825	14,667	69,462
		10 ~ 50	291	5,156	106,152	689,175
		50 ~ 100	91	3,543	83,595	644,457
		100 ~ 300	55	4,174	108,177	916,440
		300 ~ 400	5	X	X	X
		400 ~ 1000	13	1,990	61,634	832,728
		1000 ~ 5000	0	0	0	0
		5000억 이상	1	…	…	…
철강 주조업	D2731	계	392	12,320	302,712	2,317,748
		5억 미만	37	307	4,302	11,858
		5 ~ 10	50	480	9,054	37,373
		10 ~ 50	187	3,654	75,989	455,094
		50 ~ 100	66	2,711	64,658	469,661
		100 ~ 300	40	3,108	82,715	693,521
		300 ~ 400	4	526	16,752	135,948
		400 ~ 1000	8	1,534	49,242	514,293
		1000 ~ 5000	0	0	0	0
		5000억 이상	0	0	0	0
선철주물 주조업	D27311	계	217	7,566	189,230	1,470,348
		5억 미만	21	189	3,159	6,877
		5 ~ 10	25	262	4,821	19,131
		10 ~ 50	99	1,985	40,326	241,400
		50 ~ 100	39	1,524	35,506	262,599
		100 ~ 300	23	1,698	45,374	386,418
		300 ~ 400	3	452	14,177	101,305
		400 ~ 1000	7	1,456	45,867	452,618
		1000 ~ 5000	0	0	0	0
		5000억 이상	0	0	0	0
강주물 주조업	D27312	계	175	4,754	113,482	847,400
		5억 미만	16	118	1,143	4,981
		5 ~ 10	25	218	4,233	18,242
		10 ~ 50	88	1,669	35,663	213,694
		50 ~ 100	27	1,187	29,152	207,062
		100 ~ 300	17	1,410	37,341	307,103
		300 ~ 400	1	…	…	…
		400 ~ 1000	1	…	…	…
		1000 ~ 5000	0	0	0	0
		5000억 이상	0	0	0	0

보조셀 감추기 경로 없음! 산업분류 통합 필요!

보조셀 감추기 경로 없음! 산업분류 통합 필요!

보조셀 감추기 경로 없음! 산업분류 통합 필요!

〈부표 1-2〉의 계속

산업별		출하액규모별	2006			
			사업체수	종사자수	급여액	출하액
비철금속 주조업	D2732	계	228	4,777	104,238	1,652,445
		5억 미만	33	238	2,858	9,662
		5 ~ 10	44	345	5,613	32,089
		10 ~ 50	104	1,502	30,163	234,081
		50 ~ 100	25	832	18,937	174,796
		100 ~ 300	15	1,066	25,462	222,919
		300 ~ 400	1	…	…	…
		400 ~ 1000	5	456	12,392	318,435
		1000 ~ 5000	0	0	0	0
		5000억 이상	1	…	…	…
알루미늄주물 주조업	D27321	계	138	2,985	63,530	699,581
		5억 미만	15	91	1,269	4,741
		5 ~ 10	32	254	3,862	22,914
		10 ~ 50	59	897	17,567	134,095
		50 ~ 100	16	566	12,539	114,814
		100 ~ 300	11	644	15,695	159,553
		300 ~ 400	1	…	…	…
		400 ~ 1000	4	364	9,311	229,838
		1000 ~ 5000	0	0	0	0
		5000억 이상	0	0	0	0
동주물 주조업	D27322	계	39	849	19,960	210,691
		5억 미만	6	41	409	1,712
		5 ~ 10	5	32	508	3,711
		10 ~ 50	19	229	5,402	42,758
		50 ~ 100	6	172	4,329	40,691
		100 ~ 300	2	…	…	…
		300 ~ 400	0	0	0	0
		400 ~ 1000	1	…	…	…
		1000 ~ 5000	0	0	0	0
		5000억 이상	0	0	0	0
기타 비철금속 주조업	D27329	계	51	943	20,748	742,173
		5억 미만	12	106	1,180	3,209
		5 ~ 10	7	59	1,243	5,464
		10 ~ 50	26	376	7,194	57,228
		50 ~ 100	3	94	2,069	19,291
		100 ~ 300	2	…	…	…
		300 ~ 400	0	0	0	0
		400 ~ 1000	0	0	0	0
		1000 ~ 5000	0	0	0	0
		5000억 이상	1	…	…	…

보조셀 감추기 경로 없음! 산업분류 통합 필요!

보조셀 감추기 경로 없음! 산업분류 통합 필요!

보조셀 감추기 경로 없음! 산업분류 통합 필요!

보조셀 감추기 경로 없음! 산업분류 통합 필요!

〈부표 1-2〉의 계속

산업별		출하액규모별	2006			
			사업체수	주요생산비	부가가치	유형자산
제 1차 금속산업	D27	계	3,173	64,609,305	25,282,343	30,990,026
		5억 미만	282	41,453	52,022	116,485
		5 ~ 10	435	165,844	158,271	192,179
		10 ~ 50	1,342	2,051,180	1,196,434	1,296,553
		50 ~ 100	408	1,984,482	911,558	884,715
		100 ~ 300	373	4,699,446	1,767,107	1,680,906
		300 ~ 400	74	1,935,759	663,142	739,204
		400 ~ 1000	162	7,425,694	2,401,935	2,468,939
		1000 ~ 5000	68	9,535,974	3,297,505	2,889,028
		5000억 이상	29	36,769,473	14,834,369	20,722,017
제 1차 철강산업	D271	계	1,805	42,037,894	19,157,092	25,132,337
		5억 미만	160	24,975	29,804	70,232
		5 ~ 10	264	103,168	94,507	133,457
		10 ~ 50	756	1,117,804	674,054	778,416
		50 ~ 100	209	1,023,066	473,636	498,338
		100 ~ 300	200	2,534,414	900,606	962,410
		300 ~ 400	49	1,310,743	421,336	459,146
		400 ~ 1000	102	4,546,556	1,486,613	1,509,710
		1000 ~ 5000	47	6,208,858	2,244,494	2,136,169
		5000억 이상	18	25,168,310	12,832,042	18,584,459
제철, 제강 및 합금철 제조업	D2711	계	47	21,229,727	12,102,779	15,648,650
		5억 미만	3	581	2,084	17,662
		5 ~ 10	4	1,722	2,023	449
		10 ~ 50	12	14,836	12,127	17,376
		50 ~ 100	4	13,408	21,207	5,297
		100 ~ 300	3	34,465	3,181	5,481
		300 ~ 400	0	0	0	0
		400 ~ 1000	5	223,002	64,700	67,128
		1000 ~ 5000	6	X	X	X
		5000억 이상	10	X	X	X

〈부표 1-2〉의 계속

산업별		출하액규모별	2006			
			사업체수	주요생산비	부가가치	유형자산
제철 및 제강업	D27111	계	29	20,908,681	11,999,689	15,500,760
		5억 미만	1	X	X	X
		5 ~ 10	3	X	X	X
		10 ~ 50	7	9,288	7,894	12,334
		50 ~ 100	1	X	X	X
		100 ~ 300	1	X	X	X
		300 ~ 400		…	…	…
		400 ~ 1000	1	X	X	X
		1000 ~ 5000	5	X	X	X
		5000억 이상	10	19,773,825	11,556,410	14,898,793
합금철 제조업	D27112	계	18	321,046	103,090	147,890
		5억 미만	2	X	X	X
		5 ~ 10	1	X	X	X
		10 ~ 50	5	5,548	4,233	5,042
		50 ~ 100	3	X	X	X
		100 ~ 300	2	X	X	X
		300 ~ 400	0			
		400 ~ 1000	4	X	X	X
		1000 ~ 5000	1	X	X	X
		5000억 이상	0	0	0	0
철강압연·압출 및 연신제품	D2712	계	359	7,200,275	2,042,077	3,788,933
		5억 미만	17	1,680	3,430	4,820
		5 ~ 10	40	17,064	12,478	19,693
		10 ~ 50	141	241,951	109,788	127,063
		50 ~ 100	46	228,293	95,713	114,923
		100 ~ 300	45	577,334	229,396	243,978
		300 ~ 400	20	517,821	186,153	161,796
		400 ~ 1000	31	1,338,073	475,849	695,702
		1000 ~ 5000	16	1,867,387	701,253	581,950
		5000억 이상	3	2,410,672	228,017	1,839,008
열간 압연 및 압출제품 제조업	D27121	계	147	1,748,217	531,855	620,688
		5억 미만	8	760	1,536	1,114
		5 ~ 10	24	9,287	8,121	10,878
		10 ~ 50	51	73,811	44,274	41,727
		50 ~ 100	18	77,728	38,040	48,960
		100 ~ 300	22	305,595	112,071	109,148
		300 ~ 400	9	219,445	93,057	74,514
		400 ~ 1000	9	418,826	90,659	161,373
		1000 ~ 5000	6	642,765	144,097	172,974
		5000억 이상	0	0	0	0

〈부표 1-2〉의 계속

산업별		출하액규모별	2006			
			사업체수	주요생산비	부가가치	유형자산
냉간 압연 및 압출제품 제조업	D27122	계	77	2,481,761	806,258	1,300,475
		5억 미만	3	X	X	X
		5 ~ 10	5	2,237	1,070	1,596
		10 ~ 50	30	58,993	27,724	43,234
		50 ~ 100	9	44,133	25,810	26,751
		100 ~ 300	10	130,273	39,218	52,071
		300 ~ 400	6	173,563	46,866	50,446
		400 ~ 1000	6	234,432	101,474	160,430
		1000 ~ 5000	6	768,304	394,150	176,457
		5000억 이상	2	X	X	X
철강선 제조업	D27123	계	135	2,970,297	703,964	1,867,770
		5억 미만	6	X	X	X
		5 ~ 10	11	5,540	3,287	7,219
		10 ~ 50	60	109,147	37,790	42,102
		50 ~ 100	19	106,432	31,863	39,212
		100 ~ 300	13	141,466	78,107	82,759
		300 ~ 400	5	124,813	46,230	36,836
		400 ~ 1000	16	684,815	283,716	373,899
		1000 ~ 5000	4	456,318	163,006	232,519
		5000억 이상	1	X	X	X
철강관 제조업	D2713	계	695	4,656,569	2,229,242	1,975,669
		5억 미만	72	11,108	14,305	30,140
		5 ~ 10	91	35,003	33,697	28,038
		10 ~ 50	311	441,815	282,811	287,076
		50 ~ 100	89	427,081	211,489	195,913
		100 ~ 300	88	1,018,440	423,283	406,293
		300 ~ 400	8	184,043	102,746	106,600
		400 ~ 1000	27	1,117,955	425,138	425,088
		1000 ~ 5000	7	X	X	X
		5000억 이상	2	X	X	X
주철관 제조업	D27131	계	33	151,655	125,626	147,407
		5억 미만	0	0	0	0
		5 ~ 10	6	2,124	2,134	2,254
		10 ~ 50	14	19,403	18,128	14,164
		50 ~ 100	7	33,828	19,853	14,916
		100 ~ 300	5	54,135	34,163	53,895
		300 ~ 400	0	0	0	0
		400 ~ 1000	1	…	…	…
		1000 ~ 5000	0	0	0	0
		5000억 이상	0	0	0	0

보조셀 감추기
경로 없음!
산업분류
통합 필요!

〈부표 1-2〉의 계속

산업별		출하액규모별	2006			
			사업체수	주요생산비	부가가치	유형자산
제 1차 비철금속산업	D272	계	748	19,970,901	4,751,212	4,649,814
		5억 미만	52	7,762	9,003	22,554
		5 ~ 10	77	26,353	29,942	30,565
		10 ~ 50	295	524,495	238,197	257,731
		50 ~ 100	108	562,312	190,378	182,298
		100 ~ 300	118	1,625,800	494,329	387,915
		300 ~ 400	20	X	X	X
		400 ~ 1000	47	2,343,599	619,631	682,509
		1000 ~ 5000	21	3,327,116	1,053,011	752,859
		5000억 이상	10	X	X	X
비철금속 제련, 정련 및 합금	D2721	계	217	9,486,869	1,979,394	1,478,966
		5억 미만	11	X	X	X
		5 ~ 10	18	4,691	7,269	6,053
		10 ~ 50	79	145,332	69,927	56,863
		50 ~ 100	39	200,461	66,327	69,672
		100 ~ 300	38	527,333	143,906	81,468
		300 ~ 400	8	X	X	X
		400 ~ 1000	13	742,680	155,259	104,775
		1000 ~ 5000	6	800,341	187,003	98,627
		5000억 이상	5	X	X	X
동 제련, 정련 및 합금 제조업	D27211	계	22	4,466,477	842,382	477,180
		5억 미만	2	X	X	X
		5 ~ 10	1	X	X	X
		10 ~ 50	7	12,161	5,037	3,001
		50 ~ 100	4	17,461	7,467	12,077
		100 ~ 300	1	X	X	X
		300 ~ 400	1	X	X	X
		400 ~ 1000	2	X	X	X
		1000 ~ 5000	2	X	X	X
		5000억 이상	2	X	X	X
알루미늄 제련, 정련 및 합금	D27212	계	128	1,804,046	403,992	319,313
		5억 미만	3	X	X	X
		5 ~ 10	8	2,826	3,152	3,909
		10 ~ 50	43	80,075	38,038	36,764
		50 ~ 100	27	142,134	45,431	47,591
		100 ~ 300	31	435,786	112,445	68,021
		300 ~ 400	5	108,964	54,476	27,319
		400 ~ 1000	7	481,753	65,554	59,877
		1000 ~ 5000	4	X	X	X
		5000억 이상	0	0	0	0

〈부표 1-2〉의 계속

산업별		출하액규모별	2006			
			사업체수	주요생산비	부가가치	유형자산
알루미늄 압연,압출 및 연신제품	D27222	계	240	3,565,817	942,702	1,426,851
		5억 미만	16	2,681	2,622	4,217
		5 ~ 10	29	11,279	10,618	11,323
		10 ~ 50	105	195,916	67,491	85,371
		50 ~ 100	35	178,573	68,197	68,184
		100 ~ 300	28	378,785	143,123	169,747
		300 ~ 400	7	X	X	X
		400 ~ 1000	13	658,294	162,909	199,488
		1000 ~ 5000	5	639,167	183,662	253,150
		5000억 이상	2	X	X	X
기타 비철금속 압연,압출 및 연신제품	D27229	계	45	574,182	113,813	85,843
		5억 미만	6	1,114	1,380	1,086
		5 ~ 10	6	2,318	2,286	2,452
		10 ~ 50	18	23,354	13,390	10,560
		50 ~ 100	4	22,795	8,472	11,233
		100 ~ 300	8	133,315	34,942	17,035
		300 ~ 400	1	X	X	X
		400 ~ 1000	0	0	0	0
		1000 ~ 5000	2	X	X	X
		5000억 이상	0	0	0	0
기타 제 1차 비철금속산업	D2729	계	64	380,889	190,976	158,275
		5억 미만	3	X	X	X
		5 ~ 10	14	4,794	5,766	6,116
		10 ~ 50	26	37,550	28,504	36,666
		50 ~ 100	5	21,540	7,759	6,096
		100 ~ 300	11	135,894	55,064	29,455
		300 ~ 400	1	X	X	X
		400 ~ 1000	4	153,326	86,930	72,666
		1000 ~ 5000	0	0	0	0
		5000억 이상	0	0	0	0
기타 제 1차 비철금속산업	D27290	계	64	380,889	190,976	158,275
		5억 미만	3	X	X	X
		5 ~ 10	14	4,794	5,766	6,116
		10 ~ 50	26	37,550	28,504	36,666
		50 ~ 100	5	21,540	7,759	6,096
		100 ~ 300	11	135,894	55,064	29,455
		300 ~ 400	1	X	X	X
		400 ~ 1000	4	153,326	86,930	72,666
		1000 ~ 5000	0	0	0	0
		5000억 이상	0	0	0	0

〈부표 1-2〉의 계속

산업별		출하액규모별	2006			
			사업체수	주요생산비	부가가치	유형자산
연 및 아연제련, 정련 및 합금	D27213	계	32	2,492,086	660,536	626,071
		5억 미만	2	X	X	X
		5 ~ 10	5	X	X	X
		10 ~ 50	13	23,434	9,053	5,065
		50 ~ 100	4	20,112	6,342	5,899
		100 ~ 300	4	X	X	X
		300 ~ 400	0	0	0	0
		400 ~ 1000	2	X	X	X
		1000 ~ 5000	0	0	0	0
		5000억 이상	2	X	X	X
기타 비철금속제련,정련 및 합금	D27219	계	35	724,260	72,484	56,402
		5억 미만	4	551	1,186	2,835
		5 ~ 10	4	1,052	923	1,220
		10 ~ 50	16	29,662	17,799	12,033
		50 ~ 100	4	20,754	7,087	4,105
		100 ~ 300	2	X	X	X
		300 ~ 400	2	X	X	X
		400 ~ 1000	2	X	X	X
		1000 ~ 5000	0	0	0	0
		5000억 이상	1	X	X	X
비철금속압연,압출 및 연신제품	D2722	계	467	10,103,143	2,580,842	3,012,573
		5억 미만	38	5,931	6,585	7,655
		5 ~ 10	45	16,868	16,907	18,396
		10 ~ 50	190	341,613	139,766	164,202
		50 ~ 100	64	340,311	116,292	106,530
		100 ~ 300	69	962,573	295,359	276,992
		300 ~ 400	11	X	X	X
		400 ~ 1000	30	1,447,593	377,442	505,068
		1000 ~ 5000	15	2,526,775	866,008	654,232
		5000억 이상	5	X	X	X
동 압연, 압출 및 연신제품 제조업	D27221	계	182	5,963,144	1,524,327	1,499,879
		5억 미만	16	2,136	2,583	2,352
		5 ~ 10	10	3,271	4,003	4,621
		10 ~ 50	67	122,343	58,885	68,271
		50 ~ 100	25	138,943	39,623	27,113
		100 ~ 300	33	450,473	117,294	90,210
		300 ~ 400	3	X	X	X
		400 ~ 1000	17	789,299	214,533	305,580
		1000 ~ 5000	8	X	X	X
		5000억 이상	3	X	X	X

〈부표 1-2〉의 계속

산업별		출하액규모별	2006			
			사업체수	주요생산비	부가가치	유형자산
비철금속 주조업	D2732	계	228	1,245,489	412,218	376,327
		5억 미만	33	4,053	5,916	8,107
		5 ~ 10	44	17,946	14,902	18,548
		10 ~ 50	104	153,248	83,358	87,132
		50 ~ 100	25	117,503	60,143	63,785
		100 ~ 300	15	139,380	83,289	81,983
		300 ~ 400	1	…	…	…
		400 ~ 1000	5	220,543	96,162	81,667
		1000 ~ 5000	0	0	0	0
		5000억 이상	1	…	…	…
알루미늄주물 주조업	D27321	계	138	483,570	220,496	231,450
		5억 미만	15	2,090	3,019	5,952
		5 ~ 10	32	12,875	10,812	15,444
		10 ~ 50	59	87,976	48,712	51,588
		50 ~ 100	16	77,851	38,802	45,563
		100 ~ 300	11	103,671	56,817	38,825
		300 ~ 400	1	…	…	…
		400 ~ 1000	4	175,838	52,375	59,981
		1000 ~ 5000	0	0	0	0
		5000억 이상	0	0	0	0
동주물 주조업	D27322	계	39	121,492	86,652	60,589
		5억 미만	6	821	859	935
		5 ~ 10	5	2,199	1,536	747
		10 ~ 50	19	29,078	13,160	9,393
		50 ~ 100	6	27,030	13,913	6,427
		100 ~ 300	2	…	…	…
		300 ~ 400	0	0	0	0
		400 ~ 1000	1	…	…	…
		1000 ~ 5000	0	0	0	0
		5000억 이상	0	0	0	0
기타 비철금속 주조업	D27329	계	51	640,427	105,070	84,288
		5억 미만	12	1,142	2,038	1,220
		5 ~ 10	7	2,872	2,554	2,357
		10 ~ 50	26	36,194	21,486	26,151
		50 ~ 100	3	12,622	7,428	11,795
		100 ~ 300	2	…	…	…
		300 ~ 400	0	0	0	0
		400 ~ 1000	0	0	0	0
		1000 ~ 5000	0	0	0	0
		5000억 이상	1	…	…	…

보조셀 감추기 경로 없음! 산업분류 통합 필요!

보조셀 감추기 경로 없음! 산업분류 통합 필요!

보조셀 감추기 경로 없음! 산업분류 통합 필요!

보조셀 감추기 경로 없음! 산업분류 통합 필요!

〈부표 1-2〉의 계속

산업별		출하액규모별	2006			
			사업체수	주요생산비	부가가치	유형자산
금속 주조업	D273	계	620	2,600,510	1,374,039	1,207,875
		5억 미만	70	8,716	13,215	23,699
		5 ~ 10	94	36,323	33,822	28,157
		10 ~ 50	291	408,881	284,183	260,406
		50 ~ 100	91	399,104	247,544	204,079
		100 ~ 300	55	539,232	372,172	330,581
		300 ~ 400	5	X	X	X
		400 ~ 1000	13	535,539	295,691	276,720
		1000 ~ 5000	0	0	0	0
		5000억 이상	1	…	…	…
철강 주조업	D2731	계	392	1,355,021	961,821	831,548
		5억 미만	37	4,663	7,299	15,592
		5 ~ 10	50	18,377	18,920	9,609
		10 ~ 50	187	255,633	200,825	173,274
		50 ~ 100	66	281,601	187,401	140,294
		100 ~ 300	40	399,852	288,883	248,598
		300 ~ 400	4	79,899	58,964	49,128
		400 ~ 1000	8	314,996	199,529	195,053
		1000 ~ 5000	0	0	0	0
		5000억 이상	0	0	0	0
선철주물 주조업	D27311	계	217	873,378	595,152	556,097
		5억 미만	21	1,917	4,984	11,652
		5 ~ 10	25	9,301	9,735	4,667
		10 ~ 50	99	134,199	108,321	97,212
		50 ~ 100	39	152,698	108,451	80,974
		100 ~ 300	23	225,565	157,622	138,437
		300 ~ 400	3	62,241	41,381	41,194
		400 ~ 1000	7	287,457	164,658	181,961
		1000 ~ 5000	0	0	0	0
		5000억 이상	0	0	0	0
강주물 주조업	D27312	계	175	481,643	366,669	275,451
		5억 미만	16	2,746	2,315	3,940
		5 ~ 10	25	9,076	9,185	4,942
		10 ~ 50	88	121,434	92,504	76,062
		50 ~ 100	27	128,903	78,950	59,320
		100 ~ 300	17	174,287	131,261	110,161
		300 ~ 400	1	…	…	…
		400 ~ 1000	1	…	…	…
		1000 ~ 5000	0	0	0	0
		5000억 이상	0	0	0	0

보조셀 감추기 경로 없음! 산업분류 통합 필요!

보조셀 감추기 경로 없음! 산업분류 통합 필요!

보조셀 감추기 경로 없음! 산업분류 통합 필요!

제2장

그리드별 통계서비스의 비밀보호 방안

박민정 · 신우람* · 박천영* · 서영오**

요 약

통계청은 SGIS 시스템을 통해 센서스 자료의 빈도, 비율 등의 통계를 제공하고 있다. 이 시스템은 이용자의 변수별 범주 선택에 따라 지도 위에 해당 통계량을 표현한다. 이처럼 하나의 마이크로데이터에서 대량의 빈도표를 생성할 때는, 일반적으로 통용되는 빈도표에 대한 매스킹 기법을 적용하는 것이 불가능하다. 본 과제는 이러한 경우의 비밀보호 처리를 다룬다. 대량의 빈도표 제공을 위한 새로운 알고리즘 BSCA를 제안하고 R패키지를 구축해서 실무에 활용할 수 있도록 했다. 더불어 노출위험-정보손실을 이론적으로 논했다.

제안된 패키지는 SGIS에 바로 적용할 수 있도록 맞춤형으로 구축되었고, 결과로 SGIS에 사용할 데이터베이스를 반환한다. 또한 향후 빅데이터센터에서도 활용할 수 있도록 비밀보호 처리된 전체 빈도표 대신에 이용자가 선택한 일부 빈도표를 반환하는 함수도 제공한다. 제안된 BSCA 알고리즘의 이론적 타당성은 관련 논문 발간으로 점검하였다. 이론적 타당성 검증과 더불어 BSCA 적용 전후 빈도값 변화, 즉 정보손실의 분포를 집계구 단위 빈도표 및 그리드별 빈도표 각각에 대해 제시했다.

주요 용어 : 빈도표, 작은 셀 조정, 노출위험

✍ 이 연구는 2018년 하반기 통계개발원 연구보고서에 수록된 자료입니다.

* 자료협조

** 패키지 구현

제1절 서 론

통계청에서는 통계지리정보서비스(SGIS)를 통해 그리드별 통계서비스를 제공하고 있다. 이 서비스는 100m, 1km 등의 격자로 구분된 지도 위에 통계를 제공하는 것으로, 최대한의 정보를 이용자에게 제공하려는 통계청의 노력을 보여주는 것이기도 하다. 따라서 그리드별 통계서비스에서는 자료 유용성 확보가 중요한 문제가 된다. 이때, 자료 유용성을 최대한으로 확보한다는 것은 제공하는 자료가 원래 자료와 차이가 거의 없다는 의미를 가지므로, 정보손실을 최소화한다는 것과 같은 말이 된다.

그러나 정보손실이 없는 자료를 제공할수록 개별정보 노출위험은 증가할 수밖에 없다. 나아가 개별정보 노출위험이 크다면 자료제공이 가능하지 않게 된다. 즉, 자료 유용성과 개별정보 노출위험은 상충관계(trade off)를 가지므로, 이를 극복하여 적절한 비밀보호 처리 방안을 찾기 위해서는 어렵고 인내심을 요구하는 연구 과정이 필요하다.

본 과제는 센서스, 행정자료 등의 전수 자료를 제공할 때, 개인 혹은 사업체 등의 개별정보는 보호하면서 자료 유용성을 극대화하기 위한 방안을 모색하기 위해 시작되었다. 특히 전수 자료를 이용해 만들 수 있는 다양한 종류의 빈도표를 제공할 때, 통계적 비밀보호(Statistical Data Confidentiality)를 위한 자료 처리 방안을 다루고 있다.

빈도표에서 어느 셀의 값이 크면, 같은 조건을 가지는 개체들이 그만큼 많은 것이므로 충분한 익명성(k-anonymity)이 확보되어 해당 셀이 안전하다고 판단한다. 그러나 셀 값이 1이면 해당 셀에 속하는 개체는 노출위험이 최대라고 판단하며, 2일 때는 해당 셀에 속하는 두 개체 중 어느 한 개체가 다른 개체를 식별할 수 있어 노출위험이 높다고 판단한다. 이처럼 빈도값이 어떤 기준값보다 작아서 익명성이 충분히 확보되지 않은 셀들은 노출위험이 높은 것으로 판단한다. 이때 노출위험이 있다고 판단하는 작은 셀에 대한 기준값은 노출위험에 대한 사회 구성원의 합의에 따라 결정할 수 있으나, 통계분석에 오류를 일으키지 않는 범위도 고려해야 한다. 국내외 통계작성기관들은 사업체는 3, 개인이나 가구는 5를 기준값으로 사용하는 경향이 있다.

단순한 이차원 빈도표에서는 주어진 기준값 이상으로 익명성을 확보하여 셀 값을 제공하는 것이 어렵지 않다. 작은 셀들을 NA[1]처리(local suppression)하고, 행의 합이나 열의 합은 작은 셀을 제외해서 계산하면 된다. 또는 행의 합이나 열의 합을 제공하면서 몇 개의 셀을 추가로 NA처리(secondary suppression)해서 작은 셀들의 값을 역산할 수 없게 만들면 된다. 하지만 복잡한 위계 구조를 가지는 대량의 빈도표를 비교적 정확하게 제공하면서 노출위험을 일정 수준 이하로 유지하는 해법은 아직 제시되어 있지 않았다. 따라서

1) not available의 약자

위계 구조를 가지는 대량의 빈도표를 제공할 때 정보손실과 노출위험을 동시에 축소하기 위한 새로운 방안을 연구할 필요가 있다.

빈도표를 제공할 때 개별정보가 노출될 위험을 일정 수준 이하로 제한(Statistical Disclosure Limitation/Control)하면서도 정보손실이 적어 원자료 분석 결과와 유사한 값을 제공하는 것은 쉬운 일이 아니다. 효율적인 방법을 찾기 위해서, 연구 과정에서 먼저 전통적인 빈도표의 매스킹 기법들을 정리하고 해외 사례를 검토했다. 이러한 과정을 통해 만족스러운 방안을 찾아 도입하기는 어렵다는 결론에 도달하여, 본 연구에서는 새로운 알고리즘을 제안한다. 제안된 알고리즘은 BSCA(Bounded Small Cell Adjustments)라고 부르기로 한다. 본 보고서에는 알고리즘 설명 및 이 알고리즘을 실제 자료에 적용한 결과를 함께 제시한다. BSCA의 아이디어는 2016년 통계개발원에서 관련 연구 과제를 수행하는 과정 중에 제안되었고, 본 연구 과제를 수행하면서 알고리즘, 이론적 근거, 프로그램 등이 완성되었다. 한편 BSCA 알고리즘은 통계적 비밀보호 분야의 주요 학회 중 하나인 PSD 2018[2]에서 발표 및 발간되었다.

보고서의 구성은 다음과 같다. 2절에서는 빈도표의 통계적 비밀보호 문제를 이해하기 위해 하나의 마이크로데이터에서 대단히 많은 개수의 빈도표를 제공할 때 빈도표 사이의 위계 구조를 설명한다. 대전 지역 및 그리드 '다바' 지역 인구주택총조사(인총) 자료를 예제로 사용했다. 대량의 빈도표를 제공하는 시스템인 SGIS에 대한 자세한 소개는 <부록> 1.에 제시되어 있다. 3절에서는 빈도표를 위한 매스킹 기법들과 노출 발생을 설명한다. 매스킹 기법들 중 영국 인총자료를 위해 심층 연구되었던 작은 셀 조정 기법의 장단점도 함께 논의한다. 관련 영국 사례의 자세한 사항은 <부록> 2. 및 3.에서 다룬다. 4절에서는 새로 제안하는 알고리즘인 BSCA를 설명하고, 실제로 SGIS 시스템에 적용하기 위한 절차를 보여준다. 복잡한 증명은 Park(2018)을 참고할 수 있다. 이론적인 내용을 모두 이해하지 않더라도, 4절 마지막 부분의 R패키지 활용을 위한 간단한 코드 사용법을 익히면 BSCA를 사용하는데 무리가 없다. 본 보고서는 5절의 맺음말로 마친다.

제2절 위계 구조를 가지는 대량의 빈도표 제공

1. 빈도표의 위계 구조

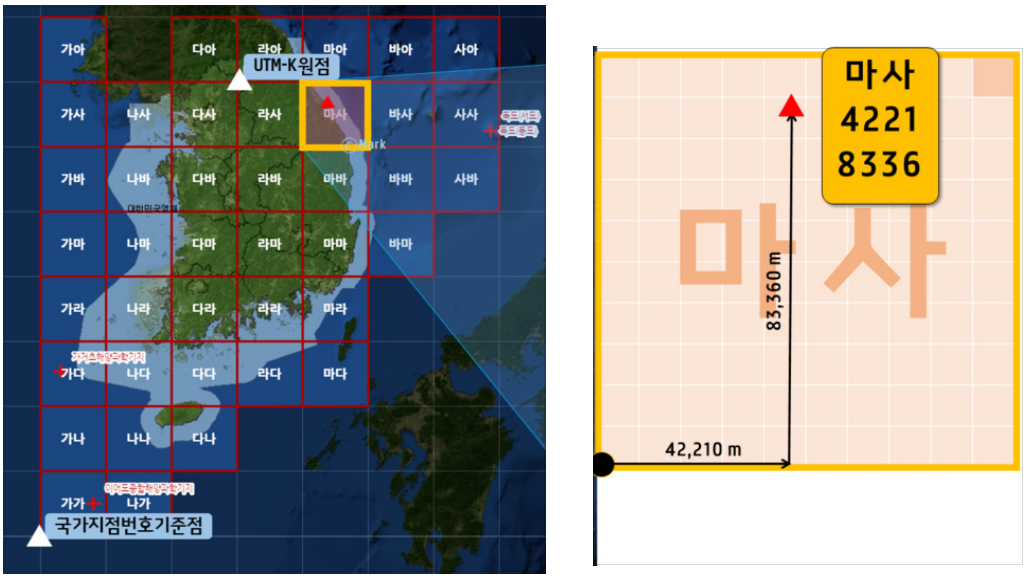
SGIS가 제공하는 통계에는 빈도, 비율, 지수 등이 있다. 이용자가 선택한 수준의 지역 단위별로 SGIS는 주어진 마이크로데이터를 집계하여 얻은 통계를 제공한다. 집계를

2) Privacy in Statistical Databases 2018 (https://unescoprivacychair.urv.cat/psd2018)

위한 자료처리 과정에서 고려해야 하는 것은 데이터가 보유한 변수의 개수 P와 지역 수준의 개수가 된다.

행정 구역에 근거한 지역 수준은 보통 집계구(OA), 읍면동(LA3), 시군구(LA2), 시도(LA1)로 구성되며, 그리드 서비스에서는 지역 수준이 격자단위 100m(OA), 1km(LA3), 10km(LA2), 100km(LA1)로 이루어진다. 약자 OA는 'Output Area'를 LA는 'Local Area'를 의미하며, 편의상 기존 지역단위와 격자단위 모두 동일한 약자를 사용하도록 하겠다. 전국은 LA0로 표현하도록 한다.

다음 [그림 2-1]은 각 그리드에 기호를 지정하는 방법을 설명하고 있다. 먼저 국가지점번호 기준점에서 가로와 세로를 100km 단위로 동쪽과 북쪽으로 이동하며 가나다 순으로 기호를 부여한다. 각 100km×100km 그리드 안에서는 서남쪽 모서리부터 거리를 미터 단위로 나타내고 이 숫자를 가로축, 세로축 순서로 표시해서 그리드 기호로 사용한다. 이는 1m×1m격자를 나타내는 것이며, 1m×1m격자에 주어진 가로축 세로축 숫자를 100으로 나눈 값의 정수 부분 3자리 숫자를 사용해 100m×100m 격자의 기호로, 1,000으로 나눈 값의 정수 부분 2자리 숫자를 1km×1km 격자 기호로, 10,000으로 나눈 값의 정수부분 1자리 숫자를 10km×10km 격자 기호로 사용한다. [그림 2-1]의 오른쪽 부분은 기준점에서 가로(서→동)로 다섯 번째, 세로(남→북)로 일곱 번째 그리드 '마사'에 대해서, 서남쪽 모서리에서 동쪽으로 42,210m, 북쪽으로 83,360m 이동한 지점의 10m×10m 수준의 그리드 기호가 '마사 4221 8336'이 되는 것을 예시한다.



[그림 2-1] 국가지점번호 부여 방법

대전 지역의 2010년 인총 '인구' 자료 및 그리드 '다바' 지역의 2016년 인총 '가구' 자료에 대해 각각의 지역 수준은 다음과 같이 정리할 수 있다. 인총자료는 인구, 가구, 주택자료로 구성되어 있으며, 자료의 양이 방대하므로 본 연구에서는 행정지역별로는 2010년[3] '인구' 자료, 그리드별로는 2016년 '가구' 자료 총 2개 마이크로데이터 파일을 이용해 BSCA 알고리즘을 테스트했다. 다음 <표 2-1> 및 <표 2-2>는 대전 지역 및 그리드 '다바' 지역에 대해 각 지역 단위의 개수와 크기를 요약하고 있다. 이때 크기란 인구수 혹은 가구수를 나타낸다. 참고로, <표 2-2>에서 그리드 LA1에 대해서 100개 LA2, 10,000개 LA3, 1,000,000개 OA가 존재해야 하지만 빈도수가 0인 지역 단위는 제외했다.

〈표 2-1〉 대전 행정 지역 : 지역 수준별 규모

지역 수준	OA (집계구)	LA3 (읍면동)	LA2 (시군구)	LA1 (시도)
지역 개수	2,506	78	5	1
평균 인구수	586	18,841	293,920	1,469,599
인구수 범위	58~4,125	2,578~47,734	200,371~491,320	1,469,599

〈표 2-2〉 그리드 '다바' 지역 : 그리드 수준별 규모

지역 수준	OA (100m)	LA3 (1km)	LA2 (10km)	LA1 (100km)
지역 개수	144,473	8,259	98	1
평균 가구수	13	232	19,541	1,914,981
가구수 범위	1~1,268	1~13,287	188~269,844	1,914,981

SGIS가 집계하는 '인구' 마이크로데이터는 행정지역 변수를 포함하여 다음 <표 2-3>과 같은 형식을 가진다. 먼저 4개($P=4$) 변수를 가지는 인총의 '인구' 자료를 대상으로 빈도표의 위계 구조(hierarchical structure)를 살펴보자. '인구' 데이터의 4개 변수는 성별(G, 2), 나이(A, 18), 혼인상태(M, 5), 교육수준(E, 8)이며 괄호 안의 숫자는 범주의 개수를 나타낸다. 지역변수는 총 13자리로 구성되며, 시도(2자리), 시군구(2자리), 읍면동(3자리), 집계구(6자리)를 이어 붙여 만든다. 다음 표의 지역 변수에서는 대전을 나타내는 25는 생략되어

3) 보다 많은 변수의 처리 여부를 테스트하고 알고리즘 개선 여부도 점검하기 위해, 기존 통계개발원 연구과제(박민정과 성주현, 2016)에서 사용했던 것과 동일한 자료를 사용했다.

있으며, 변수 맨 앞부분의 시군구를 나타내는 01~05는 0을 생략해 나타냈다. 한편 대전시 총인구는 $N=1,469,599$이며, 이는 마이크로데이터에서 행의 개수이기도 하다.

〈표 2-3〉 행정지역 변수를 가지는 '인구' 자료 마이크로데이터 형식

id	G	A	M	E	area
1	1	1	1	1	1068020001
...	2	17	5	1	2057040003
...	1	15	1	3	3070020402
...			...		
N	2	18	5	8	5060020479

마이크로데이터를 집계하여 제공할 때 변수 개수에 따라 다양한 빈도표가 만들어진다. 총 4개 변수로 구성된 빈도표는 다음 <표 2-4>와 같다. 마이크로데이터와 빈도표가 얼핏 유사해 보일 수도 있으나, <표 2-3>의 마이크로데이터는 $N(=1,469,599)$개의 열을 가지는데 비해, <표 2-4>의 빈도표는 $K(=2\times18\times5\times8=1,440)$개의 열로 구성된다. 각 지역마다 이러한 빈도표를 만들 수 있으며, 빈도표에 나타난 빈도의 합은 해당 지역 개체수의 총계 N_{LA}가 된다.

참고로 빈도표라면 보통 2차원을 생각하기 쉬운데 변수가 많은 경우 다음 <표 2-4>와 같은 펼친 형태를 이용해 빈도표를 표현할 필요가 있다. 이러한 빈도표를 flattened frequency table이라고 부르며, 유럽 지역에서는 cube 또는 hypercube라고 부르기도 한다. 참고로 R 프로그램에서는 ftable이라는 명령어를 사용해 이러한 형식의 빈도표를 얻을 수 있다.

〈표 2-4〉 행정지역별 '인구' 자료 빈도표 형식 (변수 4개, 각 지역단위마다 작성)

case	A	G	M	E	frequency
1	1	1	1	1	$f_{1,1,1,1}=f_1$
2	2	1	1	2	$f_{1,1,1,2}=f_2$
...		...			...
K	2	18	5	8	$f_{2,18,5,8}=f_K$
total					N_{LA}

한편, 동일한 마이크로데이터로부터 변수의 개수가 P개인 빈도표뿐만 아니라, 보다 적은 $p(<P)$개 변수에 대한 빈도표들도 만들 수 있다. 총 4개 변수로 구성된 데이터에서 3개 변수로 구성된 빈도표는 다음과 같이 4종류가 있다. 이러한 빈도표들 역시 각 지역마다 존재한다.

〈표 2-5〉 행정지역별 '인구' 빈도표 형식 (변수 3개, 각 지역단위마다 작성)

case	G	A	M	frequency
1	1	1	1	$f_{1,1,1}=f_1$
...		...		...
K (180)	2	18	5	$f_{2,18,5}=f_K$

case	G	A	E	frequency
1	1	1	1	$f_{1,1,1}=f_1$
...		...		...
K (288)	2	18	8	$f_{2,18,8}=f_K$

case	G	M	E	frequency
1	1	1	1	$f_{1,1,1}=f_1$
...		...		...
K (80)	2	5	8	$f_{2,5,8}=f_K$

case	A	M	E	frequency
1	1	1	1	$f_{1,1,1}=f_1$
...		...		...
K (720)	18	5	8	$f_{18,5,8}=f_K$

이제 SGIS가 제공하는 '인구' 자료에 대해 빈도표를 유형별로 정리하면 다음 <표 2-6>과 같이 표현할 수 있다. 지역 수준은 하위 첨자로, 변수의 개수는 상위 첨자로 표현했다. $P=4$의 경우, 각 지역마다 1개의 T^{v4}, 4개의 T^{v3}, 6개의 T^{v2} 및 4개의 T^{v1}의 빈도표가 존재한다. 일반적으로 ${}_PC_P=1$개의 T^{vP}, ${}_PC_{P-1}=P$개의 $T^{v(P-1)}$, ..., ${}_PC_1=P$개의 T^{v1}유형의 빈도표가 만들어지므로, 각 지역마다 총 2^P-1개의 빈도표가 있게 된다. 이상의 내용을 종합하면 다음과 같은 수식으로 모든 유형의 빈도표를 표현할 수 있다.

$$T_{i.LA}^{j.vk},\ k=1,...,P,\ j=1,...,{}_pC_k,\ i=1,...,|LA|$$

〈표 2-6〉 '인구' 자료에 대한 빈도표 유형

T	OA	LA3	LA2	LA1
$p=4$	T_{OA}^{v4}	T_{LA3}^{v4}	T_{LA2}^{v4}	T_{LA1}^{v4}
$p=3$	T_{OA}^{v3}	T_{LA3}^{v3}	T_{LA2}^{v3}	T_{LA1}^{v3}
$p=2$	T_{OA}^{v2}	T_{LA3}^{v2}	T_{LA2}^{v2}	T_{LA1}^{v2}
$p=1$	T_{OA}^{v1}	T_{LA3}^{v1}	T_{LA2}^{v1}	T_{LA1}^{v1}

이제 마이크로데이터를 그대로 집계하여 얻는 빈도표는 T로 표기하도록 하고, 빈도표 T에 통계적 비밀보호 처리를 하여 바꾼 빈도표는 $\tilde{T}$로 표기하도록 한다. 통계작성기관은 T 대신에 $\tilde{T}$를 제공하여 개별정보 노출위험을 제어하고자 한다.

가장 세부적인 빈도표는 T_{OA}^{vP} 이며, 이 빈도표는 제공하는 모든 변수들 조합에 따른 빈도를 모든 세부지역에 대해 가지고 있다. 이제 이 빈도표를 기저 빈도표(full table)이라고 부르도록 하겠다. 다른 모든 빈도표는 마이크로데이터를 집계하여 얻을 수도 있지만 기저 빈도표를 집계하여 얻을 수도 있다. 기저 빈도표를 집계하여 얻을 수 있는 빈도표들을 상위 수준 빈도표라고 부르도록 하겠다. 변수의 개수가 적어지면서 상위 수준 빈도표를 얻을 수도 있고, 지역이 넓어지면서 상위 수준 빈도표를 얻을 수도 있다.

참고로 4개의 변수를 대전 지역에서 제공할 때 다루어야 하는 빈도의 개수를 정리하면 다음 <표 2-7>과 같다. 총 빈도의 개수는 7,969,430로 약 800만 개가 된다. 만약 변수 범주별로 중복 선택을 허용하면 다루어야 할 빈도의 개수는 $2^{(2+18+5+8)}-1$로 85억 개가 넘는다. 본 연구에서는 중복 선택은 고려하지 않고 알고리즘을 개발했으며, 중복 선택에 관한 문제는 정보손실을 수용하는 것으로 마무리한다. 이에 관해서는 6절에서 자세히 설명한다.

〈표 2-7〉 빈도표 유형별로 생성되는 빈도의 개수

	OA (2506개 지역)	LA3 (78개 지역)	LA2 (5개 지역)	LA1 (1개 지역)
4개 변수 (1440개 빈도)	1440×2506 3608640	1440×78 112320	1440×5 7200	1440×1 1440
3개 변수 (1268개 빈도)	1268×2506 3177608	1268×78 98904	1268×5 6340	1268×1 1268
2개 변수 (336개 빈도)	336×2506 842016	336×78 26208	336×5 1680	336×1 336
1개 변수 (33개 빈도)	33×2506 82698	33×78 2574	33×5 165	33×1 33

하나의 마이크로데이터에서 다양한 유형의 빈도표를 지역별로 서비스할 때는 지금까지 설명한 빈도표 사이에 발생하는 위계(hierarchy) 구조를 이해할 필요가 있다. 이는 개별정보 노출위험을 제어하기 위해서 뿐만 아니라 SGIS 시스템의 작동 원리를 이해하기 위해서도 필요하다.

한편 SGIS가 집계하는 '가구' 마이크로데이터는 그리드 지역 변수를 포함하여 다음 <표 2-8>과 같은 형식을 가진다. 인총 '가구' 자료에 대해서는 SGIS가 그리드에 따른 세대구성 정보만을 서비스[4]한다. 이 경우 총 변수 개수 $P=1$이 된다. 제공해야 하는 빈도표의 종류는 다음 <표 2-9>와 같다.

〈표 2-8〉 그리드별 '가구' 자료 마이크로데이터 형식 (변수 1개)

id	H	grid
1	3	다바699952
...	2	다바011926
...	5	다바048202
...		
N	7	다바020673

〈표 2-9〉 '가구' 자료에 대한 빈도표 유형

T	OA	LA3	LA2	LA1
$p=1$	T_{OA}^{v1}	T_{LA3}^{v1}	T_{LA2}^{v1}	T_{LA1}^{v1}

2. SGIS 시스템의 자료 구조

SGIS는 이용자가 요청할 때마다 마이크로데이터를 집계하지는 않는다. 그렇게 시스템을 구축하면 빠른 시간에 통계를 제공하기 어렵기 때문이다. 효율적으로 자료를 제공하기 위해 대부분의 경우에 대해 집계된 자료인 빈도표를 지역단위마다 데이터베이스로 구축해두고, 이용자 요청에 따라 집계된 자료를 재집계하여 제공한다. 즉, 마이크로데이터 대신에 사전에 데이터베이스로 구축해 둔 기저 빈도표를 이용해 빠르게 이용자의 쿼리에 응답한다.

SGIS에서는 제공하는 빈도표들을 데이터마트(datamart)라고 부른다. 데이터마트는 앞에 설명한 T_{LA}^{vp}에 해당한다. 자료제공 속도 향상을 위해 SGIS 내부에서는 여러 유형의 데이터마트를 마이크로데이터와 함께 데이터베이스로 사용할 수 있다. 예를 들면 T_{OA}^{vP}

4) 실제 SGIS의 서비스는 총 6개 가구 유형에 따른 빈도값을 그리드 단위로 지도상에 표시한다. 여기서는 대량의 빈도표에 대한 노출제어 연구를 목적으로 데이터를 사용하므로, 마이크로데이터에 표기된 대로 7개 범주에 대해 빈도표를 생성하도록 한다.

및 T_{LA2}^{vP}를 모두 데이터마트로 사용하고, 다른 유형의 빈도표는 T_{OA}^{vP} 또는 T_{LA2}^{vP}를 집계하여 제공할 수 있다. 또한 4개 지역 수준별로 기저 빈도표들을 모두 데이터베이스로 구축할 수도 있다.

본 연구는 이미 사용하고 있는 데이터마트 대신에 통계적으로 비밀보호 처리된 빈도표를 데이터마트로 사용하도록 제공하고, 상위 수준 빈도표를 집계할 때 SGIS 시스템 내부에 약간의 코드를 추가하여 전체적으로 통계적 비밀보호 처리가 이루어지게 하는 방안을 수립하는 것을 목표로 하고 있다. 따라서 큰 폭으로 시스템을 개선할 필요 없이, 현재 구축된 시스템을 효율적으로 활용할 수 있다.

3. 이용자가 생성하는 빈도표

이용자는 SGIS를 통해 얻은 빈도표들을 합계하여 새로운 빈도표를 만들 수 있다. 특정 변수 조합에 대해서 지역별로 얻은 빈도들을 더해서 상위 수준 빈도값을 계산할 수도 있고, 한 지역에서 변수들을 줄여 나가면서 상위 수준 빈도값을 얻을 수도 있다. 이용자가 생성하는 빈도표를 다음과 같이 표현할 수 있다. 변수를 줄여 상위 수준 빈도표를 만들면 윗첨자에 ag를, 지역별 빈도표를 취합해 상위 수준 빈도표를 만들 때는 아래첨자에 ag를 표기한다. 예를 들어 $\widetilde{T}_{i.LA3.ag}^{p4}$는 i번째 LA3 지역에 속한 하위 집계구들의 빈도표 $\widetilde{T}_{OA\in i.LA3}^{p4}$에 나타난 빈도들을 더해서 얻는다.

다음 <표 2-10> 및 <표 2-11>은 이용자가 집계하여 얻을 수 있는 빈도표 유형들을 정리한 것이다. SGIS는 이 빈도표들을 제공하지 않는다. 이용자는 SGIS가 제공하는 $\widetilde{T}$와 자신들이 만들어낸 $\widetilde{T}_{ag}$또는 $\widetilde{T}^{ag}$를 비교해서 참값 T를 알아내려 시도할 수 있다.

〈표 2-10〉 이용자가 생성하는 빈도표 : 상위 지역 수준으로 집계하는 경우

$\widetilde{T}_{ag}$	OA	LA3	LA2	LA1
$p=4$		$\widetilde{T}_{LA3.ag}^{p4}$	$\widetilde{T}_{LA2.ag}^{p4}$	$\widetilde{T}_{LA1.ag}^{p4}$
$p=3$		$\widetilde{T}_{LA3.ag}^{p3}$	$\widetilde{T}_{LA2.ag}^{p3}$	$\widetilde{T}_{LA1.ag}^{p3}$
$p=2$		$\widetilde{T}_{LA3.ag}^{p2}$	$\widetilde{T}_{LA2.ag}^{p2}$	$\widetilde{T}_{LA1.ag}^{p2}$
$p=1$		$\widetilde{T}_{LA3.ag}^{p1}$	$\widetilde{T}_{LA2.ag}^{p1}$	$\widetilde{T}_{LA1.ag}^{p1}$

〈표 2-11〉 이용자가 생성하는 빈도표 : 변수의 수를 줄여 집계

$\widetilde{T}^{ag}$	OA	LA3	LA2	LA1
$p=4$				
$p=3$	$\widetilde{T}_{OA}^{v3.ag}$	$\widetilde{T}_{LA3}^{v3.ag}$	$\widetilde{T}_{LA2}^{v3.ag}$	$\widetilde{T}_{LA1}^{v3.ag}$
$p=2$	$\widetilde{T}_{OA}^{v2.ag}$	$\widetilde{T}_{LA3}^{v2.ag}$	$\widetilde{T}_{LA2}^{v2.ag}$	$\widetilde{T}_{LA1}^{v2.ag}$
$p=1$	$\widetilde{T}_{OA}^{v1.ag}$	$\widetilde{T}_{LA3}^{v1.ag}$	$\widetilde{T}_{LA2}^{v1.ag}$	$\widetilde{T}_{LA1}^{v1.ag}$

4. 대용량 빈도표 제공 시스템의 고려 사항

다양한 지역 수준에 대해 대량의 빈도표를 제공할 때는 위계 구조가 발생하므로, 같은 변수 조합에 대한 빈도값을 여러 경로로 만들 수 있다. 빈도표를 서로 다른 경로로 합계하여도 값들이 일치하는 경우, 가산성(additivity)을 만족한다고 한다. 즉, $\widetilde{T}=\widetilde{T}^{ag}=\widetilde{T}_{ag}$가 되면, 시스템이 가산성을 갖추었다고 한다.

SGIS에서 빈도를 제공할 때 시스템은 이용자의 동일한 요청에 대해 동일한 값을 제공하는 것이 바람직하다. 이를 셀 값에 대한 일치성(consistency)라고 부른다. 예를 들어 노출위험을 줄이기 위해서 이용자가 요청할 때마다 값을 랜덤하게 조금씩 바꾸어 제공하는 경우 시스템은 일치성을 갖추지 못하게 된다. 뿐만 아니라 반복적인 요청으로 얻은 값들의 분포를 살피면 참값을 알아낼 수도 있으므로 노출위험을 줄이지도 못하면서 일치성도 잃어버릴 수 있다.

한편 시스템은 정보손실(information loss, IL)을 적게 가져야 한다. 이것은 $\widetilde{T}-T$로 표현할 수 있다. 참 빈도값과 제공하는 빈도값의 차이가 적은 자료를 제공하는 것이 이용자 입장에서 바람직하다.

마지막으로 시스템은 노출위험(disclosure risk, DR)을 일정 수준 이하로 유지해야 한다. 노출위험은 여러 방법으로 측정할 수 있겠지만, 정해진 수준의 노출위험을 지키면서 자료를 제공할 필요가 있다. 이는 자료제공기관 입장에서 반드시 필요한 조건이라고 할 수 있다.

지금까지 언급한 가산성, 일치성, 정보손실 최소화, 노출위험 제어를 모두 만족하며 자료를 제공하는 시스템을 구축한 사례는 없다. 따라서 이들 사이의 상충관계를 잘 살피고, 우선순위를 두어 자료제공과 비밀보호 전략을 세울 필요가 있다. BSCA알고리즘은 가산성을 만족하지 못하지만 일치성을 만족하고, 정보손실과 노출위험 제어가 가능하다. 이에

관한 자세한 설명은 4절에서 다룬다. 빈도표에 대한 다른 비밀보호 방법들 역시 이렇게 4가지 측면에서 평가할 수 있다.

대용량 빈도표의 비밀보호 처리를 본격적으로 다루기에 앞서, 다음 절에서는 빈도표에서 노출이 무엇이며 어떠한 매스킹 기법들을 사용해 왔는지 살펴보자.

제3절 빈도표를 위한 매스킹 기법

1. 빈도표에서 개별정보 노출

빈도표에서 노출은 여러 경로를 통해 일어날 수 있다. 먼저 제공되는 빈도표에서 셀 값이나 주변합이 1이면 개체가 식별된다. 이외에도 제공하는 빈도표들을 종합적으로 고려하면 파악할 수 있는 노출 유형들이 있다. 다음 <표 2-12>에 그러한 노출의 종류들을 정리했다(Hundepool 외, 2012).

먼저 그룹 속성 노출의 경우는, 빈도표에서 값이 작은 셀들만 점검해서는 간과하기 쉬운 노출 유형이다. 예를 들어 질병 종류에 대해 작성된 연령·지역별 빈도표에서 특정 범주에만 빈도가 몰려 있을 경우, 해당 그룹이 모두 걸린 질병의 종류가 노출되고 이는 각 개체의 개별 정보가 노출됨을 의미한다. 얼핏 보아서는 빈도값이 크기 때문에 노출이 없다고 생각할 수 있지만, 다른 셀들이 모두 값이 0이라면 아무리 큰 빈도를 가지는 그룹이라고 하더라도 해당 속성이 노출될 수밖에 없다.

다음으로 포함 관계에 있는 빈도표들을 서로 빼면 통계기관이 제공하지 않는 빈도표들이 추가로 생성되어 노출이 일어날 수 있다. 이를 차이에 의한 노출이라고 한다. 예를 들어 소수의 시설 거주 노인들의 정보를 보호하기 위해 해당 빈도표를 제공하지 않고, 전체 노인의 빈도표와 개별 주택에 거주하는 노인의 빈도표만을 제공할 수 있다. 하지만 두 빈도표의 빈도값들의 차이를 통해 시설에 거주하는 노인에 관한 빈도표를 작성할 수 있다.

마지막으로 통계기관이 제공한 빈도표들을 서로 연결하고 유추하여 새로운 정보를 만들 수 있다. 이를 표 연계에 의한 노출이라고 한다. 하나의 마이크로데이터에서 여러 빈도표들을 만들어 제공하는 경우에는 각 빈도표들의 변수가 일치하는 경우가 많으므로 빈도표들을 연결하는 것이 더 쉬워지고, 연계에 의한 노출이 일어날 확률이 높아진다.

〈표 2-12〉 빈도표에서 발생하는 노출의 종류

개별 속성 노출 (individual attributes disclosure)	- 빈도표의 어떤 부분 합(margin)이 1인 경우
그룹 속성 노출 (group attributes disclosure)	- 열이나 행이 대부분 0이거나 작은 값일 때 예) 특정 지역 65세 모두 특정 건강 상태
차이에 의한 노출 (disclosure by difference)	- 포함관계인 두 빈도표의 차이로 새 빈도표 발생 예) 전체 노인 - 개인 주거 노인 = 시설 주거 노인
표 연계에 의한 노출 (disclosure by linking tables)	- 하나의 마이크로데이터에서 여러 빈도표가 만들어지는 경우, 공통 셀 및 부분합들이 많아 연계가 더욱 용이

예를 통해 빈도표에서 개별정보 노출을 좀 더 자세히 알아보자. 어느 지역 A에서 다음 <표 2-13>과 같은 빈도표가 작성되었다고 하자. 이 표를 통해 해당 지역에서 미혼(S)인 여성(F)은 1명이라는 사실이 드러나며, 이처럼 셀 값이 1일 때 식별이 일어난다고 한다.

〈표 2-13〉 빈도표를 통한 개체 식별 예제

혼인상태	성별		합계
	M	F	
M	38	17	55
D	7	4	11
S	3	***1***	4
합계	48	22	70

한편 셀 값이 1인 셀이 다른 빈도표에서 행 또는 열의 합으로 주어질 때, 속성이 노출된다. 다음 <표 2-14>를 살펴보자. 유일했던 미혼 여성의 근무시간 유형은 3번에 속한다는 사실이 노출됨을 알 수 있다. 즉, 유일하여 식별된 개체의 근무시간 속성이 노출되고, 이를 속성 노출이라고 한다.

다음으로 한 셀을 제외한 모든 셀에서 값이 0이면, 그 한 셀에 속하는 모든 사람의 속성이 노출된다. 이를 그룹 속성 노출이라고 한다. 앞의 <표 2-14>에서는 혼인 상태가 D인 여성 4명의 근무시간 유형은 모두 1번이라는 속성이 노출된다. 즉 나와 같은 특징을 가지는 사람이 4명 있어서 익명성이 확보되었지만, 근무시간이라는 속성에 대해 모두 같은 범주에 속하므로 그룹 속성 노출이 일어난다. 질병번호와 같이 민감한 변수에 대해서는 그룹 속성 노출을 특히 더 주의할 필요가 있다.

〈표 2-14〉 빈도표를 통한 속성 및 그룹 속성 노출 예제

성별	혼인상태	근무시간	빈도수	빈도수
F	M	1	14	17
		2	3	
		3	0	
	D	1	4	4
		2	0	
		3	0	
	S	1	0	***1***
		2	0	
		3	***1***	

추가로 다음과 같은 예를 통해 지금까지 설명한 노출에 대해 다시 한 번 정리해보자. 다음 <표 2-15>와 같은 나이 및 교육수준에 대한 빈도표가[5] 제공된다고 하자.

〈표 2-15〉 빈도표를 통한 그룹 속성 노출 예제

교육 수준	나이						총합
	16-25	26-49	40-59	60-69	70-79	80-	
1	17	16	13	8	4	1	59
2	23	14	6	2	0	0	45
총합	40	30	19	10	4	1	104

나이가 80 이상인 경우의 부분합이 1이기 때문에 식별이 일어나고, 교육수준 1에 해당하는 셀값도 1이므로 이 개체의 교육수준이 1이라는 새로운 사실이 파악되어 속성 노출이 일어난다. 또한 60-69세 및 교육수준 2인 셀의 경우, 이 셀에 속한 응답자는 다른 한 명을 식별할 수 있게 된다. 한편, 70-79세의 경우, 모든 개체가 교육수준 1에 모여 있기 때문에 그룹의 속성이 노출된다고 할 수 있다. 이와 같이 한 셀을 제외하고 행이나 열이 0으로 구성될 경우, 그룹 속성 노출이 일어나게 된다.

빈도표를 제공할 때 의도치 않게 행이나 열의 합이 주어져서 차이에 의해 셀의 값이 파악되는 경우도 있다. 이를 차이에 의한 노출(disclosure by difference)라고 부른다. 차이에

5) Sholomo, N(2015), Statistical disclosure control: A risk-utility framework, lecture material for the 1st EMOS Spring School.

의한 노출은 구역 변경 등으로 인해 노출이 일어나는 것과 각 빈도표가 제공하는 변수의 범주 등의 차이에 의해 노출이 일어나는 것으로 나누어 생각할 수 있다. 먼저 통계 제공 구역 변경으로 인해 노출이 일어나는 사례를 살펴보면 다음 <표 2-16>과 같다. 기준년 1에서 기준년 2가 되면서 지역 C의 구역이 미세하게 조정되었다고 하자. 지역C에 새로 포함된 작은 구역에 사는 미혼 여성 1명의 근무시간 유형이 1번에 해당함을 알 수 있다.

〈표 2-16〉 지역 기준 변경으로 발생하는 차이에 의한 노출 예제

미혼(S)		빈도수		지역 변경에 의한 빈도값 차이
성별	근무시간	기준년 1	기준년 2	
		지역C	지역C	
M	1	50	52	2
	2	128	130	2
	3	39	39	0
F	1	54	55	1
	2	140	140	0
	3	49	49	0

또한 범주가 조금 다른 빈도표를 각각 제공할 때 차이에 의한 노출이 발생할 수 있다. <표 2-15>의 사례에 대해서 다음 <표 2-17>과 같은 또 다른 빈도표가 공표되었다고 한다면, 두 빈도표의 나이 범주에 대해서 서로 포함 관계가 성립되므로 이 지역에 거주하는 15세는 한 명이라는 정보를 파악할 수 있게 된다. 이러한 노출은 차이에 의한 노출이 된다. 이런 경우가 생기지 않으려면 통계기관은 제공하는 다양한 통계에 대해 동일한 범주를 사용하려는 노력을 지속할 필요가 있다.

〈표 2-17〉 변수 범주 변화로 발생하는 차이에 의한 노출 예제

나이	15-25	26-49	40-59	60-	총합
총합	41	30	19	15	105

마지막으로 표 연계에 의한 노출을 알아보자. 주어진 정보가 다음 <표 2-18> 및 <표 2-19>와 같다고 하자.

〈표 2-18〉 안전하게 제공된 빈도표 1

지역A 근무	빈도
O	49
X	

지역A 거주	빈도
O	102
X	

지역A 취업	빈도
O	85
X	17

〈표 2-19〉 안전하게 제공된 빈도표 2

지역A 근무 지역A 거주	남성		
	지역A 근무 지역A 거주	타지역 근무 지역A 거주	지역A 근무 타지역 거주
22	21	58	23

이러한 정보는 단순한 집계 결과로 안전하다고 판단할 수 있다. 그러나 이렇게 주어진 정보를 합으로 사용하면서, 두 빈도표를 연계하면 다음 <표 2-20>과 같이 노출이 발생할 수 있다. 기울어짐체로 표시된 1, 5, 4, 63은 원래 빈도표가 제공하지 않지만 이용자가 계산해 낼 수 있는 값들이다. 특히 1의 값이 드러나면서 통계기관이 의도하지 않은 노출이 발생한다.

〈표 2-20〉 안전하게 제공된 빈도표 1과 빈도표 2를 연계했을 때 발생하는 노출

지역A 거주	지역A 취업	지역A 근무	성별	빈도수			
O	O	O	M F	21 ***1***	22	85	102
		X	M F	58 ***5***	***63***		
	X	O	M F	- -	-	17	
		X	M F	- -	-		
X	O	O	M F	23 ***4***	27	-	-
		X	M F	- -	-		
	X	O	M F	- -	-	-	
		X	M F	- -	-		

2. 빈도표를 위한 매스킹 기법

빈도표에서 개별정보 노출 방지를 위한 매스킹 방법들은 표 작성 이전에(pre-tabular) 마이크로데이터를 변형하는 기법들 및 표 작성 이후에(post-tabular) 빈도들을 변형하는 기법들, 두 유형으로 나누어진다.

먼저 통계표 작성 이전에 수행되는 매스킹 기법들은 자료 맞교환(record swapping), 과다대체(over-imputation) 등이 있다. 이들 기법을 마이크로데이터에 적용하여 미이크로데이터를 변형한 이후에 빈도표를 생성할 수 있으며, 이렇게 만든 빈도표를 통해서는 어떤 개체의 정보를 특정하기 어려우므로 노출위험이 감소한다고 판단한다. 마이크로데이터에서 자료 변조가 일어나고 이 마이크로데이터에서 만든 빈도표의 값은 변조하지 않아 일정하게 유지되므로, 하나의 마이크로데이터에서 다수의 빈도표를 생성할 때 수반되는 가산성이나 차이에 의한 노출에 관한 복잡한 문제를 고려하지 않아도 되는 장점이 있다.

그러나 이용자에게는 자료가 어떻게 변조되었다는 정보에 대한 투명성이 확보되지 않는 단점이 있다. 우리나라의 경우, 투명성 확보는 필수적인 문제라고 판단하여 마이크로데이터 변조를 통한 빈도표의 매스킹 처리는 고려하지 않도록 한다. 자료 맞교환을 사용할 때 매스킹 효과에 관해서는 <부록> 2.에 정리되어 있는 영국 사례 관련 논문(Shlomo 외, 2015)에 자세히 논의되어 있으니 관심 있는 독자는 참고하기 바란다.

다음으로 마이크로데이터는 변조하지 않고 빈도표에 매스킹 처리를 하여 개별정보가 노출되지 않도록 할 수 있다. 이 경우 이용자는 통계표에 어떠한 조치가 취해졌는지 비교적 투명하게 알 수 있다. 구체적인 기법들로는 감추기(suppression), 라운딩(rounding) 등이 있다. 자세한 내용은 Hundepool 외(2012)의 5장에 설명되어 있으며, 여기서는 이 문헌의 내용을 재구성해 각 기법을 소개하고 노출이 발생하는 문제를 예시를 통해 논의하도록 한다.

가. 감추기

감추기(suppression)는 널리 사용되는 매스킹 기법 중 하나이다. 개별정보의 노출위험을 지나치게 높이는 변수를 제공하지 않는 것을 감추기라고 한다. 예를 들면, 지역 변수 중 위도와 경도를 나타내는 변수를 제공하지는 않는다. 마이크로데이터를 제공할 때 희귀병 여부를 나타내는 변수를 감추고 제공하지 않을 수도 있다. 이러한 감추기를 통해 노출위험을 획기적으로 감소시킬 수 있으나 정보손실도 매우 커서 자료의 유용성이 없어질 수 있다. 어느 변수를 감춘다는 것은 자료제공 범위에 관한 문제이기도 하므로 본 연구

에서는 이러한 감추기를 고려하지 않도록 한다. 본 보고서에서는 제공범위가 결정된 상태에서 노출위험을 제어하는 문제를 논의한다.

참고로 표 재디자인(re-design)을 통해 노출위험을 낮출 수도 있다. 그룹화 등을 통해 변수들의 범주를 폭넓게 결정하는 것 등이 여기에 속하는 방법이다. 표 재디자인은 세부정보를 손실시켜 자료 유용성을 크게 감소시키므로, 감추기와 유사하게 제공범위 결정과 관련된 사항으로 판단하여 본 연구에서는 고려하지 않도록 한다. 즉, 변수제공 여부 문제와 표 재디자인은 어떻게 할지 결정되어 주어진 것으로 간주한다.

변수를 모두 감추는 대신에 민감한 셀만 감추는 것을 국소 감추기(local suppression)라고 한다. 크기가 B 미만인 셀들을 작은 셀이라 부르면, 작은 셀들은 위험하다고 판단하여 작은 셀들의 값을 NA 등으로 표기하여 통계표를 제공할 수 있다. 만약 감춘 셀이 하나라면 통계표의 행의 합과 열의 합을 통해 감추어진 값을 손쉽게 알아낼 수 있다. 때문에 그러한 계산이 가능하지 않도록 안전한 셀들 중에서 하나를 더 감추는 작업이 필요하다. 작은 셀을 감추는 것을 주요 감추기(primary suppression)이라고 부르고, 한 셀을 더 감추는 것을 2차 감추기(secondary suppression)라고 한다. 이차원 빈도표에서는 2차 감추기가 용이하지만, 다차원 빈도표에서는 2차 감추기할 셀을 결정하는 것이 매우 복잡한 문제가 된다. 따라서 국소 감추기는 통계작성기관에서 단일한 통계표를 제공할 때 흔히 사용하는 방법이지만,[6] 대량의 빈도표를 제공하기에는 적합하지 않다고 판단된다.

나. 라운딩

어떤 값 B를 기준으로 라운딩(conventional rounding)한다는 것은 가까운 B의 배수로 숫자를 반올림하는 것을 말한다. 라운딩을 하면 빈도표에는 모두 B의 배수들만 존재하게 된다. 다음 <표 2-21>이 보여주는 $B=5$인 경우의 예를 살펴보면, 다양한 값들이 모두 5의 배수로 바뀌어 있음을 볼 수 있다. 라운딩은 이용자의 요청마다 동일한 값을 제공하므로 일치성을 유지하나, 가산성이 없고, 최대 정보손실은 $B/2$보다 작거나 같게 된다. 또한 차이에 의해서 원래 값을 유추할 수도 있다. 따라서 원래 의도한 대로 노출위험을 제어하기 어렵다.

6) 2차원의 단일한 통계표를 제공할 때 감추기 처리 등에 관한 문제는 정동욱과 김경미(2008)에 잘 정리되어 있으므로 관심 있는 독자는 참고할 수 있다.

〈표 2-21〉 $B=5$를 기준으로 한 라운딩 예제

원래 빈도표

	M	F	
A	1	0	1
B	3	3	6
C	12	20	32
	16	23	39

변형된 빈도표

	M	F	
A	0	0	0
B	5	5	5
C	10	20	30
	15	25	40

기존 라운딩이 빈도값을 가장 가까운 B의 배수로 바꾸는데 비하여, 랜덤 라운딩(random rounding)은 가까운 두 개의 B의 배수 값 중 하나를 랜덤하게 선택해 빈도값을 바꿔준다. 다만 변하는 값을 확률적으로 선택하여 전체 합이 유지되도록 한다. 다음 <표 2-22>는 $B=3$일 때의 랜덤 라운딩의 과정을 보여주고 있다.

원래 값이 1 또는 2이면 0 또는 3으로 변하고, 원래값이 1, 4 또는 5인 경우 3 또는 6으로 바뀌는 것을 알 수 있다.

〈표 2-22〉 랜덤 라운딩의 원리 ($B=3$일 때)

원래 빈도값	조정 결과	
1 (4, ...)	2/3 확률로	0 (3, ...)
	1/3 확률로	3 (6, ...)
2 (5, ...)	1/3 확률로	0 (3, ...)
	2/3 확률로	3 (6, ...)

하나의 빈도표에 이와 같은 랜덤 라운딩을 실시하면 총합의 기댓값은 원래와 같게 유지된다는 것은 중요한 성질이다. 참고로 랜덤 라운딩을 B를 사용하여 일반적인 수식으로 표현하면 다음과 같고, 이때, $a_k \in Z$, $b=1,\ldots,(B-1)$이다.

$$f_k = a_k \times B + b \rightarrow \tilde{f}_k = a_k \times B \text{ with probability } (B-b)/B$$
$$f_k = a_k \times B + b \rightarrow \tilde{f}_k = a_k \times B + B \text{ with probability } b/B$$

한편, $B=5$일 때, 랜덤 라운딩의 예는 다음 <표 2-23>과 같다. 5의 배수는 그대로 유지되며 다른 값들은 가까운 5의 배수 중 하나로 변환된다. 빈도표를 한 번만 제공할 때는 무슨 값인지 특정할 수 없도록 불확실성을 발생시켜 익명성이 보장된다. 그러나 여러 번 같은 빈도를 제공할 때는 얻어진 결과값의 분포를 통해 원래 값을 특정할 수도 있다.

예를 들면 스무 번을 요청했더니 4번 0이 나오고 16번 5가 나온다면 참값은 1이라고 짐작할 수 있기 때문이다. 또한 이렇게 사용자가 요청할 때마다 랜덤 라운딩을 해서 값을 제공한다면 일치성도 상실하게 된다. 랜덤 라운딩에는 기존 라운딩과 마찬가지로 가산성이 없으며, 차이에 의한 노출이 발생할 수 있다. 정보손실은 B 미만이다. 한편, 사용자에게 비밀보호 방법에 대한 투명성이 확보되고 작은 값들은 모두 사라지기 때문에 노출 관련 인지적 위험 요소도 줄어드는 장점이 있다.

〈표 2-23〉 $B=5$를 기준으로 한 랜덤 라운딩 예제

원래 빈도표

	M	F	
A	1	0	1
B	3	3	6
C	12	20	32
	16	23	39

변형된 빈도표

	M	F	
A	0 or 5	0	0 or 5
B	0 or 5	0 or 5	5 or 10
C	10 or 15	20	30 or 35
	15 or 20	20 or 25	35 or 40

라운딩과 랜덤 라운딩은 가산성이 확보되지 않으며 역계산(unpicking)이 가능한 경우가 많다. 이를 보완하기 위해 제어된 라운딩(controlled rounding) 기법이 제안되었다. 이를 구현하기 위해서는 선형계획법(linear programing)을 사용한다. 예를 들면 다음 <표 2-24>와 같다. 모든 빈도값이 5의 배수로 바뀌었지만 행의 합, 열의 합, 총합이 모두 일치해서 가산성이 있다. 가산성을 유지하는 해를 선형계획법을 이용해 찾는 것이 제어된 라운딩의 핵심이며, 경우에 따라 해가 없을 수 있고 빈도표에 범주 개수가 많으면 해를 찾는데 많은 계산이 필요하다.

〈표 2-24〉 $B=5$를 기준으로 한 제어된 라운딩 예제

원래 빈도표

	M	F	
A	1	0	1
B	3	3	6
C	12	20	32
	16	23	39

변형된 빈도표

	M	F	
A	5	0	5
B	0	5	5
C	10	20	30
	15	25	40

이처럼 제어된 라운딩 기법은 구현이 복잡한 문제가 있다. 따라서 센서스처럼 규모가 크고 많은 빈도표를 생성하는 경우나 연계된 통계표에 적용하기 어려운 단점이 있다.

제어된 라운딩에 관한 자세한 설명과 이와 개념적으로 유사하지만 총계표(Magnitude Data)를 대상으로 적용하는 CTA(Controlled Tabular Adjustments) 기법에 관한 자세한 설명은 정동욱과 김경미(2008)를 참고할 수 있다.

마지막으로 라운딩 기법 중 준-제어 라운딩(Semi-Controlled Rounding) 기법이 있다. 가산성을 포기하여 단위 빈도표마다 제어 라운딩 기법을 사용하는 것이 기본 아이디어이며, 선형 계획법을 사용해 빈도표 전체합을 유지한다. 기존 라운딩 기법들보다 자료 유용성도 증가하며 연계된 통계표들 사이에서 일치성을 확보하는 것으로 알려져 있다. 대용량 빈도표를 제공하는 시스템(flexible table generator)에서 사용 가능하다. 그러나 계산 시간이 길고 최적해가 없는 경우도 있다.

준-제어 라운딩 기법들 중 특화된 기법으로 작은 셀 조정(Small Cell Adjustments, 이하 SCA) 및 Banardization이 있다. Banardization은 0이 아닌 셀 값에 -1, 0, +1을 일정한 확률로 더하고, 바뀐 값들을 더하여 주변합을 결정하는 것이다. 이 방법은 표들 사이에 총합이 일치하지 않고, SGIS와 같은 시스템(flexible table generators)에서는 구현하기 어려운 단점이 있다.

본 연구에서는 국내 실무 적용 가능성을 염두에 두고, 빈도표를 위한 매스킹 기법들에 관한 여러 해외 사례를 검토했다. 우리나라는 자료의 보이지 않는 변형을 대체로 허용하지 않는 정서가 있기 때문에 마이크로데이터 수준에서 자료 변형을 통한 노출제어 처리는 검토하지 않기로 했다. 마이크로데이터를 집계한 자료에 대하여서도, 자료 교환과 같은 자료 변형 기법은 적용에서 배제한다. 남은 방법들로는 호주 ABS의 방식과 영국 ONS에서 사용한 SCA가 있다. 본 연구에서는 호주 ABS의 기법을 개괄적으로 검토하였으나 증명된 효율성에 비하여 적용과 관리가 복잡하여 더 이상 고려하지 않기로 했다. 따라서 영국에서 많은 연구가 수행되었으면서 센서스 빈도표 비밀보호에 실제로 사용되었던 작은 셀 조정(SCA) 기법을 국내에서 활용하기 위해 자세한 검토를 진행했다.

다. 작은 셀 조정

영국 ONS가 센서스 빈도표의 공표를 위해 활용했던 SCA(small cell adjustments) 기법은 제 2절에서 다루었던 빈도표의 노출제어 기법 중, 일부 합을 보존하는 라운딩(semi-controlled rounding)에 해당한다. 영국(스코틀랜드 제외)에서 마이크로데이터 수준의 자료교환(swapping)과 함께 사용되었고, 호주 ABS에서도 센서스 빈도표 공표를 위해 사용되었다. SCA는 상위 수준 빈도표를 합계하는 방법에 따라 두 버전을 고려할 수 있는데, 이를 각각 SCA1 및 SCA2라고 하자. SCA1의 절차는 다음과 같다.

1. 각 OA마다 P개 변수로 구성된 기저 빈도표를 생성한다.
2. T_{OA}^{vP} 각 빈도표마다 작은 셀들을 랜덤 라운딩해서 $\widetilde{T}_{OA}^{vP}$들을 얻는다.
3. $\widetilde{T}_{OA}^{vP}$를 집계하여 $\widetilde{T}_{LA}^{vp}$를 얻는다.

SCA1의 절차 2번 및 3번을 다음 <표 2-25>와 같이 표현할 수도 있다. 두꺼운 글씨체는 랜덤 라운딩을, 화살표는 집계(aggregation)을 나타낸다. <표 2-25>의 화살표는 $\widetilde{T}_{OA}^{v4}$를 집계하고 있음을 나타낸다.

〈표 2-25〉 SCA1에서 상위수준 빈도표의 집계방식

$\widetilde{T}$	OA	LA3	LA2	LA1
p=4	$\widetilde{\boldsymbol{T}}_{\boldsymbol{OA}}^{\boldsymbol{v4}}$	$\rightarrow \widetilde{T}_{LA3}^{v4}$	$\rightarrow \widetilde{T}_{LA2}^{v4}$	$\rightarrow \widetilde{T}_{LA1}^{v4}$
p=3	$\downarrow \widetilde{T}_{OA}^{v3}$	$\searrow \widetilde{T}_{LA3}^{v3}$	$\searrow \widetilde{T}_{LA2}^{v3}$	$\searrow \widetilde{T}_{LA1}^{v3}$
p=2	$\downarrow \widetilde{T}_{OA}^{v2}$	$\searrow \widetilde{T}_{LA3}^{v2}$	$\searrow \widetilde{T}_{LA2}^{v2}$	$\searrow \widetilde{T}_{LA1}^{v2}$
p=1	$\downarrow \widetilde{T}_{OA}^{v1}$	$\searrow \widetilde{T}_{LA3}^{v1}$	$\searrow \widetilde{T}_{LA2}^{v1}$	$\searrow \widetilde{T}_{LA1}^{v1}$

실제로 대전 행정 지역 자료에 SCA1을 적용하면 어떻게 되는지 살펴보자. 총 4개 변수로 이루어진 OA 수준의 빈도표에서 1의 값을 가지는 셀은 166,218개 있고, 2의 값을 가지는 셀은 66,243개 있으므로 작은 셀은 총 232,461개이다. <표 2-7>을 참조하면 OA 수준에서 전체 셀들 개수는 1,440×2,506 = 3,608,640개이므로 총 6.4%가 작은 셀이고, 이 값들이 SCA1에 의해 바뀌는 것이다. 1 또는 2의 참값 빈도수들이 0 또는 3의 노출제어 처리된 빈도수로 바뀌는 내용은 다음 <표 2-26>과 같다. 작은 셀 값 1 또는 2가 주어진 확률 1/3 또는 2/3에 따라 0 또는 3으로 랜덤하게 바뀌었음을 확인할 수 있다.

〈표 2-26〉 행정지역별 '인구' 자료에서 기저 빈도표 작은 셀에 SCA를 적용한 결과

$T_{OA}^{v4} \setminus \widetilde{T}_{OA}^{v4}$		SCA 처리 이후 빈도		합계
		0	3	
원래 빈도	1	110,834	55,384	166,218
	2	22,082	44,161	66,243

총 6.4%의 작은 셀들의 값이 0 또는 3으로 바뀌는 것은 작은 변화가 아니다. OA 수준의 셀들 중 89.2%가 값이 0이기 때문이다. 값이 3인 셀이 1.06%이고, 3을 초과하는

크기를 가지는 셀들은 전체의 3.29%에 불과하다. 따라서 OA 수준의 작은 셀 변화를 그대로 상위 수준 빈도표들에 반영하면 정보손실이 클 수 있다. 참고로 다음 <표 2-27>은 4개 변수를 사용했을 때 각 지역 수준별로 기저 빈도표들에서 셀 값의 분포를 보여준다.

〈표 2-27〉 각 지역 수준별 기저 빈도표들의 빈도수 분포

		셀의 값(빈도수)					합계
		0	1	2	3	4+	
LA1	셀 개수	547	49	27	21	796	1,440
	셀 비율(%)	37.99	3.40	1.88	1.46	55.27	100
LA2	셀 개수	3,271	353	222	164	3,190	7,200
	셀 비율(%)	45.43	4.90	3.08	2.28	44.31	100
LA3	셀 개수	73,355	8,422	4,448	2,816	23,279	112,320
	셀 비율(%)	65.31	7.50	3.96	2.51	20.72	100
OA	셀 개수	3,218,982	166,218	66,243	38,262	118,935	3,608,640
	셀 비율(%)	89.20	4.61	1.84	1.06	3.29	100

한편, SCA2는 SCA1에서 발생하는 정보손실을 줄이기 위해 각 지역 수준마다 기저 빈도표에 대해 랜덤 라운딩을 실시하고, 변수 개수를 줄여 얻는 상위 수준 빈도표는 해당 지역 수준에서 각 기저 빈도표를 집계해서 얻는 방식을 말한다. 이렇게 집계하면 $\tilde{T}^{v1}_{LA1}$과 같은 경우, SCA1에 비해 작은 셀들에서 생기는 정보손실이 훨씬 적게 된다. 다만 이러한 SCA2 방식은 영국 ONS에서 사용한 것이 아니며, 노출위험과 정보손실을 검토하는 과정에서 비교 대상으로서 검토하는 알고리즘이다.

〈표 2-28〉 SCA2에서 상위수준 빈도표의 집계방식

$\tilde{T}$	OA	LA3	LA2	LA1
p=4	$\tilde{T}^{v4}_{OA}$	$\tilde{T}^{v4}_{LA3}$	$\tilde{T}^{v4}_{LA2}$	$\tilde{T}^{v4}_{LA1}$
p=3	$\downarrow \tilde{T}^{v3}_{OA}$	$\downarrow \tilde{T}^{v3}_{LA3}$	$\downarrow \tilde{T}^{v3}_{LA2}$	$\downarrow \tilde{T}^{v3}_{LA1}$
p=2	$\downarrow \tilde{T}^{v2}_{OA}$	$\downarrow \tilde{T}^{v2}_{LA3}$	$\downarrow \tilde{T}^{v2}_{LA2}$	$\downarrow \tilde{T}^{v2}_{LA1}$
p=1	$\downarrow \tilde{T}^{v1}_{OA}$	$\downarrow \tilde{T}^{v1}_{LA3}$	$\downarrow \tilde{T}^{v1}_{LA2}$	$\downarrow \tilde{T}^{v1}_{LA1}$

3. 작은 셀 조정에서 노출위험과 정보손실

가. 정보손실 발생

작은 셀을 정의하는 기준이 $B=3$일 때, SCA가 원래 의도하는 정보손실은 다음 <표 2-29>와 같다. 그러나 OA 수준에서 얻은 빈도표를 SCA1 방식으로 합계하면 <표 2-30>과 같이 의도하지 않은 3 이상의 정보손실을 얻게 되고, 상위 수준에서 집계될수록 정보손실이 더 많이 발생한다. |IL|은 원래 빈도값과 SCA1 처리 이후 빈도값 차이의 절대값을 의미한다. <표 2-30>은 각 빈도표 유형별로 |IL|값에 따른 셀들의 비율을 나타내고 있다. SCA가 의도하지 않은 B=3 이상의 정보손실이 발생함을 알 수 있다.

〈표 2-29〉 SCA가 의도하는 정보손실

$T/\tilde{T}$	$\tilde{f}$=0	$\tilde{f}$=3
$f=1$	-1	2
$f=2$	-2	1

〈표 2-30〉 빈도표 유형별 정보손실 비율

	OA				LA3			
|IL|	0	1	2	3+	0	1	2	3+
p=4	93.6%	4.3%	2.1%	0.0%	73.0%	16.7%	8.1%	2.2%
p=3	81.8%	10.4%	5.9%	1.8%	54.4%	18.9%	12.8%	13.9%
p=2	55.7%	19.1%	12.8%	12.4%	28.5%	17.5%	14.2%	39.7%
p=1	19.0%	19.2%	15.9%	45.9%	5.7%	10.2%	9.0%	75.1%

	LA2				LA1			
|IL|	0	1	2	3+	0	1	2	3+
p=4	50.1%	11.9%	10.3%	27.8%	41.1%	8.2%	7.7%	43.0%
p=3	36.2%	9.4%	8.3%	46.2%	30.8%	5.0%	4.8%	59.4%
p=2	17.4%	5.7%	6.1%	70.8%	15.8%	1.2%	2.1%	81.0%
p=1	1.8%	1.2%	3.6%	93.3%	3.0%	0.0%	0.0%	97.0%

물론 아주 큰 값을 가지는 셀에서 3 정도의 정보손실은 신경 쓸 필요가 없을 수도 있지만, 중간 규모의 셀들에서 발생하는 정보손실은 심각한 자료 유용성 저하를 일으킬 수 있다. 중간 규모의 셀들을 $(B, 3B] = [4, 9]$라고 하고, 정보손실의 범위를 정리하면

다음 <표 2-31>과 같다. 표를 살펴보면 예를 들어 참값은 4인데 바뀐 값은 12가 되고 따라서 정보손실이 참값의 두 배인 8이 되는 경우가 발생할 수도 있음을 알 수 있다.

〈표 2-31〉 빈도표 유형별 정보손실 범주(중간 규모 셀 대상)

range(IL)	OA	LA3	LA2	LA1
p=4	[0,0]	[-6,6]	[-8,12]	[-9,9]
p=3	[-9,12]	[-9,13]	[-9,10]	[-5,7]
p=2	[-9,12]	[-8,10]	[-5,8]	-
p=1	[-9,13]	-	-	-

중간 규모의 셀들에 대해서 정보손실 절대값이 셀 크기의 50%를 초과하는 경우에 대하여 그 비율을 정리하면 <표 2-32>와 같다. 예를 들어 $\tilde{T}^{\nu 3}_{LA3}$에 속하며 값이 4에서 9사이인 셀들 중 10%이상은 셀 크기의 50%를 초과하는 정보손실을 가진다. 이러한 빈도표의 부정확성은 이용자가 받아들이기 어려울 수 있다.

〈표 2-32〉 정보손실 절대값이 셀 크기의 50%를 초과하는 비율(중간 규모 셀 대상)

	OA	LA3	LA2	LA1
p=4	0.0%	3.3%	27.6%	23.5%
p=3	3.6%	10.7%	26.6%	14.6%
p=2	11.3%	19.0%	24.1%	-
p=1	20.7%	-	-	-

나. 노출위험 발생

SCA1에 의해 발생하는 정보손실을 줄이기 위해 SCA2를 이용할 수 있다. 각 지역 수준마다 기저 빈도표를 랜덤 라운딩하고, 변수의 개수가 줄어든 상위 빈도표는 기저 빈도표를 집계하여 작성하므로 SCA1보다 정보손실이 줄어들 것은 자명하다. 그러나 정보손실이 아무리 줄어든다고 해도, SCA2 방식으로 빈도표를 생성하면 가리고자 했던 작은 셀들에 대해 노출이 발생할 수 있다. 다음 <표 2-33>과 같은 상황을 생각해보자. 모든 변수를 가지는 빈도표에서 하위 지역 6개 셀들이 모여서 상위 지역 1개의 셀을 구성한다고 하자. 또 하위 지역 각 셀의 참값은 모두 1이면서 우연히 모두 $B=3$으로 변환되었다고 가정하자. 이때 각 지역 수준마다 작은 셀에 대해 랜덤 라운딩을 실시하므로, 작은 셀이 아닌 상위 지역의 참값 6은 라운딩되지 않고 참값을 그대로 제공하게 된다.

또한 자료 이용자는 시스템이 제공한 값들이 작은 셀 조정되었음을 알고 있다. 즉, 주어진 값 3에 대해 참값이 1, 2, 3 중 하나인 것을 알고 있다. 또 SCA2를 적용한 시스템은 이들의 합계가 6이라는 것을 알려준다. 따라서 이용자는 시스템이 제공하지 않으려 했던 참값이 모두 1인 것을 손쉽게 알게 된다.

〈표 2-33〉 SCA2를 적용할 때 작은 셀 노출 예제

	$LA3_1$	$LA3_2$	...	$LA3_6$	$LA2$
변환된 셀 값 $\tilde{f}$	3	3	...	3	6
참값 f의 후보들	1,2,3	1,2,3	...	1,2,3	6
$\tilde{f}$의 추정값	1	1	...	1	6

이렇듯 간단한 자료 분석과 사례를 통해서 작은 셀 조정으로는 정보손실을 줄이면서 노출위험을 제어하기가 어려움을 확인했다. 따라서 하나의 마이크로데이터로부터 위계구조가 있는 빈도표들을 대량으로 제공하면서, SCA가 원래 의도한 대로 노출위험을 제어하고, 정보손실을 줄이고, 이용자에게 투명한 새로운 자료처리 방법을 찾아야 한다.

참고로 SCA1 방식으로는 정보손실이 크고, SCA2 방식으로는 노출이 발생할 수 있다는 것은 해외의 통계청 자료제공 관련 전문가들에게는 이미 알려져 있는 사실이기도 하다. 때문에 해외의 통계기관들은 사전에 마이크로데이터에서 자료교환을 하거나 CTA 알고리즘을 사용하는 등의 방안들을 주로 연구해왔다. 그러한 방법들은 대체로 정보손실이 비교적 많이 발생하고, 이용자들에게 처리 과정이 투명하지 않은 단점을 가진다. 정보보호나 정보손실에 대한 인식은 나라마다 다르므로 우리나라에 적합한 방법을 새로 만들 필요가 있다.

제4절 개선된 작은 셀 조정(BSCA) 알고리즘 제안

1. 정보손실과 노출위험 관계에 대한 새로운 해석

통계별 비밀보호 분야의 문헌들을 살펴보면, 자료 제공을 위해 비밀보호 처리를 할 때 주요 관심은 얼마나 노출위험을 잘 억제하는가에 있었다고 볼 수 있다. 통계적 비밀보호 처리에 관한 권위 있는 학술 대회인 PSD(Privacy in Statistical Databases)에서 발표된 논문 중 하나에는 다음과 같은 내용이 있다(Hernandez-Carcia와 Salazar-Gonzalez, 2014).

"A widely accepted paradigm is that protection has priority over utility. This means that a minimum level of protection is a-priori decided and set in the optimization problem through constraints. Then an output maximizing the utility is searched among all solutions with an acceptable level of protection."

즉, 비밀보호 분야 방법론 연구는 노출위험을 먼저 제어하고, 정해진 노출위험 수준을 만족하면서 정보손실을 최소화하는 방안을 찾는 방향으로 진행되어 왔다고 할 수 있다.

그러나 앞 절에서 여러 매스킹 기법들을 살펴봤을 때, 의도한 대로 노출위험을 제어하지 못하는 것을 알 수 있었다. 그러므로 이제 새로운 대안을 찾고자 하는 상황에서, 아예 정반대로 정보손실 축소를 우선순위에 두고 생각해보고자 한다. 특히 우리나라와 같이 정보손실에 민감한 경우, 정보손실을 줄이는 방안을 찾는 것이 노출위험을 줄이는 것보다 어쩌면 더 중요한 문제일 수도 있다. 때문에 순서를 바꾸어 생각해보는 노력도 의미가 있다.

물론 결론적으로는 전통적인 접근과 동일하게, 노출위험의 허용 가능한 수준이 결정되면 발생하는 정보손실 정도를 설명해야 할 것이다. 다만 알고리즘 개발 단계에서는 반대로 생각해보는 것이 새로운 방법을 찾는 데 도움이 될 수 있기 때문에, 이 절에서는 노출위험과 정보손실 우선순위에 대해 일종의 패러다임 전환을 논의해 보고자 한다. 이러한 맥락에서 본 연구에서는 다음과 같이 정보손실을 줄이는 것을 먼저 고려하고, 다음으로 노출위험을 제어하는 순서로 새로운 알고리즘을 고안했다.

1. 수용 가능한 정보손실 정도를 결정한다.
2. 적절한 노출위험 측도를 고려하고, 노출의 원인을 파악하여 제어 방안을 찾는다.

먼저 정보손실에 대해 생각해보자. 전체 레코드의 개수가 N인 마이크로데이터에서 빈도표를 만들어 제공할 때, 모든 매스킹 처리에 대해서 각 셀에서 발생할 수 있는 정보손실은 $\{0,1,2,...,N\}$ 중에 있게 된다고 할 수 있다. 만약 SCA를 적용한다면 크기가 큰 셀들은 그대로 두고, 작은 셀의 $1,2,...,(B-1)$이라는 값들을 모두 0 또는 B로 바꾸므로 정보손실은 $0,1,...,(B-1)$ 중에 있게 된다. 따라서 SCA는 정보손실의 최대값이 $(B-1)$이도록 디자인된 알고리즘이라고 할 수 있다. 즉, SCA는 $max\,IL=(B-1)$이도록 의도되었지만, 빈도표들 사이에서 위계 구조가 있어 정보손실이 늘어난다고 이해할 수 있다.

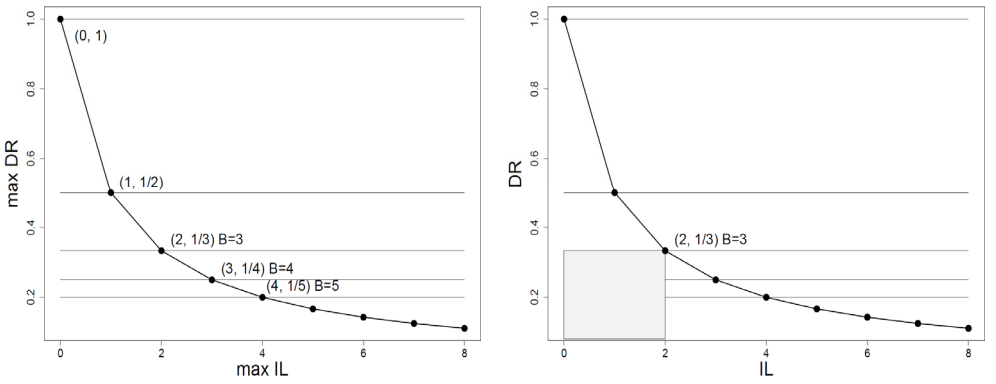
본 연구에서는 우리나라 상황에서는 기존에 제안된 기법들 중 SCA를 가장 효율적이라고 판단하고, SCA의 장점을 활용해 새로운 대안을 모색하고 있다. 따라서 새로 제안하는 알고리즘은 수용 가능한 정보손실 수준을 B 미만이 되도록 하는 것을 목표로 했다. 수식으로 표현하면 $max\,IL=(B-1)$이 된다.

다음으로 SCA가 의도하는 노출위험에 대해 생각해보자. 주어진 기준 B에 대해서, 어느 셀의 크기가 B 이상이면 해당 셀은 자연스럽게 익명성 B를 확보한다. SCA는 크기가 B보다 작은 셀을 0 또는 B로 변화시키는데, 0으로 변한 셀은 $\{0,1,2,...,(B-1)\}$ 중에 참값이 있으므로 가능한 값이 B개 존재하여 익명성 B를 확보한다고 본다. 또한 B로 변한 작은 셀 역시 참값이 $\{1,2,...,B\}$ 중에 있고 참값으로 가능한 값이 B개 존재하므로 익명성 B를 확보한다. 따라서 새로 제안하는 알고리즘에서는 노출위험을 k-익명성(k-anonymity)을 이용해 측정하도록 하고 최대 노출위험이 $1/B$이도록 제어하는 것, 즉 $max\,DR = 1/B$을 목표로 했다.

이제 수용가능한 정보손실과 노출위험 제어 목표를 모두 B를 통해 표현할 수 있으므로 다음과 같은 간단한 수식을 통해 정보손실과 노출위험 사이의 관계를 정리할 수 있다. 이는 단순한 이차원 분할표에서와 같이 SCA를 원래 제안한 의도대로 사용할 수 있다면 도달 가능한 관계이다.

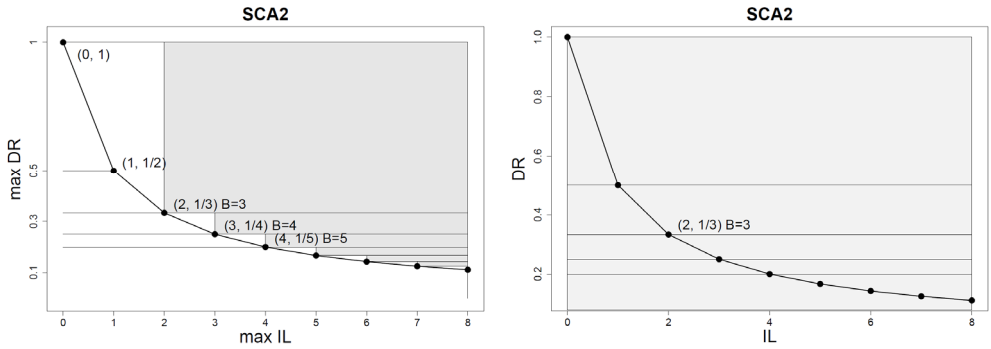
$$y = \frac{1}{x+1}, \qquad x = max\,IL, \qquad y = max\,DR$$

이를 그림으로 표현하면 [그림 2-2]와 같다. 왼쪽은 이상적인 노출위험-정보손실 관계를 나타낸다. 정보손실과 노출위험은 주어진 기준 B에 따라 그림에 나타나는 곡선을 따라가며 최대값을 가지게 된다. 예를 들어 $B=3$이면 빈도표 각 셀에서 최대 정보손실은 2, 최대 노출위험은 $1/3$인 것을 의미한다. 오른쪽 그림에서는은 주어진 $B=3$에 대해서 정보손실과 노출위험이 존재하는 구간을 회색으로 표현했다. 회색으로 표현된 구역은 이상적인 경우 모든 셀에서 정보손실은 2 이하, 노출위험은 $1/3$ 이하에서 제어됨을 나타낸다.



[그림 2-2] 이상적인 노출위험-정보손실 관계

반면, 앞 절에서 논의한 대로 SCA2는 노출위험이 최대치인 1까지 증가할 수 있고, 정보손실은 상당히 큰 값까지 커질 수 있다. 이를 나타낸 것인 다음 [그림 2-3]이다. 왼쪽 그림은 SCA2를 적용한 결과 노출위험-정보손실 최대치 사이의 관계를 나타낸다. 예를 들어 $B=3$이면 원래 의도와 다르게, 작은 빈도들을 합계하여 얻어지는 상위 수준의 어느 셀에서 최대 정보손실은 2 이상의 값을 가질 수도 있고, 최대 노출위험은 1/3 이상일 수도 있다는 것을 회색 영역으로 표현했다. 오른쪽은 주어진 $B=3$에 대해서 정보손실과 노출위험이 존재하는 구간을 회색으로 표현했는데, 의도대로 곡선에 나타난 점 이하에서 제어되지 않고 정보손실이나 노출위험이 모든 값을 가질 수 있음을 나타낸다.

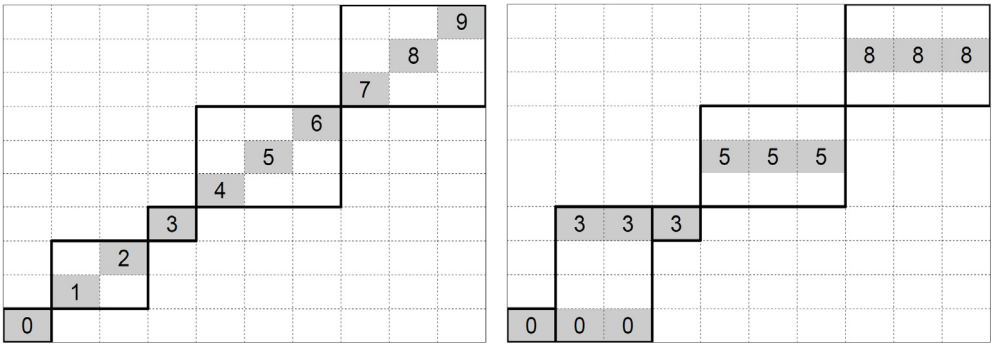


[그림 2-3] SCA2를 적용하는 경우 노출위험-정보손실 관계

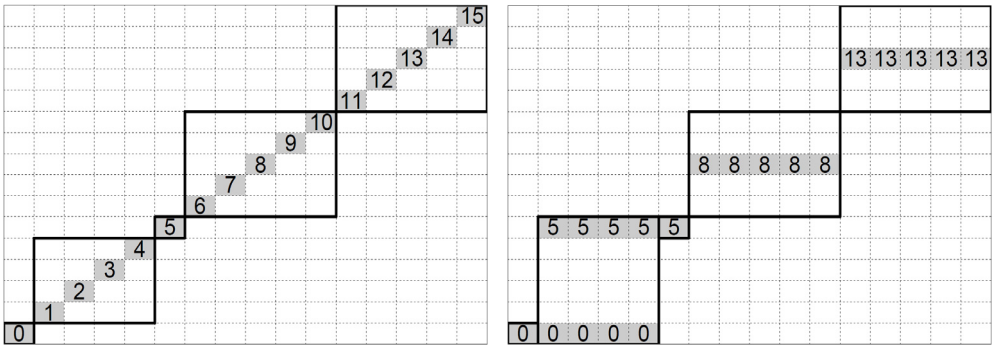
2. 정보손실 축소 방안

지금까지 이상적인 노출위험-정보손실 관계를 정리했다. 이제 새로운 알고리즘을 만들기 위해 정보손실을 축소하는 방법, 노출위험을 축소하는 방법을 차례로 생각해보자.

먼저 정보손실을 축소하기 위해 고정 그리드 규칙(Fixed Grid Rule, 이하 FGR)을 만들어 적용하기로 했다. 이는 특정한 값 대신에 그 값이 포함된 범위의 대표값을 제공하는 것으로, 본질적으로 라운딩이나 국소통합(microaggregation)과 같은 개념이라고 할 수 있다. 이를 그림으로 표현하면 다음 [그림 2-4] 및 [그림 2-5]와 같다. 물론 직선에서 구간을 사용해서 고정 구간 규칙(Fixced Interval Rule)이라고 부를 수도 있으나, 그림으로 좀 더 잘 이해되도록 평면에서 그리드로 나타냈다. FGR을 적용한다는 의미는 다음과 같다. B보다 작은 값에 대해서는 SCA를 적용하지만, B보다 큰 값에 대해서는 고정된 B단위로 구간을 나누어, 구간의 중앙값을 사용한다. [그림 2-4]는 $B=3$일 때를, [그림 2-5]는 $B=5$일 때를 시각적으로 표현했다.



[그림 2-4] $B=3$일 때, 고정 그리드 규칙(Fixed Grid Rule)



[그림 2-5] $B=5$일 때 고정 그리드 규칙(Fixed Grid Rule)

FGR이 국소통합과 다른 점은 데이터에 따라 대표값이 바뀌지 않는다는 점이다. 국소통합은 데이터마다 그룹으로 묶이는 자료들이 바뀌면서 각 그룹의 대표값도 바뀌지만, FGR은 B에 의해 사전에 구간이 정해져 있어 주어진 자료 분포의 영향을 받지 않는다. 한편, FGR이 기존의 라운딩 기법과 다른 점은 대표값으로 B의 배수 대신에 구간의 중앙값을 사용하므로 $B+1+[B/2]$형식의 값을 가지게 된다는 것이다. 최대 정보손실은 라운딩과 마찬가지로 $[B/2]$가 된다. 본 연구에서는 편의상 B가 홀수인 경우만을 고려한다.

이제 SCA에 FGR을 적용하기 위해 자료를 다음과 같이 표현하자. 위계 구조상 여러 개의 빈도를 합해서 상위 수준의 빈도를 결정할 때, 기저 수준을 LA_S, 상위 수준을 LA_L라 하겠다. 이 기호는 지역 수준이 넓어지는 경우나, 변수의 개수가 줄어드는 경우를 모두 일반화한 것이다. 이제 i번째 상위 빈도 f_i가 K_i개의 하위 빈도들을 가진다고 하면 빈도값들은 다음 <표 2-34>와 같이 정리된다.

〈표 2-34〉 상위 수준 빈도값을 구성하는 기저 빈도표의 빈도값

			LA_S			LA_L
id	1	...	j	...	K_i	i
frequency	f_{i1}	...	f_{ij}	...	f_{iK_i}	f_i

편의상 모든 f_{ij}는 B 이하의 값을 가진다고 가정한다. SCA는 $\{f_{ij}\}$들 중에서 값이 큰 셀들을 바꾸지 않는다는 것을 기억할 필요가 있다. 큰 값들은 최종 단계에서 더해주면 되므로 알고리즘에서 고려할 필요가 없다. 따라서 일반성을 잃지 않고 LA_S에 있는 모든 f_{ij}들은 B 이하의 값을 가진다고 가정하도록 한다.

의도하지 않은 B 이상의 정보손실이 일어나는 경우는 SCA 처리로 변형된 작은 셀들을 집계해서 상위 수준 빈도를 얻을 때이다. 따라서 정보손실을 줄이기 위해서 작은 빈도들의 합으로 이루어진 상위 셀을 계산할 때 정보손실이 크면 f_i에 FGR을 적용하도록 한다. 상위 수준 셀에서 빈도값은 $f_i = \sum f_{ij}$(이하 $\sum = \sum_{j=1}^{j=K_i}$)이다. SCA로 조정된 $\{\tilde{f}_{ij}\}$에서 얻어지는 상위 수준 셀에서의 빈도값은 $\sum \tilde{f}_{ij}$이다. 두 값의 차이인 정보손실이 B 미만 일 때는 FGR을 적용할 필요가 없이 $\sum \tilde{f}_{ij}$을 $\tilde{f}_i$로 사용하면 된다. 그러나 정보손실, $\sum \tilde{f}_{ij}$와 $\sum f_{ij}$ 사이의 차이가 B 이상이면 FGR을 적용하여 $\tilde{f}_i$로 사용하고자 한다. 이를 위해 f_i를 다음과 같이 표현한다.

$$f_i - 1 = a_i B + b_i, \quad a_i \in Z, \quad b_i \in \{0,1,2,...,(B-1)\}$$

이는 단순한 모드 함수이다. 다만 SCA를 적용할 때 구간이 $B+1$부터 시작하므로 위와 같이 f_i 대신에 $f_i - 1$에 모드 함수를 사용해서 표현하는 것이 적절하다. 이제 f_i를 대신해서, 제공할 $\tilde{f}_i$로 해당 구간의 대표값 $a_i B + [B/2] + 1$을 고려할 수 있다. 즉, 참 빈도값 f_{ij} 또는 f_i 대신에 SCA를 적용해 바꾼 $\tilde{f}_{ij}$ 또는 FGR을 적용한 $\tilde{f}_i$를 제공할 값으로 검토할 수 있다. f_{ij}는 SCA를 적용했기 때문에 정보손실이 $(B-1)$ 이하이며, f_i는 FGR을 적용했기 때문에 정보손실이 $[B/2]$ 이하가 된다. 따라서 전체적으로는 SCA에서 의도한 대로 정보손실이 $(B-1)$ 이하로 제한된다. 다만 이렇게 FGR만을 적용하면 SCA2와 마찬가지로 노출이 발생할 수 있으며, 이에 관한 자세한 증명은 Park(2018)을 참고하도록 한다. 이제 노출을 방지하기 위한 알고리즘의 다음 단계를 만들도록 한다.

3. 노출위험 제어 방안

SCA가 단일 빈도표에서는 익명성 확보에 효율적이었으나, 대량의 빈도표를 제공하게 되면 빈도표 위계 구조 때문에 노출이 발생했다. 이러한 노출을 제어하기 위해서는 시스템을 통해 이용자가 얻게 되는 정보를 검토할 필요가 있다.

이용자는 시스템을 통해 $\{\tilde{f}_{ij},\ j=1,2,...,K_i\}$ 및 $\tilde{f}_i$를 제공받는다. $\tilde{f}_{ij}$가 0의 값을 가지면 이용자는 $\tilde{f}_{ij}$의 값이 0에서 $(B-1)$ 사이의 값이라 추측하고, $\tilde{f}_{ij}$가 B라는 값을 가지면 이용자는 $\tilde{f}_{ij}$의 값이 1에서 B 사이의 값이라고 추측할 수 있다. 이제 $\{\tilde{f}_{ij}\}$ 중에서 B의 개수를 k_i라고 하자. 이용자는 $\{\tilde{f}_{ij}\}$를 통해 $\sum f_{ij}$의 범위를 다음과 같이 계산할 수 있다.

$$f_i = \sum f_{ij} \in [k_i, (K_i - k_i)(B-1) + k_i B] = [k_i,\ K_i(B-1) + k_i] = [k_i, u_i]$$

이상의 내용을 정리하면 다음 <표 2-35>와 같다.

〈표 2-35〉 SCA와 빈도표 위계 구조로부터 알아낼 수 있는 빈도 참값에 대한 정보

$\tilde{f}_{ij}$	f_{ij}	#(cells)	$\sum \tilde{f}_{ij}$	$f_i = \sum f_{ij}$
0	0, ... , (B-1)	$K_i - k_i$	$k_i B$	$[k_i, u_i]$
B	B, ..., B	k_i		
관측	추측	관측	관측	추측

주: $u_i = (K_i - k_i)(B-1) + k_i$

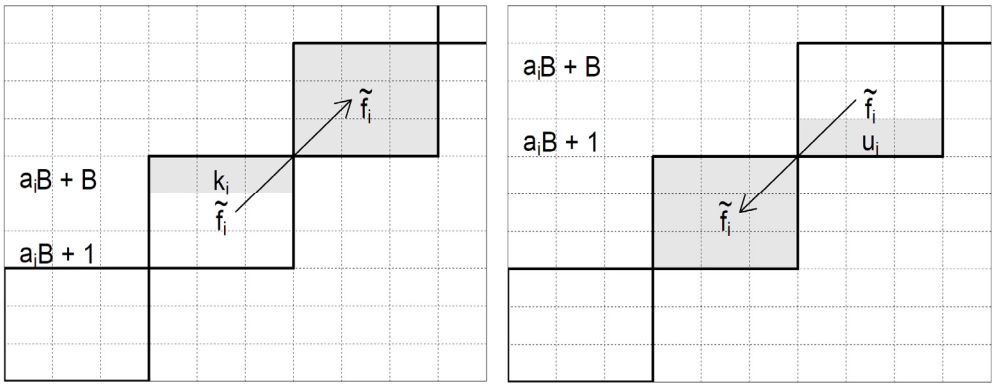
지금까지의 내용을 종합하면, 정보손실과 노출위험을 동시에 축소하며 $\tilde{f}_i$를 결정하기 위한 알고리즘은 다음 <표 2-36>과 같은 세 가지 제약조건을 가진다.

〈표 2-36〉 상위 수준 빈도값을 결정하기 위한 고려 사항

원인	$\tilde{f}_i$을 결정하기 위한 제약 조건
SCA	$\tilde{f}_i = 0$ or $\tilde{f}_i \geq B$
정보손실 제어	$\lvert \tilde{f}_i - \sum f_{ij} \rvert < B$
노출위험 제어	$k_i \leq \tilde{f}_i \leq u_i$

앞 절에서 $\tilde{f}_i$를 결정하기 위해 FGR로 정보손실을 축소했다. 이렇게 결정된 $\tilde{f}_i$가 제공하는 범위 $C=[a_iB+1,\ a_iB+B]$ 및 시스템이 기저 빈도표에 제공하는 빈도값들 $\{\tilde{f}_{ij}\}$를 분석하여 얻은 범위 $D=[k_i,\ u_i]$를 비교하여 이용자가 얻는 추가 정보가 있다면 노출이 발생할 수 있다. 즉, C가 D에 포함되어 있으면 노출이 발생하지 않지만, 그렇지 않은 경우 이용자가 시스템이 제공하는 정보 C보다 좁은 범위의 정보를 얻게 되어서 노출이 발생한다.

이를 방지하기 위해 새로 제안하는 알고리즘에서는 다음과 같이 구간을 하나씩 옮기는 단계를 제시한다. 즉, $a_iB+1<k_i$이면 $\tilde{f}_i$에 B를 더해서 다음 구간으로 옮겨 $C\subset D$을 유지한다. 또한 $a_iB+B>u_i$이면 $\tilde{f}_i$에서 B를 빼서 이전 구간으로 옮겨서 $C\subset D$를 만족하도록 한다. 그러면 이용자는 더 이상 시스템을 통해 추가 정보를 얻을 수 없게 된다. 다음 [그림 2-6]은 이러한 과정을 보여준다.



[그림 2-6] 노출제어를 위한 빈도값 이동

이 때 정보손실은 $[B/2]$에서 $(B-1)$만큼 추가로 늘어날 수 있다. 이는 노출위험을 제어하기 위해 추가로 지불해야 하는 정보손실이 된다.

이러한 과정을 $B=3$에 대해 예를 들어 살펴보면 다음 <표 2-37>과 같다. 총 6개의 빈도가 상위 수준 빈도를 구성한다고 하자. 또 이 셀들의 참값이 모두 1이고, SCA에 의해 이 참값이 모두 우연히 3으로 변했다고 하자. 앞 절에서 기존의 SCA1을 적용하면 $\tilde{f}_i$로 $\sum\tilde{f}_{ij}$인 18을 제공해야 하며, SCA2를 적용하면 $\tilde{f}_i$로 f_i의 참값 6을 제공하게 된다. SCA1에 의한 18은 정보손실이 너무 커서 제공하기 곤란하고, SCA2에 의한 6은 참값 1에 대해 익명성이 1이 되어 즉, 노출위험이 1이 되어 제공하기 어렵다.

〈표 2-37〉 BSCA를 통한 상위 수준 빈도표에서의 빈도값 결정 예제

	OA_1	OA_2	...	OA_6	LA3
f	1	1	...	1	6
SCA	3	3	...	3	18 or 6
1단계 (FGR)	3	3	...	3	5
추측값	1,2,3	1,2,3	...	1,2,3	[4,6]∩[6,18]=6
2단계 (Moving Grid)	3	3	...	3	8
추측값	1,2,3	1,2,3	...	1,2,3	[5,11]

이를 극복하기 위해, 제안한 알고리즘에 따라 값을 결정하는 과정은 다음과 같다. $\sum\tilde{f}_{ij}$와 f_i의 차이가 B보다 크므로 FGR을 적용하여 $\tilde{f}_i$로 5를 제공한다. (이는 정보손실 축소 과정에 해당한다.) 따라서 이용자는 제공받은 값을 통해 f_i가 $[4,6]$에 존재함을 알 수 있다. 또한 이용자는 제공받은 SCA 처리된 6개의 3을 이용해 f_i는 $[k_i,u_i]=[6,18]$에 존재함도 알 수 있다. 그러면 이용자는 f_i는 두 정보의 교집합인 6이 참값이라는 결론을 얻는다.

하지만 제안한 알고리즘의 마지막 단계는 이렇게 이용자가 얻는 두 정보의 집합 사이의 관계를 점검한다. 즉, $\tilde{f}_i$가 제공하는 구간 $C=[4,6]$와 $\{\tilde{f}_{ij}\}$가 제공하는 구간 $D=[k_i,u_i]=[6,18]$을 비교할 때, $C\subset D$가 아니므로 구간을 하나 옮기는 조치를 취한다. 이제 알고리즘 마지막 단계를 통해 최종 결정되는 $\tilde{f}_i$의 값은 8이 된다. 이용자는 참값이 $C=[7,9]\pm 2$에 속하는 것을 알지만 D와 비교해도 참값을 특정할 수 없게 된다. $C\subset D$가 유지되기 때문이다. 한편 정보손실은 1이 아닌 2가 되어 정보손실이 약간 증가하게 된다.

4. BSCA 알고리즘

이제 기본적인 SCA, 정보손실을 줄이기 위한 FGR, 노출이 발생하는 경우에 구간을 옮기는 단계를 가지는 알고리즘을 정리하면 다음과 같다. 작은 셀의 기준은 B라 하고 $b=1,\ldots,B-1$이라 표현하자. 또한 주어진 변수의 개수는 P라 하고 $p=1,\ldots,P$라 하자. 다음의 1~3은 SCA1과 동일한 과정이다.

BSCA 알고리즘

1. 각 OA마다 P개 변수로 구성된 기저 빈도표를 생성한다.

2. T_{OA}^{vP} 각 빈도표마다 작은 셀들을 다음과 같은 랜덤 라운딩 기법을 이용해 조정하고, 이렇게 얻은 빈도표를 $\tilde{T}_{OA}^{vP}$라 하자.

> $(B-b)/B$의 확률로 셀 값 b를 0으로 바꾼다.
> b/B의 확률로 셀 값 b를 B로 바꾼다.

3. $\tilde{T}_{OA}^{vP}$를 집계하여 T_{LA}^{vp}를 얻는다. (즉, $\tilde{f}_i = \sum \tilde{f}_{ij}$)

4. $|\tilde{f}_i - f_i| \geq B$인 모든 셀에 대해 $a_i > 0$인 경우에는

   ① 다음과 같이 fixed grid rule을 적용해 $\tilde{f}_i$값을 업데이트한다.

   A. $f_i - 1 = a_i B + b_i, \quad a_i \in Z, \quad b_i \in \{0,1,2,...,(B-1)\}$

   B. $\tilde{f}_i - 1 = a_i B + [B/2]$ 로 표현한다. 즉 $\tilde{f}_i = a_i B + [B/2] + 1$로 업데이트한다.

   ② 노출을 피하기 위해 다음과 같은 경우 B를 더하거나 빼어 $\tilde{f}_i$값을 업데이트한다.

   A. $\tilde{f}_i - [B/2] < k_i$이면 $\tilde{f}_i = \tilde{f}_i + B$

   B. $u_i < \tilde{f}_i + [B/2]$이면 $\tilde{f}_i = \tilde{f}_i - B$

5. $|\tilde{f}_i - f_i| \geq B$인 모든 셀에 대해 $a_i = 0$인 경우에는

   ① $\tilde{f}_i = B$ 로 $\tilde{f}_i$값을 업데이트한다.

   ② 노출을 피하기 위해 다음과 같은 경우 $[B/2]$를 더하거나 빼어 $\tilde{f}_i$값을 업데이트한다.

   A. $1 < k_i$이면 $\tilde{f}_i = \tilde{f}_i + [B/2]$

   B. $u_i < B$이면 $\tilde{f}_i = \tilde{f}_i - [B/2]$ (일어나지 않음)

어느 경우이든 BSCA를 통해 정보손실은 $B + [B/2]$ 이내에서 결정된다. SCA가 원래 의도한 정보손실이 B 이내인 것에 비해서 오직 $[B/2]$만큼만 증가하는 것이다. 한편, 노출위험은 SCA가 원래 의도한 대로 모든 경우에서 익명성 B를 보장한다.

SGIS시스템을 통해 빈도표를 제공할 때 T_{LA}^{vP}를 데이터베이스로 구축하고 T_{LA}^{vp}는 T_{LA}^{vP}를 집계해서 서비스한다는 것을 앞에서 설명했다. 비밀보호 처리를 위해서 할 일은 T_{LA}^{vP}를 데이터베이스로 구축할 때, 본 연구과제에서 구현한 R패키지 BSCA를 사용하고, 위 알고리즘 중 단계 4 및 단계 5를 SGIS 시스템에서 구현하는 것이다. 따라서, 해외 사례들에 비하여 매우 간단하게 정보손실과 노출위험을 의도대로 제어할 수 있다. BSCA에서 사용한 노출위험 측도는 k-익명성이며, 정보손실은 BSCA 적용 전후 빈도값의 차이이다.

5. 적용 결과

인총 인구 자료 총 4개 변수에 대해 대전 행정지역별로 아무런 처리를 하지 않은 빈도표 T및 BSCA를 적용해 변경한 빈도표 $\tilde{T}$를 얻고, 정보손실을 계산한 결과는 다음 <표 2-38>에 정리되어 있다. $B=1$을 적용했으므로 SCA가 의도하는 정보손실은 최대 2인데, BSCA를 적용해서 $[B/2]=3$만큼 정보손실이 추가로 증가하여 3이 될 수 있다. 하지만 <표 2-38>은 이런 빈도값은 전체의 1% 미만임을 보여준다.

〈표 2-38〉 '인구' 자료 행정지역별 빈도표 정보손실 비율

	OA				LA3			
|IL|	0	1	2	**3**	0	1	2	**3**
p=4	93.6%	4.3%	2.1%	**0.0**%	73.2%	18.1%	8.4%	**0.4**%
p=3	82.0%	11.5%	6.3%	**0.2**%	57.5%	28.8%	13.3%	**0.4**%
p=2	58.2%	27.3%	14.0%	**0.5**%	39.7%	45.4%	14.7%	**0.3**%
p=1	31.5%	51.8%	16.5%	**0.2**%	30.0%	61.0%	9.0%	**0.0**%

	LA2				LA1			
|IL|	0	1	2	**3**	0	1	2	**3**
p=4	57.6%	31.7%	10.3%	**0.4**%	55.4%	36.8%	7.6%	**0.1**%
p=3	50.7%	40.7%	8.4%	**0.2**%	49.6%	44.4%	5.8%	**0.2**%
p=2	39.4%	55.1%	5.4%	**0.1**%	44.4%	53.3%	2.4%	**0.0**%
p=1	34.6%	63.6%	1.8%	**0.0**%	39.4%	60.6%	0.0%	**0.0**%

인총 가구자료 가구유형 변수에 대해 그리드별로 아무런 처리를 하지 않은 가구수 빈도표 T 및 BSCA를 적용해 변경한 빈도표 $\tilde{T}$를 얻고, 정보손실을 계산한 결과는 다음 <표 2-39>에 정리되어 있다. $B=3$을 적용했으므로 SCA가 의도하는 정보손실은 최대 2인데, BSCA를 적용해서 $[B/2]=1$만큼 정보손실이 추가로 증가하여 3이 될 수 있다 하지만 <표 2-39>는 이런 빈도값이 거의 없음을 보여준다. 하나의 LA3는 100개의 OA로 구성되므로 BSCA의 단계②에 해당하는 그리드 이동이 일어나는 경우들이 발생하지만, LA2는 10,000개의 OA로 구성되어 랜덤하게 발생한 정보손실들이 거의 대부분 서로 상쇄되기 때문에 BSCA의 단계②에 해당하는 그리드 이동이 거의 일어나지 않는다.

〈표 2-39〉 '가구' 자료 그리드별 빈도표에서 정보손실 비율

	OA				LA3			
|IL|	0	1	2	**3**	0	1	2	**3**
p=1	80.1	13.3	6.6	-	53.0	33.1	13.8	0.003

	LA2				LA1			
|IL|	0	1	2	**3**	0	1	2	**3**
p=1	29.3	63.0	7.7	-	14.3	71.4	14.3	-

참고로 BSCA를 적용할 때 빈도값들이 어떻게 변화하는지 살펴보면 <표 2-40>과 같다. 작은 셀 기준은 $B=3$이다. 'X99763'과 같은 기호는 그리드 아이디를 나타낸다. 기저 빈도표는 100m 그리드별 빈도표이다. 원래 빈도표의 1 또는 2가 BSCA 적용 빈도표에서 0 또는 3으로 변화되었다. 상위 수준 빈도표, 1km, 10km, 100km 그리드별 빈도표는 기저 빈도표를 집계해서 얻는다. BSCA 적용 전후 빈도값들이 조금씩 달라져 있음을 볼 수 있다.

기존의 국소 감추기를 사용하면 빈도표의 위계 구조를 이용해 NA 처리한 작은 셀의 값들을 알아낼 수 있다. 그러나 BSCA를 사용하면 값을 특정할 수 없으며, 여전히 후보가 되는 값들 중 하나라는 것만을 알 수 있다. 예를 들면 국소 감추기에서 빈도값 1을 특정해낼 수 있는 경우가 많지만, BSCA를 적용하면, 0, 1, 2 중 하나 혹은 1, 2, 3 중 하나라는 사실만을 알아낼 수 있다.

또한 영국 ONS가 사용했던 SCA를 사용하면 상위 수준 빈도표에서 정보손실이 크다. 그러나 BSCA를 적용하면 <표 2-40>과 같이 정보손실이 제어된다.

〈표 2-40〉 BSCA 적용 사례 : '가구' 자료 7개 범주에 대한 빈도값 변화

OA : 100m 그리드

	X99763	X99762	X99759	X99756	X99745	X99742
원래 빈도표	0	0	0	0	0	0
	3	0	4	1	1	0
	4	1	5	0	0	1
	2	0	2	0	0	0
	0	0	0	0	0	0
	8	0	0	1	0	0
	0	0	0	0	0	0

	X99763	X99762	X99759	X99756	X99745	X99742
BSCA 적용 빈도표	0	0	0	0	0	0
	3	0	4	3	0	0
	4	0	5	0	0	3
	3	0	3	0	0	0
	0	0	0	0	0	0
	8	0	0	0	0	0
	0	0	0	0	0	0

LA3 : 1km 그리드

	X5678	X5688	X5698	X5699	X5672	X5682
원래 빈도표	0	1	0	0	2	0
	20	9	0	23	8	51
	20	33	5	28	8	88
	6	7	0	9	2	19
	0	0	0	0	0	1
	20	13	3	42	10	82
	1	3	0	3	0	1

	X5678	X5688	X5698	X5699	X5672	X5682
BSCA 적용 빈도표	0	0	0	0	3	0
	22	9	0	23	8	52
	21	31	6	27	8	87
	5	6	0	8	0	18
	0	0	0	0	0	0
	19	14	4	42	12	81
	0	3	0	4	0	0

LA2 : 10km 그리드

	X87	X107	X76	X86	X97	X96
원래 빈도표	42	27	21	39	24	21
	8361	3802	1446	11842	3509	1328
	27577	9919	2274	38525	10069	2792
	3073	1220	379	4007	1384	418
	33	17	4	46	18	9
	14633	5395	2282	20274	7473	2354
	1096	404	150	1247	620	139

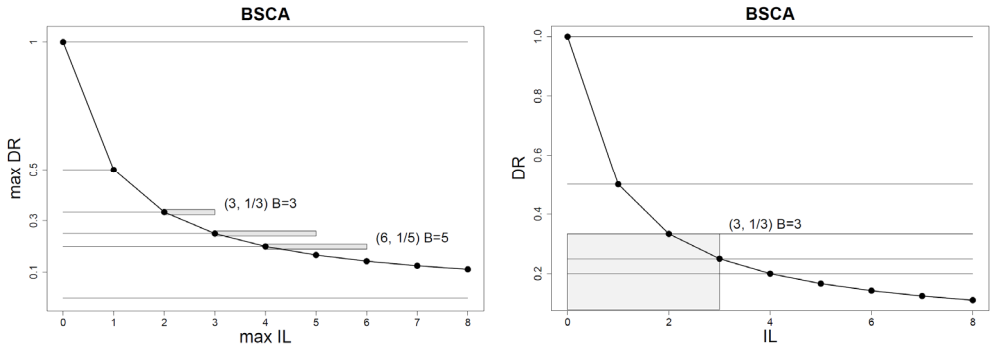
	X87	X107	X76	X86	X97	X96
BSCA 적용 빈도표	41	26	20	38	23	20
	8361	3801	1445	11841	3508	1329
	27577	9918	2273	38524	10068	2791
	3074	1219	380	4007	1384	419
	32	17	6	47	17	8
	14632	5394	2281	20275	7474	2355
	1096	404	148	1246	621	140

LA1 : 100km 그리드

	X1
원래 빈도표	1504
	319781
	891598
	96703
	1171
	573418
	30806

	X1
BSCA 적용 빈도표	1505
	319782
	891599
	96702
	1173
	573418
	30805

BSCA 알고리즘을 위험-유용성 지도에서 표현하면 [그림 2-7]과 같이 나타낼 수 있다. $B=3$일 때, 노출위험은 SCA가 의도한 대로 모두 1/3 이하에서 제어되고, 정보손실 역시 대부분의 경우 3 미만에서 제어된다. 노출위험을 제어하기 위해 빈도값들의 정보손실은 3까지 증가할 수 있다. '인구' 자료의 경우 1% 미만의 셀에서 정보손실이 증가했고, '가구' 자료의 경우 0.01% 미만의 셀에서 정보손실이 증가했다.



[그림 2-7] BSCA를 적용하는 경우 노출위험-정보손실 관계

6. BSCA R패키지 개발

SGIS에 서비스용 데이터베이스를 구축할 때, 다음 <표 2-41>과 같이 R패키지를 사용할 수 있다. 마이크로데이터에서 빈도표를 만들려는 대상 변수들을 선택해서 'key'라는 변수로 저장한다. 또 지역 수준을 나타내는 변수들을 모아서 'area'라는 변수로 저장한다. 이들을 인수로 'make.full.table'명령어를 실행한다. 그러면 출력된 result[[1]]은 참값 빈도를 가지는 빈도표 T를, result[[2]]는 BSCA 알고리즘이 적용된 빈도표 $\tilde{T}$를 반환한다. 기존에 SGIS 시스템의 서비스용 데이터베이스로 result[[1]]를 사용했다면, 이제는 result[[2]]를 서비스용 데이터베이스로 사용해서 노출위험과 정보손실이 제어된 빈도표를 제공할 수 있다. 또한 기저 빈도표 생성 후 전체 빈도표를 생성하지 않고, 원하는 지역에 대해서만 빈도표를 생성할 수도 있다.

〈표 2-41〉 BSCA를 적용해서 전체 빈도표를 제공하는 명령어 예제

```
library(BSCA)

B = 3
key = data.frame(microdata[, c('hh_type')])
```

```
area = data.frame(microdata[, c('LA1', 'LA2', 'LA3', 'OA')])

result = make.full.table(area, key, B)
```

전체 빈도표는 필요하지 않고, 일부 지역의 빈도표만을 얻고 싶은 경우 다음 <표 2-42>와 같이 코딩하면 된다. R패키지 BSCA는 향후 세부 형식들을 수정하는 등의 추가 검토 과정을 거쳐 향후 일반에 공개할 예정이다. 본 보고서에서는 현재의 입출력 형식을 정리했다.

〈표 2-42〉 일부 빈도표에 BSCA를 적용하는 R패키지 명령어 예제

```
library(BSCA)

OA_result = sca.table(microdata, area.column, key.column, B=5)
final_table = bsca.table('둔산 1동', B=5, OA_result)
```

제5절 맺음말

2016년 SGIS에서 제공하는 대량의 빈도표에 대한 통계적 노출제어 연구가 수행된 바 있다. 기존 연구는 행정지역에 대해 수행되었고, BSCA의 아이디어가 제시되었었다. 하지만 새로운 알고리즘을 적용할 때 계산 시간 등의 프로그래밍 관련 문제를 해결하기 어려웠고, 이미 서비스되고 있는 빈도표에 대해 변화된 값을 제공하는 것 역시 쉬운 일이 아니었다. 원래 제공되고 있는 값과 약간이라도 다른 값을 제공받는 것에 대한 이용자의 거부감이 클 수 있기 때문이다.

본 연구는 SGIS에서 그리드별로 빈도표를 제공할 때 통계적 노출제어 처리를 하는 방안을 다루고 있다. 현재 SGIS는 전국 단위를 기준으로 산출된 구간을 그리드별로 제공하고 있으며, 이는 지나치게 정보손실이 큰 값이다. BSCA를 적용한다면 참값과 다른 값이기는 하지만, 현재 제공되는 값보다 훨씬 정확한 정보를 제공할 수 있다. 따라서 원래 제공되고 있는 값에 변화가 있더라도 이용자에게 거부감은 없을 것이라 예상된다.

BSCA 알고리즘의 중요한 특징은 다음과 같다. 첫째, 작은 셀을 정의하는 B를 기준으로 SCA가 원래 의도대로 노출위험을 $1/B$ 이하로 제어할 수 있다. 또한 대부분의 셀에서 정보손실을 B 미만으로 제어한다. 따라서 수식으로 노출위험과 정보손실이 어느 정도인지 명확하게 표현할 수 있다. 이는 최근 통계적 비밀보호 분야가 지향하는 'formal privacy'(Vilhuber와 Schmutte, 2017)를 구현한 것이기도 하다.

둘째, BSCA를 이용하는 시스템은 비상호적 조건(non-interactive setting)에서 자료를 서비스하고 있다는 것을 가정한다. 이는 시스템이 요청받을 쿼리를 모두 알고 있고, 모든 답을 미리 가지고 있을 수 있다는 것을 의미한다. 실제로 SGIS는 이용자가 요청할 대부분의 쿼리를 데이터베이스로 구축해서 가지고 있고, 이는 자료제공 속도 향상을 위해 필수적이다. BSCA 알고리즘 및 R패키지는 이런 환경에서 기존 시스템을 거의 변경할 필요 없이 활용할 수 있다는 장점이 있다.

우리나라 SGIS와 유사한 서비스인 미국 OnTheMap에서는 여러 정보를 지도 위에 서비스하기 위해 재현자료와 차등정보보호가 연구되고 있다(박민정 외, 2018). 이러한 방법들은 노출위험을 더욱 낮추는 대신 정보손실을 많이 발생시키는 경향이 있는 것으로 파악된다. 노출위험을 확실하게 억제하는 것이 우선순위이기 때문이다. 향후 재현자료 생성이나 차등정보보호 구현과 같은 새로운 기법들이 더욱 활발하게 연구된다면, 노출위험은 강하게 억제하면서도 정보손실을 BSCA처럼 줄여주는 효율적인 방안들이 개발될 수 있을지도 모른다. BSCA는 대부분 셀의 정보손실을 B 미만으로 제한하며 모든 셀의 노출위험을 $1/B$ 이하로 억제하지만, 차등정보보호 수준으로 안전한 방안이라 볼 수는 없기 때문이다.[7] 다만 궁극적으로 효율적인 대안이 도출될 때까지는, 위계구조를 가지는 대량의 빈도표를 제공할 때 BSCA 알고리즘을 이용할 수 있다. 안전하고, 유용하고, 편리한 대용량 자료제공을 위해 앞으로 통계적 비밀보호 분야의 발전이 지속되기를 기대한다.

7) 물론 빈도표에서 차등정보보호 적용이 가능하지 않을 수도 있다(Duchi 외, 2018).

참고문헌

박민정, 성주현(2016), 「통계지리정보(인총) 집계구 단위 빈도표의 통계적 노출제어」, 통계개발원(내부보고서).

정동욱, 김경미(2008), “매크로데이터의 비밀보호방법에 대한 고찰”, 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, 182-201.

Hernandez-Carcia, M.-S., Salazar-Gonzalez, J.-J.(2014), Further development with perturbation techniques to protect tabular data, In: Domingo-Ferrer, J. (ed.) PSD 2014. LNCS, vol. 8744, pp. 24-35. Springer, Cham.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Nordhort, E. S., Seri, G., and de Wolf P. P.(2010), ESSNet Handbook on Statistical Disclosure Control, ver 1.2.

Park, M-J.(2018), Bounded small cell adjustments for flexible frequency table generators, In: Domingo-Ferrer, J. and Montes, F. (eds.) PSD 2018. LNCS, vol. 11126, pp. 13-27. Springer, Cham.

Shlomo, N., Antal, L., and Elliot, M.(2015), Measuring disclosure risk and data utility for flexible table generators, Journal of Official Statistics, 31, 2, pp. 305-324.

Vilhuber, L. and Schmutte, I. M.(2017), Proceedings from the 2016 NSF Sloan Workshop on Practical Privacy, Digital Commons @ ILR.

<부 록>

1. 통계지리정보서비스(SGIS) 소개

통계지리정보서비스(SGIS, Statistical Geospatial Information Service)는 소지역 통계를 이용하여 정책개발 및 의사결정에 활용하는 것을 목표로 서비스의 이용편의성 향상과 다양화를 추진하고 있다. 그간 추진된 서비스 내역은 다음과 같다.

〈부표 2-1〉 통계지리정보서비스 연혁

연 도	추 진 내 용
2001~2002년	- 2000 인구센서스 결과에 대한 거처 포인트 입력 시작 - 거처 포인트 담을 구역인 기초단위구 및 집계구 설정
2003년	- 서비스를 위한 소지역통계시스템 구축 및 시연(서비스는 미실시)
2005년	- 통계지리정보과 출범(7월)
2006년	- 인구부문에 대한 소지역통계시스템 구축 및 내부망 서비스(7월) - “통계내비게이터” 시범 서비스 시작(대전광역시, 12월)
2007년	- 특별시/광역시를 대상으로 통계내비게이터 본격 서비스 실시
2009년	- 통계내비게이터 전국 서비스 실시 - ’00, ’05년 인구총조사 결과 및 ’04~’07년 사업체조사 결과 수록
2010년	- ’05년 농림어업총조사 및 ’00~’03년 전국사업체조사 결과 수록 - 전국사업체조사 ’08년 결과 서비스 및 ’09년 결과 수록
2011년	- ’00년 농어업총조사 결과 및 ’10년 사업체조사 결과 수록 - 통계지리정보 시계열서비스 및 위치기반 모바일 서비스 실시
2012년	- 행정안전부『국가기초구역』설정 및 검증 지원('10년 기준 공간통계자료(집계구별 인구, 가구, 주택, 사업체) 제공)
2013년	- 공간정보 표준과의 정합성을 강화한 Open API v2.0 서비스 실시 - 고령화 서비스 개편 및 서비스 개시 - SGIS 오픈플랫폼 구축 정보화전략계획(ISP) 수립
2014년	- SGIS 오픈플랫폼 1단계 서비스 구축 · 포털 서비스시스템 및 시범과제(생활업종통계지도) 개발 · 위치기반 플랫폼 관리 및 운영시스템 개발, Open API 34종 개발 · 인총(‘00～‘10년), 2013년 전국사업체조사 등 센서스 공간DB 구축

연 도	추 진 내 용
2015년	- SGIS 오픈플랫폼 1단계 서비스 개시 - SGIS 오픈플랫폼 2단계 서비스 구축 · 국민생활과 밀접한 오픈플랫폼 활용콘텐츠 개발(3종) · 데이터관리 및 분석시스템 개발, 위치기반 데이터관리시스템 고도화 · 포털 및 공간통계검색시스템 고도화, 모바일 포털 개발 등
2016년	- SGIS 오픈플랫폼 2단계 서비스 개시 - SGIS 오픈플랫폼 3단계 서비스 구축 · 플랫폼 활용 정책지원 서비스 2종 및 사용자 참여 서비스 기획·개발 * 정책통계지도, 기술업종 통계지도, 통계갤러리 · 통계조사지원을 위한 'GIS기반 조사지원시스템' 개발 등
2017년	- SGIS 오픈플랫폼 3단계 서비스 개시 - SGIS 오픈플랫폼 4단계 서비스 구축 - SGIS 스마트플랫폼 구축을 위한 정보화전략계획(ISP) 수립

통계지리정보서비스의 핵심은 소지역 통계 생산과 활용에 기반을 두고 있다. 여기서 소지역의 의미는 행정구역의 최소단위인 읍면동보다 더 작은 공간적 영역을 의미한다. 통계청은 2001년부터 소지역 통계공표구역인 집계구를 도입하기 위한 연구에 착수하여 2006년 대전광역시를 시범지역으로 인구 500명을 최적 기준으로 설정한 집계구를 기초로 통계지리정보서비스를 실시하였다. 이후 특별시, 광역시를 대상으로 서비스 지역을 확대하여 2009년부터는 전국 서비스를 실시하고 있다.

집계구는 통계청에서 공표하는 전수조사 결과를 담고 있는데, 대표적으로 인구주택총조사, 농림어업총조사, 전국사업체조사 결과를 집계구 단위로 집계하여 통계지리정보서비스를 통해 조회가 가능하도록 지원하고 있다. 또한 대화형통계지도서비스에서는 집계구별로 인구, 가구, 주택, 사업체수, 종사자 수 등에 관한 상세한 정보를 조회하고 조회 결과를 지도에서 직관적으로 확인할 수 있도록 다양한 시각화 서비스를 시행하고 있다.

집계구는 2017년 기준으로 102,024개가 설정되어 있으며, 행정동 평균 1/29 크기이다. 집계구 획정에 사용되는 기초단위구는 도로, 철도, 하천, 능선 등 지형지물을 이용하여 구획단 최소공간단위로 정의할 수 있는데, 기초단위구에 인구수를 집계한 후, 최적 인구기준(500명)과 사회경제적 동질성, 획정된 형태의 안정성을 고려하여 집계구로 획정된다. 평균적으로 4~5개의 기초단위구가 하나의 집계구로 병합되는 형태로 획정된다. 2010년까지는 인구주택총조사의 조사주기에 따라 5년 주기로 획정되었으나, 등록센서스가 도입된 2015년부터는 매년 집계구 획정 업무가 발생하고 있다.

한편 집계구는 지역의 인구이동이나 지형지물의 변화 등의 이유로 획정 시마다 형태가 달라지는 특징을 가진다. 따라서 특정 시점의 소지역 현황을 파악하기에는 유리하지만, 시계열적 비교의 어려움이 발생한다는 단점이 있어 그리드 서비스를 도입하게 되었다. 그리드 서비스는 2012년 통계지리정보서비스에 시범적으로 적용되었다가 2017년부터 행정안전부 국가지점번호체계를 도입하여 100m, 1km, 10km 단위의 격자를 이용한 그리드 서비스가 시행되고 있다.

2. 빈도표 매스킹 영국 사례

빈도표의 노출제어 문제는 네덜란드, 호주, 미국, 영국 등에서 연구되고 있지만 본 연구에서 국내 환경을 고려하여 영국 센서스에 적용되고 학술 저널에 발표된 방법론을 주로 검토했다. 이에 부록에서 영국의 관련 연구 배경을 정리하여 소개한다(Hundepool 등, 2010).

통계적 노출제어의 목적은 통계 작성 목적과는 상관없는 개인이나 사업체 등의 사적인 정보는 보호하면서도, 동시에 이용자들에게는 최대한 많은 가치를 제공하는데 있다. 통계표를 제공할 때 개별 정보를 보호하기 위해서는 주로 통계표 내에서 작은 셀의 개수와 속성 노출 여부를 살핀다. 통계표의 어떤 셀이 한 두 개의 개체로 계산된 작은 셀일 때, 각 개체를 식별할 수 있기 때문이다. 나아가 작은 셀들과 더불어 통계표에서 0의 값을 가지는 셀들의 개수나 배치를 이용해 통계표가 제공되지 않을 때는 알 수 없었던 각 개체에 대한 정보를 추론할 수 있다면, 개체별 속성노출, 즉 사적인 정보 노출이 일어날 수도 있어서 적절한 조치가 필요하다. 그러나 개별 정보 노출을 제어하기 위한 조치에는 정보손실이 수반될 수밖에 없으므로, 이용자들에게 제공되는 자료의 가치 역시 함께 감소되게 마련이다. 자료의 유용성을 최대한 지키면서 의도하지 않은 개별 정보의 노출을 제어하기 위해, ONS에서는 레코드 맞교환(swapping), 임계치 적용, 작은 셀 조정 등과 같은 방법들을 검토하였다. 각 기법의 내용을 간략히 설명하면 다음과 같다.

먼저, 레코드 맞교환은 통계표를 만들기 이전의 마이크로데이터에서 어떤 레코드를 다른 지역의 유사한 레코드와 바꾸는 것을 말한다. 교환 비율은 모든 지역에서 동일하며 역추적 위험을 막기 위해 교환 비율 값은 제공되지 않는다. 다음으로, 한 지역의 거주자 및 가구수 최소값을 설정하여 노출위험을 줄이는 것을 임계치 적용이라고 부른다. 임계치는 통계표마다 다른데, 표준적인 통계표들에 대해서는 1,000명과 400가구 기준을 적용한다. 센서스 지역 단위 통계(CAS; Census Area Statistics)에 대해서는 100명과 40가구가 기준이다. 행정 구역(parish/community) 프로파일은 50명과 20가구가 기준이며, 이보다 작은 단위에서는 거주자, 남, 여, 가구수의 전체합만을 공표한다.

통계표 작성 이전의 레코드 맞교환 처리만으로는 원하는 수준의 노출제어 효과를 거두기 어렵고, 노출제어를 위해 너무 많은 레코드 맞교환 처리 시 정보손실이 커지게 된다. 이에 ONS에서는 통계표 작성 이후의 조치인 작은 셀 조정(small cell adjustment)을 함께 적용하였다. 작은 셀

조정의 구체적인 방법은 다음과 같다. 주어진 통계표에서 셀 값이 작은 경우는 반올림 기법을 활용해 수정한다. 작은 셀들이 수정된 이후에, 수정된 값을 그대로 이용하여 총합 및 중간합(부분합)을 계산한다. 따라서 각 통계표는 내부적으로 가법성을 유지할 수 있다. 한편, 각 통계표는 독립적으로 조정되므로, 같은 인구와 서로 다른 변수를 가지는 두 개 통계표의 인구수 합이 반드시 같을 필요는 없게 된다. 같은 이유로 하위 지역 통계표들의 합이 상위 지역 통계표의 결과와 같지 않을 수 있다.

지금까지 소개한 세 종류의 기법들을 2001 센서스를 대상으로 실험한 결과, ONS는 노출제어 각 방법 단독으로는 적합한 보호를 제공하지 못한다고 판단하였다. 따라서 2011년 센서스를 위해서는 몇몇 방법들을 결합하여 활용하기로 하고, 각 방법들에 대해 일련의 평가 기준을 가지고 장·단점을 비교하였다. 영국 ONS가 고려한 노출제어 기법 및 평가 기준은 아래 정리되어 있다.

〈부표 2-2〉 노출제어 처리 기법

노출제어 처리 기법	
Record swapping	Conventional rounding
Over-imputation	Random rounding
Data switching	Controlled rounding
Post randomization(PRAM)	Semi-controlled rounding
Sampling	Small cell adjustment
Suppression	Barnardisation
ABS cell perturbation	

〈부표 2-3〉 노출제어 처리 기법 평가 기준

노출제어 처리 기법 평가 기준
• 전반적으로 이용자가 수용할만한가?
• 가법적이고 일관된 통계표를 제공할 수 있는가?
• 지역별, 범주별로 통계표를 차분할 때 발생하는 차이에 의해 노출이 일어나는가?
• 실무적이며 구현 가능한 방법인지, 현재까지 유사한 적용사례가 있는지?
• 해당 기법 적용으로 마이크로데이터 제공이 제한되지는 않는지?
• 해당 기법이 이해하기 간단하고, 설명하기 쉬운지?
• 기법 적용 전·후, 추정량의 편의, 분산, 합계 및 기타 통계 분석 결과에 미치는 영향

제공되는 통계의 정확도와 자료의 보이지 않는 변형에 민감한 국내 정서상, ONS의 기준이나 원칙을 바로 수용하기는 어렵다. 하지만 ONS에서 이러한 기준을 가지고 각 기법들을 평가한 결과와 향후 적용을 위해 고려한 세부 내용들을 살펴보는 것은 의미가 있으므로 이를 살펴보았다.

부정적으로 평가된 기법들은 다음과 같다. 라운딩(rounding)은 정보손실이 크며, 데이터 스위칭(data switching), PRAM, 샘플링 및 Barnardisation은 너무 복잡하거나 이해하기 어렵다고 판단되었다.

최종 후보로 검토된 기법들에서 논의된 내용들을 살펴보자. 레코드 맞교환은 편의를 최소화하기 위해 제어 변수로 지역, 성별, 연령 그룹 등을 고려하였다. 특히 보통 센서스에 포함된 변수들이 지역 변수에 대해 독립적이라고 가정할 수 있으므로, 가구들 사이에 지역 변수 교환도 긍정적으로 평가되었다.

한편, 맞교환 비율과 대체(imputation) 비율을 균형 있게 운영하기 위해, 대체되지 않은 레코드를 위주로 맞교환하거나 대체율이 높은 지역에서는 맞교환 비율을 줄이는 등의 조치를 고려하였다. 또한, 과대체(over-imputation) 기법은 캐나다 통계청에서 고안한 패키지인 CANCEIS를 이용하여 지역 변수 이외의 변수에 대해 활용되었다.

다음으로, 호주 통계청에서 고안한 셀 변조 기법(ABS cell perturbation)에 대한 ONS의 설명과 평가는 다음과 같다. 이 방법은 두 단계 처리 과정을 가지는데, 첫 번째 단계에서는 노출이 일어나지 않도록 값들이 변조되고, 두 번째 단계에서는 통계표의 가법성을 회복하도록 값들이 변조된다. 이러한 기본적인 프로시저를 바탕으로 자료 유용성을 향상시키도록 IACP(Invariant ABS Cell Perturbation)이라는 기법이 제안되기도 하였다(Spicer와 Tudor, 2009). 그러나 IACP 기법은 2011년 센서스 노출제어에서는 제외되었는데, 이는 다른 방법들과 비교하여 통계표들 사이의 일관성을 유지하지 못하고 구현 등 실무 적용에서 어려움이 예상되기 때문이었다.

마지막으로 작은 셀 조정(Small Cell Adjustment)은 작은 셀들에 대해 편의가 발생하지 않도록 랜덤 라운딩 처리를 하고, 주변합을 구해 제공되는 통계표들을 완성하는 것을 말한다. 주변합이나 전체합은 변형된 셀을 이용해서도, 변형되지 않은 셀을 이용해서도 구해질 수 있다. 이 기법은 잉글랜드, 웨일즈, 북아일랜드의 2001 센서스 통계표를 보호하기 위해 레코드 맞교환과 함께 사용된 사례가 있기도 하다. 스코틀랜드에서는 레코드 맞교환의 효과가 부족하다고 판단된 직장별 인원수에 대한 빈도표에서만 활용되었다.

이처럼 영국에서는 통계표 작성 이전 및 이후에 적용할 수 있는 몇몇 기법들을 검토하고 함께 활용하는 방안들을 심층 검토하였다. 이러한 사례를 참고하고, 국내 실정을 고려하여 본 연구를 진행했다. 한편, 영국 사례에서 사용한 노출위험 측도를 다음 절에 정리했으니 관심 있는 독자는 참고하기 바란다.

3. 엔트로피 기반 빈도표의 노출위험 측정 방법

보다 종합적인 노출위험 측도를 만들기 위해서는 엔트로피 개념을 도입한다. 이는 빈도표에서 어떤 행이나 열이 대부분 0으로 채워져 있고 소수의 셀에만 빈도가 집중되어 있는 정도를 측정하기 위해서이다. 실제로 영국 ONS에서는 2011년 센서스를 준비하면서 많은 빈도표를 작성해보고 한 두 셀에만 빈도가 집중된 경우들의 비율을 계산하였다고 한다. 그러나 이렇게 사례별로 노출위험을 검토하는 것은 복잡하기만 할 뿐이므로, 일반적으로 다양한 빈도표에 대한 계량적인 노출위험 측도를 만들기 위해 정보 이론의 엔트로피 개념을 도입한다(Shlomo 등, 2015).

본 연구에서는 노출위험 측도로 k-익명성을 선택했다. 그러나 엔트로피 기반 노출위험 측도도 연구된 바 있어 <부록>에 소개한다.

엔트로피는 이산 확률 변수 X 가 확률 분포 $P=(p_1,p_2,...,p_K)$를 따를 때 다음과 같이 정의된다.

$$H(X)=H(P)=-\sum_{i=1}^{K}p_i\cdot\log p_i$$

엔트로피 $H(P)$가 작다는[1] 것은 어느 한 셀에 빈도가 집중적으로 분포되어 있다는 의미이므로 속성 노출이 일어나기 쉬워진다. 반대로 관심 변수의 분포가 균일 분포를 따라서 노출위험이 최소가 되는 경우, 즉 $P=(1/K,1/K,...,1/K)$인 경우, 엔트로피는 최대값을 가져 $H(P)=\log K$가 성립한다. 이를 정리하면 <부표 2-4>와 같다. 엔트로피 정의를 사용하면 빈도표의 분포로 보아 노출위험이 높을 때 엔트로피는 0의 값을, 노출위험이 낮을 때 엔트로피는 최대값을 가지게 된다. 따라서 엔트로피를 이용하여 노출위험 정도를 표현하려는 노력은 타당성을 얻게 된다.

〈부표 2-4〉 빈도표의 분포에 따른 엔트로피와 노출위험

X 의 확률분포	대부분 셀이 0이고 빈도가 한 셀에 집중되어 있는 경우	대부분 셀에서 빈도가 고르게 분포하는 경우
엔트로피 $H(P)$	0	$\log K$
노출위험	최대	최소

1) 만약 p_i가 0의 값을 가지면 $\lim_{x\to 0} x\log x=0$이므로 해당 항을 0으로 간주하고, p_i가 1일 경우도 해당 항의 값은 0이 된다. 또한 $-p_i\cdot\log p_i\geq 0$이여서 엔트로피는 음수가 될 수 없으므로, 엔트로피가 0의 값을 가지는($H(P)=0$) 경우는 확률이 어느 한 지점에 집중되어 있음을 가리키는 것이 된다.

엔트로피를 이용하여 노출위험 측도를 고안하기에 앞서 먼저 앞에서 보았던 <표 2-3>과 같은 빈도표를 일반적으로 표현하면 다음 <부표 2-5>와 같다.

〈부표 2-5〉 일반적인 빈도표

셀 번호	1	2	…	K	총합
빈도수	F_1	F_2	…	F_K	N

또한 빈도표의 노출위험 측도가 가져야 할 바람직한 성질은 다음과 같다.

(a) 값이 0인 셀들이 많을수록 노출위험이 커져야 한다.
(b) 빈도표에서 셀 값들이 균일한 분포를 보일수록 노출위험이 작아야 한다.
(c) 셀의 값이 작으면 큰 값일 때보다 노출위험이 높아야 한다.
(d) 손쉬운 비교를 위해 노출위험은 0과 1사이의 값을 가져야 한다.
(e) 노출제어 처리 이후 노출위험 값은 작아져야 한다.

이제 일반적인 빈도표에 대해 엔트로피 정의를 이용하여, 위에 언급한 바람직한 성질을 모두 가지는 노출위험 측도를 고안하는 과정을 살펴보자. 일반적인 빈도표에 대해 엔트로피 식을 정리하면 다음과 같다.

$$0 \le H(P) = H(\frac{F}{N}) = -\sum_{i=1}^{K} \frac{F_i}{N} \cdot \log\frac{F_i}{N} = \frac{N \cdot \log N - \sum F_i \cdot \log F_i}{N}$$

노출위험은 0에서 1사이의 값을 가지는 성질 (d)를 만족하기 위해 다음과 같이 변형한다.

$$1 - \frac{H(F/N)}{\log K}$$

엔트로피는 정의로부터 균일한 분포를 걸러내고 더불어 위의 변형과정을 거치면, 노출위험이 낮은 균일한 분포에 대해 0에 가까운 값을, 노출위험이 높은 한 지점에 집중된 분포에 대해 1에 가까운 값을 가지므로 성질 (b) 및 (d)를 만족한다.

다음으로 성질 (c)와 관련된 빈도표의 전체적인 규모와 성질 (a)와 관련된 값이 0인 셀들의 비율을 반영하는 노출위험 측도를 만들기 위해 위의 측도를 다음과 같이 확장한다.

$$R(F, w_1, w_2) = w_1 \cdot \left[\frac{|C_0|}{K}\right] + w_2 \cdot \left[1 - \frac{N \cdot \log N - \sum F_i \cdot \log F_i}{N \cdot \log K}\right] + (1 - w_2 - w_3) \cdot \left[-\frac{1}{\sqrt{N}} \cdot \log \frac{1}{e\sqrt{N}}\right]$$

여기서 C_0는 0의 값을 가지는 셀들의 집합을, $|C_0|$는 그러한 셀들의 개수를, w_1 및 w_2는 $0 \le w_1 + w_2 \le 1$을 만족하면서 위 측도의 각 항목에 대한 가중값을 나타낸다. 이제 첫 항은 0의 값을 가지는 셀들을 다루어 성질 (a)를, 둘째 항은 엔트로피를 가져온 것이므로 성질 (b)를, 셋째 항은 빈도표의 크기를 반영하므로 성질 (c)를 반영한다.

나아가 각 항목에 대한 가중값을 통해 특정 유형의 노출을 강조하여 노출위험 측도를 사용할 수 있는데, 작은 셀들에 대해 노출위험을 더욱 강조하고자 다음과 같이 두 번째 항을 빈도가 6보다 작거나 같은 경우와 큰 경우로 나누어 가중값을 달리 적용할 수 있다.

$$w_{2,1} \cdot \left[1 - \frac{H(F/N)}{\log K}\right]_{[F_i \le 6]} + w_{2,2} \cdot \left[1 - \frac{H(F/N)}{\log K}\right]_{[F_i \ge 7]}$$

영국 ONS에서는 시뮬레이션을 통해 경험적으로 얻은 $w_1 = 0.1$, $w_{2,1} = 0.7$, $w_{2,2} = 0.1$ 및 $w_3 = 0.1$을 사용하여 2011년 센서스의 빈도표들에 대해 노출제어 처리 전후 노출위험을 측정한 바 있다.

지금까지 성질 (a)~(d)를 만족하는 노출위험 측도가 만들어지는 과정을 소개하였다. 적절한 노출위험 측도를 찾는 것은 쉬운 일이 아니어서 다양하고 반복적인 검토가 필요하다. 예를 들어, 노출위험을 낮추기 위해 3을 기반으로 랜덤 반올림을 시행했다고 하자. 그러면 셀 값이 1인 경우 0 혹은 3의 값을 가지도록 변환되고, 이는 0인 셀의 수를 늘리면서 분포를 균일하지 않은 방향으로 바꾸게 되므로 위에 제안된 노출위험 측도의 값을 커지게 하는 결과를 가져온다. 즉, 이 측도는 노출제어 처리가 오히려 노출위험을 높이는 효과를 가진다고 표현하게 되는 것이다. 하지만 실제로는 0인 셀들은 실제로 0인지 변환되어 0인지를 모르므로 불확실성(uncertainty)이 증가되어 노출위험이 낮아졌다고 할 수 있다. 이를 보완하여 성질 (e)를 만족시키기 위해 위 노출위험 측도의 첫 항을 다음과 같이 수정한다.

$$(|C_0|/K)^{|C_0 \cup C_0'|/|C_0 \cap C_0'|}$$

이 때 C_0 는 원래 빈도표에서 0인 셀들의 집합을, C_0'는 노출제어 처리 이후 빈도표에서 0인 셀들의 집합을 나타낸다. 수정된 항에서 $|C_0|/K$는 항상 1보다 작거나 같은 값을 가지며, 지수 부분인 $|C_0 \bigcup C_0'|/|C_0 \bigcap C_0'|$은 노출제어 처리 후 항상 1보다 크거나 같은 값을 가지므로 수정된 항은 노출제어 처리 이후 항상 작아지게 된다. 예를 들어 원래 빈도표에서 10%가 값이 0인 셀들이고, 노출제어 처리 이후 20%가 값이 0인 셀들이라고 하자. 그러면 이 항의 값은 0.1에서 노출제어 처리 이후 0.01로 줄어들게 된다.

2부

재현자료 생성 기법 및 활용

3. 재현자료 작성 방법론 검토

박민정 · 김정연

제3장

재현자료 작성 방법론 검토

박민정 · 김정연

요 약

일반국민에게 마이크로데이터를 제공할 때 개별정보가 식별되어 민감한 정보가 노출되지 않도록 사전에 처리하는 것은 통계기관이 마땅히 해야 할 일이다. 이를 위해 각국의 통계기관은 전통적으로 매스킹 기법들을 사용해 왔으나, 노출위험 감소를 위해 자료를 매스킹 처리할 경우 정보손실이 커지는 문제가 발생한다. 노출위험과 정보손실이 동시에 축소되지 않는 한계를 극복하기 위한 노력이 지속되었지만 적절한 해결책은 아직 없어 보인다. 다만, 유력한 대안으로 차등정보보호와 재현자료 작성이 연구되고 있다.

본 보고서는 재현자료 작성 방법론을 소개한다. 재현자료가 국내에는 아직 낯선 개념이지만 해외에서는 자료제공에 활발히 이용되는 사례들이 발견되고 있다. 이에 재현자료 활용 가능성을 판단하기 위한 기초연구로서 작성 방법론을 검토하였다. 재현자료 생성은 다중대체 기법을 근간으로 하므로 이를 이해하기 위해 베이지안 통계를 먼저 소개하고, 다중대체, 완전 재현자료, 부분 재현자료 그리고 재현자료 생성 R패키지 소개로 보고서를 구성하였다.

주요용어 : 마이크로데이터, 비밀보호, 매스킹, 재현자료, 노출위험, 정보손실

✍ 이 연구는 2017년 하반기 통계개발원 연구보고서에 수록된 자료입니다.

제1절 서 론

국가통계 작성기관으로서 통계청은 생산한 통계(각종 공식 집계자료)와 공식 통계를 생산하는데 사용된 표본 조사 마이크로데이터를 일반 국민에게 제공하고 있다. 보통 각종 집계자료는 KOSIS를 통해, 각 응답 개체들의 정보를 쌓아 행렬 형태로 만든 마이크로데이터는 MDIS를 통해 누구나 쉽게 접근할 수 있도록 제공한다. 나아가 4차 산업혁명과 빅데이터가 주요 키워드인 요즘, 센서스, 행정자료와 같은 대용량 자료에 대한 분석 수요가 증가하고 있으므로, 통계청은 이러한 대용량 전수 자료를 통계데이터센터를 방문한 인증된 사용자에게 제공하고 있다.

통계기관의 자료제공은 사회발전과 부가가치 창출을 위하여 끊임없이 장려되어야 한다. 그러나 적극적인 자료제공이 실현되기 위해서는 개인정보보호라는 커다란 난관을 극복해야 한다. 개인정보보호 역시 사회발전에 있어 반드시 지켜져야 할 중요한 가치이므로, 통계청은 양립하기 어려운 중요한 두 가지 목표, 개인정보보호와 효율적인 정보제공을 동시에 달성하기 위해 적극적인 노력을 해야 한다. 그러나 이는 정말로 해결하기 쉽지 않은 문제이다.

각국의 통계기관은 자료를 제공하기 전에 개별정보가 노출될 위험을 점검하고, 노출이 일어나지 않도록 많은 노력을 한다. 안타깝게도 충분한 정보를, 안전하고, 편리하게 제공할 수 있는 완전한 방법은 아직 찾지 못했으며, 이는 노출위험과 정보손실 사이의 상충관계(trade-off) 문제가 해결되지 않고 있음을 의미한다. 노출위험을 줄이기 위해 자료를 변형하는 모든 노력들은 정보손실을 일으키기 때문이다.

이에 통계기관 입장에서는 제공하는 자료와 이용자의 유형을 구분하여 노출위험을 제어하는 방법들을 다각도로 연구하고 실험하고 있는 상황이다. 그러한 노력들을 정리하면 아래 <표 3-1>과 같으며, 이를 통해 이용자를 구분하여 적절한 자료제공 전략을 수립하는 것이 노출제어에 있어 효율적인 대안이 될 수 있음을 알 수 있다. 다만 각 방법들은 장단점을 가지며, 별다른 한계 없이 범용적으로 사용할 수 있는 방법은 아직 없다고 판단된다.

〈표 3-1〉 이용자 유형에 따른 노출제어(비밀보호) 전략

종류	물리적 노출제어	통계적 노출제어(SDC)[1]			
이용자	심층 이용자	불특정 다수			
배포 형식	인가 파일 가공 통계표 데이터 센터 원격 접속	공공이용파일		OAS[2]	
전략 기법	접근제어 및 결과통제 ➢ 결과물에 SDC 처리 필요	매스킹 기법[3]	재현자료[4]	mOAS[5]	차등정보 보호[6]

통계개발원은 그동안 일련의 보고서 발간을 통해 통계적 노출제어에 관해 지속적으로 연구해 왔으며, 본 보고서에서는 그 마지막 단계로 재현자료를 검토한다. 재현자료는 개인정보보호라는 제약조건을 완전히 극복하기 위해서 제안되었다고 할 수 있다. 재현자료 생성 목적이 원 자료에 있는 어떠한 개별 자료도 포함하지 않지만, 원 자료가 가지고 있는 변수들 사이의 관계를 온전히 보존하는 새로운 마이크로데이터를 생성하는 것이기 때문이다. 물론 복잡한 실제 자료에 대해서는 이러한 재현자료 작성이 손쉽게 실현 가능한 작업이라 말하기는 어렵다. 참고로 초창기에는 원 자료를 대신하여 공공이용파일로 배포하기 위한 목적으로 재현자료에 대한 연구가 진행되었으나, 요즘은 원 자료를 분석하기 전에 자료 탐색 및 모형 구축을 위한 용도로 재현자료를 사용하려는 움직임이 더 많이 발견되고 있다.

조사된 실제 자료는 전혀 포함되지 않으며, 변수값 전체를 재현하는 것을 완전 재현자료(fully synthetic data)라고 하며, 이러한 개념과 방법론은 Rubin(1993)에 의해 처음 제안되었다. 완전 재현자료는 추정량의 편의(bias)가 너무 큰 경우가 많으므로 정보손실이 매우 큰 방법이라 볼 수 있다. 이를 극복하기 위해 노출위험이 높은 일부 값들만 재현하여 만든 자료를 부분 재현자료(partially synthetic data)라고 한다. 부분 재현자료에서는 노출위험이 발생하는 대신, 정보손실이 완전 재현자료보다 훨씬 적게 발생한다.

재현자료 원리를 체계적으로 이해하여 재현자료를 생성하기 위해서는 베이지안 통계학, 자료 생성에 필요한 여러 통계 계산 알고리즘, 다중대체 기법, 완전 재현자료, 부분 재현자료, 결합규칙, 부분 재현자료 생성 기법 개선을 위한 비모수 모형들, 노출위험과 정보손실 측도, 재현자료 작성 해외사례 분석, 프로그램 구축 또는 R 패키지 활용에 대한 검토가 필요하다. 한정된 연구기간 및 인력으로 이러한 세부 주제들을 모두 정리하여 보고서에 담기는 어려우므로, 본 보고서는 현재 사용가능한 R패키지를 이용해 재현자료를 생성하는 과정을 전반적으로 이해하는 것을 목표로 한다. 이에 통계계산 알고리즘, 비모수 모형, 노출위험과 정보손실 측도 및 해외 사례는 간략한 소개로 대체한다. 따라서 본 보고서는

---

1) Statistical Disclosure Control/Limitation이라는 용어는 통계적 노출제어/노출제한/노출관리 등으로 번역되어 사용되고 있으며, 개별정보노출을 방지하기 위해 노출위험과 정보손실을 측정하는 방법, 자료를 변형하는 다양한 방법 등을 통칭한다. 통계적 자료 비밀보호(Statistical Data Confidentiality)라는 용어도 유사한 의미로 사용된다.

2) Online Analytic System의 약자로 자료를 분석하는 시스템을 통해 자료를 이용할 수 있도록 제공하는 것을 말한다.

3) 지금까지 각국 통계기관에서 주로 사용되어온 SDC 기법들을 지칭하며, (micro)aggregation, recoding, swapping, (local) suppression, adding noise 등을 예로 들 수 있다. 박민정(2015)을 참고하라.

4) Synthetic Data에 대한 박민정과 김항준(2016)의 번역을 따른다.

5) Krenzke 외(2013)가 제안하고 있는 매스킹 기법을 사용한 온라인 분석 시스템을 말한다. 박민정(2016)을 참고하라.

6) Differential Privacy에 대한 이용희(2013)의 번역 표현을 따른다.

베이지안 통계의 이해, 다중대체의 개념, 완전 재현자료 생성, 부분 재현자료 생성 및 R패키지 활용법 소개로 구성된다. 베이지안 통계 및 다중대체에 익숙한 독자는 2절과 3절을 생략해도 좋다.

제2절 베이지안 통계의 이해

일반적으로 재현자료 연구에서는 자료를 관찰된 부분과 관찰되지 않은 부분으로 나누어 생각한다. 완전 재현자료 생성 과정은 관찰된 자료를 바탕으로, 관찰되지 않은 부분에 대한 조건부 확률분포를 고려하고, 이 분포로부터 관찰되지 않은 부분의 값들을 랜덤하게 생성한 뒤, 생성된 값으로 구성된 자료를 공개하는 것으로 이루어진다. 이러한 논리적 흐름은 베이지안 통계를 기반으로 하고 있기 때문에, 우도함수, 사전분포, 사후분포, 사후예측분포 등과 같은 베이지안 통계에서의 중요한 개념 및 성질들을 살펴볼 필요가 있다.

이번 절에서는 통계추론에 있어 고전적 통계(frequentist statistics)와 베이지안 통계(Bayesian statistics)가 어떤 차이점이 있는지 살펴보고, 베이지안 추론이란 무엇인지 설명한다. 이를 위해 통계추론에 필요한 기본 개념과 추론의 바탕이 되는 우도함수에 대해 살펴본 후, 베이지안 추론에 사용되는 사전분포, 사후분포, 예측분포 및 MCMC 등의 통계계산 알고리즘을 소개한다. 마지막으로 복잡한 사후분포 추정을 대신하여 실제 재현자료 생성에 중요하게 사용되는 베이지안 붓스트랩을 설명한다. 이 절을 읽기에 앞서 통계의 기본 개념을 정리하고 싶은 독자를 위해, 베이지안 통계의 이해에 필요한 기본적인 확률 개념을 <부록> 1.에 소개하였다. 이 절의 내용에 관해 보다 깊은 이해를 원하는 독자는 이 절의 주요 참고문헌인 오만숙(2012)을 공부하기를 권한다.

1. 통계적 추론

가. 기본개념

관찰의 대상이 되는 집단 전체로부터 얻어진 특성값들의 모임을 모집단(population)이라 한다. 이 모집단의 특성을 나타내는 수치를 모수(parameter)라 한다. 표본(sample)은 모집단의 일부를 추출한 것을 가리킨다. 통계적 추론의 목적은 표본 자료에 주어진 정보를 이용해 미지의 관심모수 θ의 후보값들 중에서 실제 θ값일 가능성이 높은 것을 추정하고, 이 추정에서 발생하는 오차가 얼마인지 알아내는 것이다.

표본 자료 $x=(x_1, x_2, \cdots, x_n)$가 주어졌을 때, 이 자료들은 미지의 매커니즘에서 얻어지며 모수 θ에 대한 정보를 가진다고 할 수 있다. 통계적 추론을 위해 미지의 매커니즘에 대해 자료를 생성하는 확률모형을 가정하면, 표본 자료 x는 확률변수 $X=(X_1, X_2, \cdots, X_n)$에 대한 관측값이 된다. 확률변수 X는 미지의 관심모수 θ에 의존하는 확률밀도함수 $f(x|\theta)$를 갖는다고 가정한다.

이제 자료 x가 주어졌을 때 임의의 θ에 대하여 변수 X의 x에서의 조건부 결합확률밀도함수 $f(x|\theta)$를 계산할 수 있다. 만약 X가 이산 변수라면 $f(x|\theta_0)$는 $\theta=\theta_0$인 경우 주어진 자료 x가 발생할 확률이다. 두 개의 θ값 θ_a 및 θ_b에 대해서, $f(x|\theta_a)>f(x|\theta_b)$라면 $\theta=\theta_a$일 때 주어진 자료 x가 나타날 확률이 $\theta=\theta_b$일 때보다 크다는 것을 의미한다. 따라서 θ의 추정값으로 θ_b보다는 θ_a를 선택하는 것이 바람직하다. 이처럼 $f(x|\theta)$를 이용하여 여러 θ값에 대하여 추정값으로서의 가능성을 타진할 수 있으므로 $f(x|\theta)$를 θ의 우도함수(likelihood function, 가능도 함수)라고 부른다. 자료 x는 이미 관측되어 숫자로 주어지므로, $f(x|\theta)$는 θ만의 함수가 된다. 따라서 우도함수를 $f(x|\theta)$가 아니라 $L(\theta|x)$ 또는 $L(\theta)$로 표현할 수 있다.

예를 들어, 동전 던지기 실험을 한다고 가정하자. 확률 변수 X에 대해, 앞면이 나오면 $X=1$, 뒷면이 나오면 $X=0$이라 하자. 독립적인 동전 던지기를 n번 실행할 때, 각 회에서 앞면이 나올 확률(즉, $X=1$이 될 확률)이 θ로 일정하다고 하자. $x=(x_1, x_2, \cdots, x_n)$ ($x_i=0$ 또는 1, $i=1, \cdots, n$)가 동전 던지기를 n번 실행한 결과 얻은 관측값이라고 할 때, 우도함수는 다음과 같다.

$$f(x|\theta)=\prod_{i=1}^{n} f(x_i|\theta)=\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}=L(\theta|x)$$

이는 각 관측값에서 계산될 확률밀도함수 값들의 곱으로, 이를 미지의 모수 θ의 함수로 본 것이 우도함수이다.

나. 고전적 통계추론과 베이지안 통계추론의 차이점

고전적 통계에서는 추론의 대상인 모수 θ가 고정된 상수라고 가정한다. 반면 관측을 통해 얻어진 자료 $x=(x_1, x_2, \cdots, x_n)$는 확률변수 $X=(X_1, X_2, \cdots, X_n)$의 확률분포를 따르며, 확률변수 X에서 얻어질 수 있는 많은 관측값들 중 하나가 된다. 고전적 통계에서 통계적 추론은 이 특정 관측값 x에 의존하는 것이 아니라 발생할 수 있는 모든 가능한 관측값들을 고려하여 이루어진다. 이를 위해 결합확률밀도함수, 즉 우도함수를 최대로 만드는 θ가 추정량이 된다.

이에 반해 베이지안 통계에서 자료 $x=(x_1, x_2, \cdots, x_n)$는 이미 관측되었으므로 주어진

고정값으로 간주한다. 대신 모수 θ가 확률변수이고 확률분포를 가진다. 즉 베이지안 통계에서는 발생할 수 있는 모수의 모든 값을 고려하고자 한다. 모든 통계분석 이전에 모수 θ에 대해 가정하는 확률분포를 사전분포(prior distribution)라고 부른다. 이 사전분포와 우도함수를 바탕으로 관측된 자료 $x=(x_1, x_2, \cdots, x_n)$가 주어졌을 때, 모수 θ의 조건부 분포 $\pi(\theta|x_1, x_2, \cdots, x_n)$를 사후분포(posterior distribution)라 한다. 베이지안 관점에서 통계추론의 과정은 사후분포로부터 모수 θ의 특성을 알아내는 것이므로, 사후분포를 유도하는 것이 베이지안 추론의 핵심이 된다.

2. 베이지안 통계

이제 사전분포, 사후분포, 사후예측분포, MCMC 등 베이지안 통계에서 중요한 개념을 살펴보자.

가. 사전분포

모수 θ가 확률변수이고 확률분포를 가진다고 가정하는 베이지안 통계에서는 추론을 위해 모수 θ에 대한 분포가 필요하다. 따라서 자료와 무관하게 또는 자료를 얻기 전에 θ에 대하여 가지고 있는 정보를 내포하는 분포를 가정하며 이를 사전분포 $\pi(\theta)$라 부른다. 베이지안 관점에서 추론은 어떤 사전분포에서 시작하여 관측된 자료를 바탕으로 사후분포 $\pi(\theta|x)$를 추정해 나가는 것이기 때문에 보통은 사전분포의 영향력이 크지 않기를 바란다. 참고로 이항분포, 정규분포 등 분포가 수식으로 표현되는 많은 경우 고전적 통계추론 관점에서 구한 추정량과 베이지안의 사후추정량은 일치한다. 예를 들어, (1) 확률 변수 X가 이항분포를 따를 때 $X=1$이 될 확률 θ에 대해 무정보 사전분포(noninformative prior distribution)인 균일분포 $U(0, 1)$를 가정하거나 (2) 확률변수 X가 정규분포를 따를 때, X의 평균 θ에 대해 무정보-부적절 사전분포(noninformative and improper prior distribution)인 균일분포 $U(-\infty, \infty)$를 가정한 경우, 고전적 통계추론 관점에서 구한 추정량과 베이지안의 사후추정량은 일치한다.

베이지안 통계에서 사전분포를 고려할 때는 관측치가 주어져 있지 않은 상황이다. 따라서 어떤 사전분포를 사용할지 결정하는데 일정한 법칙이 있는 것이 아니며 다양한 선택이 가능하다. 한편 이런 다양성 때문에 때로는 베이지안 추론의 결과가 사전분포에 얼마나 민감하게 변하는지 조사하는 민감성 분석(sensitivity analysis)이 필요한 경우도 있다. 다양한 사전분포의 사례들은 <부록> 2.를 참고하기 바란다.

나. 사후분포

관측값 $x=(x_1, x_2, \cdots, x_n)$에 대한 확률밀도함수 $f(x|\theta)$가 주어지고, θ의 사전분포 $\pi(\theta)$가 정해지면 베이즈 정리에 의해 다음과 같은 조건부 분포를 구할 수 있다. 이 θ의 조건부분포는 자료 $x=(x_1, x_2, \cdots, x_n)$를 얻고 난 후의 θ에 대한 정보를 요약하는 것이므로 사전분포에 대응되는 개념으로 사후분포(posterior distribution)라 부른다. θ가 연속인 경우 사후분포는 다음과 같다.

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f(x)}$$

$$= \frac{f(x|\theta)\pi(\theta)}{\int_{\theta\in\Theta} f(x|\theta)\pi(\theta)\,d\theta} \qquad (1)$$

단, Θ는 모수 θ의 가능한 값을 모두 포함하는 모수 공간(parameter space)을 의미한다. 만약, θ가 이산인 경우는 위 식의 적분이 합으로 바뀐다.

베이지안 추론은 이 사후분포를 근거로 한다. 예를 들면, θ를 추정하기 위해 평균, 중앙값, 모드 등을 구하기 원할 경우, 베이지안 추론에서는 사후분포를 이용해 계산한 사후평균, 사후 중앙값, 사후 모드 등을 사용할 수 있으며 수식으로 표현하면 다음과 같다.

(1) 사후평균: $\hat{\theta} = \int_{\theta\in\Theta} \theta\,\pi(\theta|x)\,d\theta$

(2) 사후 중앙값: $\pi(\theta \le \hat{\theta}|x) = \pi(\theta \ge \hat{\theta}|x)$를 만족하는 $\hat{\theta}$

(3) 사후 모드: $\hat{\theta} = argmax_\theta\,\pi(\theta|x)$

다. 사후예측분포

베이지안 추론의 중요한 특징 중의 하나는 관측값 $x_1, x_2, \cdots, x_n$이 주어졌을 때, 미래의 관측값 X_{n+1}에 대한 사후예측분포(posterior predictive distribution)를 모수의 추정값에 의존하지 않고 단지 이전의 자료에만 의존하여 구할 수 있다는 것이다. 재현자료는 관측된 자료를 바탕으로 관측되지 않은 부분의 값을 생성하여 얻어지므로, 사후예측분포는 재현자료 생성에서 핵심적인 역할을 한다. 관측값 $x=(x_1, x_2, \cdots, x_n)$이 주어졌을 때 모수 θ의 사후분포가 $\pi(\theta|x)$라고 하자. 미래의 관측값 X_{n+1}의 예측분포는 다음과 같다.

$$f(x_{n+1}|x) = \int_{\theta \in \Theta} f(x_{n+1}|\theta)\,\pi(\theta|x)\,d\theta$$

사후예측분포에 대한 보충 설명은 <부록> 3.을 참고하라.

라. 사후분포의 근사

모수 θ에 대한 베이지안 추론과정은 θ의 사후분포를 구하고, 사후분포로부터 θ의 추정값들을 구하는 것이다. 따라서 사후분포의 유도는 베이지안 추론의 핵심이다. 그러나 실제로는 θ의 분포를 정확하게 구할 수 없는 경우가 빈번히 발생하는데, 그 이유는 모형이 복잡하거나 모수가 많아서 식 (1)의 분모 부분 계산이 어렵기 때문이다. 이럴 경우 사후분포를 유도하는 대신 사후평균, 사후분산, 특정 사건에 대한 사후확률 등을 근사적으로 계산하는 것이 필요하다.

예를 들어 $f(X)$를 확률밀도함수로 갖는 확률분포의 성질을 알고 싶지만, 수식으로 유도되지 않아서 이론적으로(analytically) 파악하기가 어려운 경우를 가정하자. 만약 $f(X)$를 밀도함수로 갖는 독립표본 $X_1, X_2, \cdots, X_N$을 생성할 수 있다면, 이 독립표본을 바탕으로 표본평균, 표본분산, 표본의 분포 등을 추정할 수 있다. 또한 표본의 크기 N을 크게 하면 추정의 정확성을 원하는 수준까지 올릴 수도 있다. 이처럼 원하는 분포로부터 독립표본을 생성한 뒤, 생성한 표본을 바탕으로 추론하는 것을 몬테칼로 방법이라 한다. 나아가 독립표본이 아니라, 정상분포(stationary distribution) $f(X)$로 수렴하는 마코프 연쇄 $\{X_n : n \geq 0\}$로부터 표본을 생성한 뒤에 몬테칼로 방법을 적용하는 것을 마코프 연쇄 몬테칼로(Markov Chain Monte Carlo, MCMC) 방법이라 한다.

사후분포에 대한 근사방법으로 많이 사용되는 것도 몬테칼로 표본 생성방법이다. θ의 사후분포가 $\pi(\theta|x)$일 때, $\pi(\theta|x)$로부터 표본 $\theta_1, \theta_2, \cdots, \theta_N$을 생성할 수 있다면, 이 표본의 표본평균, 표본분산, 경험적 분포(empirical distribution)들이 사후평균, 사후 분산, 사후 확률분포들에 대한 추정값 역할을 할 수 있다. 또한 표본의 크기 N을 크게 하여 추정의 정확성도 확보할 수 있다. 자세한 설명은 <부록> 4.에 서술되어 있으니 참고하기 바란다.

3. 베이지안 붓스트랩

재현자료 작성에서 자주 사용되는 베이지안 붓스트랩(Bayesian bootstrap)에 대해 살펴보자. 베이지안 붓스트랩은 모수추정과정 없이 몬테칼로 방식으로 자료를 생성할 때 사용될 수 있다.

먼저 Efron(1979)이 제안한 붓스트랩(bootstrap, 재표집)을 간단히 설명하면 다음과 같다. 주어진 표본 $x = (x_1, x_2, \cdots, x_n)$에서 동일한 확률로 복원추출을 통해 크기 n인

새로운 표본을 얻는다. 이를 $x^* = (x_1^*, x_2^*, \cdots, x_n^*)$라 하자. 관심모수 Q를 추정할 때는 이러한 재표집 표본을 R개 세트 생성하고 각 재표집 표본에서 얻은 점추정값들을 사용한다. 예를 들어 표본평균의 분산을 알고 싶은 관심모수 Q라고 하자. 또 활용할 수 있는 자료는 크기 n인 표본 한 세트라고 하자. 기존의 통계학에서 표본 한 세트로 표본평균의 분산을 알기는 어렵다. 하지만 붓스트랩으로 재표집 표본을 R개 세트 생성하고, 각 재표집 표본에서 표본평균을 구하고, 이 표본평균들의 분산을 위에 말한 관심모수 Q의 추정값으로 사용할 수 있다. 이와 같이 붓스트랩은 수식으로 유도하기 어려운 통계량들을 추정할 수 있는 길을 열어주었다고 평가받는다.

이와 비교하여 베이지안 붓스트랩(Rubin, 1981)을 살펴보자. 기존 붓스트랩과의 가장 큰 차이점은 복원추출 시 사용하는 가중값에 분포를 가정하는 것이다. Efron의 붓스트랩에서는 재표집 표본을 생성할 때 원래 자료에 $(w_1, w_2, \cdots, w_n) = \left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right)$의 가중값을 주었다고 할 수 있다. 붓스트랩에서 동일한 확률로, 즉 가중값 $1/n$을 사용하여 원 자료를 재표집하는데 비하여, 베이지안 붓스트랩에서는 가중값을 $Dirichlet(1, 1, ..., 1)$ 사후분포를 가지는 확률변수(random variable)로 생각한다. 가중값의 사후분포를 구하는 과정에 대해서는 <부록> 5.를 참고하라.

실제로 베이지안 붓스트랩을 구현한 절차는 다음과 같다.

(1) 먼저, 균일분포 $U(0,1)$에서 $(n-1)$개의 난수를 발생시키고, 오름차순으로 정렬한 것을 $a_1, a_2, \cdots, a_{n-1}$이라 하자.

(2) 여기에 $a_0 = 0,\ a_n = 1$을 추가한다.

(3) 균일분포 $U(0,1)$로부터 n개의 난수 $u_1, u_2, \cdots, u_n$를 생성한다.

(4) 다음과 같이 크기 n인 재표집 표본 X^*를 얻는다.

$$X^* = (x_k^* \mid x_k^* = x_j,\ a_{j-1} < u_k \le a_j,\ k = 1, 2, \cdots, n)$$

위의 과정 (1)-(4)를 R번 반복하여 재표집 표본을 R개 세트 얻는다. 각 재표집 표본에서 얻은 점추정값들을 사용하여 관심모수 Q에 관한 추론을 한다.

관심 모수 Q에 대한 추정값이 $\hat{Q} = \frac{1}{R}\sum_{i=1}^{R} \hat{Q}_i$라고 하자. Rubin(1981)은 다음을 증명하여 베이지안 붓스트랩이 효율적인 방법임을 보였다.

$$E_B(\hat{Q}) = E_{BB}(\hat{Q}),\ \ V_B(\hat{Q}) = V_{BB}(\hat{Q}) \times \frac{n+1}{n}$$

단, E_B 및 V_B는 각각 붓스트랩 재표집에서 기댓값과 분산을, E_{BB} 및 V_{BB}는 베이지안 붓스트랩 재표집에서의 기댓값과 분산을 나타낸다.

제3절 다중대체

재현자료는 Rubin이 제안한 다중대체(multiple imputation) 기법을 활용하여 생성한다. 대체(imputation)는 관측된 마이크로데이터에서 발생한 결측값(missing value)을 적절한 값으로 바꾸는 것을 말한다. 대체값으로 결측값이 채워진 자료를 완전 자료(complete data)라 하고, 데이터 행렬 안의 모든 값이 전부 관측된 경우를 가정하는 일반적인 통계 분석 방법을 적용하여 분석할 수 있다. 완전 자료를 구축하기 위해서는 결측값을 한 값으로 대체할 수도 있고 여러 값으로 대체할 수도 있다. 여러 개의 대체값을 제공하는 것을 다중대체라 하며, 그 목적은 일종의 점 추정값인 대체값의 불확실성을 고려함으로써 추론의 정확성을 높이기 위함이다.

이번 절에서는 결측값의 발생 원인을 살펴보고, 결측값 대체 방법으로 단순대체(single imputation)와 다중대체를 설명한다. 이를 통해 재현자료 생성에 핵심 역할을 하는 다중대체 방법을 이해하도록 한다. 재현자료를 통해 통계량을 추정할 때 분산 수식이 다중대체에서 분산을 계산하는 수식과 달라지지만, 기본적인 아이디어는 다중대체와 동일하므로 다중대체를 바르게 이해하는 것이 중요하다.

1. 결측값과 대체

결측값이란 관측되지 않은 자료값들을 가리킨다. 결측값이 생성되는 원인을 다음과 같이 세 가지로 나누어 볼 수 있다.

(1) 완전임의결측(missing completely at random, MCAR): 결측 여부가 우연에 의해 결정된다. 결측이 자료 안의 어떤 변수와도 관련이 없으므로 결측값의 분포는 다른 모든 관측값들과 무관하다.

(2) 임의결측(missing at random, MAR): 결측의 원인이 관측된 자료 내에 있지만, 결측이 일어난 변수 자체와는 관련이 없다. 즉, 결측이 일어난 변수는 결측에 영향을 주지 않으나, 다른 변수들이 결측에 영향을 끼친다.

(3) 비임의결측(missing not at random, MNAR): 결측값과 결측 여부 간에 연관이 있다. 때문에 nonignorable nonresponse라고도 한다.

예를 들어, 우울 정도에 대한 조사를 시행했다고 하자. 우울 정도를 나타내는 변수에 결측값이 존재한다고 할 때, 이 결측이 완전 우연에 의해 생겼다고 한다면 이 경우는 완전임의결측에 해당한다. 만약 우울 상태의 정도에 상관없이 남자들은 우울 정도에

대한 응답을 꺼려하는 경향이 있다면, 성별은 결측에 영향을 끼친다고 할 수 있다. 하지만 성별에 관한 정보가 주어진 경우, 우울 정도를 나타내는 변수는 결측에 영향을 주지 않으므로 임의결측이 된다. 마지막으로 우울한 상태가 심한 사람들이 우울 정도에 대한 응답을 꺼려한다면, 이는 결측이 일어난 변수가 결측에 영향을 미치는 것이므로 비임의 결측에 해당한다.

결측값 처리에 가장 대표적인 방법은 마이크로데이터에서 행을 삭제(listwise deletion)하는 것으로, 이는 결측값이 하나라도 있는 개체의 자료 전체를 분석에서 제외시키는 방법이다. 하지만 결측값이 비임의결측에 의해 발생되었다면 결측된 개체를 제외한 자료를 분석한 결과는 편향되어(biased) 있을 수 있다. 또한 결측값이 완전임의결측에 의한 것이라 가정할지라도, 표본 크기의 감소로 인한 검정력(power) 저하가 초래될 수 있다. 이러한 단점을 보완하고자 제안된 것이 이 절에서 소개하는 결측값 대체 기법들이다.

2. 단순대체

단순대체란 각 결측값에 하나의 대체값을 제공하는 것을 말한다. 결측이 일어난 변수가 이산적일 경우 각 관측값들의 구성 비율을 유지하도록 결측이 일어난 값들을 대체할 수 있다. 나아가 완전하게 관측된 다른 이산 변수를 고려하여 얻는 부분 집합마다 얻어진 구성 비율을 유지하도록 대체값을 할당할 수도 있다. 예를 들어, 주어진 자료가 $Y=\{1,1,1,2,2,2,2,3,NA,NA\}$로 두 개의 결측값을 가진다고 하자. 이때 $P(Y=1)=3/8$, $P(Y=2)=4/8$, $P(Y=3)=1/8$이므로 확률 $3/8$으로 값 1을, 확률 $4/8$로 값 2를, 확률 $1/8$로 값 3을 발생시켜 NA로 표기된 결측값들을 대체할 수 있다.

연속 변수의 경우, 결측이 일어난 변수만 고려하여 평균(unconditional mean)으로 결측값들을 대체할 수도 있고, 다른 변수들과의 회귀분석 결과(regression mean)를 이용해 결측값을 대체할 수도 있다. 또한 회귀 예측값과 임의로 생성한 오차를 합하여 결측값을 대체(stochastic regression imputation)하여 예측값의 불확실성을 보완하는 방법을 생각할 수도 있다.

종단자료의 경우는 마지막으로 관측된 값으로 결측값을 대체(last observation carried forward)하기도 하고, 표본조사의 경우 자료 내의 관측된 다른 표본 개체를 임의로 골라(hot deck) 결측값을 대체하기도 한다. 이때 결측이 발생한 자료 내의 정보를 이용하지 않고, 이전의 표본조사나 다른 자료에서 비슷한 값을 이용(cold deck)하는 경우도 있다(송주원과 안형진, 2009).

단순대체의 장점은 다음과 같이 세 가지로 요약될 수 있다. ① 일반적인 분석 방법들을 그대로 사용할 수 있다. ② 자료를 수집한 사람이 가용한 모든 자료들을 바탕으로 대체값을

추정할 수 있기 때문에 자료의 활용성이 높아진다. ③ 일단 대체값이 결정되면 이후의 자료 분석에서 일관된 자료가 활용된다.

그러나 단순대체는 추정량의 표준편차가 과소 추정되는 문제점을 가지고 있다. 이는 대체된 값을 실제 관측된 값처럼 사용하므로, 실제 표본수보다 더 많은 표본수를 사용하기 때문이다. 즉, 단일대체는 결측값 때문에 발생하는 불확실성을 고려하지 않는 단점을 가지며, 이를 보완하기 위해서 최대우도 추정법을 사용하거나 다음에 살펴볼 다중대체를 활용할 수 있다.

3. 다중대체

Rubin(1987)이 제안한 다중대체의 핵심은 앞에서 언급한 결측으로 인한 불확실성을 고려하기 위해 하나의 결측값에 대해 여러 개의 대체값을 제공한다는 것이다. 대체값으로 여러 개의 값을 고려하기 때문에, 결측값이 (완전)임의결측에 의한 경우뿐만 아니라 비임의 결측인 경우에도 적용 가능하다. 또한, 결측값들이 대체된 이후에 이 대체 자료에 대한 새로운 분석방법이 요구되는 것이 아니라, 일반적인 통계분석 방법들을 동일하게 적용될 수 있다는 장점이 있다.

[그림 3-1]과 같이 *로 표현되어 있는 결측값을 포함하는 자료가 존재한다고 가정했을 때, 다중대체는 다음과 같이 세 단계를 거쳐 수행된다.

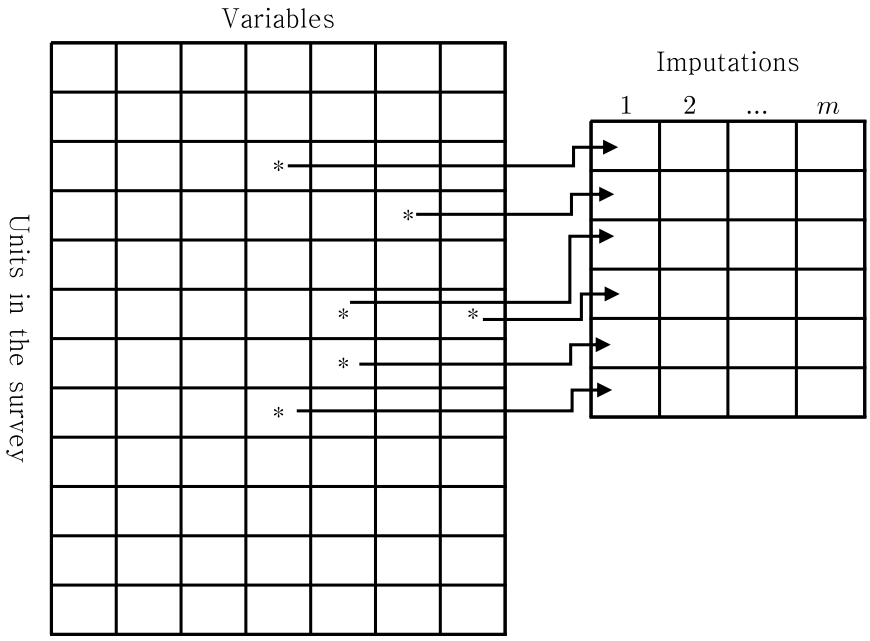
(1) Imputation: 각 결측값에 대해 적절한 대체 모형을 선택하여 결측된 자료의 예측분포를 구한다. 이 예측분포로부터 각 결측값에 대한 대체값들을 m번 랜덤하게 생성한다. 대체값들로 채워진 자료를 각각 $D^{(1)}, D^{(2)}, \cdots, D^{(m)}$라 하자.

(2) Analysis/Estimation: 각 자료 $D^{(i)}\,(i = 1, \cdots, m)$에 대해 원하는 통계분석을 시행한다. 총 m번 분석하게 된다. 각 자료로부터 얻어진 관심 모수 Q에 대한 점추정값을 $\hat{q}_i$이라 하고, $\hat{q}_i$에 대한 분산 추정량을 v_i이라 하자.

(3) Pooling: 위에서 얻어진 m개의 추정값들을 Rubin이 제시한 결합방법을 이용해 다음과 같이 합해 최종 추정값 및 분산을 계산한다.

$$\overline{Q}_m = \frac{1}{m}\sum_{m=1}^{m} \hat{q}_i,\ \ T_m = b_m/m + b_m + \overline{v}_m$$
$$(b_m = \textstyle\sum_i (q_i - \overline{q}_m)^2/(m-1),\ \ \overline{v}_m = \textstyle\sum_i v_i/m)$$

Rubin(1987)은 $\overline{Q}_m$은 Q의 불편 추정량이며 근사적으로 t-분포를 따른다는 것을 증명하여 다중대체된 자료를 종합한 통계분석의 타당성을 확보하였다.



[그림 3-1] 결측값과 다중대체(Rubin, 1987)

위의 다중대체 과정을 수행하기 위해서는 단계 (1)에서 사후예측분포를 구해야 한다. 모집단의 특성을 나타내는 모수를 θ, 자료에서 결측값이 없는 변수를 X, 결측값이 있는 변수를 Y라 하자. 변수 Y에서 관측된 부분을 Y_{obs}라고 표현하자. 변수 Y가 k개의 변수를 가진다고 가정하여 $Y=(Y_1,\cdots,Y_k)$라고 표현하면 사후예측분포는 다음과 같다.

$$P(Y_1,\cdots,Y_k|X,Y_{obs})=\int P(Y_1,\cdots,Y_k|X,Y_{obs},\theta)\,P(\theta|X,Y_{obs})\,d\theta$$

이때 $P(\theta|X,Y_{obs})$는 (X,Y_{obs})가 주어졌을 때, θ의 사후분포이다. 또 $P(Y_1,\cdots,Y_k|X,Y_{obs},\theta)$는 (X,Y_{obs},θ)가 주어졌을 때, $Y=(Y_1,\cdots,Y_k)$의 조건부 결합분포이다.

먼저 위의 사후예측분포에서 재료를 생성하기 위한 알고리즘으로 결합모형(joint modeling) 다중대체 알고리즘에 대해 살펴보자. 이론적으로 모형에 근거하여 모수를 추정하는 과정을 통해 위의 사후예측분포에서 자료를 생성하기 위한 알고리즘은 다음과 같다. 이 알고리즘은 $Y=(Y_1,\cdots,Y_k)$의 결합분포로부터 한 번에 k개의 값들을 랜덤 생성하여 동시에 대체하는 접근법이다.

(1) $P(\theta|X,Y_{obs})$로부터 표본 θ^* 생성

(2) $P(Y_1,\cdots,Y_k|X,Y_{obs},\theta^*)$로부터 표본 $Y^*=(Y_1^*,\cdots,Y_k^*)$ 생성

(3) 원하는 크기의 표본을 얻을 때까지 (1)-(2)의 과정을 반복한다.

참고로 사후분포 계산에 대해 좀 더 살펴보자. 초창기 베이지안 이론에서는 보통 사후분포를 수식으로 유도하였다. 때문에 <부록> 2.에 소개되어 있듯이 사후분포의 업데이트를 쉽게 해주는 공액사전분포의 활용이 중요하게 부각되기도 한다. 하지만 모든 분포들이 수식으로 유도되기는 어렵고 표본 생성 등의 계산과정이 많으므로, 베이지안 통계는 철학적으로, 이론적으로 주목받는데 그치고, 실제 활용에서까지 크게 각광받지는 못했었다. 하지만 컴퓨터의 발전과 더불어 기존의 방법들이 풀지 못하는 문제들을 푸는데 도움이 된다는 것이 알려지면서 베이지안 통계가 다시 주목받고 있다. 이는 통계계산 알고리즘인 MCMC 등을 이용해 사후분포를 수식으로 유도하지 않고도 구하는 것이 가능해졌기 때문이다.

다시 사후분포를 수식으로 유도하여 구하는 문제를 생각하면, 위에 설명한 알고리즘에서는 모든 변수를 한 번에 고려하여 자료를 생성한다. 이를 결합모형(joint modeling)에 의한 자료생성 방식이라 한다. 그러나 이러한 방식으로는 문제를 풀 수 없는 경우가 많아진다. 연속형 변수와 범주형 변수가 섞여 있는 경우만을 생각해도 그 어려움을 쉽게 짐작할 수 있다. 이에 한 변수씩 자료를 생성하는 순차회귀 다중대체 방식이 제안되었다. 이는 결합모형 다중대체의 한계를 뛰어넘어 베이지안 방법론을 활용할 수 있게 해준다. 이에 대해 알아보자.

4. 순차회귀 다중대체

베이지안 통계에서 모수적으로 사후분포를 추정하기 위한 방법은 크게 결합모형(joint modeling) 및 조건부모형(fully conditional specification)으로 나눌 수 있다. 결합모형은 변수 $Y=(Y_1, \cdots, Y_k)$의 결합분포로부터 한 번에 k개의 값들을 랜덤하게 생성하여 동시에 대체하는 방법이다. 순차회귀 다중대체(Sequential Regression Multiple Imputation, SRMI)는 결합모형과 달리 한 변수씩 순차적으로 표본을 생성하는 방법으로(Raghunathan 외, 2001), 조건부모형에 속하며, Multiple Imputation by Chained Equations(MICE)라고도 한다.

앞 절에서와 마찬가지로 모집단의 특성을 나타내는 모수를 θ, 자료에서 결측값이 없는 변수를 X, 결측값이 있는 k개의 변수를 Y라 하자. 또 변수 Y에서 관측된 부분을 Y_{obs}라고 표현하자. 만약 변수들 사이에 순차적 의존성이 알려져 있다고 가정하면, (X, Y_{obs}, θ)가 주어졌을 때, $Y=(Y_1,\cdots, Y_k)$의 조건부 결합분포는 다음과 같은 조건부 분포들의 결합으로 나타낼 수 있다.

$$P(Y_1, Y_2, \cdots, Y_k|X, Y_{obs}, \theta)$$

$$= P(Y_1|X, Y_{obs}, \theta)P(Y_2|X, Y_{obs}, \theta, Y_1) \cdots P(Y_k|X, Y_{obs}, \theta, Y_1, Y_2, \cdots Y_{k-1})$$

위의 조건부 확률을 바탕으로 한 SRMI 알고리즘은 다음과 같다.

(1) $P(\theta|X, Y_{obs})$로부터 표본 θ^* 생성

(2-1) $P(Y_1|X, Y_{obs}, \theta^*)$로부터 표본 Y_1^* 생성

(2-2) $P(Y_2|X, Y_{obs}, \theta^*, Y_1)$로부터 표본 Y_2^* 생성

(2-3) $P(Y_3|X, Y_{obs}, \theta^*, Y_1^*, Y_2^*)$로부터 표본 Y_3^* 생성

.....

(2-k) $P(Y_k|X, Y_{obs}, \theta^*, Y_1^*, Y_2^*, \cdots, Y_{k-1}^*)$로부터 표본 Y_k^* 생성

(3) 원하는 크기 n인 표본을 얻을 때까지 (1)-(2)의 과정을 반복한다.

SRMI는 한 번에 한 변수씩 표본을 생성하기 때문에 여러 변수를 고려하더라도 각 변수의 특성에 맞는 분포로부터 표본을 생성할 수 있다. 또한, 변수들 사이의 위계적 관계도 각 단계에서 충분히 반영될 수 있다.

만약 변수들 간의 의존성이 순차적이지 않다면, 해당 변수를 제외한 나머지 모든 변수에 대한 조건부 분포를 고려하여 표본을 생성할 수 있다. 즉, j번째 변수 Y_j에 대해 Y_j를 제외한 나머지 변수들을 $Y_{-j} = (Y_1, \cdots, Y_{j-1}, Y_{j+1}, \cdots, Y_k)$라 하면, 표본 생성은 다음의 조건부 분포를 이용한다.

$$P(Y_1|X, Y_{obs}, Y_{-1}, \theta)$$

$$P(Y_2|X, Y_{obs}, Y_{-2}, \theta)$$

$$\vdots$$

$$P(Y_k|X, Y_{obs}, Y_{-k}, \theta)$$

관련 예제가 <부록> 6.에 수록되어 있으니 관심 있는 독자는 참고하기 바란다.

제4절 완전 재현자료

이제 Rubin이 제안한 재현자료 생성 방법을 살펴보기로 하자. 먼저 완전 재현자료가 제안된 배경과 기본적인 아이디어를 설명하고, 완전 재현자료 생성 방법론과 결합규칙을 소개한다. 2절과 3절의 내용을 바탕으로 이번 절에서 설명하는 재현자료 생성의 큰 흐름을 이해하도록 한다.

1. 아이디어

1993년 국가통계방법론 분야에서 권위 있는 저널인 Journal of Official Statistics에서는 비밀보호(data confidentiality)에 대한 특집호를 기획하여 발간했다. 기존의 매스킹 기법들로는 노출위험과 정보손실 사이의 상충관계[7]를 극복하여 자료를 제공하기 어렵기 때문에 새로운 대안을 모색하기 위한 시도였다고 생각된다. 이 특집호에 실린 두 개의 논문(Rubin, 1993; Little, 1993)은 매스킹 처리를 대체하는 비밀보호 대안으로서 현재 연구되고 있는 재현자료의 기초를 놓았다(Drechsler, 2011).

먼저 Rubin(1993)은 자신의 다중대체 기법(Rubin, 1987)을 기반으로 훗날 완전 재현자료라 불리는 방안을 제시했다. 이는 ① 표본틀(모집단)에서 조사되지 않은 모든 값들을 결측값으로 취급하여 다중대체하고, ② 대체되어 채워진 재현 모집단에서 단순랜덤추출로 표본을 추출하여 제공하는 두 단계로 이루어진다. 두 번째 주요 논문인 Little(1993)의 아이디어는 기본적으로 Rubin(1993)의 제안과 유사하며, 다른 점은 표본 자료의 일부만을 대체하는 것이었다. 이는 부분 재현자료 작성 방법론으로 발전한다. 부분 재현자료에 관한 자세한 내용은 다음 절에서 다루도록 한다.

Rubin(1993)에 따르면, 제공되는 재현자료는 실제 데이터를 포함하지 않으므로 민감한 정보가 노출되기 대단히 어렵다. 또한 대체 모형을 잘 선택하여 예측력을 높인다면 원래 자료에 담긴 정보 대부분이 재현자료에 보존될 수 있다고 보았다. 이를 통해 기존 매스킹 기법들이 가지는 한계를 극복하고, 아래와 같은 자료제공에 있어 바람직한 세 조건을 만족시킬 수 있는 방향으로서 재현자료를 제시했다.

(1) 개체별 정보 노출 방지 등, 안전성 확보
(2) 타당한 추론을 가능하게 하는 적은 정보손실
(3) 일반적인 통계기법을 쉽게 적용할 수 있다는 의미에서 이용자의 편의성

---

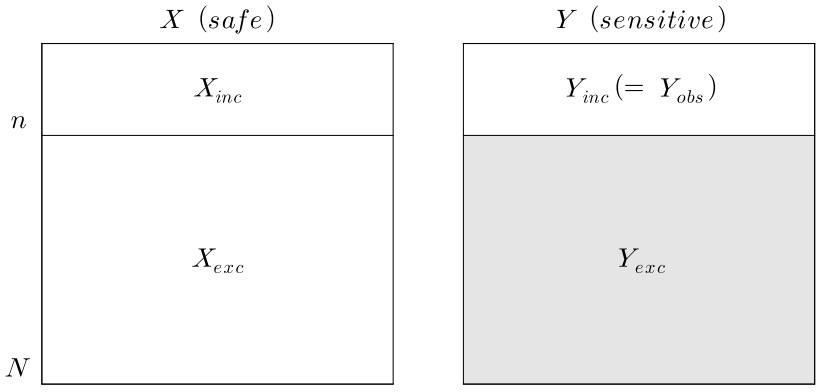
7) 간단한 예를 들면 노출위험을 줄이기 위해 연령의 각 세를 5세 단위로 재코딩(global recoding, 전통적인 매스킹 기법 중 하나)하여 자료를 변형하면 노출위험은 줄어들지만 정보손실이 증가한다. 더 자세한 내용은 박민정(2014)을 참고하라.

이러한 아이디어는 기존의 매스킹 처리에 견주어 획기적인(groundbreaking) 것으로 여겨졌다. 그러나 모형 의존성이 크기 때문에 생성된 자료를 이용해 얻은 통계 추정량의 편의가 큰 경우가 빈번하여, 실제로 통계기관에서 완전 재현자료가 원 자료를 대신해 자료배포에 이용된 적은 없는 것으로 파악된다. 다만 완전 재현자료 작성 방법론의 주요 내용은 부분 재현자료 생성으로 이어지며, 이론적인 발전과 여러 분야의 실제 자료에 적용하는 연구들이 지금까지 지속되고 있다.

2. 완전 재현자료 생성

가. 자료구조

먼저 자료의 구조를 살펴보자. 성별, 연령, 지역 등, 조사 및 응답 여부에 상관없이 모든 개체에게 관측될 수 있는 p개 변수에 대해 N개 개체로 구성된 마이크로데이터 행렬을 $X=(X_i,\ i=,2,...,N)$라 하자. 이때 X_i는 p개의 값들로 이루어진 행벡터이며, p개 변수에 대한 i번째 개체의 정보를 담고 있다. X는 표본틀이나 행정 자료 등에서 통계기관이 얻을 수 있고 노출되어도 문제가 없는 안전한 자료라 가정한다. 한편 조사를 통해 관측하고자 하는 k개 변수, N개 개체로 구성된 마이크로데이터 행렬을 $Y=(Y_i,\ i=1,...,N)$라 하자. X_i와 마찬가지로, Y_i는 k개 변수들에 대한 i번째 개체의 응답값들로 이루어진 행벡터이다. Y는 소득, 특정 질병의 유무, 정치적 성향 등과 같이 개별정보가 노출되면 곤란한 민감한 변수들로 구성되어 있다고 가정한다.



[그림 3-2] 자료구조

이제 모집단의 크기를 N이라 하고 표본으로 추출되어 조사에 응답한 개체수, 표본의 크기를 n이라 하자. 편의상 무응답은 없다고 가정한다. 이제 모집단의 Y 중에서 표

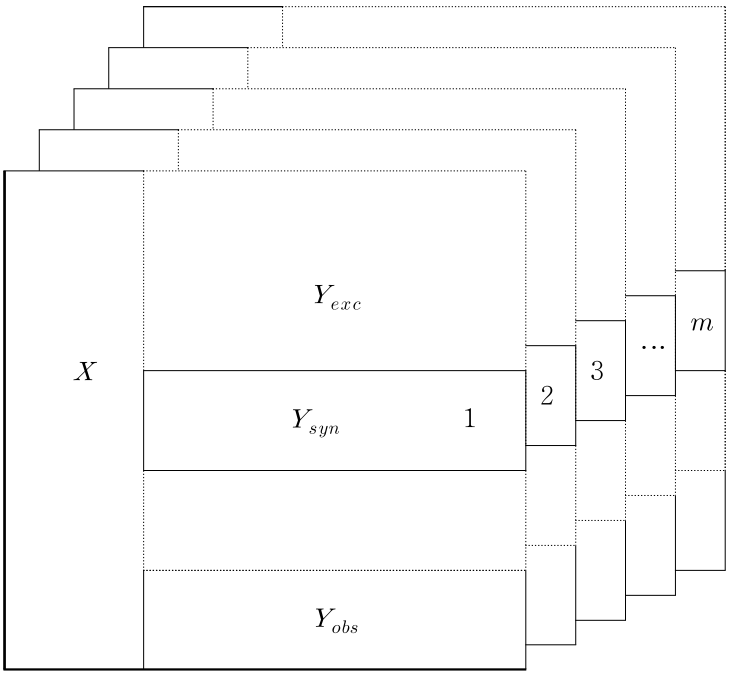
본에 포함되어 관측된 자료가 존재하는 개체들의 모임을 $Y_{inc}(=Y_{obs})$, 표본에서 제외된 개체들의 모임을 Y_{exc}로 표현하자. 변수 X에 대해서도 $Y_{inc}(=Y_{obs})$ 및 Y_{exc}에 대응하도록 모집단의 개체를 나누어 X_{inc} 및 X_{exc}라 표현하자. 즉, $X=(X_{inc}, X_{exc})$가 된다. 위의 [그림 3-2]는 표본 포함 여부에 따른 자료구조를 나타내며 회색으로 처리된 Y_{exc}는 자료값이 존재하지 않는 부분에 해당한다.

나. 모수적 재현자료 생성

완전 재현자료 생성은 다중대체 방법을 그대로 사용한다. 완전 재현자료를 작성을 두 단계로 정리하면 다음과 같다.

(1) 결측값 대체: Y_{exc} 전체를 결측값으로 간주하고, (X, Y_{obs})을 바탕으로 추정한 사후예측분포로부터 Y_{exc}의 대체값을 생성한다.

(2) 랜덤 추출: 재현된 모집단, 즉 조사에 참여하지 않은 사람들의 자료 (X_{exc}, Y_{exc})로부터 랜덤하게 추출한 표본을 배포한다.



출처: J. Reiter and J. Drechsler, Lecture notes in a professional development continuing education course "Synthetic data sets for statistical disclosure limitation", JSM 2017.

[그림 3-3] 완전 재현자료 생성을 이해하기 위한 자료구조

다중대체와 마찬가지로 위의 과정을 m번 반복하여 m개의 샘플을 제공한다. 이러한 내용을 그림으로 표현하면 앞의 [그림 3-3]과 같다. 그림의 관측된 Y_{obs}대신에 재현 모집단에서 추출된 재현자료 Y_{syn}을 m개 세트 제공한다고 이해할 수 있다.

한편, 위의 과정에서 X를 안전하다고 가정했으므로 Y_{exc}만을 생성하였으나, X에도 노출위험이 있어 제공하기 어려운 경우 모집단의 X와 Y를 모두 생성한다.

구체적으로 사후예측분포를 어떻게 추정하고 자료를 생성하는지 이해하기 위해 위의 두 단계를 수식으로 살펴보자. 먼저, X가 안전하여 공개해도 상관없는 경우, X_{exc}을 재현 모집단에서 그대로 사용한다. 관측된 자료 X, Y_{obs} 및 결측값 대체 모형에 대한 가정을 바탕으로 다음과 같은 사후예측분포를 얻는다.

$$P(Y_{exc}|X, Y_{obs}) = \int P(Y_{exc}|X, Y_{obs}, \theta)\, P(\theta|X, Y_{obs})\, d\theta$$

변수 Y에 대해 공통으로 적용할 다변량 모형이 있다면 3절 다중대체에서 설명한 결합 모형에 의한 재현자료 생성이 가능하고, 변수들이 서로 다른 분포를 따른다면 3절의 순차회귀 다중대체 등을 사용하여 재현자료 생성이 가능하다.

한편, X가 노출위험이 높은 변수로 구성되어 있어서 공개가 불가능한 경우 완전 재현자료 생성은 다음과 같은 과정을 거친다. 관측된 자료 X, Y_{obs} 및 모형에 대한 가정을 바탕으로 얻어진 사후예측분포 $\Pr(X_{exc}, Y_{exc}|X_{inc}, Y_{obs})$로부터 재현 모집단 $\{(X^{(i)}, Y^{(i)}) : i = 1, \cdots, m\}$을 생성한다. 이때 재현모집단의 변수들 (X_f, Y_f)을 추출할 사후예측분포는 다음과 같다.

$$P(X_f, Y_f|X_{inc}, Y_{obs}) = \int P(X_f, Y_f|X_{inc}, Y_{obs}, \theta)\, P(\theta|X_{inc}, Y_{obs})\, d\theta$$

이제 m개의 재현 모집단 $\{Pop^{(i)} = (X^{(i)}, Y^{(i)}) : m = 1, \cdots, m\}$이 생성되었다고 하자. 각 재현 모집단 $Pop^{(i)} = (X^{(i)}, Y^{(i)})$으로부터 크기가 n^*인 표본을 랜덤 추출하고 이를 다음과 같이 표현한다.

$$D^{(i)} = (x_j^{(i)},\ y_j^{(i)},\ j = 1, 2, \cdots, n^*)$$

이제 이 표본들의 집합 $\{D^{(i)} : i = 1, \cdots, m\}$을 공개하는 것이 단계 (2)랜덤 추출에 해당한다.

이해를 돕기 위해 Raghunathan 외(2003)에 소개된 모의실험을 따라하여 완전 재현자료를 직접 만들어보자. 관련 R코드는 <부록> 7.에 제시되어 있다. 여기서는 모수적인 방법으로 재현자료를 생성한다. 이 모의실험은 변수들이 다변량 정규분포를 따르는

경우로 사후 분포를 수식으로 유도하여 계산하고 결합모형 방식으로 재현자료를 생성하는 가장 단순한 경우에 해당한다.

모의실험에서 모집단은 $k=5$개 변수를 가지고, 이 확률변수들은 다음과 같은 다변량 정규분포를 따른다고 가정한다. 아래의 다변량 정규분포에서 크기 $N=1000$인 자료를 생성하여 모집단이라 하고, 이 모집단에서 크기 $n=100$인 표본을 추출하여 Y_{obs}라고 생각하자.

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix} \right)$$

이 모의실험에서 모수적 방법으로 완전 재현자료 생성하는 과정은 다음과 같이 요약될 수 있다.

(1) 표본에서 표본평균 $\overline{x}$과 표본 분산 S을 구한다.

(2) Wishart 분포 $W_p(S^{-1}/(n-1), n-1)$으로부터 난수 W 생성 후 $\Sigma^* = W^{-1}$이라 한다.

(3) 다변량 정규분포 $N_p(\overline{x}, \Sigma^*/n)$로부터 난수 μ^*를 생성한다.

(4) 다변량 정규분포 $N_p(\mu^*, \Sigma^*)$로부터 크기 $N=1000$인 재현 모집단을 생성한다.

(5) 재현 모집단으로부터 크기 $n^* = 250$인 표본을 무작위 추출 후 저장한다.

(6) 주어진 표본에 대하여 (2)-(5)을 $m=5$회 반복한다.

Wishart 분포

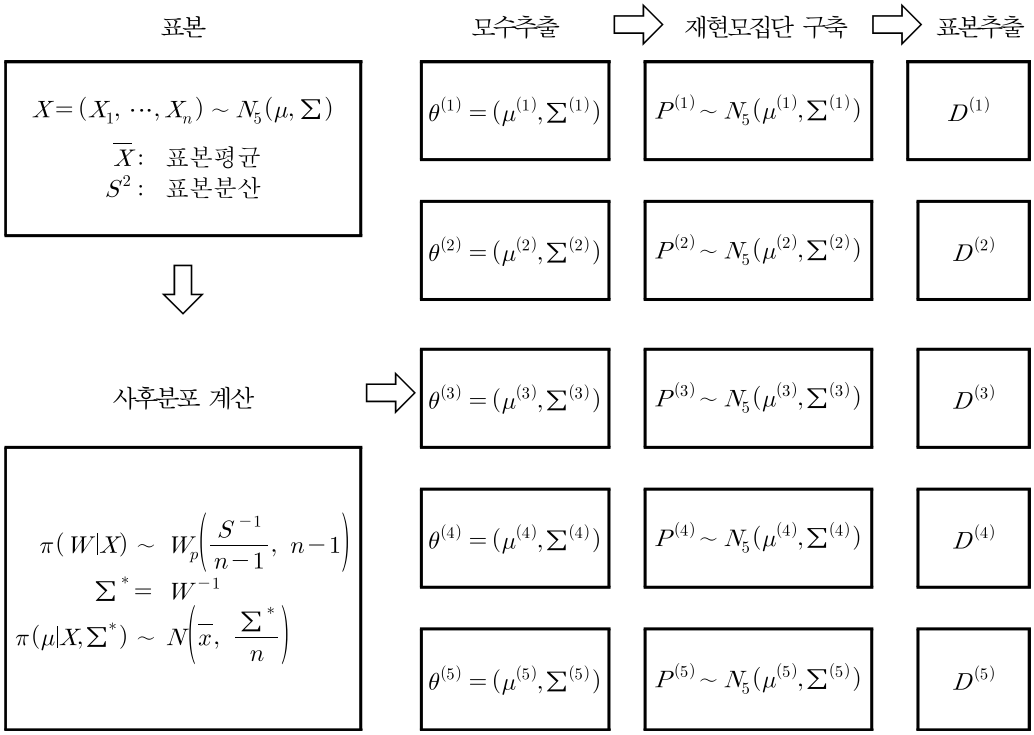
다음과 같이 평균이 $(0, 0, \cdots, 0)$, 공분산이 V인 p차원 다변량 정규분포를 따르는 크기 n인 표본 $X_1, X_2, \cdots, X_n \sim N_p(0, V)$을 가정하자. 이때 $S = \sum_{i=1}^{n} X_i^T X_i$는 다음과 같은 Wishart 분포를 따른다(Wishart, 1928).

$$S \sim W_p(V, n), \; E(S) = n V.$$

앞의 2절과 3절에서 살펴본 베이지안 이론에 따라 위의 과정을 설명해 보자. 먼저 우리는 이 모집단이 다변량 정규분포를 따른다고 가정했다. 그러면 모수 θ는 평균과 분산이 된다. (이 모수에 대해 무정보 사전분포를 사용하면 $\pi(\mu, \Sigma) \propto |\Sigma|^{-1/2}$가 된다.

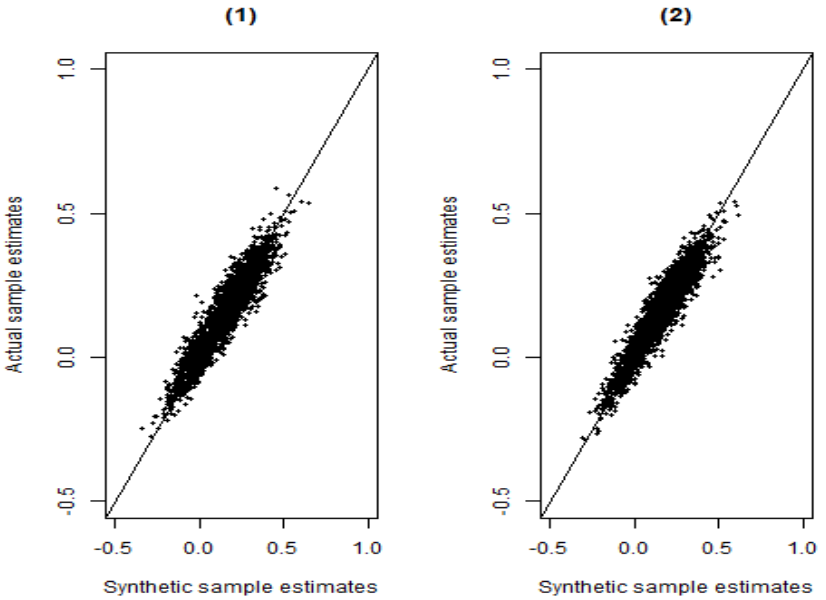
이는 가정한 모형에 대해 공액사전분포가 되므로 2개 모수에 대한 사후분포는 Wishart 분포와 다변량 정규분포가 된다.) 과정 (1)-(3)에서는 주어진 표본자료 Y_{obs}를 기반으로 이론적으로 수식을 통해 계산된 사후분포에서 모수 $\theta^* = (\mu^*, \Sigma^*)$를 생성한다. 생성된 모수를 통해 모집단의 분포를 효과적으로 추정했다 판단하고, 과정 (4)에서는 추정된 모집단의 분포에서 자료를 생성하여 재현 모집단을 구축한다. 즉 (1)-(4)의 과정이 Rubin이 제안한 결측값 다중대체의 단계를 구성하는 것이다.

한편, 과정 (5)는 이 재현 모집단에서 단순랜덤추출로 표본을 추출하여 재현자료를 생성하는 것을 보여준다. 이때 재현자료(표본)의 크기는 원래 표본자료의 크기 n과 동일할 필요가 없다. 또한 단순랜덤추출로 재현자료를 생성하였으므로 이용자는 복잡한 표본추출 기법이나 가중값을 고려하지 않고, 일반적인 통계분석을 바로 시행할 수 있다. 이러한 재현자료 생성 프로시저는 아래 [그림 3-4]에 표현되어 있다. 참고로, 최종 통계량을 구하기 위해서는 $m = 5$개 재현자료 세트에서 원하는 통계량을 각각 계산하여 평균한다. 이 통계량 추정값의 분산은 다음 소절에 소개하는 결합규칙에서 제시하는 수식을 이용해 결정한다.



[그림 3-4] 다변량 정규분포를 이용한 재현자료 생성

이렇게 재현자료를 생성하는 것이 타당한지 알아보기 위해 위 모의실험에서 처음에 생성된 하나의 모집단에 대해 하나의 표본을 추출하고 5개의 재현자료를 만드는 작업을 총 500회 반복하였다. 재현자료를 통해 원래 표본자료와 동일한 통계량을 얻을 수 있는지 알아보기 위해, 첫 번째 변수 Y_1을 반응변수로 나머지 네 변수를 설명변수로 회귀모형을 적합하여 얻은 회귀계수를 구했다. 즉, 표본자료$(n=100)$에 대해 회귀계수 $\hat{\beta}$을 구하고, 5개 재현자료$(n^* = 250)$에 대해 회귀계수를 각각 구해 평균하여 $\hat{\beta}^*$라 하였다. 아래 [그림 3-5]에서 (1)은 모수적 방법으로 500번 반복하여 얻은 회귀계수 $\hat{\beta}^*$을 x축에, $\hat{\beta}$을 y축에 그렸다. 오차가 있으나 전반적으로 원 자료와 재현자료가 생성하는 통계량이 유사한 경향이 있음을 알 수 있다. 참고로 [그림 3-5]에서 (2)는 뒤에 소개하는 비모수적 방법을 적용한 결과이다. 세부 통계량들은 <부록> 8.에 제시되어 있다.



[그림 3-5] 원 자료와 재현자료를 이용해 얻은 통계량 비교 실험

다. 비모수적 재현자료 생성

비모수적으로 완전 재현자료를 생성한다는 것은 사후분포를 추정하지 않고 관측값의 일부를 랜덤 추출한 표본으로 재현자료를 구성하는 것을 의미한다. 앞의 소절에서 보았다시피 사후분포를 모수적으로 추정하기 위해서는 수식을 유도하는 과정이 필요하나,

모든 경우에 이러한 수식 유도가 가능하지는 않다는 단점이 있다. 때문에 최근 연구에서도 비모수적 재현자료 생성이 주요 관심사인 경우가 많다. 비모수적 완전 재현자료 작성에서는 베이지안 붓스트랩을 사용한 랜덤추출을 주로 사용한다. 베이지안 붓스트랩은 2절에 설명되어 있다.

앞에서 살펴보았던 다변량 정규분포 모의실험과 같은 상황에서 비모수적으로 재현자료를 생성해보자. 모수적 방법을 사용할 때는 모집단이 다변량 정규분포를 따른다고 가정하여 사후분포를 추정하고 추정된 사후분포를 이용하여 재현 모집단을 생성하였다. 하지만 베이지안 붓스트랩을 사용하면 모집단 분포에 대한 가정 없이 비모수적으로 재현 모집단을 생성할 수 있다. 관련 R코드는 <부록>에 첨부되어 있으며, 베이지안 붓스트랩을 이용한 재현자료 생성 과정을 설명하면 다음과 같다. 앞의 모의실험과 같이 크기 $N=1000$인 모집단과 추출된 크기 $n=100$인 표본이 있다고 하자. 이 표본을 $x_1, \cdots, x_n$라고 하자.

(1) 균일분포 $U(0,1)$로부터 $(n-1)$개의 난수를 생성하고 오름차순으로 정렬하면 다음과 같다.

$$a_0 = 0,\ a_1,\ a_2,\ \cdots,\ a_{n-1},\ a_n = 1$$

(2) 균일분포 $U(0,1)$로부터 새로운 N개의 난수 $u_1, u_2, \cdots, u_N$을 생성한다. 각 u_i $(i=1, \cdots, N)$에 대하여, $a_{j-1} < u_i \le a_j$를 만족하는 j를 찾은 뒤, 이에 대응하는 표본 x_j을 재현 모집단에 포함시킨다.

(3) (2)에서 구한 재현 모집단으로부터 크기 $n^* = 250$인 표본을 무작위 추출 후 저장한다.

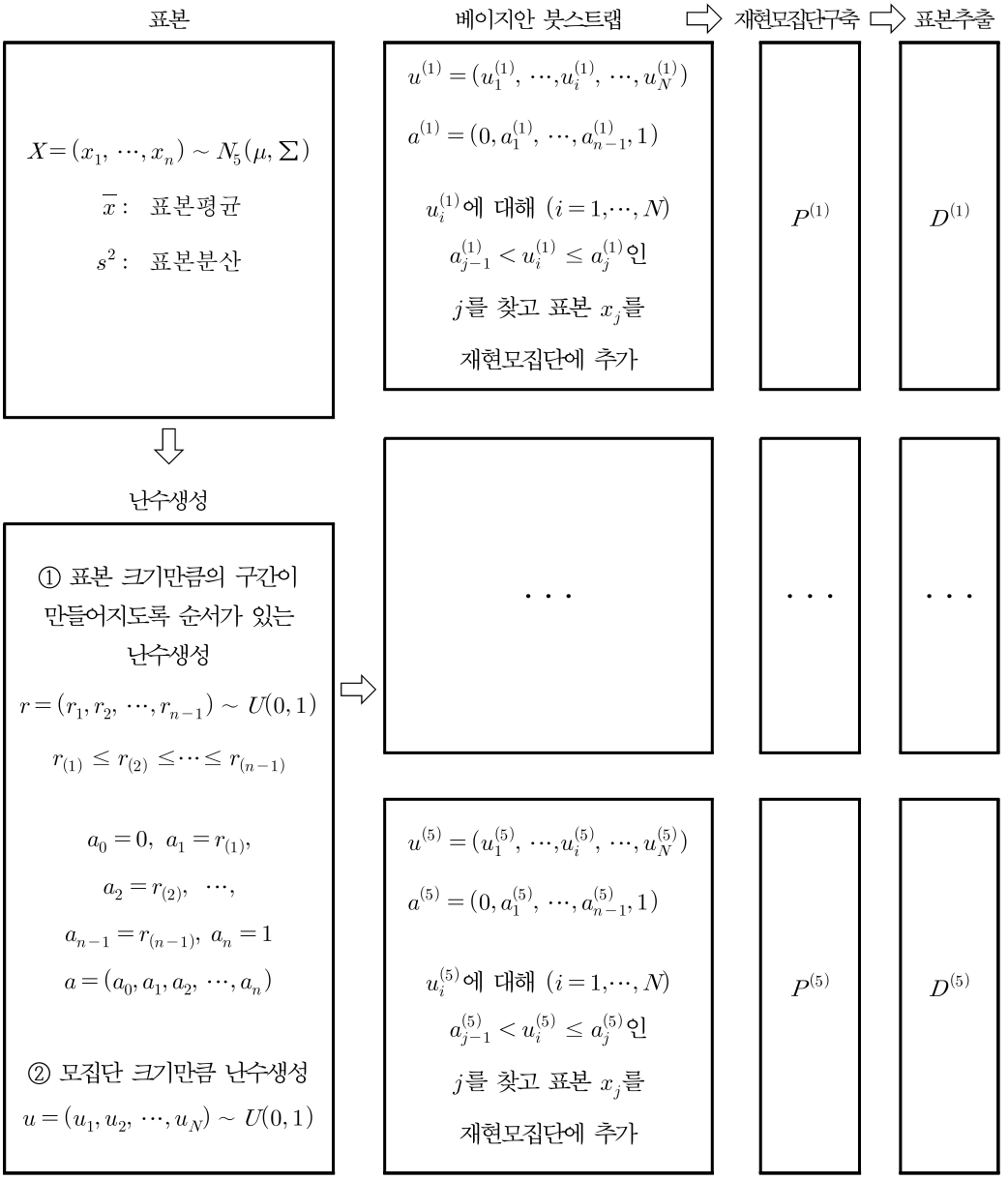
(4) 주어진 표본에 대하여 (1)-(3)의 과정을 $m=5$회 반복한다.

이를 도식화 한 것이 다음 [그림 3-6]이다. 비모수적으로 재현모집단을 생성할 때는 실제 자료를 사용하므로 완전 재현자료 생성만큼 노출위험이 완전히 없어지지는 않는다.

3. 결합규칙

사용자 입장에서 재현자료를 활용하는 문제를 생각할 때, 가장 낯선 사항은 제공받은 자료가 여러 세트라는 것이다. 재현자료 이론에서는 각 세트에서 얻은 이용자가 원하는 통계량을 결합하는 규칙을 제공한다. 단일대체의 단점을 보완하기 위해, 즉 적절한 분산 추정량을 얻기 위해 다중대체를 사용하는데 동의한다면, 여러 재현자료 세트를 제공받아 결합해야 하는 이유도 같은 맥락에서 이해할 수 있다.

이제 이용자가 재현자료를 바탕으로 관심 모수 $Q = Q(X, Y)$에 대해 추론한다고 하자. 원 자료를 바탕으로 얻어지는 추정량을 q, 분산을 v로 표현하자. 또한 i번째 완전 재현자료 $D^{(i)}$를 바탕으로 얻어지는 추정량과 그것의 분산을 각각 q_i 및 v_i라고 하자. 이때 모수 Q에 대한 추정량과 분산은 다음과 같이 계산한다.



[그림 3-6] 베이지안 붓스트랩을 기반으로 한 재현자료 생성

$$\bar{q}_m = \sum_{i=1}^{m} q_i/m,\ \ T_f = b_m/m + b_m - \bar{v}_m$$

$$(b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2/(m-1),\ \ \bar{v}_m = \sum_{i=1}^{m} v_i/m)$$

Raghunathan 외(2003)에서는 $\bar{q}_m$이 Q에 대한 불편(unbiased) 추정량이며 T_f 역시 $\bar{q}_m$의 분산에 대한 불편 추정량임을 증명하였다. 또한, $\bar{q}_m$의 분포가 다음과 같은 t-분포로 근사됨을 보였다.

$$\sqrt{T_m}(\bar{q}_m - Q) \sim t_{\nu_f}$$

단, 자유도는 $\nu_f = (m-1)(1 - \bar{u}_m/((1+1/m)b_m))^2$이다.

제5절 부분 재현자료

앞 절에서 완전 재현자료가 제안된 배경, 생성 방법, 이용자가 최종 통계량을 얻기 위해 m개 재현자료를 활용하여 얻은 통계량들을 결합하는 규칙에 대해 살펴보았다. 모수적 접근으로 생성한 완전 재현자료는 먼저 모집단을 재현하고, 재현한 모집단에서 표본을 추출하는 원리를 따르므로 실제 자료가 전혀 공개되지 않는다. 따라서 노출제어 측면에서 매우 우수하며 단순랜덤추출로 얻은 표본을 제공하므로 표본설계를 고려하지 않고 일반적인 분석방법을 그대로 적용할 수 있어 사용자에게 편리하다. 하지만 결측값을 대체할 때 모집단에 가정하는 모형에 매우 의존적이고, 현실적으로 많은 변수를 대체할 모형이 존재하지 않을 수도 있다는 단점을 가진다. 즉, 모집단에 대해 모형을 잘못 가정하는 경우, 얻어진 통계량에서 편의가 무척 크다는 문제가 있다. 이는 노출위험이 작은 대신에 정보손실이 매우 큰 자료를 제공하는 셈이 된다. 이제 이러한 정보손실을 완화하고자 발전해 온 부분 재현자료 생성 방법과 결합 규칙에 대해 살펴보자. 부분 재현자료는 정보손실을 줄이기 위해 노출위험 증가하는 것을 수용하는 측면에서 이해하는 것이 바람직하다.

1. 아이디어

부분 재현자료 작성은 Little(1993)에서 비롯되었다. 자료의 모든 정보가 민감하다고

보기는 어려운 경우도 많으므로, 모든 변수가 아니라 노출제어 처리가 필요한 일부 변수만 다중대체하자는 것이 핵심 아이디어이다. 대체를 고려하는 변수는 다음과 같은 두 가지 유형으로 나누어 볼 수 있다.

(1) 민감변수 : 소득, 특정 질병의 유무, 정치적 성향 등, 응답자가 노출을 꺼리는 변수

(2) 키변수 : 응답자의 특성을 나타내어 조합하면 유일한 응답자를 식별해 낼 수 있는 변수

부분 재현자료를 생성할 때, 노출위험이 높다고 판단되는 개체들의 키변수와 민감변수 모두에서 다중대체를 시행할 수도 있다. 아니면 대체하기로 선택된 변수의 모든 값을 대체할 수도 있다. 또한 대체하기로 선택된 변수에 대해서 노출위험이 높은 일부 개체에 대해서만 값을 대체하는 것도 가능하다. 즉 부분 재현자료 생성에서는 자료의 일부만을 결측값으로 간주하고 다중대체 기법을 적용한다. 이때 대체되지 않은 값들은 원래의 관측값이므로 자료 고유의 성질이 대부분 유지된다. 이는 정보손실을 줄일 수 있는 가장 큰 이유이기도 하다.

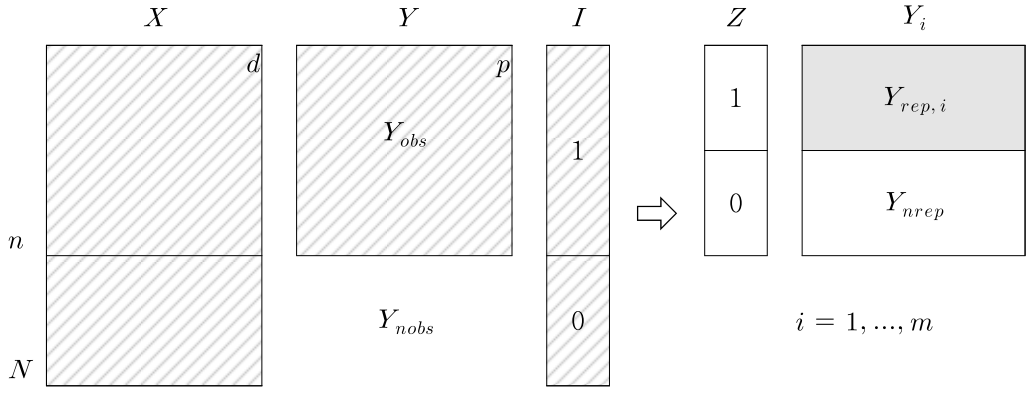
2. 부분 재현자료 생성

가. 자료구조

모집단의 크기는 N이라 하자. 노출위험이 낮아 대체가 필요 없는 d개 변수에 대해 N개 개체로 구성된 마이크로데이터 행렬을 $X=(X_i,\ i=1,2,\cdots,N)$라 하자. 이때 X_i는 d개의 값들로 이루어진 행벡터이며, d개 변수에 대한 i번째 개체의 정보를 담고 있다. 보통 X는 층화 변수 등으로 표본틀이나 행정 자료를 통해 통계기관이 모집단 전체에 대한 대략적인 정보를 가지고 있는 자료이다. 한편 노출위험이 높아 대체 필요성이 존재하는 p개 변수, N개 개체로 구성된 마이크로데이터 행렬을 $Y=(Y_i,\ i=1,\cdots,N)$라 하자. X_i와 마찬가지로, Y_i는 p개 변수들에 대한 i번째 개체의 응답값들로 이루어진 행벡터이다. 표본에 포함되어 관측된 n개 개체의 자료는 Y_{obs}로 표현하고, 정보가 없는 나머지 개체에 대한 자료를 Y_{nobs}로 표현한다. 편의상 Y_{obs}에는 무응답이 없다고 가정한다.

자료구조를 도식화하면 아래 [그림 3-7]과 같다. 변수 I는 모집단의 개체가 표본에 포함된 여부를 나타낸다. 회색으로 표시된 부분이 대체가 일어나는 부분을 나타내고, 빗금으로 표현된 부분은 $D=(X, Y_{obs}, I)$로 표현하여 재현자료를 생성하기 위해 이용할 수 있는 정보들의 집합을 표현한다. I는 표본에 포함된 여부를 나타내는 변수이다. 변수 Z는 Y_{obs}에 속하는 개체들 중, 노출위험이 높아 대체되는 경우는 1의 값을, 노출

위험이 낮아 대체될 필요가 없는 경우는 0의 값을 가지는 변수이다. 보통 변수 Z의 값은 D의 정보를 이용하여 결정한다. 이제 $Z=1$인 개체들에 대해서 재현자료 Y_{rep}을 m개 세트 생성하여 부분 재현자료로 사용하면 된다. Y_{nrep}는 $Z=0$인 개체들의 원래값을 의미하며 모든 재현자료 세트에서 동일하게 제공된다. 최종 제공되는 부분 재현자료는 $d_i=(X,\ Y_{rep,i},\ Y_{nrep},\ I,\ Z),\ i=1,\ \cdots,\ m$로 표현할 수 있다.



[그림 3-7] 부분 재현자료 생성을 이해하기 위한 자료구조

나. 재현자료 생성

완전 재현자료의 경우와 마찬가지로, [그림 3-7]로 표현되는 자료에 대해 사후예측분포를 나타내면 아래와 같다.

$$P(\ Y_{nobs}|X,\ Y_{obs}\ )=\int P(\ Y_{nobs}|X,\ Y_{obs},\ \theta\ )\,P(\theta|X,\ Y_{obs}\ )\,d\theta$$

이 사후예측분포를 따르는 자료들을 생성하기 위해 이론적으로(analytically) 위의 수식을 계산하여 분포식을 구하고 그 식으로부터 원하는 통계량을 계산하거나 자료를 생성할 수도 있다. 그러나 반대로 비모수적으로 혹은 몬테칼로 기법을 사용하여 자료를 생성하여 사후예측분포를 근사적으로 만들고 원하는 통계량을 얻을 수도 있다. 재현자료에서는 일반적인 베이지안 통계와 달리 사후분포를 구하는 것보다 생성된 자료를 필요로 하므로 이론적인 풀이보다는 몬테칼로 방법이 유용하다. 먼저 완전 재현자료 생성은 다음과 같은 절차로 간단히 요약될 수 있다.

$$\theta_i\ \sim\ P(\theta|X,\ Y_{obs}\ )$$

$$Y_{nobs}|\,\theta_i\ \sim\ P(\ Y_{nobs}|X,\ Y_{obs},\ \theta\ )$$

부분 재현자료를 생성할 때는 완전 재현자료 생성에 사용되는 사후분포 $P(\theta|X, Y_{obs})$ 대신에, $P(\theta|X, Y_{rep})$을 이용하여 모수를 생성한다. 따라서 모형을 추정하는 대상 집단이 축소되고 편의를 줄일 수 있게 된다. 예를 들어 Y가 소득 변수이고 Y값이 1억원 이상인 고소득을 노출위험이 높다고 판단하여 Y_{rep}으로 정한다고 하자. 완전 재현자료 생성에서는 사후분포를 추정하거나 모수값을 생성하기 위해 사용되는 자료에 모든 소득의 값이 반영된다. 하지만 부분 재현자료 생성에서는 고소득 자료만을 이용하여 사후분포를 추정하거나 모수값을 생성하므로 편의가 많이 줄어들게 된다. 한편 결측값 대체 모형으로는 조건부 예측분포, 회귀분석 등과 같은 기존의 일반적인 대체 방법들을 적용할 수 있다.

완전 재현자료의 과정과 유사하게 대체모형을 이용하여 대체된 자료 $Y_{rep,i}$와 원 자료와 동일한 Y_{nrep}를 합친 대체자료를 $Y^{(i)}$라 하자. 이제 m개의 대체자료 세트 $\{Y^{(i)}: i=1, \cdots, m\}$을 생성한 뒤, X와 결합한 m개의 자료 $\{D^{(i)}=(X, Y^{(i)}): i=1, \cdots, m\}$을 공개한다. 한편 완전 재현자료를 사용할 때는 단순랜덤추출된 것으로 가정할 수 있는 것에 반하여 부분 재현자료는 원래의 표본 설계를 따른다고 가정하고 분석한다.

3. 결합규칙

재현자료를 바탕으로 관심 모수 $Q=Q(X, Y)$에 대해 추론한다고 하자. 원 자료를 사용하여 얻는 Q에 대한 추정량을 q, 분산은 v라 하자. 한편 i번째 부분 재현자료 $D^{(i)}$로부터 얻어진 추정량과 그것의 분산을 각각 q_i 및 $v_i (i=1, \cdots, m)$라 하자. 이때 모수 Q에 대한 추정량과 분산은 다음과 같다. 다중대체, 완전 재현자료, 부분 자료에서 사용하는 분산식 T_m, T_f, T_p가 조금씩 다름을 주시할 필요가 있다.

$$\bar{q}_m = \sum_{i=1}^{m} q_i/m, \quad T_p = b_m/m + \bar{v}_m$$

$$(b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2/(m-1), \quad \bar{v}_m = \sum_{i=1}^{m} v_i/m)$$

Reiter 외(2005)에서는 $\bar{q}_m$이 Q에 대한 불편(unbiased) 추정량이며 T_p 역시 $\bar{q}_m$의 분산에 대한 불편 추정량임을 증명하였다. 또한, $\bar{q}_m$의 분포가 다음과 같은 t-분포로 근사됨을 보였다.

$$\sqrt{T_p}(\bar{q}_m - Q) \sim t_{\nu_p}$$

단, 자유도는 $\nu_p = (m-1)(1+\bar{u}_m/(b_m/m))^2$이다.

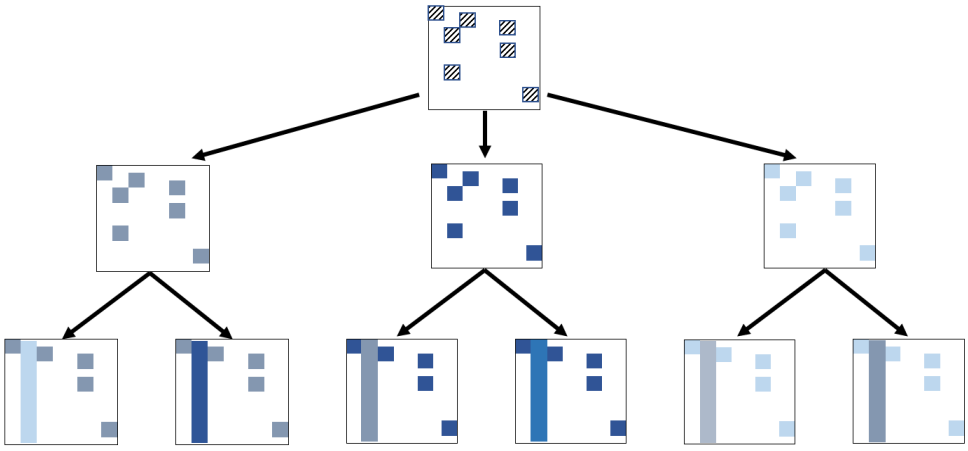
4. 부분 재현자료 생성: 결측값이 존재하는 경우

부분 재현자료를 생성할 때 결측값을 함께 처리하는 연구 결과가 있어 간단히 소개한다. 주요 아이디어는 완전 재현자료 생성 프로세스처럼 다중대체와 재현자료 생성을 두 단계로 나누어 수행하는 것이다. 결측값이 존재하는 경우 부분 재현자료 생성은 다음과 같이 제안되었다.

(1) 다중대체를 통해 r개의 데이터 세트를 생성

(2) (1)에서 생성된 데이터 세트 각각에 대해 m개의 부분 재현자료를 생성

다음 [그림 3-8]은 이를 도식화 한 것으로, 먼저 $r=3$개의 다중대체 데이터 세트를 생성하고 각 세트에서 $m=2$개의 부분 재현자료를 생성하는 과정을 나타낸다.



출처: J. Reiter and J. Drechsler, Lecture notes in a professional development continuing education course "Synthetic data sets for statistical disclosure limitation", JSM 2017.

[그림 3-8] 부분 재현자료 생성 - 결측값이 존재하는 경우

위의 과정을 거쳐 부분 재현자료는 총 $M=mr$개 세트가 생성된다. 이용자가 이를 이용하여 관심 모수 $Q=Q(X, Y)$에 대한 추론을 하고자 할 때 평균 및 분산을 얻는 결합규칙은 다음과 같다.

$$\bar{q}_M = \sum_{l=1}^{m}\sum_{i=1}^{r} q_i^{(l)}/(mr) = \sum_{l=1}^{m} \bar{q}^{(l)}/m, \quad T_M = (1+m^{-1})B_M - \bar{b}_M/r + \bar{u}_M$$

단, $B_M = \sum_{l=1}^{m} (\bar{q}^{(l)} - \bar{q}_M)^2/(m-1)$,

$$\overline{b}_M = \sum_{l=1}^{m}\sum_{i=1}^{r}(q_i^{(l)} - \overline{q}^{(l)})^2/(m(r-1)) = \sum_{l=1}^{m} b^{(l)}/m,$$

$\overline{u}_M = \sum_{l=1}^{m}\sum_{i=1}^{r} u_i^{(l)}/(mr)$이다.

한편 Reiter (2004)는 $\overline{q}_M$이 Q에 대한 불편 추정량이며 다음과 같은 근사분포를 갖는다는 것을 증명하였다.

$$\sqrt{T_M}(\overline{q}_M - Q) \sim t_{\nu_M}$$

단, 자유도는 $\nu_M = \left(\frac{((1+1/m)B_M)^2}{(m-1)T_M^2} + \frac{(\overline{b}_M/r)^2}{m(r-1)T_M^2} \right)^{-1}$ 이다.

제6절 R패키지 활용

영국은 잉글랜드와 웨일즈, 스코틀랜드, 북아일랜드, 세 지역별로 종단 연구 마이크로데이터를 구축했다. 이 마이크로데이터는 개인과 직계가족에 대해 수십 년간의 센서스와 행정자료를 연계하여 만든 것으로 잠재적 활용 가치가 매우 높으나, 개별정보 노출위험이 높으므로 승인된 연구자와 자료 담당 직원들만 물리적으로 제한된 공간에서 자료를 분석하도록 허용하고 있다. 이에 노출위험을 제한하여 좀 더 폭넓은 이용자들이, 자료를 좀 더 많이 사용할 수 있도록 제공하는 것을 목표로, 이 자료에 대한 재현자료를 만드는 프로젝트(synthetic data estimation for UK longitudinal studies, SYLLS)가 수행되었다.

생성된 재현자료는 이용자의 개인 컴퓨터에서 분석될 수 있고, 이용자가 재현자료를 사용하여 만든 최종 모형을 구현한 코드를 담당 직원에게 보내면 담당 직원은 이용자의 코드로 원래 자료를 대신 분석해 줄 수 있다. 이렇게 재현자료를 활용하는 경우, 이용자는 자료 분석을 위해 데이터센터를 방문하지 않을 수 있다. 이러한 방식의 자료제공을 위해서는, 재현자료를 만드는 담당자들이 각 이용자에게 적합하도록 재현자료를 쉽게 생성하는 것이 요구된다. 이를 위해 R패키지 synthpop이 만들어졌다. 이 절에서는 재현자료를 생성하기 위한 최근 방법론들을 비모수적 기법들 위주로 간단히 소개하고, R패키지 synthpop을 사용하여 재현자료를 생성하는 과정을 소개한다.

참고로 재현자료를 생성하는데 사용할 수 있는 다른 R패키지로 simPopulation(Alfons 외,

2011; 2013) 및 simPop(Meindl 외, 2016)이 있다. 이 R패키지들은 가구 조사 자료에 대한 재현자료를 모수적으로 생성하도록 구현되어 있으나, 다중대체보다는 마이크로 시뮬레이션과 유사한 방식으로 알려져 있다(Nowok 외, 2016).

1. 재현자료 생성 방법론의 발전

앞의 4절과 5절에서 설명한 것처럼 재현자료 생성 아이디어가 Rubin(1993) 및 Little(1993)에서 제안된 이후, 재현자료 생성의 이론적 타당성과 최종 통계량을 만들기 위한 결합규칙이 Raghunathan 외(2003), Reiter(2003) 및 Reiter과 Raghunathan(2007) 등에서 연구되었다. 또한, 제안된 방법들을 활용해 실제 자료의 재현자료를 작성해 본 다양한 연구 사례들도 존재한다. 미국 CPS(Current Popultation Survey), LBD(Longitudinal Business Data), SIPP(Survey of Income and Progame Participation)에 대해 연구한 사례들과(Reiter, 2005; Drechsler와 Reiter, 2010; Abowd와 Woodcock, 2004; Kinney 외, 2011; Abowd 외, 2006), 독일 IAB establishment 패널 자료를 대상으로 재현자료 작성을 연구한 사례(Drechsler 외, 2008a,b) 등이 있다.

한편 베이지안 통계학의 역사를 생각할 때, 침체기를 겪었던 베이지안 방법론이 활기를 찾게된 것은 사후분포를 수식으로 계산하지 않고 자료를 생성하여 사후분포를 근사할 수 있게 되면서였다. 이와 유사하게 재현자료 생성 방법론의 발전을 살펴보면, 초기에는 사후예측분포를 계산하기 위해 사후분포를 수식으로 유도하고, 전체 변수를 결합모형 방식으로 생성하는 등 비효율적인 측면이 많았다. 방법론이 발전하면서 순차회귀 다중대체 방식으로 한 변수씩 자료를 대체는 방법도 제안되고, 자료를 생성할 때 베이지안 붓스트랩을 사용하는 방식도 제안되어 널리 활용되고 있다. 나아가 비모수적으로 CART (Classification and Regression Trees)를 사용하여 자료를 생성하는 방법이 제안(Reiter, 2005)되었고, 이후에 다른 비모수 기법의 활용으로 연구가 확장되어 가는 양상을 보인다. 최근에는 기계학습 분야의 알고리즘들인 랜덤 포레스트(random forests), 배깅(bagging), SVM(Support Vector Machine) 등을 사용하는 방식들도 발표되고 있다.

2. R패키지 synthpop 소개

R패키지 synthpop은 원 자료를 이용해 재현자료를 최대한 쉽게 생성할 수 있도록 구성되어 있다. 예를 들면 syn()이라는 하나의 함수를 사용하여, 주어진 자료에 대해 입력한 옵션에 따라 재현자료를 생성하도록 구현되어 있다. 원 자료는 R에서 data frame이나 matrix 형식이면 된다.

먼저 함수 syn()는 아무런 옵션을 주지 않으면 사후분포 추정과정을 거치지 않고 재현자료를 한 세트만 생성한다. 즉 가정한 모형의 모수를 관측된 자료를 이용해 추정하고, 추정된 모수와 가정한 모형을 이용해 원 자료와 같은 크기의 자료를 한 세트 생성해내며, 모수의 사후분포를 추정하지 않는다. 물론 같은 방식에서 옵션을 주어 자료를 m개 세트 생성할 수도 있다. 이러한 방식을 synthpop 패키지에서는 간편 생성(simple synthesis)이라고 한다. 한편 옵션을 주어 추정된 사후예측분포를 통해 자료를 m개 세트 생성할 수 있는데, 이를 정밀 생성(proper synthesis)이라고 한다.

좀 더 자세히 설명하자면 다음과 같다. 안전하다고 판단하여 값을 바꾸지 않고 그대로 제공하는 변수들을 X, 재현자료를 생성하여 대체하는 변수들을 Y라 하자. 또 Y_{obs}는 $f(Y|X_{obs},\theta)$에서 얻어진 표본이라고 가정한다. 이 때 Y의 결합분포는 앞 절들에서 설명한 대로 Y 각 변수들에 대한 일련의 조건부 분포들로 구성된다고 보고, 반복적으로 한 변수씩 재현자료를 생성해 나간다. 바로 이 조건부 분포에서 원 자료를 이용해 모수를 추정하고 바로 자료를 생성하면 간편 생성 절차가 되고, 사후예측분포를 추정해 자료를 생성하면 정밀 생성 절차가 된다.

재현자료 생성에 활용할 수 있는 기법은 여러 가지가 있다. R패키지 synthpop은 앞에서 설명한 최근 기법들의 많은 부분을 구현하고 있다. 다음 표는 이러한 기법들을 구현한 명령어들을 보여준다. 아래의 표에서 *로 표시된 부분은 모수적 방식을 선택하는 경우 변수 유형에 따른 디폴트 기법들이다.

Method option		Description	Data type
Non-parametric	ctree, cart	Classification and regression trees	Any
	surv.ctree	Classification and regression trees	Duration
Parametric	norm	Normal linear regression	Numeric
	normrank*	Normal linear regression preserving the marginal distribution	Numeric
	logreg*	Logistic regression	Binary
	polyreg*	Polytomous logistic regression	Factor, > 2 levels
	polr*	Ordered polytomous logistic regression	Ordered factor, > 2levels
	pmm	Predictive mean matching	Numeric
Other	sample	Random sample from the observed data	Any
	passive	function of other synthesised data	Any

재현자료를 쉽게 생성해보기 위해서 패키지가 제공하고 있는 데이터인 SD2011을 이용하자. 이 데이터 세트는 35개 변수, 5000개의 개체로 구성되어 있다. 변수들은 다양한 유형으로 구성되어 있으며 여기서 재현자료 생성을 설명하는데 이용할 주요 변수들은 다음 표와 같다.

Variable name	Description	Data type
sex	Sex	Binary
age	Age	Numeric
edu	Highest educational qualification	Factor, > 2 levels
marital	Marital status	Factor, > 2 levels
income	Personal monthly net income	Numeric
ls	Overall life satisfaction	Factor, > 2 levels
wkabint	Plans to go abroad to work in the next two years	Factor, > 2 levels

이제 간편 생성 절차를 통해 재현자료를 생성해보자. 디폴트대로 원 자료(ods)를 이용해 재현자료(sds)를 생성하는 명령어와 결과는 아래 그림과 같다. 모수적 방법으로 재현자료를 생성하고 싶으면 syn(ods) 대신에 syn(ods, method = “parametric”)이라고 입력하면 된다. 그러면 변수 유형에 따라 앞의 표에 디폴트로 표시된 방법들을 적용하여 생성된 재현자료가 얻어진다. 재현자료 생성 명령어 syn()를 실행하면 명령어가 실행되는 과정 중에 재현되는 변수를 아래와 같이 순서대로 보여준다.

재현자료 생성

```
> library(synthpop)
> ods = SD2011[, c("sex", "age", "edu", "marital", "income", "ls", "wkabint")]
> sds = syn(ods)
syn  variables
1    sex age edu marital income ls wkabint
```

재현자료 생성 결과를 담은 sds는 list 형식이며 총 23개 성분을 가지고 있다. sds를 실행하면 sds의 대표적인 성분들을 보여준다. 각 성분들을 따로따로 실행해 볼 수도 있다. 예를 들면 sds$m에는 생성된 재현자료 세트 개수가, sds$syn에는 생성된 자료가 할당되어 있다. 특히 sds$method는 각 변수에 대해 재현자료를 생성하는데 사용한 모형이 무엇인지를 보여준다.

실행 결과

```
> sds
Call:
($call) syn(data = ods)

Number of synthesised data sets:
($m)  1

First rows of synthesised data set:
($syn)
     sex age                  edu            marital income               ls wkabint
1   MALE  63            SECONDARY            MARRIED   2650 MOSTLY SATISFIED       NO
2 FEMALE  51            SECONDARY            MARRIED   1800           PLEASED       NO
3 FEMALE  49            SECONDARY            MARRIED     NA           PLEASED       NO
4   MALE  37 PRIMARY/NO EDUCATION              SINGLE   2000 MOSTLY SATISFIED       NO
5 FEMALE  82            SECONDARY            WIDOWED   1200 MOSTLY SATISFIED       NO
6 FEMALE  61            SECONDARY DE FACTO SEPARATED    700            MIXED       NO
...

Synthesising methods:
($method)
     sex      age      edu  marital   income       ls  wkabint
"sample"   "cart"   "cart"   "cart"   "cart"   "cart"   "cart"

Order of synthesis:
($visit.sequence)
    sex     age     edu marital  income      ls wkabint
      1       2       3       4       5       6       7

Matrix of predictors:
($predictor.matrix)
        sex age edu marital income ls wkabint
sex       0   0   0       0      0  0       0
age       1   0   0       0      0  0       0
edu       1   1   0       0      0  0       0
marital   1   1   1       0      0  0       0
income    1   1   1       1      0  0       0
ls        1   1   1       1      1  0       0
wkabint   1   1   1       1      1  1       0
```

한편, 15세 이하 혼인상태는 기혼이 나타나지 않도록 하는 등, 재현자료를 생성할 때 각 변수마다 고려할 사항이 있을 수 있다. 이처럼 각 변수마다 조건을 설정하거나, 자료를 생성하는 순서를 지정하거나, 자료를 생성하는 모형을 정해주는 등 다양한 옵션들의 실행이 가능하다. 자세한 내용은 Nowok 외(2016)을 참고하기 바란다.

제7절 맺음말

응답 개체들의 프라이버시는 보호하면서 마이크로데이터를 불특정 다수에게 이용하기 편리하도록 제공할 수 있다면, 통계기관은 본연의 임무 중 하나를 매우 성공적으로 수행하는 것으로 평가받을 수 있다. 하지만 비밀보호와 자료 유용성 확보를 위해 성공적인 통계적 노출제어 방안을 마련하는 것은 학문적으로도 무척 어려운 일로, 각국의 다양한 분야의 연구자들이 이 문제를 해결하기 위해 노력하고 있다. 본 보고서에서는 그런 노력의 하나인 재현자료 생성을 위한 배경 및 방법론을 간단히 소개했다.

재현자료 생성의 본래 목적은 원 자료와 완전히 다르지만 원 자료의 분포를 보존하는 자료를, 원 자료 대신에 제공하는 것이었다. 그러면 노출위험이 0인 상태가 보장되므로 더 이상 노출위험은 고려하지 않을 수 있다. 이에 따라 정보손실을 줄이는데 집중하여 모집단 분포를 잘 추정하는 모형을 개발하려는 노력이 이루어졌다. 하지만 현실적으로 실제 자료에 대해 그러한 모형을 구축하여 재현자료를 성공적으로 생성해내기가 쉽지 않았다. 즉, 정보손실이 너무 큰 재현자료가 생성되는 문제가 있었다. 따라서 일부 노출위험이 높은 자료만을 재현자료로 바꾸어 정보손실을 줄이는 연구가 지속되어 오고 있다. 다만 이렇게 부분 재현자료를 활용하는 경우, 매스킹 처리에서 발생했던 노출위험과 정보손실 상충(trade-off) 문제가 여전히 존재한다. 또한 부분 재현자료에서도 정보손실은 0이 아니기 때문에 아무리 원 자료와 유사하다 해도 이용자가 만족하지 않는 문제가 있을 수 있다. 나아가 베이지안 다중대체 기법을 사용하기 때문에 여러 세트의 자료를 제공받는 형식에 대해서도 이용자에게 심리적 거부감 등의 불편함이 있을 수 있다.

재현자료 생성의 목적이 원래 마이크로데이터를 완전히 대신하는 공공이용파일을 배포하는 것이라면 보다 많은 연구가 필요하며, 만족스러운 활용 결과를 얻는 것이 현재로서는 어려울 수 있다. 하지만 캐나다 통계청의 경우처럼 데이터센터를 방문하기 전에 이용자가 재현자료를 받아 사전 분석용으로 활용하고, 센터 방문 시간을 줄일 수 있다면 데이터센터 운영에 큰 도움을 줄 수 있다. 또한 영국의 경우처럼 재현자료를 이용해 자료를 분석한 코드를 데이터센터 직원에게 보내고 결과를 수령한다면 센터 방문 자체를 하지 않고 자료를 분석하는 효과를 도모할 수도 있다. (참고로 이때 분석 결과물에 대한 통계적 노출제어 처리를 수행하는 것은 별도의 문제이다.) 한편 미국의 경우 재현자료를 분석하는 코드를 담당 직원에게 보내면 직원은 원 자료 분석결과와 재현자료 분석결과를 비교하여, 원자료 분석결과 대신에 이 비교 결과를 이용자에게 제공한다. 이처럼 각 나라별로, 상황별로 다양한 방식으로 재현자료가 활용되고 있다.

본 보고서에서는 재현자료 생성을 이해하기 위해 재현자료 개념, 필요한 배경지식, 기본적인 생성 방법론을 소개했다. 나아가 간단히 재현자료를 생성해 볼 수 있는 R패키지도

소개하였다. 그러나 실제로 재현자료를 활용하는 전략을 수립하기 위해서는 해외 사례의 심층 검토, 재현자료의 노출위험 측정에 관한 이론, 최신 재현자료 생성 방법론에 대한 이해 및 여러 방법론의 비교 검토가 필요하다. 매스킹의 대안으로 거론되고 있는 재현자료와 차등정보보호를 비교 · 검토하는 연구도 필요해 보인다. 모쪼록 자료제공 관계자들이 재현자료를 이해하는데, 나아가 효율적인 자료제공을 위한 연구가 순조롭게 진행되는데 이번 기초연구가 도움이 되기를 바란다.

참고문헌

박민정(2014), 「가계금융 · 복지조사 마이크로데이터 매스킹 방안 평가」, 통계개발원.

박민정(2015), 「마이크로데이터 매스킹 실무 활용 안내서」, 통계개발원.

박민정(2016), 「통계적 노출제어 최신동향 검토 - 차등정보보호를 중심으로-」, 통계개발원.

박민정 · 김항준(2016), 「마이크로데이터 공표를 위한 통계적 노출제어 방법론 고찰」, 응용통계연구, 39(6), pp. 1041-1059.

송주원, 안형진(2009), 『무응답 자료처리 및 분석』, 통계교육원.

오만숙(2012), 『R 몬테칼로와 함께하는 베이지안 통계추론』, 자유아카데미.

이용희(2013), 「정보공개 환경에서 개인정보 보호와 노출 위험의 측정에 대한 통계적 방법」, 한국데이터정보과학회지, 24(5), pp. 1029-1041.

Abowd, J.M., Woodcock, S.D.(2004), "Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data", In J Domingo-Ferrer, V Torra (eds.), Privacy in Statistical Databases, 2004, pp. 290-297, Springer-Verlag.

Drechsler, J. and Reiter, J.P.(2010), "Sampling with Synthesis: A New Approach for Releasing Public Use Census Microdata." Journal of the American Statistical Association, 105(492), pp. 1347-1357.

Drechsler, J.(2011), Synthetic Datasets for Statistical Disclosure Control, Springer.

Krenzke, T., Gentleman, J.F., Li, Jianzhu and Moriarity, C.(2013), "Addressing Disclosure Concerns and Analysis Demands in a Real-Time Online Analytic System", Journal of Official Statistics, 29, pp. 99-124.

Little, R.J.A.(1993), "Statistical Analysis of Masked Data", Journal of Official Statistics, 9(2), pp.407-426.

Nowok, B., Raab, G.M. and Dibben, C.(2016). "synthpop: Bespoke Creation of Synthetic Data in R", Journal of Statistical Software, 74(11), pp. 1-26.

Raghunathan, T.E., Lepkowski, J.M., Hoewyk, J.V. and Solenberger, P.(2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models", Survey Methodology, 27(1), pp. 85-95.

Raghunathan, T.E., Reiter, J.P. and Rubin, D.B.(2003), "Multiple Imputation for Statistical Disclosure Limtation", Journal of Official Statistics, 19, pp. 1-16.

Reiter, J.P.(2003), "Inference for Partially Synthetic, Public Use Microdata Sets", Survey methodology, 29(2), pp. 181-188.

Reiter, J.P.(2004), "Simultaneous use of multiple imputation for missing data and disclosure limitation", Statistics Canada, 30(2), pp. 235-242.

Reiter, J.P.(2005), "Releasing Multiply Imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study", Journal of the Royal Statistical Society A, 168(1), pp. 185-205.

Reiter, J.P. and Raghunathan, E.(2007), "The Multiple Adaptations of Multiple Imputation", Journal of the American Statistical Society, 102(480), pp. 1462-1471.

Rubin, D.B.(1987), The Bayesian Bootstrap, Annals of Statistics, 9, pp. 130-134.

Rubin D.B.(1993), "Discussion: Statistical Disclosure Limitation", Journal of Official Statistics, 9(2), pp. 461-468.

<부 록>

1. 확률의 기본개념

베이지안 통계의 핵심은 사후분포이며 이는 자료가 관찰된 이후에, 즉 자료가 주어졌을 때 관심 모수가 가지게 되는 조건부 확률을 의미한다. 따라서 베이지안 통계에서는 조건부 확률에 대한 이해가 필수적이다. 통계학의 기본 지식을 되짚어 보고싶은 독자를 위해 본 절에서는 베이지안 통계의 기본 틀을 제공해주는 확률, 확률분포 및 확률변수와 같은 주요 개념들의 정의와 성질을 살펴본다.

가. 확률

확률은 불확실성을 나타내는 측도로 그 불확실성의 정도를 0과 1 사이의 값으로 수량화한 것이다. 통계적 실험에서 모든 가능한 실험결과들의 집합을 표본공간(sample space, S)이라고 하며, 발생 가능한 결과를 사건(event)이라고 한다. 사건 A의 확률 P(A)는 사건 A가 발생할 가능성을 측정하는 측도로 다음 세 가지 조건을 만족해야 한다.

(1) 표본공간 S에서 임의의 사건 A에 대하여 $0 \le P(A) \le 1$이다.

(2) $P(S) = 1$이다.

(3) 서로 배반인 사건 $A_1, A_2, \cdots$에 대하여 다음을 만족한다.

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$$

예를 들어, 동전을 던졌을 때 윗면을 관찰 하는 실험을 한다고 하자. 동전을 던진 결과로는 “앞면” 또는 “뒷면”만 가능하므로 표본 공간은 다음과 같다.

$$S = \{\text{앞면}, \text{뒷면}\}$$

앞면이 관찰되면 뒷면이 관찰될 수 없고, 뒷면이 관찰되면 앞면이 관찰될 수 없으므로, 앞면이 관찰되는 것과 뒷면이 관찰되는 것은 서로 배반인 사건이다. 이때, 앞면이 나올 확률을 p_1, 뒷면이 나올 확률을 p_2라 하자. 동전을 한 번 던졌을 때, 앞면 또는 뒷면 이외에 다른 경우가 발생할 수 없으므로 $p_1 + p_2 = 1$이다.

나. 조건부 확률

두 사건 A, B에 대하여 $P(B) > 0$일 때, 사건 B가 참인 조건 하에서 사건 A가 일어날 조건부 확률 $P(A|B)$은 다음과 같이 정의한다.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

조건부 확률은 표본공간이 S에서 B로 바뀌었다는 것을 의미하며, 조건부 확률 역시 앞에서 살펴본 확률의 성질을 만족한다.

예를 들어, 주사위를 던져서 제일 윗면에 나오는 숫자를 관찰하는 실험을 한다고 하자. 가장 윗면에 2가 위치한 경우 사건 A가 발생했다고 하자. 또한, 주사위를 던졌을 때 짝수 $(2,4,6)$가 가장 윗면에 위치한 경우 사건 B가 발생했다고 하자. 만약, 모든 면이 윗면에 위치할 가능성이 동일하다고 가정하면, 사건 A가 발생할 확률과 사건 B가 발생할 사건의 확률은 각각 다음과 같다.

$$P(A) = \frac{1}{6},\ P(B) = \frac{3}{6} = \frac{1}{2}$$

다음으로, 사건 B가 발생했다는 조건하에, 사건 A 역시 발생할 조건부 확률 $P(A|B)$를 살펴보자. 이 확률은 주사위를 던져서 가장 윗면에 관찰된 숫자가 짝수 $(2,4,6)$인데, 그 중에서도 2일 확률을 의미하고, 다음과 같이 계산된다.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{3/6} = \frac{1}{3}$$

반대로, 사건 A가 발생했을 때, 사건 B도 발생할 조건부 확률 $P(B|A)$는 다음과 같다.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{1/6}{1/6} = 1$$

다. 베이즈 정리

베이즈 정리는 확률의 성질로서, 조건부 확률에서의 사건의 순서를 바꿀 수 있게 해주는 유용한 공식이다. k개의 사건들의 집합 $\{A_1, A_2, \cdots, A_k\}$가 다음 두 성질을 만족하면, $\{A_1, A_2, \cdots, A_k\}$를 표본공간 S의 분할(partition)이라 부른다.

(1) $A_1, A_2, \cdots, A_k$ 가 서로 배반이다. 즉, 임의의 $i,\ j,\ i \neq j$에 대하여 $A_i \cap A_j = \varnothing$ 이다.
(2) $P(\bigcup_{i=1}^{k} A_i) = 1$이다.

$\{A_1, A_2, \cdots, A_k\}$가 표본공간 S의 분할일 때 임의의 사건 E에 대하여 다음과 같은 베이즈 정리를 유도할 수 있다.

$$P(A_i|E) = \frac{P(E|A_i)\,P(A_i)}{P(E)} = \frac{P(E|A_i)\,P(A_i)}{\sum_{j=1}^{k} P(E|A_j)\,P(A_j)}$$

예를 들어 주사위 던지기 실험을 살펴보자. 앞에서 살펴본 바와 같이 사건 A는 2가 가장 윗면에 오는 경우, 사건 B는 짝수가 가장 윗면에 오는 경우라 하자. 앞에서 $P(A|B) = 1/3$, $P(B|A) = 1$이 되는 것을 확인하였다. 베이즈 정리를 사용하여 $P(B|A) = 1$임을 확인해보겠다. 단, B의 여사건을 B^c라 하고, 이는 가장 윗면의 숫자가 홀수인 경우에 해당한다.

$$\begin{aligned} P(B|A) &= \frac{P(A|B)\,P(B)}{P(A)} \\ &= \frac{P(A|B)\,P(B)}{P(A|B)\,P(B) + P(A|B^c)\,P(B^c)} \\ &= \frac{1/3 \times 3/6}{1/3 \times 3/6 + 0 \times 3/6} = 1 \end{aligned}$$

만약 사건 A가 주어졌을 때 사건 B의 조건부 확률 $P(B|A)$를 구하고 싶지만 현실적으로 반대 순서의 조건부 확률 $P(A|B)$를 구하기가 더 쉽다면, 베이즈 정리를 사용하여 $P(B|A)$를 구할 수 있게 된다. 베이즈 정리는 베이지안 통계에서 사후분포를 유도하는 바탕이 되는 공식이다.

라. 확률변수와 확률분포

표본공간의 원소를 실수 값으로 변환하는 함수를 확률변수(random variable, 또는 변수)라 한다. 확률변수 X가 이산(discrete)인지 연속(continuous)인지에 따라 다음과 같이 두 종류로 나눌 수 있다.

(1) 이산 확률변수: 확률변수 X의 가능한 값들 $\chi = \{x_1, x_2, \cdots\}$이 셀 수 있는 변수

(2) 연속 확률변수: 확률변수 X가 어떤 구간에 속하는 임의의 값을 취할 수 있는 변수

예를 들어, 동전던지기 실험에서 확률변수 X는 다음과 같이 정의될 수 있다. 이 때, 확률변수 X의 값으로 가능한 값들은 $\{0, 1\}$의 두 가지이므로 X는 이산 확률변수이다.

$$X = \begin{cases} 1: & \text{앞면이 관찰된 경우} \\ 0: & \text{뒷면이 관찰된 경우} \end{cases}$$

확률분포란 확률변수 X의 값에 따라 확률이 어떻게 흩어져 있는지를 합이 1인 양수로써 나타낸 것이다. 동전 던지기 실험에서 앞면이 나올 확률과 뒷면이 나올 확률이 같다고 가정하자. 이때, $X=0$일 확률은 0.5, $X=1$일 확률은 0.5이다.

마. 확률밀도함수

확률변수의 종류에 따라 다음과 같이 확률밀도함수를 정의한다.

이산 확률변수 경우

이산 확률변수 X의 임의의 값 $x\in\chi$에 대해, $f(x)=P(X=x)$인 함수 $f(x)$를 X의 확률밀도함수라 한다. $f(x)$의 성질은 다음과 같다.

(1) 모든 x에 대해서 $0\le f(x)\le 1$이다.

(2) $\sum_{x\in\chi} f(x)=1$을 만족한다.

(3) 서로 겹치지 않는 임의의 부분집합 A와 B에 대하여 다음이 성립한다.

$$P(X\in A\cup B)=P(X\in A)+P(X\in B)=\sum_{x\in A} f(x)+\sum_{x\in B} f(x)$$

연속형 확률변수 경우

연속 확률변수 X에 대해 다음을 만족시키는 함수 $f(x)$를 X의 확률밀도함수라 한다.

(1) 모든 $x\in\chi$에 대하여 $f(x)\ge 0$이다.

(2) $\int_{x\in\chi} f(x)\,dy=1$

(3) 서로 배반인 임의의 부분집합 A와 B에 대하여 다음이 성립한다.

$$P(X\in A\cup B)=P(X\in A)+P(X\in B)=\int_{x\in A} f(x)\,dx+\int_{x\in B} f(x)\,dx$$

앞에서 살펴본 동전 던지기 실험에서 앞면이 나올 확률, 즉 $X=1$일 확률을 θ, $X=0$이 될 확률을 $1-\theta$라 하자. 이때 확률변수 X의 확률밀도함수는 다음과 같다.

$$f(x|\theta)=\theta^x(1-\theta)^{1-x},\ x=0,1$$

바. 결합확률분포

두 개의 확률변수 X와 Y에 대하여, 이들이 취할 수 있는 모든 값들의 순서 짝에 확률이 어떻게 흩어져 있는지를 합이 1인 양수로 나타낸 것을 확률변수 X와 Y의 결합확률분포 또는 결합분포라고 한다. 결합확률은 결합밀도함수로부터 구해지는데 결합밀도함수는 다음의 성질을 가진다. 결합확률분포와 결합밀도함수의 성질은 임의의 유한개의 확률변수 $X_1, X_2, \cdots, X_n$에 대해서도 마찬가지로 성립한다.

(1) $0 \le f(x,y) \le 1,\ \int_{\text{모든 } x}\int_{\text{모든 } y} f(x,y)\,dx\,dy = 1$

(2) $P(a < X \le b,\ c < Y \le d) = \int_{a < x \le b}\ \int_{c < y \le d} f(x,y)\,dx\,dy$

예를 들어, 확률변수 X는 동전을 한 번 던졌을 때 앞면이 나오면 1, 뒷면이 나오면 0의 값을 갖고, 앞면이 나올 확률을 θ, X의 확률밀도함수를 f_X라 하자. 반면, 확률변수 Y는 주사위를 던졌을 때 윗면에서 관찰되는 숫자라고 하자. 이때, Y의 값으로 가능한 값들은 $1, 2, 3, 4, 5, 6$이 된다. 모든 면이 나올 가능성이 동일하다고 가정하면, Y의 확률밀도함수 f_Y와 확률변수 X와 Y의 결합확률밀도함수는 다음과 같다.

$$f_Y(y) = \frac{1}{6},\ y = 1, \cdots, 6$$
$$f_{XY}(x,y) = \theta^x (1-\theta)^{1-x} \times \frac{1}{6}$$

2. 사전분포

가. 정보 사전분포 vs. 무정보 사전분포

추정하고자 하는 모수에 대한 사전정보나 이론적 지식이 있는 경우에 정보와 지식을 반영하는 사전분포를 주관적 사전분포(subjective prior) 또는 정보 사전분포(informative prior)라 한다. 이 경우 사전분포가 따라 분석결과에 미치는 영향에 대한 강건성 확인(robustness check)이 필요하다.

반면, ① 사전정보가 거의 없거나, ② 조금 있더라도 사용하기에는 유의하지 못하거나 부적절하다고 생각되거나, 또는 ③ 베이지안 분석결과와 고전적 분석결과를 비교하기 위하여 사전정보를 사용하지 않는 경우가 있다. 이럴 때는 객관적 사전분포(objective prior) 또는 무정보 사전분포(non-informative prior)를 사용한다. 무정보 사전분포의 간단한 예는 모수의 가능한 값에 대해 모두 같은 확률을 부여하는 균일분포(uniform distribution)를 들 수 있다.

균일분포 $U(a,b)$

확률변수 X가 균일분포 $U(a,b)$를 따른다고 할 때, 확률밀도함수는 다음과 같다.

$$f(x|a,b)=\frac{1}{b-a},\ a\le x\le b$$

X의 평균 및 분산은 다음과 같다.

$$E(X)=\frac{a+b}{2},\ \ Var(X)=\frac{(b-a)^2}{12}$$

예를 들어, 동전 던지기 실험에서 앞면이 관찰될 확률 θ가 관심모수라고 하자. 이때, θ는 0과 1사이의 값이 되므로 θ의 사전분포로 베타분포 $Beta(a,b)$가 사용될 수 있다. θ의 값에 대한 정보가 없이 0과 1사이의 값일 확률이 동일하다고 판단되면 베타분포 $Beta(1,1)$을 사용하는 것이 적절하고, 이는 균일분포로 무정보 사전분포를 사용하는 것이 된다. θ의 값이 0에 가깝다는 정보가 있는 경우 아래 그림의 베타분포 $Beta(5,1)$보다는 베타분포 $Beta(1,3)$을 사전분포로 사용하는 것이 적절하며, 이 경우 정보 사전분포를 사용한 것이 된다.

베타분포 $Beta(a,b)$

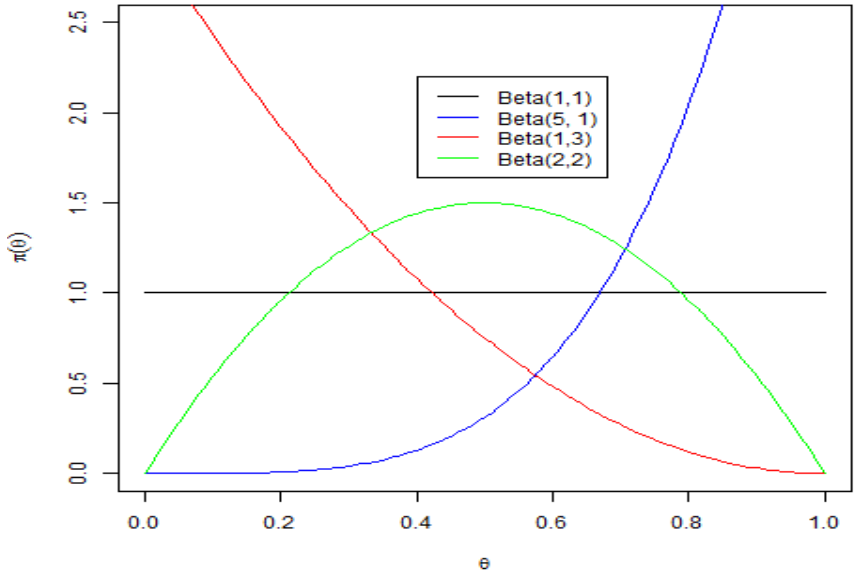
확률변수 X가 베타분포 $Beta(a,b)$를 따른다고 할 때, 확률밀도함수는 다음과 같다.

$$f(x|a,b)=\frac{\Gamma(a)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1},\ 0\le x\le 1.$$

단, $\Gamma(x)=\int_0^\infty z^{x-1}e^{-z}dz$ 이다.

X의 평균 및 분산은 다음과 같다.

$$E(X)=\frac{a}{a+b},\ \ Var(X)=\frac{ab}{(a+b)^2(a+b+1)}$$



[부그림 3-1] 베타분포 $Beta(a,b)$의 확률밀도함수

나. 적합 사전분포 vs. 부적합 사전분포

사전분포의 밀도함수가 모수공간에서 유한한 적분값을 갖는 경우에는 적합 사전분포(proper prior)라 부르고, 무한한 적분값을 갖는 경우를 부적합 사전분포(improper prior)라고 한다.

예를 들어 균일분포가 사전분포인 경우, 모수공간이 유한인 경우 적합 사전분포가 되지만, 모수공간이 무한인 경우 그 적분값이 무한대로 부적합 사전분포가 된다. 사전분포가 적합인 경우 사후 분포도 적합성을 지니게 된다. 하지만 사전분포가 부적합인 경우 사후분포의 적합성이 보장되지 않으므로, 사후분포의 적합성이 증명되는 부적합 사전분포만 사용가능하다.

다. 공액사전분포

사전분포와 사후분포가 같은 분포 형태를 갖게 하는 사전분포를 공액사전분포(invariant prior)라 한다. 이는 새로운 자료가 추가될 때마다 기존의 사후분포를 쉽게 개선(update)할 수 있는 장점이 있다. 예를 들어, 관측값이 이항분포를 따른다고 할 때, 사전분포로 베타분포를 선택할 경우 사후분포 역시 베타분포에 속하게 되어 사전분포와 사후분포가 같은 분포집합에 속하게 되므로 베타 사전분포는 이항확률분포에 대한 공액사전분포가 된다.

예를 들어, 확률변수 X가 이항분포를 따른다고 하고 이는 $X \sim Bin(n_1, \theta)$로 표현된다. 이때, X의 관찰값은 성공 확률이 θ인 베르누이 시행을 n_1번 실행했을 때 관찰된 성공의 횟수이다. 만약, 모수 θ에 대한 사전분포로 베타분포 $Beta(a, b)$를 가정하고, X의 관찰값이 x_1라고 가정하자. 이때, θ의 사후분포 $\pi(\theta|x_1)$는 베타분포 $Beta(x_1 + a,\ n_1 - x_1 + b)$임이 알려져 있다. 사전분포가 베타분포이고, 모수는 달라졌지만 사후분포 역시 베타분포가 됨을 알 수 있다. 따라서, 베타분포는 이항분포에 대한 공액사전분포가 된다.

만약, 동일한 성공확률 θ를 가정하고, n_2회의 추가적인 베르누이 실행으로부터 x_2회의 성공이 관찰되었다고 하자. 기존의 관찰값과 새로운 관찰값을 종합하면, 성공확률이 θ인 베르누이 시행을 총 $n^* = (n_1 + n_2)$번 실행할 결과 $x^* = (x_1 + x_2)$회의 성공을 관찰한 것이다. 이때 추가된 정보를 반영한 θ의 사후분포 $\pi(\theta|x^*)$는 베타분포 $Beta(x^* + a,\ n + n^* - x^* + b)$가 된다. 이와 같이 공액 사전분포는 새로운 자료가 추가될 때 마다 기존의 사후분포를 쉽게 개선(update)할 수 있는 장점이 있다.

이항분포 $B(n, \theta)$

확률변수 X가 이항분포 $B(n, \theta)$를 따른다고 할 때, 확률밀도함수는 다음과 같다.

$$f(x|\theta) = \binom{n}{x}\theta^x (1-\theta)^{(n-x)},\ x = 0, 1, \cdots, n$$

X의 평균 및 분산은 다음과 같다.

$$E(X) = n\theta,\ Var(X) = n\theta(1-\theta)$$

라. 변환불변 사전분포

확률변수 X의 밀도함수 $f(x|\theta)$가 어떤 함수 h에 대하여 $f(x|\theta) = h(x - \theta)$가 성립할 때 θ를 위치모수(location parameter)라 한다. 예를 들어, X가 평균이 θ이고 분산이 1인 정규분포 $N(\theta, 1)$를 따른다고 할 때, 밀도함수는 다음과 같다.

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$$

이 때 θ는 위치모수가 된다. 모수가 위치모수인 경우에 위치변환에 대하여 불변성을 갖는 사전밀도함수는 위치변환에 대한 변환불변 사전분포가 된다.

다음으로 확률변수 X의 밀도함수 $f(x|\theta)$가 어떤 함수 h에 대하여 $f(x|\theta)=\frac{1}{\theta}h(\frac{x}{\theta})$가 성립할 때 θ를 척도모수(scale parameter)라 한다. 예를 들어, X가 평균이 0이고 표준편차가 θ인 정규분포 $N(0,\theta^2)$를 따른다고 할 때, 밀도함수는 다음과 같다.

$$f(x|\theta)\propto\frac{1}{\theta}e^{-\frac{1}{2}\left(\frac{x}{\theta}\right)^2}$$

따라서 θ는 척도모수이다. 척도변환에 대하여 불변성을 갖는 사전밀도함수는 척도변환에 대한 변환불변 사전분포가 된다.

마. 제프리 사전분포

제프리 사전분포는 피셔 정보행렬(Fisher's information matrix)을 바탕으로 하기 때문에, 위치 혹은 척도 모수가 아닌 임의의 모수에 대하여 사용할 수 있고, 임의의 변환에 대하여 불변성을 지닌다는 장점이 있다. 제프리 사전분포에 대해 소개하면 다음과 같다.

하나의 확률변수 X의 확률밀도함수를 $f(x|\theta)$라 할 때 피셔 정보행렬은 다음과 같다.

$$I_1(\theta)=E^{X|\theta}\left[\left(\frac{d\log f(X|\theta)}{d\theta}\right)^2\right]=-E^{X|\theta}\left[\frac{d^2\log f(X|\theta)}{d\theta^2}\right]$$

만약 자료변수 $X_1, X_2, \cdots, X_n$이 서로 독립이고 같은 확률밀도함수 $f(x|\theta)$를 가지면 이때 피셔 정보행렬은 $I_n(\theta)=nI_1(\theta)$가 된다. 확률변수 $X_1, X_2, \cdots, X_n$의 관찰값 $x_1, x_2, \cdots, x_n$이 주어졌을 때, 모수 θ에 대한 무정보 사전밀도 함수 $\pi_1(\theta)=\sqrt{I_n(\theta)}$를 제프리 사전함수라 한다. 이 사전함수는 모수 θ에 대한 임의의 변환에 대해 불변성을 지닌다.

제프리 사전밀도함수는 위치 혹은 척도 모수가 아닌 임의의 모수에 대해 사용할 수 있고, 임의의 변환에 대하여 불변성을 지니므로 모수의 선택에 대하여 고민할 필요가 없다는 장점이 있어 널리 쓰이고 있다. 모수가 $\theta=(\theta_1, \theta_2, \cdots, \theta_p)$와 같이 하나가 아니라 벡터로 주어질 경우 다변량모수 θ에 대한 피셔 정보행렬의 (i,j)번째$(i,j=1, \cdots, p)$ 요소는 다음과 같다.

$$I_{ij}(\theta)=-E^{X|\theta}\left[\frac{d^2\log f(X|\theta)}{d\theta_i d\theta_j}\right]$$

다변량 모수 $\theta=(\theta_1, \theta_2, \cdots, \theta_p)$에 대한 제프리 사전밀도함수는 $\pi(\theta)=\sqrt{\det(I(\theta))}$로 주어진다.

3. 사후예측분포

사후예측분포를 계산하는 예를 들어 보자. 앞면이 나올 확률이 θ인 동전 던지기 시행을 n번 반복하여 얻은 관측값을 $x = (x_1, x_2, \cdots, x_n)$라 하자. θ의 사전분포가 베타분포 $Beta(a, b)$일 때, $n+1$번째 시행에서의 관측값이 $X_{n+1} = 1$일 예측확률은 다음과 같다.

$$\begin{aligned} P(X_{n+1} = 1|x_1, \cdots, x_n) &= \int P(X_{n+1} = 1|\theta, x_1, \cdots, x_n)\,\pi(\theta|x_1, \cdots, x_n)\,d\theta \\ &= \int P(X_{n+1} = 1|\theta)\,\pi(\theta|x_1, \cdots, x_n)\,d\theta \\ &= \int \theta\,\pi(\theta|x_1, \cdots, x_n)\,d\theta \\ &= \frac{a + \sum_{i=1}^{n} x_i}{a + b + n} \end{aligned}$$

이 예에서 θ의 사후분포가 베타분포 $Beta(a + \sum_{i=1}^{n} x_i,\ n + b - \sum_{i=1}^{n} x_i)$됨은 <부록>의 2. 다 공액사전분포에 소개되어 있다.

4. 사후분포의 근사

본 절에서는 베이지안 사후분포에서의 몬테칼로 표본 생성 방법으로 많이 알려진 마코프 연쇄 몬테칼로(Markov Chain Monte Carlo, MCMC)방법에 필요한 개념과 계산 기법들을 살펴본다. MCMC에서는 표본이 독립적이지 않은 확률과정으로부터 생성된다. 이를 이해하기 위해 필요한 기본 개념들로 확률과정, 마코프 성질, 전이확률, 기약성, 재귀성, 비주기성, 에르고딕, 정상분포를 소개한다. 이를 바탕으로 마코프 연쇄 생성에 많이 사용되는 깁스 표본기법과 메트로폴리스-헤이스팅 알고리즘에 대해 살펴본다.

가. 기본개념

1) 확률과정

순차적으로(sequentially) 관측되는 확률변수 X의 값을 모아 놓은 집합 $\{X_n : n \geq 0\}$ $(= \{X_0, X_1, X_2, \cdots\})$을 확률과정(stochastic process)이라 한다. 이 때 확률변수 $X_n (n \geq 0)$의 관측치로서 가능한 값들의 집합을 상태공간(state space)라고 한다.

2) 마코프 성질

확률과정 $\{X_n : n \geq 0\}$에 대한 다음 조건을 마코프 성질(Markov property)이라 한다.

$$\Pr(X_{n+1} = i_{n+1} | X_0 = i_0, X_1 = i_1, \cdots, X_n = i_n) = \Pr(X_{n+1} = i_{n+1} | X_n = i_n)$$

마코프 성질을 만족하는 확률과정 $\{X_n : n \geq 0\}$을 마코프 연쇄(Markov chain)라 한다. 이는 현재의 상태는 직전의 상태에만 의존한다는 것을 의미한다.

3) 전이확률

상태 i에서 한 단계 후에 상태 j로 전이(transit)할 조건부 확률 $p_{ij} = \Pr(X_{n+1} = j | X_n = i)$를 전이확률(transition probability)이라 하고, 다음과 같은 성질을 갖는다.

(1) $p_{ij} \geq 0$ for all $i, j \geq 0$

(2) $\sum_{j=0}^{\infty} p_{ij} = 1$ for $i = 0, 1, 2, \cdots$

참고로 $p_{ij}^{(n)}$은 상태 i에서 출발하여 n단계 뒤에 상태 j에 도달할 확률을 나타낸다.

예시 1: 유한상태 확률보행

예를 들어 어떤 사람이 $1, 2, 3, 4, 5, 6$ 순서로 적힌 직선상에서 무의식적으로 걷는다고 하자. 이 사람의 확률보행을 다음과 같이 가정한다.

(1) 현재 내부의 값 $2, 3, 4$ 또는 5에 머물러 있을 경우, 다음 보행에서 현재에 머무르거나 혹은 인접한 값으로 이동할 확률은 동일하다. 만일 이동했다면, 오른쪽이나 왼쪽으로 이동할 확률 또한 동일하다.

(2) 현재 양 끝값 1 또는 6에 머물러 있을 경우, 다음 보행에서 머무르거나 혹은 인접한 값으로 이동할 확률 역시 동일하다.

이 예시에서 상태공간은 $S = \{1, 2, 3, 4, 5, 6\}$이며, 전이확률은 다음과 같다.

$$P = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 & 0 & 0 \\ 0.25 & 0.5 & 0.25 & 0 & 0 & 0 \\ 0 & 0.25 & 0.5 & 0.25 & 0 & 0 \\ 0 & 0 & 0.25 & 0.5 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.5 & 0.25 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 \end{pmatrix}$$

예를 들어, 상태 1에서 상태 1로 전이될 확률은 $p_{11} = 0.5$이고, 상태 2에서 상태 3으로 전이될 확률은 $p_{23} = 0.25$이다.

4) 기약성(irreducibility)

어떤 유한한 정수 $n(\geq 0)$에 대하여 $p_{ij}^{(n)} > 0$이면 상태 i에서 상태 j로 도달 가능하다고 한다. 이를 $(i \rightarrow j)$로 나타내자. 상태 i에서 상태 j로 도달 가능하고, 상태 j에서 상태 i로 도달 가능한 경우 상호도달 가능하다고 한다. 이를 $(i \leftrightarrow j)$라 하자. 마코프 연쇄에서 모든 상태가 서로 상호도달 가능한 경우, 이 마코프 연쇄를 기약적(irreducible)이라고 한다.

예시 1의 경우 $p_{12} > 0,\ p_{21} > 0,\ p_{23} > 0,\ p_{32} > 0$ 이므로 1과 2, 2와 3은 서로 상호도달 가능하다. 또한 1에서 2를 거쳐 3으로 도달 가능하고, 3에서 2를 거쳐 1로 도달 가능하므로(즉, $1 \rightarrow 2 \rightarrow 3 \Rightarrow 1 \rightarrow 3,\ 3 \rightarrow 2 \rightarrow 1 \Rightarrow 3 \rightarrow 1$) 상태 1과 상태 3도 상호도달 가능하다. 이와 같은 방법으로 모든 상태들이 서로 상호도달 가능함을 알 수 있다. 따라서 이 전이확률을 따르는 마코프 연쇄는 기약적이다.

5) 재귀성(recurrence)

상태 i에서 출발하는 마코프 연쇄는 다시 상태 i로 되돌아올 때마다 그 시점까지의 결과와는 독립적으로 상태 i에서 출발하는 원래의 마코프 연쇄와 같은 분포를 따르는데 이를 무기억성(memoryless property)이라고 한다. 상태 i가 상태 i로 무한번 되돌아올 수 있는 상태 i를 재귀적(recurrent)이라 한다. 모든 상태가 재귀적인 마코프 연쇄를 재귀적인 마코프 연쇄라고 한다.

6) 비주기성(aperiodicity)

주어진 상태 i에 대하여 $p_{ii}^{(n)} > 0$을 만족하는 자연수 n들의 최대 공약수를 주기라고 한다. 기약적(irreducible) 마코프 연쇄에서 모든 상태가 주기 1을 갖는 경우 비주기적(aperiodic) 마코프 연쇄라고 한다.

7) 에르고딕(ergodic)

마코프 연쇄가 재귀적이고 비주기적이면 에르고딕하다고 한다.

9) 정상분포(stationary distribution)

마코프 연쇄 $\{X_n : n \geq 0\}$가 모든 상태에 대하여 기약적이고 에르고딕이라고 하자. 또한 모든 상태 j에 대하여

$$\pi_j = \lim_{n \rightarrow \infty} p_{ij}^{(n)}$$

을 만족하는 π_j가 존재하고 이는 초기상태 i에 무관하다고 하자. 이러한 확률분포 $\Pi = \{\pi_j\}$를 정상분포(stationary distribution)라고 하며, 이는 아래 방정식의 유일한 양의 해이다.

$$\pi_j = \sum_{i=0}^{\infty} \pi_i p_{ij}, \ \sum_{j=0}^{\infty} \pi_j = 1$$

이때 π_j는 마코프 연쇄 $\{X_n : n \geq 0\}$가 상태 j에 머무는 시간의 (극한)비율을 의미한다.

예시 2: 유한상태 확률보행

예시1 에서 살펴본 유한상태 확률보행과 동일한 모형을 고려하자. 이때의 확률보행은 마코프 연쇄이고, 재귀적, 기약적, 비주기적이므로 정상분포 $\pi = \{\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6\}$가 존재하고 다음을 만족시킨다.

$$\begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \\ \pi_5 \\ \pi_6 \end{pmatrix} = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 & 0 & 0 \\ 0.25 & 0.5 & 0.25 & 0 & 0 & 0 \\ 0 & 0.25 & 0.5 & 0.25 & 0 & 0 \\ 0 & 0 & 0.25 & 0.5 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.5 & 0.25 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 \end{pmatrix} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \\ \pi_5 \\ \pi_6 \end{pmatrix}$$

방정식을 푼 결과, 해는 다음과 같다.

$$(\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6) = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$$

이는 유한보행을 무한히 시행하다보면, 각 상태에 머물 확률이 $1/6$으로 동일하다 뜻이다.

나. 마코프 연쇄 몬테칼로(Markov Chain Monte Caro, MCMC) 방법

예를 들어 $f(X)$를 확률밀도함수로 갖는 확률분포의 성질을 알고 싶지만, 이를 수식으로 유도하기는 어렵다고 하자. 만약 $f(X)$를 밀도함수로 갖는 독립표본 $X_1, X_2, \cdots, X_N$을 생성할 수 있다면, 이 독립표본을 바탕으로 표본평균, 표본분산, 표본의 분포 등을 추정할 수 있다. 또한 표본의 크기 N을 크게 하면 정확도를 우리가 원하는 수준까지 올릴 수 있다. 이처럼 원하는 분포로부터 독립표본을 생성한 뒤, 표본을 바탕으로 추론하는 것을 몬테칼로(Monte Carlo) 방법이라 한다. 이와 같은 독립표본이 아니라, $f(X)$를 정상분포로 갖는 마코프 연쇄 $\{X_n : n \geq 0\}$로부터 표본을 생성한 뒤에 몬테칼로 방법을 적용하는 것을 마코프 연쇄 몬테칼로(MCMC) 방법이라 한다.

베이지안 추론에서는 모수 θ의 사후분포 $\pi(x|\theta)$로부터 θ의 사후표본 $\theta_1, \cdots, \theta_N$을 추출하는 것이 주요 목적이 된다. 따라서 여기서는 MCMC 방법 중에서 많이 사용되는 깁스 표본기법과 메트로폴리스-헤이스팅스(Metropolis-Hastings) 알고리즘을 소개한다.

1) 깁스 표본기법

크기가 n인 표본 $x=(x_1, x_2, \cdots, x_n)$을 바탕으로 다변량 모수 $\theta=(\theta_1, \cdots, \theta_p)$의 사후분포 $\pi(\theta|x)=\pi(\theta_1, \cdots, \theta_p|x)$를 구한다고 가정하자. 일반적으로 모수의 개수 p가 증가함에 따라 결합분포인 $\pi(\theta|\mathrm{x})$의 계산이 복잡해지고, 계산이 되더라도 얻어진 사후분포를 표본생성에는 사용하기는 어려운 경우가 많다. 반면 같은 조건에서 각 모수의 완전 조건부 사후분포(full conditional posterior distribution)인 $\pi(\theta_i|x, \theta_{-i})$가 결합분포보다 더 편리한 형태로 주어지는 경우가 빈번히 발생한다. 단, θ_{-i}는 i번째 원소를 제외한 나머지 모든 원소를 나타내어 $\theta_{-i}=(\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \theta_p)$라 표현할 수 있다. 깁스 표본기법은 각 모수의 완전 조건부 사후분포를 바탕으로 차례대로 각 모수의 표본을 생성하는 방법이다.

예를 들어 $p=3$으로 $\theta=(\theta_1, \theta_2, \theta_3)$일 때 완전 조건부 사후분포가 다음과 같다고 하자.

$$\pi(\theta_1|x, \theta_2, \theta_3),\ \pi(\theta_2|x, \theta_1, \theta_3),\ \pi(\theta_3|x, \theta_1, \theta_2)$$

이를 바탕으로 m개의 표본을 생성하는 깁스 표본기법 알고리즘은 다음과 같다.

깁스 표본기법 알고리즘

각 모수의 초기값 $\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}$을 정하고, 다음과 같이 모수를 생성한다.

Step 1

$$\begin{aligned}\theta_1^{(1)} &\sim \pi(\theta_1|x, \theta_2^{(0)}, \theta_3^{(0)})\\ \theta_2^{(1)} &\sim \pi(\theta_2|x, \theta_1^{(1)}, \theta_3^{(0)})\\ \theta_3^{(1)} &\sim \pi(\theta_3|x, \theta_1^{(1)}, \theta_2^{(1)})\end{aligned}$$

Step 2

$$\begin{aligned}\theta_1^{(2)} &\sim \pi(\theta_1|x, \theta_2^{(1)}, \theta_3^{(1)})\\ \theta_2^{(2)} &\sim \pi(\theta_2|x, \theta_1^{(2)}, \theta_3^{(1)})\\ \theta_3^{(2)} &\sim \pi(\theta_3|x, \theta_1^{(2)}, \theta_2^{(2)})\end{aligned}$$

$$\vdots$$

Step m

$$\begin{aligned}\theta_1^{(m)} &\sim \pi(\theta_1|x, \theta_2^{(m-1)}, \theta_3^{(m-1)})\\ \theta_2^{(m)} &\sim \pi(\theta_2|x, \theta_1^{(m)}, \theta_3^{(m-1)})\\ \theta_3^{(m)} &\sim \pi(\theta_3|x, \theta_1^{(m)}, \theta_2^{(m)})\end{aligned}$$

위의 알고리즘을 보면, θ_1의 k번째 표본 $\theta_1^{(k)}(k=1, \cdots, m)$은 가장 최근 (θ_2, θ_3)의 값인 $\left(\theta_2^{(k-1)},\ \theta_3^{(k-1)}\right)$에만 의존할 뿐 그 이전 (θ_2, θ_3)의 값에는 의존하지 않는다. 이는 $\theta_2^{(k)}$ 및 $\theta_3^{(k)}$에 대해서도 마찬가지이다. 따라서 $\left(\theta_1^{(k)},\ \theta_2^{(k)},\ \theta_3^{(k)}\right)$는 마코프 성질을 가지므로 마코프 연쇄가 된다. 또한, 조건에 들어가는 모수값을 가장 최근의 값들로 대체하면서 차례대로 $\theta=(\theta_1, \theta_2, \theta_3)$를 완전 조건부 사후분포로부터 추출하면, k가 충분히 클 때 근사적으로 $\theta^{(k)}=(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k)})$는 사후분포 $\pi(\theta|x)$를 따르게 된다. 따라서 얻어진 $\left\{\theta^{(k)}=\left(\theta_1^{(k)},\ \theta_2^{(k)},\ \theta_3^{(k)}\right):\ k=1, \cdots, m\right\}$을 $\theta=(\theta_1, \theta_2, \theta_3)$의 사후표본으로 사용할 수 있다.

한편 초기값의 영향을 줄이고자 초기 b개의 표본은 사용하지 않고 나머지 $m-b$개의 표본만 사용하기도 하는데, 이때 버려지는 초기 b개의 표본을 burn-in 자료이라 한다. 또한 표본들 사이의 상관관계가 큰 경우, 가능한 독립표본을 만들고자 모든 표본을 사용하는 대신 일정 간격을 두고 일부 표본만 사용하는 것도 가능하다. 만약 초기 b를 버린 후 100을 간격으로 표본을 추출한다면, 실제 추론에 사용하는 크기 m인 표본은 다음과 같다.

$$\left\{\theta^{(b+1)}, \theta^{(b+101)}, \theta^{(b+201)}, \cdots, \theta^{(b+(m-1)\times 100+1)}\right\}$$

2) 메트로폴리스-헤이스팅스 알고리즘

앞에서와 마찬가지로 다변량모수 $\theta=(\theta_1, \cdots, \theta_n)$의 사후분포 $\pi(\theta|x)$의 표본을 생성하고자 한다. 앞에서 살펴본 깁스 표본기법은 각 원소 θ_i들의 완전 조건부 사후분포 $\pi(\theta_i|x, \theta_{-i})$로부터 직접적인 표본생성이 가능해야 한다. 즉, 각 모수의 완전 조건부 사후분포가 표본생성이 용이한 분포여야 한다. 그러나 때로, 특히 사전분포가 공액사전분포가 아닌 경우에, 완전 조건부 사후분포로부터 표본추출이 용이하지 않아 깁스 표본기법의 사용이 불가능하다. 이에 반해 메트로폴리스-헤이스팅스 알고리즘은 사전분포의 종류, 모수의 종류 및 크기에 크게 영향 받지 않는다는 장점을 가진다. 길이가 m인 마코프 연쇄를 생성하기 위한 메트로폴리스-헤이스팅스 알고리즘은 다음과 같다.

메트로폴리스 헤이스팅스 알고리즘

Step 1

θ의 초기값 $\theta^{(0)}$을 생성 $(k=1)$

Step 2

제안분포 $q(\bullet\ |\theta^{(k-1)})$로부터 θ^*을 생성

$\theta^* \sim q(\theta|\theta^{(k-1)})$

Step 3

채택확률 R을 계산

$$R = \min\left\{1, \frac{\pi(\theta^*|x)}{\pi(\theta|x)} \times \frac{q(\theta^{(k-1)}|\theta^*)}{q(\theta^*|\theta^{(k-1)})}\right\}$$

Step 4

균일분포 $U(0,1)$로부터 난수 u를 발생

Step 5

$\theta^{(k)}$를 생성

$$\theta^{(k)} = \begin{cases} \theta^* & if\ u \le R \\ \theta^{(k-1)} & if\ u > R \end{cases}$$

Step 6

$$k = k+1$$

Step 7

$k \le m$일 때까지 Step 2~6을 반복

깁스 표본기법과 마찬가지로 k번째 표본 $\theta^{(k)}$는 이전 표본인 $\theta^{(k-1)}$에 의존해서 생성되거나 $\theta^{(k-1)}$값이 되므로 마코프 성질을 만족하고, 연쇄의 길이 k가 길어질수록 표본의 분포는 $\pi(\theta|x)$로 수렴한다.

위의 알고리즘 Step 2에서 사용된 $q(\bullet\,|\theta^{(k-1)})$에 대해 살펴보자. 이전 표본값이 주어졌을 때, k 번째 표본 $\theta^{(k)}$를 생성하기 위해 고려된 분포를 제안분포(proposal distribution, sample generating distribution)라 한다. $q(\bullet\,|\theta^{(k-1)})$는 이 제안분포의 질량밀도함수이다. 다변량 모수인 경우 제안분포는 모든 모수의 결합분포일 필요가 없다. 각 원소 θ_i별로 제안분포를 따로 고려해도 되고, 또는 임의의 모수 결합에 대해서도 제안분포 생성이 가능하다는 장점이 있다. 메트로폴리스-헤이스팅스 알고리즘은 다양한 종류의 모수가 결합된 다변량 모수의 경우에도, 각 모수별 제안분포의 질량밀도함수와 가능도함수만 계산할 수 있다면 적용할 수 있다는 것이 강력한 장점이 된다.

5. 베이지안 붓스트랩

베이지안 모형의 가정은 다음과 같다.

(1) 관측값 $x_1, x_2, \cdots, x_n$은 미지의 가중치 $(w_1, w_2, \cdots, w_n)$를$(\sum_{i=1}^{n} w_i = 1)$ 모수로 갖는 다항분포(multinomial distribution)를 따른다.

(2) 가중치$(w_1, w_2, \cdots, w_n)$는 무정보 사전분포 $\pi(q_1, w_2, \cdots, w_n) = \prod_{i=1}^{n} w_i^{-1}$를 따른다.

(3) (1)과 (2)에 의해 가중치 $(w_1, w_2, \cdots, w_n)$의 사후분포는 $Dirichlet(1, 1, ..., 1)$이다.

Dirichlet 분포

p차원 확률변수 $X=(X_1, \cdots, X_p)$가 Dirichlet 분포 $Dirichlet(n,\ \alpha=(\alpha_1, \cdots, \alpha_p))$를 따른다고 할 때, 확률밀도함수는 다음과 같다. $(B(\alpha)=\prod_{i=1}^{p}\Gamma(\alpha_i)/\Gamma(\alpha_0))$

$$f(x_1, \cdots, x_p|\alpha)=\frac{1}{B(\alpha)}\prod_{i=1}^{p}x_i^{\alpha_i-1},\ 0\le x_i\le 1(i=1.\ \cdots, p)$$

이때 X의 평균 및 분산은 다음과 같다. $(\alpha_0=\sum_{i=1}^{p}\alpha_i)$

$$E(X_i)=\frac{\alpha_i}{\alpha_0},\ Var(X_i)=\frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}$$

6. 순차회귀 다중대체 예제

구체적인 예를 생각해보자. 먼저 모집단을 크기가 $N=2500$이고 세 개의 변수 (U, Y_1, Y_2)로 구성되어 있으며, 크기 $n=100$인 표본을 생각하자. 모집단의 모형이 다음과 같다고 가정하고 모집단 자료를 생성한다.

(1) $U\sim Normal(0,1)$

(2) $Y_1\sim\Gamma(\mu_1, \mu_1/5)$ 단, $\mu_1=exp(U-1)$

(3) $Y_2\sim\Gamma(\mu_2, \mu_2/2)$ 단, $\mu_2=exp(-1+0.5\times U+0.5\times Y_1)$

여기서 (3)은 변수들 (U, Y_1, Y_2) 사이의 관계를 표현한 회귀식을 나타내며, 회귀계수의 참값은 $(-1, 0.5, 0.5)$이다. $\Gamma(a,b)$는 감마분포를 나타낸다.

감마분포 $\Gamma(a,b)$

확률변수 X가 감마분포 $\Gamma(a,b)$를 따른다고 할 때, 확률밀도 함수는 다음과 같다.

$$f(x|a,b)=x^{a-1}\frac{\exp(-x/b)}{\Gamma(a)b^a},\ x\ge 0$$

이때 X의 평균 및 분산은 다음과 같다.

$$E(X)=ab,\ Var(X)=ab^2$$

다중대체를 실험하기 위해서, 위에서 생성된 자료에 대해 결측값을 발생시켜 보자. 이를 위한 알고리즘은 아래와 같다.

(1) U에는 결측값이 없다고 가정한다.

(2) Y_1의 결측 여부는 다음 로지스틱 회귀모형을 이용해 랜덤하게 결정한다.

$$Logit(\Pr(Y_1 \ is \ missing)) = 1.5 + U$$

즉, Y_1에서 결측이 일어날 확률은 $\Pr(Y_1 \ is \ missing) = \dfrac{exp(1.5+U)}{1+exp(1.5+U)}$ 이다.

(3) Y_2의 결측 여부는 다음 로지스틱 회귀모형을 이용해 랜덤하게 결정한다.

$$Logit(\Pr(Y_2 \ is \ missing)) = 1.5 - 0.5 \times Y_1 - 0.5 \times U$$

마찬가지로, $\Pr(Y_2 \ is \ missing) = \dfrac{exp(1.5-0.5\times Y_1 - 0.5\times U)}{1+exp(1.5-0.5\times Y_1 - 0.5\times U)}$ 이다.

위의 알고리즘에 의하면, Y_1의 22%, Y_2의 29%가 결측값이 되고, (U, Y_1, Y_2) 모두가 결측이 아닌 경우는 전체의 48%에 해당한다.

이제 결측값이 생성된 자료에 대해 SRMI 방법을 적용하여 대중대체를 실시해 보자. 모의실험 알고리즘은 다음과 같다.

(1) $P(Y_1|U)$로부터 Y_1을 생성

> Y_1에 대해 Box-Cox 변환을 한 $Z_1 = \left(Y_1^{\lambda_1} - 1\right)/\lambda_1$에 대해 U를 설명변수로 하는 선형회귀 적합을 수행한다. 즉, $Z_1 = \left(Y_1^{\lambda_1} - 1\right)/\lambda_1 = a_0 + a_1 U + e$, $e \sim N(0, \sigma_1^2)$이다. 단, λ_1과 σ_1^2는 최대가능도 추정량을 사용한다.

(2) $P(Y_2|U, Y_1)$으로부터 Y_2을 생성

> Y_2에 대해 Box-Cox 변환을 한 $Z_2 = \left(Y_2^{\lambda_2} - 1\right)/\lambda_2$에 대해 U와 Z_1을 설명변수로 하는 선형회귀 적합을 수행한다. 즉, $Z_2 = \left(Y_2^{\lambda_2} - 1\right)/\lambda_2 = b_0 + b_1 U + b_2 Z_1 + e$, $e \sim N(0, \sigma_1^2)$이다. λ_2와 σ_2^2 역시 최대가능도 추정량을 사용한다.

Raghunathan 외(2001)의 모의실험 결과에 의하면, 이와 같이 생성된 자료로부터 추정된 회귀계수 (b_0, b_1, b_2)가 실제 참값과 유사한 결과를 보여준다는 것을 확인할 수 있다.

7. 완전 재현자료 생성 R코드 예제

본 절에서는 본문 제4절의 2. 완전 재현자료 생성에서 설명한 Raghunathan 외(2003)에 소개된 모의실험을 구현한 R코드를 수록하였다. 관심있는 독자는 참고하기 바란다.

```
rm(list=ls())
direc<-getwd()

#### Simulation 1 -- Parametric ###########
library(mvtnorm)
p<-5                                    # number of variables
N<-1000                                 # population size
M<-5
my_sigma<-matrix(0.5, nrow=p, ncol=p)   # Population Covariance
diag(my_sigma)<-1
my_mean<-rep(0, p)                      # Population mean

## Generate Population ##
set.seed(1)
Pop<-rmvnorm(N, mean = my_mean, sigma =my_sigma)
nRepeat<-15                             # Number of repeat

n<-100                                  # sample size
beta_syn<-list(nRepeat, type="vector")
beta_original<-NULL
se_original<-NULL

beta0_syn<-matrix(0, nrow=nRepeat, ncol=M*2)
beta2_syn<-matrix(0, nrow=nRepeat, ncol=M*2)
beta3_syn<-matrix(0, nrow=nRepeat, ncol=M*2)
```

```
beta4_syn<-matrix(0, nrow=nRepeat, ncol=M*2)
beta5_syn<-matrix(0, nrow=nRepeat, ncol=M*2)

for(i in 1:nRepeat){
    index<-sample(1:N, size=n, replace=F)              # Given a sample

    my_sample<-Pop[index,]
    sample_mean<-apply(my_sample,2,mean)
    sample_cov<-cov(my_sample)

    fit_original<-lm(my_sample[,1]~my_sample[,2]+my_sample[,3]+my_sample[,4]+my_sample[,5])
    beta_original<-rbind(beta_original, coefficients(fit_original))
    se_original<-rbind(se_original, (as.numeric(coef(summary(fit_original))[,"Std. Error"]))^2)

    # Generate Synthetic data
    beta_tmp<-NULL
    se_tmp<-NULL

    for(j in 1:M){                                     # M=5
        v<-rWishart(1, df=n-1, Sigma=solve(sample_cov)/(n-1))
        v<-v[,,1]
        sigma_star<-solve(v)
        mu_star<-rmvnorm(1, mean = sample_mean, sigma =sigma_star/n)
        syn_pop<-rmvnorm(N, mean = mu_star, sigma =sigma_star)
        tmp_index<-sample(1:N, size=250,replace=F)
        syn_sample<-syn_pop[tmp_index,]

fit_syn<-lm(syn_sample[,1]~syn_sample[,2]+syn_sample[,3]+syn_sample[,4]+syn_sample[,5])
                    beta_tmp<-rbind(beta_tmp, coefficients(fit_syn))
```

```
        tmp_tmp<-(j-1)*2+1
        tmp_loc<-tmp_tmp:(tmp_tmp+1)

beta0_syn[i,tmp_loc]<-as.numeric(c(coef(summary(fit_syn))[1,1],(coef(summary(fit_syn))[1,2])^2))
beta2_syn[i,tmp_loc]<-as.numeric(c(coef(summary(fit_syn))[2,1],(coef(summary(fit_syn))[2,2])^2))
beta3_syn[i,tmp_loc]<-as.numeric(c(coef(summary(fit_syn))[3,1],(coef(summary(fit_syn))[3,2])^2))
beta4_syn[i,tmp_loc]<-as.numeric(c(coef(summary(fit_syn))[4,1],(coef(summary(fit_syn))[4,2])^2))
beta5_syn[i,tmp_loc]<-as.numeric(c(coef(summary(fit_syn))[5,1],(coef(summary(fit_syn))[5,2])^2))
        }                                                ##end_for_j
    beta_syn[[i]]<-beta_tmp
    }                                                    ##end_for_i

plot(seq(-0.5, 1.0, by=0.1), seq(-0.5, 1.0, by=0.1), type="n", xlab="Synthetic sample estimates", ylab="Actual sample estimates", main="(1)")
for(i in 1:nRepeat){
    for(j in 1:M){
        for(k in 1:5){
        points(beta_syn[[i]][j,k], beta_original[i,k])
        }                                                ##end_for_k
    }                ##end_for_k
}                                                        ##end_for_i
abline(a=0, b=1)

### beta0 ###
tmp_col<-1
my_beta_syn<-beta0_syn
tmpTemplate<-cbind(beta_original[,tmp_col], se_original[,tmp_col],my_beta_syn,
apply(my_beta_syn[,c(1,3,5,7,9)],1,mean), apply(my_beta_syn[,c(2,4,6,8,10)],1,mean))
tmpTemplate<-round(tmpTemplate, digits=4)
write.table(tmpTemplate, sep=",", paste(direc,"beta",(tmp_col-1),".csv", sep=""), col.names=F, quote=F)
```

```
### beta2 ###
tmp_col<-2
my_beta_syn<-beta2_syn
tmpTemplate<-cbind(beta_original[,tmp_col], se_original[,tmp_col],my_beta_syn,
apply(my_beta_syn[,c(1,3,5,7,9)],1,mean), apply(my_beta_syn[,c(2,4,6,8,10)],1,mean))
tmpTemplate<-round(tmpTemplate, digits=4)
write.table(tmpTemplate, sep=",", paste(direc,"beta",(tmp_col),".csv", sep=""), col.names=F, quote=F)

### beta3 ###
tmp_col<-3
my_beta_syn<-beta2_syn
tmpTemplate<-cbind(beta_original[,tmp_col], se_original[,tmp_col],my_beta_syn,
apply(my_beta_syn[,c(1,3,5,7,9)],1,mean), apply(my_beta_syn[,c(2,4,6,8,10)],1,mean))
tmpTemplate<-round(tmpTemplate, digits=4)
write.table(tmpTemplate, sep=",", paste(direc,"beta",(tmp_col),".csv", sep=""), col.names=F, quote=F)

### beta4 ###
tmp_col<-4
my_beta_syn<-beta2_syn
tmpTemplate<-cbind(beta_original[,tmp_col], se_original[,tmp_col],my_beta_syn,
apply(my_beta_syn[,c(1,3,5,7,9)],1,mean), apply(my_beta_syn[,c(2,4,6,8,10)],1,mean))
tmpTemplate<-round(tmpTemplate, digits=4)
write.table(tmpTemplate, sep=",", paste(direc,"beta",(tmp_col),".csv", sep=""), col.names=F, quote=F)

### beta5 ###
tmp_col<-5
my_beta_syn<-beta2_syn
tmpTemplate<-cbind(beta_original[,tmp_col], se_original[,tmp_col],my_beta_syn,
apply(my_beta_syn[,c(1,3,5,7,9)],1,mean), apply(my_beta_syn[,c(2,4,6,8,10)],1,mean))
```

```
tmpTemplate<-round(tmpTemplate, digits=4)
write.table(tmpTemplate, sep=",", paste(direc,"beta",(tmp_col),".csv", sep=""), col.names=F, quote=F)


#### Simulation 2 -- Non-Parametric ##############
rm(list=ls())
library(mvtnorm)
p<-5                                          # number of variables
N<-1000                                       # population size
M<-5
my_sigma<-matrix(0.5, nrow=p, ncol=p) # Population Covariance
diag(my_sigma)<-1
my_mean<-rep(0, p)                            # Population mean

## Generate Population ##
set.seed(1)
Pop<-rmvnorm(N, mean = my_mean, sigma =my_sigma)

nRepeat<-500                                  # Number of repeat
n<-100                                        # sample size
beta_syn<-list(nRepeat, type="vector")
beta_original<-NULL
for(i in 1:nRepeat){
    # Given a sample
    index<-sample(1:N, size=n, replace=F)
    my_sample<-Pop[index,]
    fit_original<-lm(my_sample[,1]~my_sample[,2]+my_sample[,3]+my_sample[,4]+my_sample[,5])
    beta_original<-rbind(beta_original, coefficients(fit_original))
```

```
        # Generate Synthetic data
        beta_tmp<-NULL
        for(j in 1:M){                                  # M=5
                a<-sort(runif(n-1))
                a<-c(0,a,1)
                u<-runif(N)
                tmpIndex<-rep(0,N)
                for(k in 1:N){
                tmpIndex[k]<-max(which(a<u[k]))
                }

        syn_pop<-my_sample[tmpIndex,]
        tmp_index<-sample(1:N, size=250,replace=F)
        syn_sample<-syn_pop[tmp_index,]

        fit_syn<-lm(syn_sample[,1]~syn_sample[,2]+syn_sample[,3]+syn_sample[,4]+syn_sample[,5])
        beta_tmp<-rbind(beta_tmp, coefficients(fit_syn))
        }                                               ## end_for_j
beta_syn[[i]]<-beta_tmp
}                                                       ## end_for_i
plot(seq(-0.5, 1.0, by=0.1), seq(-0.5, 1.0, by=0.1), type="n", xlab="Synthetic sample estimates", ylab="Actual sample estimates", main="(2)")

for(i in 1:nRepeat){
        for(j in 1:M){
                for(k in 1:5){
                        points(beta_syn[[i]][j,k], beta_original[i,k])
                }                                       ## end_for_k
        }                                               ## end_for_k
}                                                       ## end_for_i
abline(a=0, b=1)
```

8. 추정량 비교

본문 제4절의 3. 결합규칙 부분을 이해하는데 도움이 되도록 앞 절에 수록한 R코드를 실행하여 얻은 결과 일부를 정리했다. 관심있는 독자는 참고하기 바란다.

〈부표 3-1〉 $\hat{\beta}_0$의 추정량

표본ID	q	u	q_1	u_1	q_2	u_2	q_3	u_3
1	-0.1619	0.0073	-0.0983	0.0040	-0.1972	0.0020	-0.0562	0.0029
2	-0.0590	0.0066	-0.0495	0.0028	0.0625	0.0030	-0.2089	0.0025
3	-0.0419	0.0065	-0.1381	0.0032	-0.0238	0.0033	-0.0079	0.0016
4	-0.0371	0.0073	0.0259	0.0026	-0.1127	0.0035	-0.0603	0.0026
5	-0.0229	0.0068	-0.1882	0.0033	0.0936	0.0032	-0.1193	0.0037
6	-0.1118	0.0059	-0.0016	0.0020	-0.0275	0.0024	-0.1514	0.0023
7	0.0192	0.0075	0.1197	0.0026	-0.1614	0.0027	0.0608	0.0027
8	-0.0473	0.0057	-0.0657	0.0018	-0.1942	0.0022	-0.2071	0.0025
9	-0.1434	0.0054	-0.2318	0.0028	-0.1086	0.0028	-0.2217	0.0019
10	0.0790	0.0079	0.1070	0.0032	0.2238	0.0036	-0.0796	0.0039
11	-0.2013	0.0072	-0.1367	0.0029	-0.2785	0.0031	-0.1493	0.0024
12	-0.0813	0.0068	0.0516	0.0028	0.0051	0.0024	0.0754	0.0027
13	-0.0330	0.0072	0.1468	0.0027	-0.0552	0.0026	-0.0398	0.0028
14	-0.0981	0.0064	-0.1209	0.0024	-0.1787	0.0019	-0.1350	0.0031
15	0.0287	0.0067	-0.0702	0.0031	0.0525	0.0031	0.1425	0.0024
표본ID	q	u	q_4	u_4	q_5	u_5	$\overline{q}$	$\overline{u}$
1	-0.1619	0.0073	-0.1998	0.0029	-0.1971	0.0025	-0.1497	0.0029
2	-0.0590	0.0066	-0.0361	0.0019	-0.0824	0.0024	-0.0629	0.0025
3	-0.0419	0.0065	0.0778	0.0032	-0.0724	0.0028	-0.0329	0.0028
4	-0.0371	0.0073	-0.0182	0.0033	-0.1888	0.0029	-0.0708	0.0030
5	-0.0229	0.0068	0.0285	0.0019	-0.1252	0.0020	-0.0621	0.0028
6	-0.1118	0.0059	-0.1143	0.0024	0.0695	0.0025	-0.0451	0.0023
7	0.0192	0.0075	0.0530	0.0036	-0.0209	0.0027	0.0102	0.0029
8	-0.0473	0.0057	0.0063	0.0024	-0.0873	0.0024	-0.1096	0.0023
9	-0.1434	0.0054	-0.2555	0.0021	-0.1774	0.0017	-0.1990	0.0023
10	0.0790	0.0079	0.0241	0.0026	0.1173	0.0034	0.0785	0.0033
11	-0.2013	0.0072	-0.1570	0.0028	-0.1530	0.0027	-0.1749	0.0028
12	-0.0813	0.0068	-0.1954	0.0028	-0.0467	0.0028	-0.0220	0.0027
13	-0.0330	0.0072	-0.1932	0.0040	-0.1114	0.0027	-0.0506	0.0030
14	-0.0981	0.0064	-0.1458	0.0027	0.0045	0.0034	-0.1152	0.0027
15	0.0287	0.0067	0.1147	0.0026	0.0346	0.0031	0.0548	0.0029

〈부표 3-2〉 $\hat{\beta}_2$의 추정량

표본ID	q	u	q_1	u_1	q_2	u_2	q_3	u_3
1	0.2661	0.0095	0.3209	0.0033	0.1404	0.0031	0.1386	0.0035
2	0.1272	0.0108	0.2059	0.0043	0.1862	0.0032	0.1738	0.0043
3	0.2478	0.0111	0.2742	0.0052	0.2442	0.0049	0.1857	0.0021
4	0.1266	0.0098	0.1103	0.0033	0.0863	0.0051	0.2579	0.0032
5	0.2783	0.0079	0.2704	0.0031	0.4488	0.0033	0.2350	0.0038
6	0.0471	0.0095	0.1538	0.0032	0.0167	0.0036	0.2266	0.0030
7	0.2342	0.0120	0.0984	0.0036	0.2834	0.0053	0.3459	0.0035
8	0.1448	0.0127	0.1699	0.0032	0.3666	0.0050	0.2076	0.0095
9	0.2392	0.0073	0.1890	0.0035	0.4127	0.0036	0.2522	0.0028
10	0.1583	0.0108	0.2848	0.0036	0.1459	0.0038	-0.0727	0.0056
11	0.1565	0.0105	-0.0189	0.0038	0.2162	0.0040	0.0141	0.0028
12	0.3233	0.0094	0.5424	0.0045	0.4840	0.0033	0.3168	0.0031
13	0.2315	0.0117	0.1200	0.0051	0.2092	0.0040	0.1070	0.0058
14	0.3097	0.0121	0.3097	0.0041	0.2718	0.0027	0.2400	0.0048
15	0.2898	0.0086	0.2770	0.0032	0.3854	0.0038	0.2453	0.0028
표본ID	q	u	q_4	u_4	q_5	u_5	$\overline{q}$	$\overline{u}$
1	0.2661	0.0095	0.0863	0.0035	0.1176	0.0030	0.1607	0.0033
2	0.1272	0.0108	-0.0363	0.0033	0.0859	0.0037	0.1231	0.0038
3	0.2478	0.0111	0.1669	0.0052	0.3007	0.0051	0.2344	0.0045
4	0.1266	0.0098	0.2185	0.0040	0.0415	0.0039	0.1429	0.0039
5	0.2783	0.0079	0.1048	0.0019	0.4087	0.0020	0.2935	0.0028
6	0.0471	0.0095	-0.0126	0.0039	0.0076	0.0041	0.0784	0.0036
7	0.2342	0.0120	0.3104	0.0066	0.0815	0.0059	0.2239	0.0050
8	0.1448	0.0127	0.0335	0.0039	0.2206	0.0055	0.1996	0.0054
9	0.2392	0.0073	0.2698	0.0031	0.2470	0.0017	0.2742	0.0029
10	0.1583	0.0108	0.1196	0.0028	-0.0418	0.0060	0.0872	0.0044
11	0.1565	0.0105	0.2225	0.0049	0.1577	0.0035	0.1183	0.0038
12	0.3233	0.0094	0.2359	0.0036	0.3778	0.0025	0.3914	0.0034
13	0.2315	0.0117	0.2535	0.0061	0.2859	0.0053	0.1951	0.0052
14	0.3097	0.0121	0.3858	0.0064	0.2877	0.0073	0.2990	0.0050
15	0.2898	0.0086	0.2321	0.0024	0.3454	0.0035	0.2970	0.0031

〈부표 3-3〉 $\hat{\beta}_3$의 추정량

표본ID	q	u	q_1	u_1	q_2	u_2	q_3	u_3
1	0.2493	0.0102	0.3209	0.0033	0.1404	0.0031	0.1386	0.0035
2	0.2710	0.0098	0.2059	0.0043	0.1862	0.0032	0.1738	0.0043
3	0.3051	0.0104	0.2742	0.0052	0.2442	0.0049	0.1857	0.0021
4	0.2856	0.0113	0.1103	0.0033	0.0863	0.0051	0.2579	0.0032
5	0.2045	0.0126	0.2704	0.0031	0.4488	0.0033	0.2350	0.0038
6	0.2535	0.0082	0.1538	0.0032	0.0167	0.0036	0.2266	0.0030
7	0.1504	0.0106	0.0984	0.0036	0.2834	0.0053	0.3459	0.0035
8	0.1975	0.0066	0.1699	0.0032	0.3666	0.0050	0.2076	0.0095
9	0.3489	0.0078	0.1890	0.0035	0.4127	0.0036	0.2522	0.0028
10	0.1238	0.0105	0.2848	0.0036	0.1459	0.0038	-0.0727	0.0056
11	0.2963	0.0100	-0.0189	0.0038	0.2162	0.0040	0.0141	0.0028
12	0.0327	0.0095	0.5424	0.0045	0.4840	0.0033	0.3168	0.0031
13	0.2071	0.0105	0.1200	0.0051	0.2092	0.0040	0.1070	0.0058
14	0.3228	0.0091	0.3097	0.0041	0.2718	0.0027	0.2400	0.0048
15	0.2553	0.0088	0.2770	0.0032	0.3854	0.0038	0.2453	0.0028
표본ID	q	u	q_4	u_4	q_5	u_5	$\overline{q}$	$\overline{u}$
1	0.2493	0.0102	0.0863	0.0035	0.1176	0.0030	0.1607	0.0033
2	0.2710	0.0098	-0.0363	0.0033	0.0859	0.0037	0.1231	0.0038
3	0.3051	0.0104	0.1669	0.0052	0.3007	0.0051	0.2344	0.0045
4	0.2856	0.0113	0.2185	0.0040	0.0415	0.0039	0.1429	0.0039
5	0.2045	0.0126	0.1048	0.0019	0.4087	0.0020	0.2935	0.0028
6	0.2535	0.0082	-0.0126	0.0039	0.0076	0.0041	0.0784	0.0036
7	0.1504	0.0106	0.3104	0.0066	0.0815	0.0059	0.2239	0.0050
8	0.1975	0.0066	0.0335	0.0039	0.2206	0.0055	0.1996	0.0054
9	0.3489	0.0078	0.2698	0.0031	0.2470	0.0017	0.2742	0.0029
10	0.1238	0.0105	0.1196	0.0028	-0.0418	0.0060	0.0872	0.0044
11	0.2963	0.0100	0.2225	0.0049	0.1577	0.0035	0.1183	0.0038
12	0.0327	0.0095	0.2359	0.0036	0.3778	0.0025	0.3914	0.0034
13	0.2071	0.0105	0.2535	0.0061	0.2859	0.0053	0.1951	0.0052
14	0.3228	0.0091	0.3858	0.0064	0.2877	0.0073	0.2990	0.0050
15	0.2553	0.0088	0.2321	0.0024	0.3454	0.0035	0.2970	0.0031

〈부표 3-4〉 $\hat{\beta}_4$의 추정량

표본ID	q	u	q_1	u_1	q_2	u_2	q_3	u_3
1	0.0265	0.0115	0.3209	0.0033	0.1404	0.0031	0.1386	0.0035
2	0.2011	0.0113	0.2059	0.0043	0.1862	0.0032	0.1738	0.0043
3	0.0675	0.0087	0.2742	0.0052	0.2442	0.0049	0.1857	0.0021
4	0.1278	0.0119	0.1103	0.0033	0.0863	0.0051	0.2579	0.0032
5	-0.0409	0.0116	0.2704	0.0031	0.4488	0.0033	0.2350	0.0038
6	0.2656	0.0066	0.1538	0.0032	0.0167	0.0036	0.2266	0.0030
7	0.1104	0.0122	0.0984	0.0036	0.2834	0.0053	0.3459	0.0035
8	0.2078	0.0071	0.1699	0.0032	0.3666	0.0050	0.2076	0.0095
9	0.0179	0.0083	0.1890	0.0035	0.4127	0.0036	0.2522	0.0028
10	0.2690	0.0112	0.2848	0.0036	0.1459	0.0038	-0.0727	0.0056
11	0.1695	0.0094	-0.0189	0.0038	0.2162	0.0040	0.0141	0.0028
12	0.2946	0.0094	0.5424	0.0045	0.4840	0.0033	0.3168	0.0031
13	0.2070	0.0117	0.1200	0.0051	0.2092	0.0040	0.1070	0.0058
14	0.0296	0.0088	0.3097	0.0041	0.2718	0.0027	0.2400	0.0048
15	0.1043	0.0077	0.2770	0.0032	0.3854	0.0038	0.2453	0.0028
표본ID	q	u	q_4	u_4	q_5	u_5	$\overline{q}$	$\overline{u}$
1	0.0265	0.0115	0.0863	0.0035	0.1176	0.003	0.1607	0.0033
2	0.2011	0.0113	-0.0363	0.0033	0.0859	0.0037	0.1231	0.0038
3	0.0675	0.0087	0.1669	0.0052	0.3007	0.0051	0.2344	0.0045
4	0.1278	0.0119	0.2185	0.0040	0.0415	0.0039	0.1429	0.0039
5	-0.0409	0.0116	0.1048	0.0019	0.4087	0.0020	0.2935	0.0028
6	0.2656	0.0066	-0.0126	0.0039	0.0076	0.0041	0.0784	0.0036
7	0.1104	0.0122	0.3104	0.0066	0.0815	0.0059	0.2239	0.0050
8	0.2078	0.0071	0.0335	0.0039	0.2206	0.0055	0.1996	0.0054
9	0.0179	0.0083	0.2698	0.0031	0.2470	0.0017	0.2742	0.0029
10	0.2690	0.0112	0.1196	0.0028	-0.0418	0.0060	0.0872	0.0044
11	0.1695	0.0094	0.2225	0.0049	0.1577	0.0035	0.1183	0.0038
12	0.2946	0.0094	0.2359	0.0036	0.3778	0.0025	0.3914	0.0034
13	0.2070	0.0117	0.2535	0.0061	0.2859	0.0053	0.1951	0.0052
14	0.0296	0.0088	0.3858	0.0064	0.2877	0.0073	0.2990	0.0050
15	0.1043	0.0077	0.2321	0.0024	0.3454	0.0035	0.2970	0.0031

〈부표 3-5〉 $\hat{\beta}_5$의 추정량

표본ID	q	u	q_1	u_1	q_2	u_2	q_3	u_3
1	0.2729	0.0131	0.3209	0.0033	0.1404	0.0031	0.1386	0.0035
2	0.2728	0.0103	0.2059	0.0043	0.1862	0.0032	0.1738	0.0043
3	0.1927	0.0169	0.2742	0.0052	0.2442	0.0049	0.1857	0.0021
4	0.2929	0.0113	0.1103	0.0033	0.0863	0.0051	0.2579	0.0032
5	0.3856	0.0097	0.2704	0.0031	0.4488	0.0033	0.2350	0.0038
6	0.2978	0.0096	0.1538	0.0032	0.0167	0.0036	0.2266	0.0030
7	0.3688	0.0113	0.0984	0.0036	0.2834	0.0053	0.3459	0.0035
8	0.2508	0.0114	0.1699	0.0032	0.3666	0.0050	0.2076	0.0095
9	0.1939	0.0071	0.1890	0.0035	0.4127	0.0036	0.2522	0.0028
10	0.3360	0.0132	0.2848	0.0036	0.1459	0.0038	-0.0727	0.0056
11	0.2795	0.0119	-0.0189	0.0038	0.2162	0.0040	0.0141	0.0028
12	0.2769	0.0079	0.5424	0.0045	0.4840	0.0033	0.3168	0.0031
13	0.1858	0.0117	0.1200	0.0051	0.2092	0.0040	0.1070	0.0058
14	0.2787	0.0110	0.3097	0.0041	0.2718	0.0027	0.2400	0.0048
15	0.2103	0.0110	0.2770	0.0032	0.3854	0.0038	0.2453	0.0028
표본ID	q	u	q_4	u_4	q_5	u_5	$\overline{q}$	$\overline{u}$
1	0.2729	0.0131	0.0863	0.0035	0.1176	0.0030	0.1607	0.0033
2	0.2728	0.0103	-0.0363	0.0033	0.0859	0.0037	0.1231	0.0038
3	0.1927	0.0169	0.1669	0.0052	0.3007	0.0051	0.2344	0.0045
4	0.2929	0.0113	0.2185	0.0040	0.0415	0.0039	0.1429	0.0039
5	0.3856	0.0097	0.1048	0.0019	0.4087	0.0020	0.2935	0.0028
6	0.2978	0.0096	-0.0126	0.0039	0.0076	0.0041	0.0784	0.0036
7	0.3688	0.0113	0.3104	0.0066	0.0815	0.0059	0.2239	0.0050
8	0.2508	0.0114	0.0335	0.0039	0.2206	0.0055	0.1996	0.0054
9	0.1939	0.0071	0.2698	0.0031	0.2470	0.0017	0.2742	0.0029
10	0.3360	0.0132	0.1196	0.0028	-0.0418	0.0060	0.0872	0.0044
11	0.2795	0.0119	0.2225	0.0049	0.1577	0.0035	0.1183	0.0038
12	0.2769	0.0079	0.2359	0.0036	0.3778	0.0025	0.3914	0.0034
13	0.1858	0.0117	0.2535	0.0061	0.2859	0.0053	0.1951	0.0052
14	0.2787	0.0110	0.3858	0.0064	0.2877	0.0073	0.2990	0.0050
15	0.2103	0.0110	0.2321	0.0024	0.3454	0.0035	0.2970	0.0031

3부

차등정보보호 이론과 적용

4. 통계적 노출제어 최신동향 검토 – 차등정보보호를 중심으로 –

박민정

5. 차등정보보호에 관한 연구

박민정 · 이용희 · 권성훈

6. 차등정보보호 적용 실험 연구

박민정 · 권성훈 · 정재호

제4장

통계적 노출제어 최신동향 검토

- 차등정보보호를 중심으로 -

박민정

제1절 서론

공공재로써 널리 활용될 수 있도록 통계기관이 자료를 제공하고자 할 때, 주요 걸림돌 중 하나는 개별정보보호의 문제이다. 먼 과거에는 통계기관이 집계표 위주로 통계를 제공하였고, 그럼에도 사회의 통계 수요 대처에 충분하였다. 그러나 최근에는 이용자가 직접 자료를 심층 분석하는 경우들이 늘었기 때문에, 개별 자료들로 구성된 마이크로데이터에 대한 자료 활용 요청이 증가하고 있다. 또한 요구받는 자료도 각종 조사자료뿐만 아니라 훨씬 규모가 큰 행정자료·센서스 등의 전수자료까지 그 종류가 다양해졌다. 나아가 몇 개의 마이크로데이터를 연계한 자료를 생산하고 활용하는 경우도 생겨나기 시작하였다.

이처럼 다양한 종류의 자료 활용이 활발해지는 것은 긍정적인 사회 현상으로 볼 수 있고, 통계기관은 이에 발맞추어 자료제공 범위를 확대하고자 한다. 하지만 자료제공 범위 확대에는 개별정보 노출위험 증가가 수반된다. 실제로 해외에서는 공식적인 자료제공에서 개별정보 노출 사고들[1]이 발생하기도 하였다. 따라서 자료제공 범위 확대에 따라 적절한 개별정보보호 조치가 필요성하다는 인식도 점차 확산되고 있다.

✍ 이 연구는 2016년 하반기 통계개발원 연구보고서에 수록된 자료입니다.

1) 유명한 사례를 하나 소개하면 다음과 같다. 온라인 동영상 서비스사인 미국 넷플릭스(Netflix)에서 2006년 영화 평가 점수를 예측하는 알고리즘 경연 대회(Netflix Prize)를 열었다. 대회를 위해 비식별화 조치를 취한 48만여 명의 회원별 영화 평가 자료를 공개하였으나, 이 자료를 다른 온라인 영화 평가 서비스 제공사인 IMDB가 공개한 회원 자료와 결합하면 회원의 민감한 정보가 노출될 수 있음이 밝혀졌다. 이후 소송이 제기되면서 알고리즘 경연 대회는 더 이상 열리지 않게 되었다.

그러나 안타깝게도 개별정보의 노출위험을 줄이는 조치들을 취하게 되면 자료의 유용성은 떨어질 수밖에 없게 된다. 무엇이든 개별정보보호 조치를 취하게 되면 원래 자료에 비해 정보손실이 일어날 수밖에 없으므로, 이러한 조치는 자료제공 범위를 확대하고자 하는 정책 방향과 충돌한다. 간단한 예를 들면, 연령을 각 세별로 제공하거나 지역을 소규모 단위로 제공하면 이용자가 자료 분석을 통해 얻을 수 있는 정보는 많아진다. 그러나 바람직하지 않은 의도를 가지고 자료를 공격적으로 분석하면 개인이 식별될 수 있는 위험이 증가하게 되므로, 개별정보보호를 위해서는 연령을 5세 단위로 제공하든지 지역을 큰 권역으로 묶어 제공하는 조치를 취하게 된다. 이처럼 개별정보 노출위험 축소를 위한 조치를 취하면 원래 자료에 비해서 자료 유용성이 감소, 다른 말로 정보손실이 증가하게 된다. 따라서 정보손실은 없이 자료를 제공하면서 개별정보는 보호한다는 것은 참으로 어려운 일이라고 할 수 있다.

노출위험과 정보손실을 동시에 축소하는 한 가지 가능한 방안은 자료 이용자를 물리적으로 규제하는 것이다. 통계청의 통계 데이터 센터들이나 서울시의 빅데이터 캠퍼스 등과 같이, 통계기관들은 자료의 심층적인 이용을 원하는 이용자들이 이러한 특정 장소를 방문하여 원자료를 이용할 수 있도록 하고 있다. 다만 이용자들이 센터를 이용할 때 엄격한 보안 절차를 거치도록 하며, 결과물에 대해 통계기관이 노출위험을 검토하는 절차를 취한 이후에 이용자의 분석 결과물 반출을 허용한다. 이러한 데이터 센터의 이용은 통계기관이 제공하는 일반적인 집계표 이상의 원자료 심층 분석을 가능하게 하지만, 자료 이용자의 물리적 이동 비용 및 자료 제공자의 분석 결과물 검토라는 업무 부담을 요구한다. 이동 거리가 넓은 미국이나 유럽의 경우, 데이터 센터의 이용은 다른 국가들보다 더 큰 제약 조건을 가지게 된다. 또한 자료 이용 빈도가 높아지게 되면 자료제공기관의 업무 부담은 상당히 커질 수 있다.

데이터 센터 이용과 달리 이용자의 물리적 이동은 요구하지 않는 방안들도 있다. 법적 계약 절차를 거치는 인가파일(licensed file) 제공 및 파기, 일정 시간 동안 특정 IP의 접속을 허용하여 자료를 분석하게 하는 원격 접속 서버 운영 등이 그러한 경우에 해당한다. 데이터 센터 운영, 인가파일 제공, 원격 접속 서버 운영 등의 노출제어 노력들을 통틀어 물리적 노출제어라고 부른다면, 물리적 노출제어 방식은 공통적으로 자료 이용과 제공에 행정적, 법적, 경제적 부담이 따르는 단점을 가진다. 그러나 자료의 심층적인 이용을 위해서는 물리적 노출제어를 통한 자료제공이 필수적이며, 개별정보보호를 위해 이러한 방식을 통해서만 자료를 제공할 수밖에 없는 경우들이 존재한다. 다만 통계적 노출제어(Statistical Disclosure Control)를 통해 불특정 다수를 위한 자료제공 범위를 확대하고, 물리적 노출제어를 통한 자료제공의 비중을 줄여 나갈 수 있다면, 이용자와 제공자에게 사회·경제적 부담을 줄일 수 있는 장점이 있다.

빅데이터 시대, 나아가 4차 산업혁명 시대를 맞이하여, 더 많은 자료를, 더 활발하게 사용할 수 있도록 제공하는 노력은 계속될 것이다. 분석에 활용되는 자료의 크기와 종류도 다양해지고, 자료를 제공하는 방법도 시대의 흐름에 맞추어 변해갈 수밖에 없다. 따라서 개별정보 비밀보호를 하면서도 자료를 이용하기 편리하게 제공하기 위해 통계적 노출제어를 확대해나가는 것이 바람직하지만, 안타깝게도 아직까지는 만족스럽고 적절한 방법이 있다고 여겨지지는 않고 있다.

본 보고서에서는 통계적 노출제어에 대한 최신 동향을 정리하여 바람직한 대안을 모색하는 전환점을 마련하고자 한다. 새로운 환경에서 편리하고 안전하게 자료를 제공하기 위한 방법을 구체적으로 모색하기 위한 사전 작업으로써, 본 보고서에서는 우선 전통적인 통계적 노출제어 방안인 매스킹 기법들을 정리하고, 최근 제시되고 있는 새로운 통계적 노출제어 기법들을 소개한다. 이를 통해 어떻게 통계적으로 개별정보의 노출제어를 해 나갈 수 있을지에 대해 정책적인 고민이 이루어질 수 있도록 기초 자료를 제공하고자 한다. 특히, 컴퓨터 공학 분야에서 처음 제안되기는 하였지만, 최근 빅데이터 제공에서 있어서 노출제어 대안으로 거론되고 있는 차등정보보호의 출현 배경, 개념 및 활용에 관하여 수식을 최대한 배제하고 쉽게 소개하고자 하였다.

한편, 보다 전문적인 내용은 참고문헌들을 제시하여 관심 있는 독자들은 찾아볼 수 있도록 하였으며, 통계적 노출제어에 대한 통계개발원의 기존 연구 내용을 잘 아는 독자들은 2절을 생략할 수 있음을 밝혀둔다. 2절은 매스킹 기법들의 활용과 관련된 내용들과 재현자료 소개로 구성되어 있으며, 본 연구의 주요 목적인 차등정보보호 소개는 3절과 4절에서 이루어진다.

제2절 통계적 노출제어 기법

이 절에서는 차등정보보호를 제외한 통계적 노출제어 기법들에 대하여 차례로 소개하고자 한다. 먼저 노출위험과 정보손실의 개념을 설명하고, 이어 전통적인 매스킹 기법들을 소개한다. 다음으로 자료제공 방법이 비상호적인(non-interactive) 방식에서 상호적인(interactive) 방식으로 변화해 감에 따라, 온라인 분석 서버에 매스킹 기법을 적용하는 방식에 관해 설명하겠다. 마지막으로 미국을 중심으로 활발하게 연구되고 있는 재현자료를 간략히 소개한다. 이러한 내용은 Park and Kim(2016)에도 정리된 바 있으나, 이 절에서는 보다 일반적인 독자들을 대상으로 좀 더 쉽고 자세하게 설명하도록 하겠다.

1. 노출위험과 정보손실

마이크로데이터에서 개별정보의 노출은 언제 일어나는 것이며, 어떻게 측정할 수 있을까? 이에 관한 설명은 크게 두 갈래로 이루어질 수 있다.

우선 응답자 특성을 나타내는 변수들은 키변수, 응답자의 민감한 정보를 담고 있는 변수들은 민감변수라 하자. 또 외부인은 마이크로데이터의 키변수들을 통해 특정 응답자를 식별하고, 해당 응답자의 민감변수 정보를 파악하려 한다고 가정하자.

마이크로데이터에서 개별정보의 노출은 첫째, 키변수 조합이 유일한 응답 개체들에게서 일어난다. 가상의 사례로, 제주도에 사는 51세 여성 군인이라는 특성을 가지는 개인이 전국에서 유일하며, 자산, 부채, 소득에 관한 조사에 성실히 응답했다고 하자. 이 마이크로데이터가 아무 조치 없이 공표된다면, 이 여성의 특성(지역, 연령, 성별, 직업)에 관한 정보를 가진 외부인은 마이크로데이터에서 이 여성을 식별할 수 있고 관련된 민감한 정보(자산, 부채, 소득)를 얻을 수 있다. 이렇게 키변수 조합 혹은 개체 특성의 유일성으로 인해 개별정보의 노출이 일어난다고 보고, 표본 마이크로데이터의 어떤 개체가 모집단에서 유일할 확률을 노출위험으로 정의하는 접근이 있다. 이 방식으로 통계적 모형을 사용해 각 개체별 노출위험을 측정할 수도 있고, 측정된 개체별 노출위험의 평균이나 합을 이용해 주어진 마이크로데이터 파일 단위의 노출위험을 논할 수도 있다. 유일성에 근거한 노출위험 개념과 측정은 네델란드 통계청에 소속된 저자들의 논문인 Bethlehem 외(1990)에서 시작되었다고 볼 수 있으며, 후속 연구로 Skinner와 Shlomo(2008) 등을 통해 단점을 완화시키며 발전해가는 이론들을 파악할 수 있다.

모집단 유일성 기반 노출위험 측도가 가지는 한계(Reiter, 2005)를 일부 소개하면 다음과 같다. 먼저, 유일성 기반 측도는 외부인이 가지고 있는 자료의 특징을 반영하지 못한다. 예를 들어 외부인이 어떤 개체가 표본에 있다는 것을 안다면, 그 개체가 모집단에서는 유일하지 않더라도 표본에서 유일한 것만으로도 노출될 수 있다. 다음으로, 통계기관이 민감변수에 매스킹 처리를 했다고 하더라도, 키변수만을 고려하여 측정된 유일성에 근거한 노출위험 값은 변하지 않는다. 즉, 통계기관이 실시한 노출제어 처리 효과를 반영하지 못하는 한계를 가진다. 이러한 한계를 뛰어 넘어 노출위험을 측정하는 방안으로 의사결정론 기반 노출위험 측도가 연구되어 왔다. 의사결정론 기반 노출위험 측도에 관해서 관심 있는 독자는 <부록> 1. 및 <부록> 2.의 내용과 관련 참고문헌들을 살펴보기 바란다.

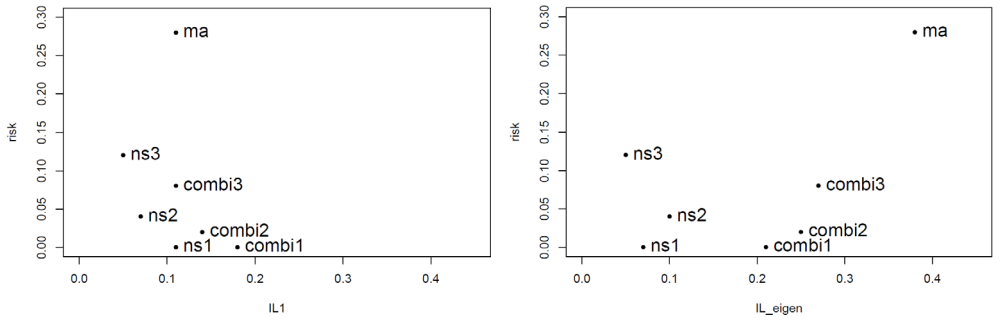
다음으로 정보손실 측정에 관하여 생각해보자. 이론적인 이해가 필요한 노출위험에 비해 정보손실은 매스킹 전후 자료의 차이를 의미하며, 매스킹 전후 자료 간 거리에

의해 비교적 간단히 측정된다고 할 수 있다. 즉, 매스킹 전후 자료값, 상관계수 혹은 주성분 행렬에 대해 평균제곱오차(Mean Square Error) 등을 계산하여 정보손실 값을 얻는다. 나아가 매스킹 전후, 주요 통계량들을 계산하고 비교하여 정보손실 측도로 활용하기도 한다. 자세한 내용은 박민정(2014)을 참고하기 바란다.

마지막으로 노출위험과 정보손실 측도를 종합적으로 고려하여 적절한 매스킹 방안을 선택하기 위해 제안된 위험-유용성 지도(Risk-Utility Map)를 소개한다. 가로축을 정보손실, 세로축을 노출위험[2]으로 하고 각 매스킹 방안별로 좌표에 측도값들을 표시하여 한눈에 매스킹 처리된 파일들을 비교할 수 있다. 이때, 원점에 가까운 파일일수록 노출위험과 정보손실이 작으므로 노출제어가 잘된 파일이라고 할 수 있다. 아래 [그림 4-1]에서 ma라고 표기된 매스킹 처리된 파일보다는 ns라고 표기된 파일들이 노출위험과 정보손실 모두 낮아 좀 더 좋은 매스킹 방안 이라고 할 수 있다. 왼쪽 그림은 매스킹 전후 자료값들 사이의 거리를 이용해 정보손실(IL1) 측도로 사용하였고, 오른쪽 그림에서는 고유값들 사이의 거리를 이용해 정보손실(IL-eigen) 측도로 사용하였다. 정보손실 측도 IL1은 자료값들의 변화를, IL-eigen은 자료 구조적 측면의 변화를 반영한다고 해석할 수 있다. 이렇게 매스킹 처리된 파일들의 평가는 노출위험과 정보손실 측도의 종류에 따라 그 결과가 달라진다.

한편, 노출위험-유용성 지도를 통해 매스킹 방안을 평가할 때, 노출위험과 정보손실 사이의 상충관계(trade off)를 고려해야 한다. 매스킹 방안 ns2와 ns3은 원점에서 거리가 유사하며, 노출위험이 낮은 대신 정보손실이 많거나 그 반대의 경우이다. 이들 사이에서는 어느 매스킹 방안이 더 우월하다고 하기 어렵다. 따라서 특정 매스킹 처리 방안을 선택하기 위해서는, 자료별로 노출위험, 정보손실, 기타 자료제공을 할 때 고려할 요소들을 종합적으로 검토하는 과정이 필요하다고 하겠다.

2) 유일성 기반 노출위험 측도는 민감변수에 적용한 매스킹 효과를 반영하지 못하고, 외부인 의사결정론 기반 노출위험 측도는 일반적인 프로그램으로 구현되어 있지 않으며 파일 단위 노출위험을 측정하기에 부적절하다는 단점을 가진다. 한편, 위험-유용성 지도에서 사용된 노출위험 측정값은 정보손실과 유사하게 매스킹 전후 거리를 이용하여 계산되었다. 자세한 방식은 박민정(2014)를 참고하면 된다.



[그림 4-1] 노출위험-정보손실 지도(박민정, 2014)

지금까지 노출위험과 정보손실 측도에 대하여 마이크로데이터 위주로[3] 설명하였다. 노출위험은 유일성이나 외부인 의사결정에 근거하여 측정할 수 있다. 정보손실은 매스킹 전후 자료들 간의 거리를 이용해 측정한다. 노출제어 방식을 결정하기 위해서는 노출위험과 정보손실을 동시에 비교하는데, 매스킹 처리를 통해서는 두 측도를 동시에 낮추기는 거의 불가능하다. 노출위험을 축소시키기 위해 자료에 어떤 변형을 가하게 되면 정보손실이 일어날 수밖에 없기 때문이다.

기존 매스킹 기법들만을 사용하면 노출위험과 정보손실이 모두 작은 방안을 찾기 위해 온갖 매스킹 기법들을 모두 적용하고 나서 위험-유용성 지도에서 원점에 가까운 방안을 찾아야 한다. 이러한 과정을 거친 후에도, 원점과 거리가 비슷한 방안들 중에서 어느 방안이 더 타당한지 결정하는 문제를 해결하는 것은 쉽지 않다. 따라서 위험-유용성 틀 안에서 적절한 통계적 노출제어 방안을 찾는 것이 과연 적절한 해법이 있는 문제인가 하는 의문을 가지게 된다.

객관적인 타당성을 가지는 적절한 해법을 찾기 위해, 연구자들은 생각하는 관점을 바꾸어 위험을 0에 가깝게 두고 정보손실을 줄이는 방법들을 찾게 되었다. 그러한 방법들로 온라인 분석 시스템을 통해 결과물을 제공할 때 매스킹 처리를 하거나, 재현자료를 제공하거나, 차등정보보호라는 새로운 노출위험 개념을 활용하는 방식들이 고려되고 있다. 그러나 통계적 노출제어 분야에서 아직은 바람직한 대안이 성숙된 상태로 제시되어 있다고 보기는 어렵고, 지속적인 연구 진행이 이루어지고 있는 과정이라고 판단된다. 어느 방향으로 연구를 하더라도 상충관계에 있는 노출위험과 정보손실을 종합적으로 고려하여 타당한 노출제어 방안을 찾아야 하므로, 자료 제공자는 노출위험과 정보손실 측도들에 대한 개념을 잘 알고 있어야 할 필요가 있다.

---

3) 집계자료의 노출위험 측정을 위해서는 k-익명성, l-다양성 등의 개념을 활용할 수 있다. 정보손실은 마이크로데이터와 마찬가지로 원자료의 값과 차이를 이용해 측정한다. 본 보고서는 차등정보보호 위주의 통계적 노출제어에 대한 최신 동향을 다루므로, 집계자료 관련 측도 및 매스킹 방법에 대하여는 연구 범위를 넘는 것으로 간주한다.

2. 매스킹 기법

마이크로데이터를 위한 매스킹(masking) 기법들은 크게 ①부분적으로 자료 값을 제공하거나(sampling), ②자료 값을 가리거나(suppression), ③구분할 수 없는 그룹으로 값을 섞어 제공하거나(aggregation), 혹은 ④자료 값을 변조(perturbative data)하는 것들, 네 가지 범주로 구분된다(Duncan 외, 2011). 자료에 적절한 변환을 하여 간접적인 식별 정보를 가리는 과정들을 통칭하여 보통 매스킹이라 부르며, 변수의 형식에 따라 사용할 수 있는 세부 기법들이 다르다. 또 각 기법별로 세부 알고리즘이 다수 존재한다. 위의 네 범주들 중 표본 추출은 일반적으로 알려진 표본 추출 방법들을 마이크로데이터에 적용하여 제공하는 것을 말한다. 예를 들면 전수자료인 인구주택총조사 자료를 1%, 2%, 10% 표본으로 제공하는 경우를 생각할 수 있다. 대용량 자료를 마이크로데이터 파일 형태로 제공할 때는 표본으로 제공할 수밖에 없기도 하다.

기본적인 매스킹 기법으로는 전반적 재코딩(global recoding)이 있으며, ②에는 국소적 감추기(local suppression), ③에 속하는 기법으로 국소통합(microaggregation), ④에는 잡음추가(noise addtion), 자료순위 교환(rank swapping), 자료섞기(shuffling)[4] 등이 있다. 이러한 매스킹 기법들을 통해 마이크로데이터를 변환하여 제공할 수 있다. 매스킹 처리되어 제공되는 마이크로데이터는 원래 마이크로데이터와 당연히 동일하지 않다. 매스킹 각 기법에 대한 간단한 세부 설명들은 <부록> 3.에 제시하였으니 참고하기 바란다.

한편, 통계기관이 각종 통계표들을 제공할 때 개별정보가 노출될 수도 있다. 통계표들을 서로 비교하고 분석하여, 즉 이용자가 제공받은 결과값들을 역으로 계산하여, 어떤 특성을 가진 개체에 대해 통계표가 직접 제공하지 않는 정보를 알아낼 수 있기 때문이다. 이를 방지하기 위한 기법들을 매크로데이터를 위한 매스킹 기법들이라 한다. 매크로데이터 매스킹 기법들로는 셀 감추기, 빈도표에 사용되는 작은 셀 조정(small cell adjustment), 총계표에 사용되는 $p\%$-규칙과 같은 경험적 규칙 등이 있다. 이러한 방법들은 통계기관이 공표하는 각종 통계표들이나 데이터 센터를 이용하는 이용자들의 분석 결과물과 같은 집계자료에 사용될 수 있다. 간단한 세부 설명들은 <부록> 3.을 참고하고, 자세한 내용들은 참고문헌을 살펴보기 바란다.

4) 전반적 재코딩, 국소적 감추기, 국소통합, 잡음추가, 자료순위 교환, 자료섞기 등은 우리말로 번역된 적절한 용어가 없어서 통계개발원에서 2013년 연구 과제를 수행할 때 번역했던 용어를 사용하였다.

3. 온라인 분석 서버와 매스킹

매스킹 기법들을 적용하여 마이크로데이터를 제공하는 경우, 노출위험을 이유로 자료제공 범위가 축소되는 경우들이 있다. 예를 들어 지역 변수를 제공하여 개인이 식별될 수 있는 위험이 현저히 증가하는 경우, 지역 변수를 제공하지 않도록 결정할 수 있다. 물론 표본의 개수가 충분하지 않아 지역 변수를 제공하기 어려운 경우도 있으나, 충분한 표본 크기를 확보하기 위해 몇 년간의 자료를 축적하여 분석하는 이용자들도 있다. 이러한 경우, 지역 변수를 제공하지 못하는 주요 원인은 개별 정보 노출위험이 된다.

노출위험 때문에 마이크로데이터 파일을 직접 제공할 수 없는 대부분의 경우, 이용자들은 데이터센터를 방문하여 자료를 이용해야 한다. 서론에서 언급했던 것처럼, 물리적 노출제어는 자료 이용자에게도 물리적 이동을 위한 불편함을, 제공자에게는 분석 결과물 검열에 대한 행정적 업무 부담을 야기하는 측면이 있다. 더욱이 연간 수백 편 이상의 논문에 사용되는 자료라면 물리적 노출제어를 하지 않고 자료를 이용할 수 있도록 할 필요성이 더욱 크게 된다.

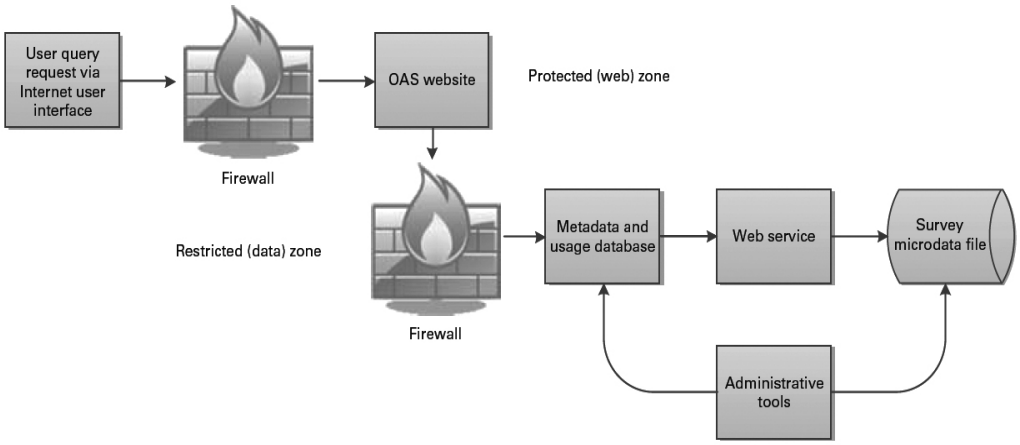
물리적 노출제어 없이 매스킹 기법들을 사용하면서도 자료제공 범위를 확대할 수 있는 방안으로, 자료를 분석할 수 있는 온라인 분석 시스템을 활용하는 방안이 연구되었다. 이는 자유로운 접속에 대해서 노출제어 가능한 요청만을 수용하고, 분석 결과물을 검열없이 제공하는 것이기 때문에 사용자의 접근을 규제하는 원격 접속과는 다르다. 또한, 마이크로데이터 파일을 이용자 개인의 컴퓨터 프로그램에서 직접 분석할 수는 없지만, 온라인 분석 시스템에서 마이크로데이터를 분석하여 기존 파일에서는 제공되지 않았던 변수를 분석할 수 있으므로, 자료제공 범위가 확대되었다고 할 수 있다. 이러한 자료제공 방안 연구 사례로 미국 NCHS(National Center for Health Statistics)의 연구 결과가 있다. 이 연구 결과를 통해 매스킹 처리 기능을 가진 온라인 분석 시스템을 활용한 자료제공을 이해해보도록 하자.

미국 NCHS에서는 1957년부터 NHIS(National Health Interview Survey)라는 가구 기반 조사를 시행하고 관련 자료를 공표해 왔다. 이 조사의 표본 설계는 전국 단위 추정을 가정하고 이루어졌다. 하지만 지역(State) 단위로 정책 수립을 하기 원하는 이용자들의 요청으로 지역 단위의 자료들을 데이터 센터를 이용해 제공하고 있다. 한편, NHIS 공공이용파일 버전은 지역 변수나 일부 민감 변수들을 제외하고 제공된다. 앞에서 설명한 것과 같이 물리적 노출제어 없이 제공하는 자료의 범위를 확대하기 위해 NCHS에서는 내부적으로 매스킹 기법을 적용한 실시간 온라인 분석 시스템(real-time Online Analytic System with Masking Methods, mOAS[5])을 만들어 지역 변수를 제공할 수 있는 방안을 연구하였다(Krenzke 외, 2013).

---

5) 본 보고서에서는 통계적 노출제어를 위해 매스킹 기법을 활용한 실시간 온라인 분석 시스템(Real-Time Online Analytic System with Masking Methods)을 편의상 mOAS라고 표기하겠다.

먼저 mOAS의 대략적인 구조를 정리하면 아래 [그림 4-2]와 같다. 그림은 사용자의 분석 요청을 받으면 방화벽과 관리 도구 등을 거쳐 자료에 접근하는 과정을 보여준다. 시스템이 메타데이터 데이터베이스, 관리도구, 마이크로데이터 파일, 분석 엔진 등으로 구성되어 있으므로, 이러한 시스템을 구축할 때는 IT 전문가와 통계 분석 및 노출제어 전문가의 협업이 필수적이다. 본 보고서에서는 mOAS를 구축하기 위해 통계 담당자가 고려해야 하는 내용들을 소개하도록 하겠다.



[그림 4-2] 온라인 분석 시스템의 일반적인 구조(Krenzke 외, 2013)

NHIS의 경우 공표용 자료가 따로 존재하므로 mOAS을 통해 자료를 분석할 때 다른 기존의 공표 자료들을 이용하는 경우와 mOAS를 통해 얻게 되는 분석 결과물 사이에 일치성이 확보되어야 하는 제약 조건이 생긴다. 이러한 제약 조건까지 고려하여 분석 엔진 구축 이전에 노출제어를 위해 결정해 두어야 하는 사항들은 아래와 같다.

① 다년간의 자료 분석을 위한 도구를 제공할지
② 분석을 허용하는 자료의 최소 개수
③ 쿼리당 요청할 수 있는 변수의 개수
④ 연령 구간화 폭, 변수 범주 결합 기능 제공 여부, 반올림 규칙 등
⑤ 공표용 자료와 데이터 센터 보안 자료들의 통합 범위
⑥ 지역별 통계값의 결합 결과와 전국 통계값의 일관성 확보 여부
⑦ 분석 엔진 속도 및 용량 관리를 위해 내부 자료 파일 구성
⑧ 사용하는 비밀보호 기법들의 공개 수준

이러한 내용들은 NHIS 자료를 이용하는 모든 상황들을 고려하기 위한 것이다. 데이터센터를 통해 제공하는 제약이 거의 없는 마이크로데이터와 공공이용파일로 제공하는 축소된 마이크로데이터를 분석해서 얻는 결과들이 mOAS를 이용해 얻는 결과와 일치성을 유지하는지, 혹은 결과물 차이에 의해서 개별정보 노출이 일어나지는 등의 세부 사항들을 고려하여 mOAS를 만들어가는 초기 과정을 보여주고 있다. 차등정보보호와 비교하자면, mOAS의 주요한 특징은 위의 8개 사항들과 같이 세부적으로 고려하는 조건들이 많으며, 각 사항들에 대해서 매스킹 기법들로 적절한 처리를 모색하여 시스템에 구현한다는 것이라 하겠다.

다양한 매스킹 처리 세부 과정 중에서 동적 부차표본추출(dynamic subsampling)을 설명하며 mOAS에 대한 소개를 마치도록 하겠다. 분석 시스템에서 통계표들을 제공하는 경우, 이용자는 자료를 공격적으로 분석하여 개체별 정보를 얻고자 할 수 있다. 특히, 반복적인 분석 요청을 통해 얻은 통계표들을 연계하여 서로 더하거나 빼서 이러한 계산들을 하는 것이 가능하다. 이러한 위험은 막으면서도 통계기관은 제공하는 통계표들 사이에 일관성을 유지해야 하는 어려움을 안고 있다.

이를 해결하기 위해 mOAS에서는 통계표를 작성할 때 전체 자료를 다 사용하는 대신, 표본 추출을 통해서 통계표를 작성하도록 하였고, 이를 동적 부차표본추출이라고 부르고 있다. 이때 같은 쿼리에 대해서 같은 결과를 제공할 수 있도록, 셀 값과 부분합들을 시드(seed)로 사용하여 난수를 생성하고, 이 난수들을 통해 표본추출을 하도록 하는 알고리즘을 구현하였다. 아래 <표 4-1>에서는 통계표에서 남성×특성2 셀에 해당하는 통계값을 작성하기 위해, 시드로 사용되는 네 개 값을 체크 표시로 표현하였다.

〈표 4-1〉 동적 부차표본추출 방식

	특성1	특성2	…	특성k	합
남성		✓			✓
여성					
합		✓			✓

4. 재현자료 활용

노출위험과 정보손실을 동시에 축소시키기 어려운 매스킹 기법의 한계를 근본적으로 뛰어넘기 위해서 재현자료(synthetic data) 생성 방안이 발전하였다. 재현자료를 생성할

때는 ①공표 자료 내 어떤 개체도 원자료 내 민감변수의 값을 가지지 않고, ②원자료에서 가능한 통계 분석/추론이 공표 자료에서도 가능할 것을 추구한다. 이것은 노출위험을 0으로 둔 상태에서 자료의 구조의 보존을 추구하는 것을 의미한다.

재현자료 생성의 기본 아이디어는 표본 조사 등을 통해 모집단을 대표하는 다수의 표본 추출이 가능한 것처럼, 조사된 표본과 구조가 유사한 재현자료 세트들을 통계적으로 생성하자는 것이다. 초기 연구로는 2001년 프랑스의 종단 연계 자료 사례가 있다. 이는 자료 수집 기관 및 표본 추출 방법이 서로 다른 여러 자료들을 종단적으로 연계하여 새로운 통계를 생산하기 위해 기존의 매스킹 기법이 아닌 새로운 방법론 도입을 시도한 것이다. 이는 완전 재현자료 생성 방법이라 하며, ①모집단에서 개체를 추출해 재현자료 세트(들)을 구성하고 ②원조사 자료를 기반으로 만든 모형을 가지고 다중대체(multiple imputation) 기법을 사용해 위의 재현자료 세트를 채우고 ③여러 개의 재현자료 세트를 공표하는 단계를 가진다.

완전 재현자료 이용은 통계적 노출제어에서 획기적인 아이디어로 판단되며 매스킹 기법만으로는 성취하기 어려웠던 목표들을 추구할 수 있는 길을 열어주었다. 일반적으로 사용자 모형의 정확도가 높을 때 노출위험이 높아지는데 인위적으로 덜 정확한 재현자료 생성 모형 사용을 통해 노출제한이 가능하며, 조사된 자료의 분포와 통계적으로 유사한 분포에서 재현자료를 생성하여 다양한 추론에 대해 통계적 타당성을 제공하고, 또한 소지역을 포함한 지리적 정보 공표 가능성도 생긴다. 나아가 비밀보호기법과 관련된 새로운 통계 이론이나 프로그램 습득이 불필요하고 원자료의 표본추출법에 관한 고려 없이 단순 추출을 가정하고 통계 분석이 가능하여 이용자의 편의성을 크게 높이기도 한다. 단점은 재현자료를 만드는 모형 및 가정에 지나치게 의존적이고 편의도 큰 것이다. 때문에 어떤 통계기관도 완전 재현자료 활용을 전적으로 받아들이지 않았다. 완전 재현자료 생성 방법론 연구는 준모수적 및 비모수적 모형 탐구, 다양한 구조를 가지는 모집단에 대한 추론 타당성 연구 및 부분 재현자료 활용 연구로 발전하는 모습을 보인다.

부분 재현자료 생성은 완전 재현자료 생성과 기본적으로 유사하나 조사 자료 중 노출위험이 큰 레코드들만을 선택해 재현자료로 변환하여 일부 정보에 대해서만 자료를 대체함으로써 모형 의존성을 낮추는 효과를 가진다. 주요 연구 사례로는 독일 IAB 기관 패널 조사 및 미국 사업체 종단자료의 비밀보호가 있다.

재현자료를 활용하는 방안은 미국을 중심으로 활발하게 연구되고 있다. 통계개발원에서는 재현자료에 대한 좀 더 자세한 연구를 진행할 계획을 가지고 있으며, 본 보고서에서는 재현자료에 관해서는 이 정도로 간략히 소개만 하도록 하고, 차등정보보호를 자세히 다루도록 하겠다.

제3절 차등정보보호

통계적 노출제어 방법들은 크게 네 그룹으로 나누어 볼 수 있다. 첫째는 전통적인 매스킹 기법들을 활용하여 매스킹 처리된 마이크로데이터 파일이나 집계자료(매크로데이터)를 제공하는 것이다. 둘째는 매스킹 기법들을 내재한 온라인 분석 시스템을 통해서 자료를 분석할 수 있도록 제공하는 방안이다. 이는 물리적 제어가 없이 제공하는 자료의 범위를 확장하는 의의를 가진다. 셋째는 원래 마이크로데이터의 재현자료를 여러 세트 작성하여 제공하고, 간단한 결합 규칙도 제공하여 이용자가 얻고자 하는 통계량을 재현자료들을 통해 계산하도록 하는 것이다. 네 번째는 자료 분석 시스템을 통해 마이크로데이터를 사용하도록 하고, 분석 결과물들이 차등정보보호를 만족하도록 제공하는 방식이 있다.

첫 번째 매스킹 기법은 각국의 통계기관들이 전통적으로 활용하여 왔으며, 세부 기법별로 많은 연구 결과들이 축적되어 있다. 그러나 매스킹 기법들로 노출위험과 정보손실을 동시에 축소시키기 불가능하므로,[6] 궁극적인 통계적 노출제어 대안을 찾기 위한 노력이 다양하게 전개되어, 다른 세 종류의 통계적 노출제어 방안들이 연구되어 왔다고 볼 수 있다.

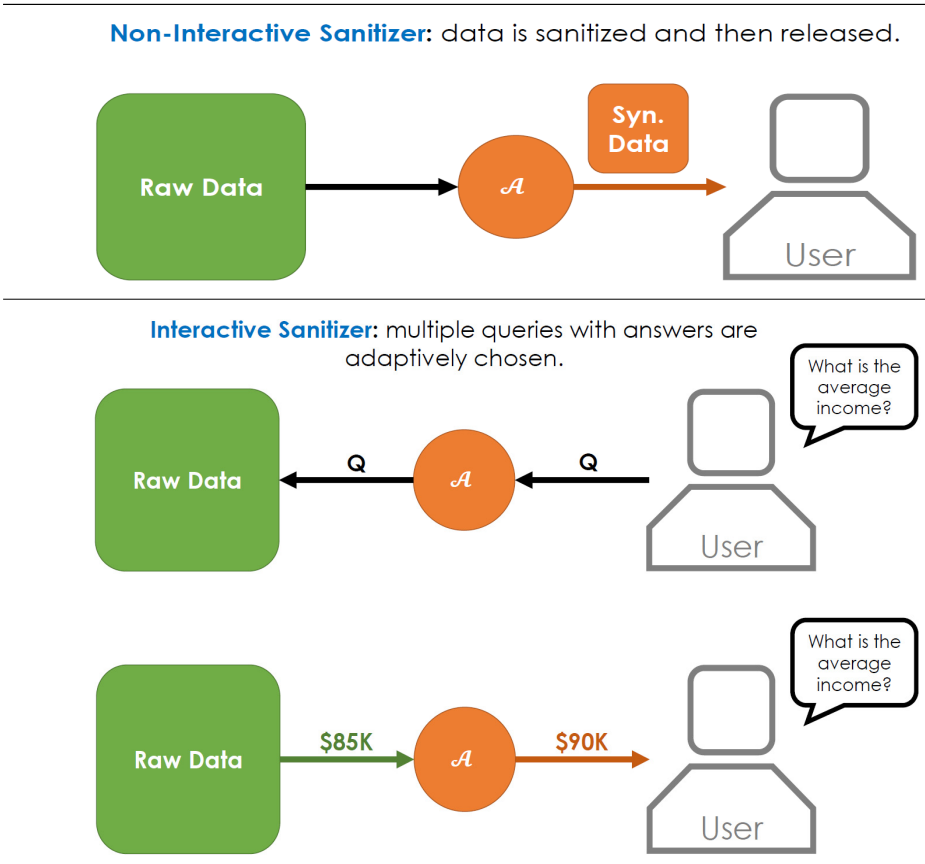
첫 번째 매스킹 방식은 자료 유형별로, 즉 마이크로데이터와 매크로데이터 각각에 대해서 사용하는 기법들이 다르다. 따라서 대다수의 경우, 각 자료별로 매스킹 방안들이 검토되어야 한다. 재현자료 역시 각 자료별로 만들어지는 것이며, 마이크로데이터의 재현자료를 제공하면 매크로데이터는 재현자료를 집계하여 얻을 수 있다. 한편, 매스킹 기법들을 내재한 온라인 분석 시스템(mOAS)의 경우 자료별로 뿐만 아니라, 분석 유형별로도 다양한 경우들을 감안하여 매스킹 기법들을 구현해야 한다. 이 방안으로는 고려해야 할, 즉 자료마다 검토되어야 하는 세부 사항들이 상대적으로 많아서 운영상 복잡함이 크다고 판단된다. 다만 매스킹이나 재현자료와 다르게 마이크로데이터를 직접 제공하지 않으며, 자료 이용자가 시스템에서 허용된 분석을 수행하고 매스킹 기법을 통해 노출제어 처리된 분석 결과물을 수령하므로 새로운 자료제공 방식이라고 할 수 있다.

분석 시스템을 통해 마이크로데이터를 제공하려는 노력을 시스템적 대안 모색이라고 명명하자. 다수의 문헌에서는 이를 상호적(interactive) 자료제공 방식이라고 표현하기도 하며, 기존의 비상호적(non-interactive) 자료제공 방식과 구분하여 설명하고 있기도 하다.

---

6) 매스킹 처리를 하면 노출위험이 감소하나, 정보손실이 증가한다. 또, 정보손실을 줄이기 위해 매스킹 처리를 적게 하면 노출위험이 증가한다. 즉, 매스킹 처리를 통해서는 노출위험과 정보손실이 항상 상충관계(trade off)에 있게 된다.

아래 그림은 2016년 북미연합통계학술대회에서 발표된 내용 일부를 인용한 것인데 자료제공 방식의 변화를 잘 표현하고 있다. 매스킹 기법을 사용하든 재현자료를 사용하든, [그림 4-3] 맨 위에서 보여주는 비상호적인 자료제공 방식에서는 통계기관이 노출제어 처리로 변형된 자료를 제공하는 역할을 했었다. 하지만 [그림 4-3] 아래편에 보이는 상호적인 자료제공 방식에서는 이용자의 요청(쿼리)을 받아서 개별정보 노출이 일어나지 않도록 변형된 분석 결과물을 제공한다. 이러한 상호적 자료제공 방식은 자료의 대용량화가 일상화되어 가고 있는 빅데이터 시대에 더욱 일반적인 자료제공 방식이 되어갈 가능성이 크다. 즉, 빅데이터 시대에는 통계적 노출제어 분야에서 시스템적 대안 모색을 추구해야 할 필요성이 커진다고 볼 수 있다.



출처: Bowen, C. M. and Liu, F.(2016), "Partitioning in Diffrentially Private Data Synthesis for Big Data", Joint Statistical Meetings 2016.

[그림 4-3] 비상호적 vs. 상호적 자료제공 방식

지금까지 차등정보보호를 제외한 통계적 노출제어 기법들을 요약하고, 빅데이터 시대를 맞이하여 자료제공 방식 변화 가능성을 살펴보았다. 차등정보보호 활용을 제외한 통계적 노출제어 기법들은 대부분 자료별로 수행되어야 한다. 예를 들어, 노출위험을 낮추기 위해 잡음을 추가하거나 재현자료를 만드는 등의 과정은 주어진 자료의 분포를 고려해 이루어진다. 노출위험 측도 역시 마찬가지로 자료의 유형에 따라 다르게 정의되는 경우가 많다. 한편, mOAS 방식은 시스템적 대안 모색에 속하기는 하지만 여전히 자료별로 여러 상황들을 함께 고려해야 하는 복잡함을 가지고 있다. 즉, 각 방안들이 공통적으로 자료별로 노출제어 처리를 해야 하는 한계를 가지고 있다고 할 수 있겠다.

차등정보보호를 이용한 자료제공 방식은 mOAS 방식처럼 온라인 분석 시스템을 통해 자료를 제공하며, 마이크로데이터를 파일 형태로 직접 제공하는 것이 아니고 이용자가 시스템 상에서 자료를 분석하므로 시스템적 대안 모색에 속한다고 할 수 있다. 차등정보보호가 가지는 기존 방식들과의 차이점은 분석 결과물을 제공할 때 대체로 자료별로 노출제어 처리를 고려하지 않으며, 분석 기법/유형별로 차등정보보호를 만족하는 결과물을 제공하고자 한다는 것이다. 따라서 구현하기가 상대적으로 단순하며, 이론적 근거를 갖추고 있는 장점이 있다. 아래 표는 각 방식의 장단점들과 공통점들을 정리하였다. 표에서 이론적 난이도가 높다는 것은 통계기관에서 해당 기업을 이해하고 활용하는데 있어 전문적 지식을 가진 담당자의 역할이 필요할 수 있다는 의미이다.

〈표 4-2〉 네 종류의 통계적 노출제어 방식

매스킹 기법(M) · 전통적으로 널리 활용 · 한계점 해결 불가능	온라인분석시스템(mOAS) · 자료제공범위 확대 가능 · 구현과정이 복잡	M 및 mOAS의 공통점 · 매스킹 기법의 활용 · 자료별 처리 필요
재현자료(SD) · 파일 제공 가능 · 이론적 난이도가 높음	차등정보보호(DP) · 상대적으로 높은 범용성 · 이론적 난이도가 높음	SD 및 DP의 차이점 · SD는 자료별 처리 · DP는 분석기법별 처리
M 및 SD의 공통점 · 마이크로데이터 파일제공	mOAS 및 DP의 공통점 · 시스템을 통한 자료분석	최신동향 · SD 및 DP 관련연구 활발

이제 이러한 통계적 노출제어의 연구 흐름 속에서 등장하게 되는 차등정보보호에 대해 살펴보자. 먼저 기존의 정보보호 개념과 차등정보보호의 의미를 비교하고, 간략한 수식을 통해 차등정보보호의 정의를 정리한다. 다음으로 차등정보보호를 구현하기 위해 등장하는 분석 함수의 민감도를 설명하여 차등정보보호 구현 방식을 이해해보도록 하겠다.

1. 정보보호의 신개념

통계적 노출제어는 개별정보보호를 구현하기 위한 방법들을 통칭하는 용어라 할 수 있다. 따라서 통계적 노출제어의 목적이 되는 개별정보보호의 정의는 무척 중요하다. 노출위험 측도가 꾸준히 학계의 관심을 받아온 이유는, 적절한 측도가 있다면 그 측도의 값을 좋게 하는 방향으로 방법론들이 개발되는 것이 자연스럽다고 볼 수 있어, 방법론 발전에 있어 일종의 기준이 될 수 있기 때문이다. 마찬가지로 통계적 노출제어 조치를 통해 달성하고자 하는 목적, 개별정보보호란 무엇인지 명확하게 아는 것은 통계적 노출제어의 발전에 무척 중요하다고 하겠다.

오랜 시간동안 각종 통계적 노출제어 기법들이 추구해야 하는 정보보호의 목적으로는 Dalenius가 1977년 제안한 아래의 명제가 여러 문헌들에서 언급되고 있다.

> "데이터베이스에 접근해서 얻을 수 있는 개인에 대한 정보는
> 데이터베이스에 접근 없이도 알 수 있어야 한다."[7]

이 명제는 절대적인 개별정보보호를 의미한다. 노출위험 측도들이 변수 종류별로, 세부 상황별로 만들어질 수밖에 없음에도 불구하고, 각종 기법들과 측도들이 달성해야 하는 목표는 매우 절대적이고 범용적인 수준의 명제로 제시되어 있었다. 적절한 자료제공 방안을 찾아야 하는 통계기관 및 연구자들에게는 길을 막고 있는 벽과도 같은 느낌을 주는 어려움이기도 하다. 더구나 통계기관은 데이터베이스를 통계적으로 활용되게 하려고 제공하는 것인데, Dalenius의 절대적 정보보호는 새로운 정보 습득에 강한 제약을 걸고 있다. 즉 절대적 정보보호는 새로운 정보를 획득하도록 하기 위한 자료제공 목적과 충돌한다고 할 수 있다.

이러한 어려움을 파악하고 새로운 정보보호 개념을 제시한 것이 마이크로소프트사의 컴퓨터 공학자인 Dwork의 2006년 연구이다(Dwork, 2006). Dwork은 Dalenius의 명제가 달성하기 불가능하다는 것을 수학적으로 증명하고,[8] 차등정보보호 개념을 제안하고 구현을 위한 이론을 제시하였다. 차등정보보호는 다음과 같은 명제로 표현해 볼 수 있다.

---

7) "Anything that can be learned from the database about a specific individual should be learned without access to the database." 이 명제는 이후에 암호학 분야에서 다음과 같이 재정의(Dwork, 2006) 되기도 하였다. "Access to a statistical database should not enable one to learn anything about an individual that could not be learned without access."

8) 이는 논문에 'Impossibility of Absolute Disclosure Prevention'이라는 제목 아래 정리되어 있다. 증명에 관심 있는 독자는 관련 논문(Dwork, 2006)을 참고하기 바란다.

> "주어진 쿼리를, 한 개체만 차이나는 두 개의 데이터베이스에 적용했을 때,
> 얻어지는 결과물의 차이를 제한하여 해당 개체의 정보가 드러나지 않게 한다."

차등정보보호는 기본적으로 자료 분석 시스템을 통해 마이크로데이터 파일을 제공하는 것을 가정한다. 또 이용자가 개체를 하나씩 제외/포함하면서 제공받은 마이크로데이터를 공격적으로 분석하고, 분석 결과물을 통해서 그 한 개체의 정보를 알아내는 상황을 가정한다. 이러한 가정 아래서 Dwork은 위의 차등정보보호로써 개별정보보호를 정의하고, 차등정보보호를 달성하도록 분석도구를 제공하는 통계적 방법을 연구할 것을 제안하였다.

복잡한 수식들을 배제하고 차등정보보호를 한마디로 설명한다면, 개인의 데이터베이스 참여 유무가 그 개인의 정보노출을 일으키지 않는다는 것이다. 즉, Dalenius의 절대적 정보보호는 외부인의 데이터베이스 접근에 초점을 맞췄다면, Dwork의 차등정보보호는 개인의 데이터베이스 참여에 초점을 맞추고 있다고 할 수 있다.

간단한 예를 통해 기존의 정보보호 개념과 차등정보보호를 비교하여 이해해보자. "Terry Gross의 키"라는 널리 알려진 예제를 사용하여 아래 표에 두 개념을 비교하였다.

〈표 4-3〉 절대적 정보보호 vs. 차등정보보호

Terry Gross 키 예제	
민감정보	어떤 개인의 키가 얼마인지
데이터베이스	성별×인종별 평균키
보조정보	Terry는 리투아니아 여성 평균키보다 2센티 작다

절대적 정보보호	차등정보보호
데이터베이스 미접속자의 정보 ➛ Terry는 좀 작다 데이터베이스 접속자의 정보 ➛ Terry의 정확한 키	Terry의 데이터베이스 미참여 ➛ Terry의 정확한 키 노출 Terry의 데이터베이스 참여 ➛ Terry의 정확한 키 노출
정보보호 실패	차등정보보호 성공

어떤 개인의 키가 알려지는 것이 민감한 사안이라고, 또한 통계기관은 데이터베이스를 통해 성별×인종별 평균키를 제공한다고 하자. 일반적으로 성별×인종별 평균키를 제공한다고 해서 민감 정보인 특정인의 키가 노출되지는 않는다. 한편, 이 예제에서는 어떤 외부인이 Terry Gross라는 여성이 리투아니아 여성 평균키보다 2센티 작다는 보조정보(auxiliary information)를 가지고 있다고 가정한다.

먼저 Dalenius의 관점에서 이를 살펴보자. 만약 외부인이 데이터베이스에 접근했다면, 그 사람은 Terry Gross의 정확한 키를 알게 된다. 외부인이 데이터베이스에 접근하지 않았다면, Terry Gross는 키가 작은 편이라는 정보만을 아는데 그친다. 따라서 통계기관이 제공한 데이터베이스 접근 유무에 따라 외부인이 얻는 정보가 달라지므로, 통계기관은 Dalenius의 절대적 정보보호를 만족하지 못하는 데이터베이스를 제공한 것이 된다.

다음으로 차등정보보호 관점에서 Terry Gross의 사례를 생각해보자. Terry가 데이터베이스에 참여했다면, 외부인은 Terry의 정확한 키를 알 수 있다. Terry가 데이터베이스에 참여하지 않았어도, 외부인은 Terry의 정확한 키를 알 수 있다. 즉, Terry의 데이터베이스 참여 유무와 상관없이, 외부인은 보조정보를 활용하여 Terry의 정확한 키를 알 수 있으므로 데이터베이스를 통해 얻게 되는 분석 결과물에는 차이가 없고, 통계기관이 제공한 데이터베이스는 차등정보보호를 만족한다.

차등정보보호를 제안한 Dwork은 보조정보의 존재로 인해 Dalenius의 명제는 이루어질 수 없는 것임을 수학적으로 증명하였다. 나아가, 차등정보보호를 정의하고, 차등정보보호를 만족시키며 자료를 제공할 수 있는 방법을 제안하였다. 차등정보보호의 정의를 수식을 통해 살펴보면 다음과 같다.[9)]

먼저 K를 임의의 랜덤화 함수라고 하자.[10)] 또한 D및 D'를 개체(레코드) 하나만 다른 데이터베이스라고 하자.[11)] 그러면 각 데이터베이스의 분석 결과물이 특정한 값 t일 확률의 비율을 일정 수준보다 작게 하는 것이 차등정보보호의 수학적 정의가 된다. 확률의 비율을 제한(ratio bounded)하므로 로그를 취하면 차이에 대해 한계를 주는 형식이 된다. 아래 네 개 수식은 모두 같은 의미를 가진다.

$$\frac{\Pr[K(D)=t]}{\Pr[K(D')=t]} \leq \exp(\epsilon)$$

9) 3절의 나머지 부분들은 차등정보보호를 수식을 사용하여 설명하고 있다. 수식을 좋아하지 않는 독자들은 이 부분을 생략하고 수록된 사례들을 통해 차등정보보호의 개념을 이해할 수 있다.
10) 함수 K의 결과값을 이용자의 쿼리에 대해 시스템이 제공하는 결과값이라고 이해해도 된다.
11) 데이터베이스를 통계기관이 보유하고 있는 마이크로데이터라고 생각할 수 있다.

$$\left|\log\frac{\Pr[K(D)=t]}{\Pr[K(D')=t]}\right| \leq \epsilon$$

$$\Pr[K(D)=t] \leq \Pr[K(D)=t] \cdot \exp(\epsilon)$$

$$\log(\Pr[K(D)=t]) - \log(\Pr[K(D)=t]) \leq \epsilon$$

이 수식들은 데이터베이스가 D에서 D'로 바뀌어도 시스템이 제공하는 분석 결과물 K의 결과값들은 ϵ에 의해 정해지는 범위 내에서만 차이를 가질 수 있다는 의미로 해석할 수 있다. 허용되는 차이의 범위는 ϵ에 의해 결정되므로 ϵ을 노출허용 수준/한계[12]를 규정하는 모수라고 해석하기도 한다. 주어진 ϵ의 값이 작으면 데이터베이스가 바뀔 때 결과값 분포의 차이가 작게 유지되므로, 두 데이터베이스 분석 결과 차이에 의해서 개별 정보가 노출될 위험이 작아지면서 강한 수준의 정보보호를 의미하게 된다. 반대로 ϵ의 값이 크면 데이터베이스에서 한 개체가 있고 없음에 따른 결과값의 차이를 더 많이 드러낼 수 있으므로 해당 개체의 값을 역으로 계산할 가능성도 커지는 것이 된다. 따라서 개별 정보가 노출될 위험이 커지고 약한 수준의 정보보호를 의미한다. 아래 <표 4-4>는 이러한 내용을 정리하였다.

〈표 4-4〉 차등정보보호 정의 및 의미

정의	$\log(\Pr[K(D)=t]) - \log(\Pr[K(D)=t]) \leq \epsilon$
의미	ϵ이 작음 ↔ 분석 결과물 분포의 차이가 작음 ↔ 강한 정보보호
	ϵ이 큼 ↔ 분석 결과물 분포의 차이가 큼 ↔ 약한 정보보호

2. 차등정보보호 구현

이제 차등정보보호를 만족시키는 분석 결과물을 제공하기 위한 세부 과정에 대해 알아보자. 앞에서 차등정보보호는 분석 유형/기법별로 적용될 수 있다고 설명하였었다.

분석도구, 즉 사용자가 요청하는 쿼리를 f라고 하자. 사용자의 쿼리에 데이터베이스를 입력하면 사용자가 원하는 분석 결과는 $f(D)$및 $f(D')$로 표현된다. 만약 f가 평균을 제공하는 분석도구이고, $f(D)$및 $f(D')$의 값들이 그대로 제공되면 두 값의 차이를 이용해 D에는 있고 D'에는 없는 한 개체의 값을 거꾸로 계산해낼 수 있다. 이러한 추적이

12) privacy leakage, privacy budget (amount of differential privacy used) 표현을 번역하였다.

가능하지 않도록 f에 적절한 조치를 취하는 메커니즘, 혹은 랜덤화 함수를 K_f라고 하자. 이 메커니즘이 차등정보보호를 만족하도록 분석 결과를 제공한다는 것은 K_f가 위의 차등정보보호 정의를 나타내는 수식을 만족한다는 것을 의미한다. Dwork은 f에 잡음을 추가하는 메커니즘 K_f를 사용하면 차등정보보호를 만족한다는 것을 증명하였다. 즉, $K_f = f + noise$이고 $noise \sim Lap(\Delta f/\epsilon)$이면 메커니즘 K_f는 차등정보보호를 만족하며, 이를 라플라스 메커니즘이라고 부른다.

참고로 차등정보보호의 잡음추가를 매스킹 기법의 잡음추가와 비교해보자. 매스킹에서는 자료의 분포와 유사한 분포를 가지는 잡음을 추가한다. 따라서 변형된 마이크로데이터의 분포가 원래 마이크로데이터의 분포와 유사하게 만들어지는 것을 추구한다. 반면에 차등정보보호의 잡음추가는 분석 결과물에 분석 기법별로 잡음을 추가하는 것이며, 자료별 분포를 고려하지 않는다.

이제 차등정보보호에서 추가되는 잡음의 의미를 ϵ 및 Δf와 관련하여 이해해보자. 라플라스 분포를 $Lap(b)$라고 표현하면 분산은 $2b^2$이므로, 차등정보보호를 만족하기 위한 라플라스 메커니즘 잡음의 분산은 $(\Delta f)^2/\epsilon^2$로 표현된다. 따라서 ϵ이 작으면 잡음의 분산이 커진다. 이는 잡음을 많이 더하여 참값을 가리고 데이터베이스 변화에 따른 결과값의 차이가 드러나지 않게 하는 것이므로 강한 정보보호를 의미한다. 반대로 ϵ이 크면 잡음의 분산이 작아지고, 참값에 추가되는 잡음이 작아지므로 데이터베이스 변화에 따른 결과값의 차이가 더 잘 드러나게 된다. 즉 약한 정보보호를 이루게 된다.

라플라스 메커니즘에서 잡음은 쿼리의 민감도 Δf에 의해서도 영향을 받는다. Dwork은 쿼리, 즉 분석 함수의 민감도를 다음과 같이 정의하였다.

$$\Delta f = \max_{D, D'} \| f(D) - f(D') \|_1$$

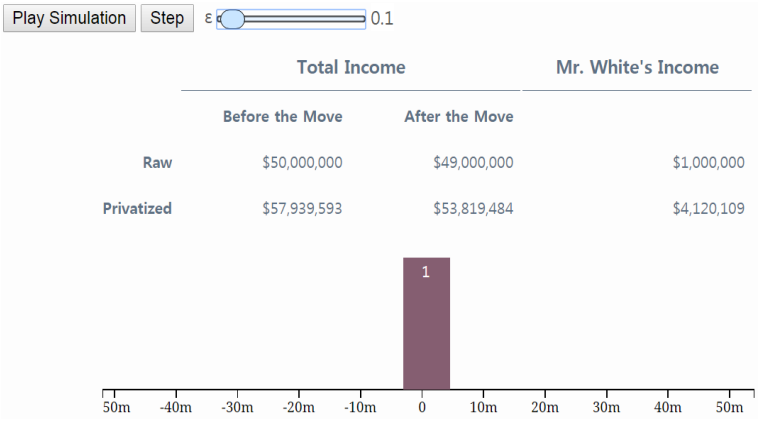
쿼리의 민감도 Δf가 잡음의 분산을 결정하는 구조이므로, 차등정보보호에서는 쿼리, 즉 분석 기법별로 더해지는 잡음의 크기가 달라진다. 쿼리가 평균인 경우 민감도가 작아 잡음의 크기가 작을 수 있지만, 쿼리가 최대값이면 민감도가 커져서 차등정보보호를 만족하기 위해 필요한 잡음의 크기가 커지게 된다.

차등정보보호를 이해하기 쉽도록 구현한 온라인 포스트[13]가 있어서 그 내용을 이용해 차등정보보호를 구현하는 과정을 좀 더 설명해보자. 이 예제에서 주어진 마이크로데이터는 각 개인의 수입액으로 구성되어 있다. 또, 이용자는 분석도구를 이용하여 전체 구성원의

13) research.neustar.biz/2014/09/08/differential-privacy-the-basics/ 다수의 논문들보다 이 온라인 자료가 차등정보보호에 대한 일반적인 이해를 도울 수 있다고 판단하여 본 보고서에서 활용하였다.

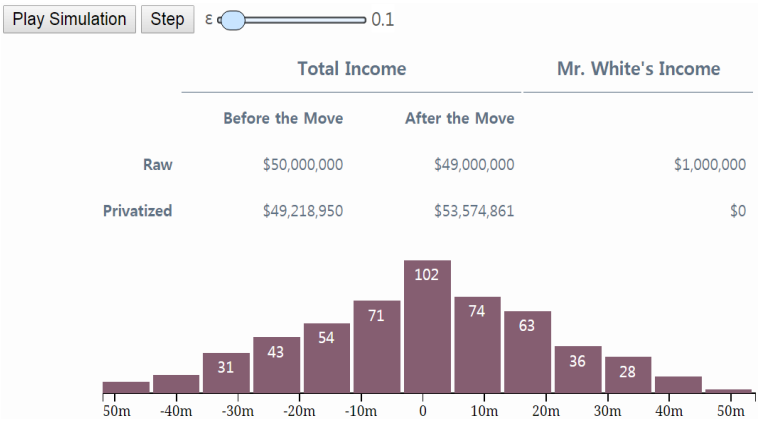
수입합을 얻을 수 있고, 전체 수입합 참값은 오천만 불이다. 분석을 위해 제공되는 마이크로데이터에서 Mr. White의 수입은 백만 불이며, 이 데이터에서 최고 수입액을 가지는 개체이다. 이제 마이크로데이터 분석을 통해 White씨의 수입이 노출되지 않도록 차등정보보호를 구현한 결과를 살펴보자. 총합 제공이라는 분석도구의 민감도는 이미 결정되어 있으므로, 노출에 관한 모수 ϵ을 바꾸어가며 분석 결과물의 변화를 그림들에 나타냈다.

먼저 아래 [그림 4-4]는 ϵ값이 작아 강한 정보보호를 하는 경우이다. 그림에는 원래 마이크로데이터(Raw) 전체 수입합 오천만 불과 차등정보보호를 만족하도록 라플라스 메커니즘을 적용한 경우(Privatized)의 전체 수입합 오천팔백만 불이 제시되어 있다. 한 개체 White씨의 수입액을 제외한 데이터베이스에 대해 전체 수입 총합 사천구백만 불이 제공된다면 자료 이용자는 White씨의 수입액 참값 백만불을 알아낼 수 있다. 하지만 라플라스 메커니즘을 적용하면 White씨가 없는 데이터베이스의 수입 총합은 오천 사백만불 가량이 되어, White씨의 수입은 약 사백만불일 것으로 엉터리로 계산된다.



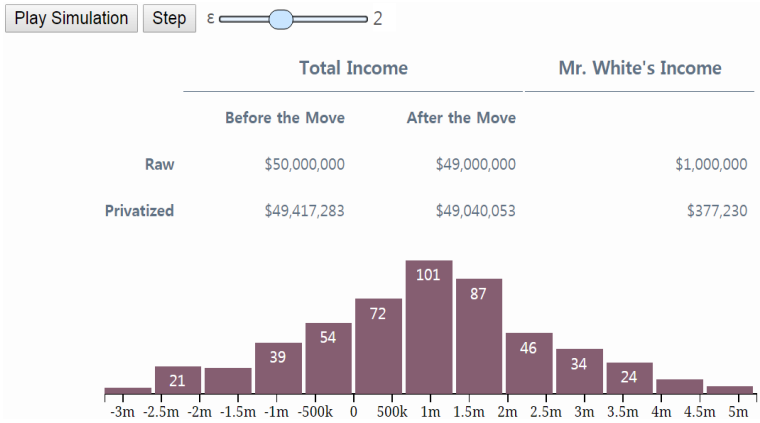
[그림 4-4] 분석도구가 총합일 때 차등정보보호 구현(ϵ=0.1, 1회 실험)

자료에 반복적으로 접근하여 모의실험을 하면 다음 [그림 4-5]와 같은 결과를 얻는다. 반복적인 계산을 통해서도 White씨의 수입은 오백만 불 범위에서 가장 많이 나타나 참값을 추정하는 것이 어려움을 알 수 있다.



[그림 4-5] 분석도구가 총합일 때 차등정보보호 구현(ϵ=0.1, 반복 실험)

다음으로 ϵ값이 커서 약한 정보보호를 하는 경우 때의 모의실험 결과를 살펴보자. 아래 [그림 4-6]의 히스토그램을 통해 자료 이용자는 ϵ값이 클 때는 White씨의 수입이 백만 불 근처라는 결론을 얻을 수도 있다.



[그림 4-6] 분석도구가 총합일 때 차등정보보호 구현(ϵ=2.0, 반복 실험)

마지막으로 차등정보보호를 구현하는 다른 메커니즘들을 소개한다. 기존의 ϵ-차등정보보호의 단점은 지나치게 보수적이라는 것이다. 위의 총합 분석도구 예제에서 살펴보아도 강한 정보보호에서는 정보손실이 너무 크다는 것을 알 수 있다. 이러한 경향은 분석 함수의 민감도가 클 때 더 드러난다.

이러한 단점을 보완하기 위하여 아래 수식으로 표현되는 (ϵ,δ)-차등정보보호가 제안되었다. 이것은 극단값들에 노출허용 한계(privacy leakage)를 좀 더 주는 것을 의미한다.

$$\Pr[K(D)=t] \leq \Pr[K(D)=t] \cdot \exp(\epsilon)+\delta$$

또한, 라플라스 메커니즘을 보완하여 지수적 매커니즘(exponential mechanism)이 제안되는 등, 차등정보보호 관련 이론들이 계속하여 발전하고 있다.

제4절 차등정보보호 적용 사례

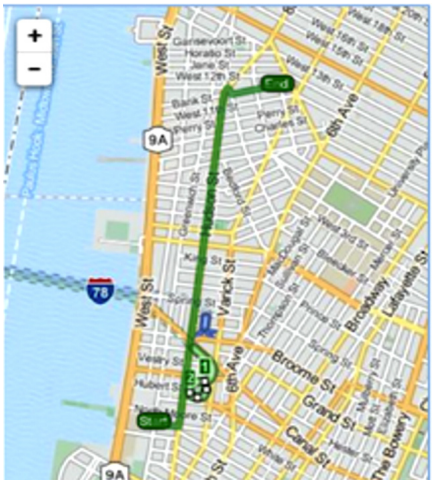
이제 차등정보보호를 적용하여 자료를 제공하는 사례를 살펴보자. 사례를 통해 차등정보보호를 적용하여 자료를 제공하는 경우의 장단점을 생각해볼 수 있다.

2013년에 축적된 약 4만명 가량의 뉴욕시 택시 운전자들의 운행 승하차 장소, 시간, 요금 및 팁으로 구성된 마이크로데이터가 익명화 처리 후 제공[14]되었다. 이 자료를 통해 개별 운전자의 연간 수입을 계산하는 것이 가능하고, 따라서 택시 운전자들의 임금 결정을 위한 근거 자료로 사용될 수 있으므로 이 마이크로데이터는 정책적으로 유용성이 큰 데이터베이스라고 할 수 있다. 하지만 승객의 입장에서도 택시 운전자의 입장에서도 이 데이터베이스를 통해 개별정보가 노출될 위험이 있다. 이에 관해 분석한 온라인 포스트[15]의 내용을 정리하여 차등정보보호를 적용하여 자료를 배포하는 것을 이해해보자.

1. 뉴욕시 택시 운전자 데이터베이스 : 승객의 개별정보보호

먼저 택시의 승객인 미국 어느 유명 배우 A씨의 사생활이 이 데이터베이스를 통해 드러날 수 있음을 확인할 수 있다. 이 자료를 분석한 외부인은 가십 블로그에 떠도는 A씨의 택시 승차 사진과 가십 내용을 통해 배포된 데이터베이스에서 다음 [그림 4-7]과 같은 A씨의 이동 경로를 찾아내었다.

14) New York City Taxi and Limousine Commission에 의해 공표되었다.
15) research.neustar.biz/author/atockar



[그림 4-7] A씨의 이동 경로

한편, A씨 개별정보 노출을 위한 분석에는 다음과 같은 포인트 쿼리가 사용되었다.

```
SELECT D.dropoff_latitude, D.dropoff_longitude, F.total_amount, F.tip_amount
FROM tripData AS D, tripFare AS F
WHERE D.hack_license = F.hack_license AND D.pickup_datetime = F.pickup_datetime
  AND pickup_datetime > "2013-07-08 19:33:00" AND pickup_datetime < "2013-07-08 19:37:00"
  AND pickup_latitude > 40.719 AND pickup_latitude < 40.7204
  AND pickup_longitude > -74.0106 AND pickup_longitude < -74.01;
```

이 쿼리를 사용해 외부인은 A씨의 승하차 위치, 시간, 요금, 팁을 알 수 있었으며, A씨가 지급한 요금은 $10.5, 팁은 $0인 것으로 발견되었다. 이에 따라 적절한 팁을 지급하는 것이 중요하게 여겨지는 미국 사회에서 A씨에 대한 가십거리가 늘었다고 할 수 있다. 따라서 이것은 익명화 처리만 하고 자료를 배포할 경우에 발생할 수 있는 개인정보 침해 사례에 해당한다고 볼 수 있다.

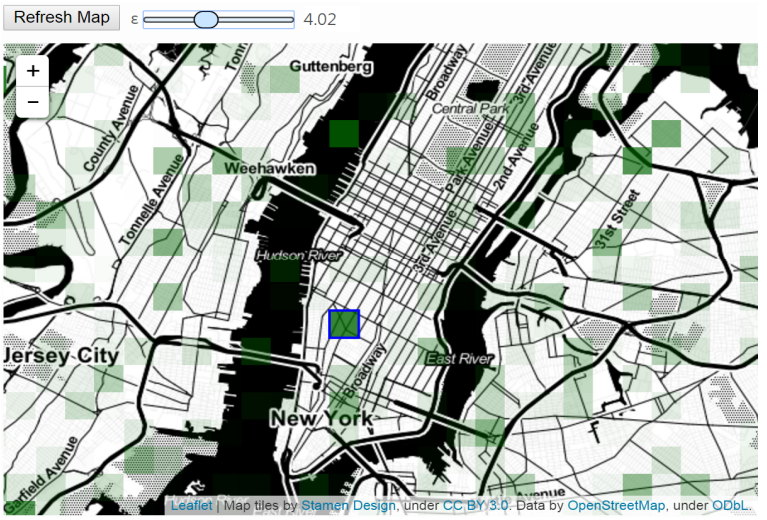
이제 차등정보보호를 적용하여 자료를 제공하면 어떻게 달라지는 알아보자. 먼저 차등정보보호를 만족하기 위해, 특정 시간 및 장소에서 얼마나 많은 택시가 승객을 태웠는지를 묻는 총합 쿼리(aggregation query)를 허용한다. 또 결과는 하나씩 제공하지 않고 격자 형태로 제공한다. 즉, 쿼리를 집계 형식으로 받고 결과물은 격자 형태로 제공하여 차등정보보호를 만족하도록 할 수 있다. 결과물의 형태는 다음 [그림 4-8]과 같게 된다.

노출허용 모수 ϵ을 0.5로 하여 어느 정도 강한 정보보호를 하도록 자료를 제공하면 A씨의 승하차 장소는 이 지도와 같이 제공된다. 색이 진할수록 승하차 확률이 높은 장소를 의미하며, 진짜 승하차 장소는 두꺼운 파란색 선으로 표시된 네모 구역이다. 추정되는 요금은 $25~$30, 팁은 $16~$18이 된다.



[그림 4-8] 차등정보보호를 적용하여 제공되는 A씨의 이동 경로 관련 정보(ϵ=0.5)

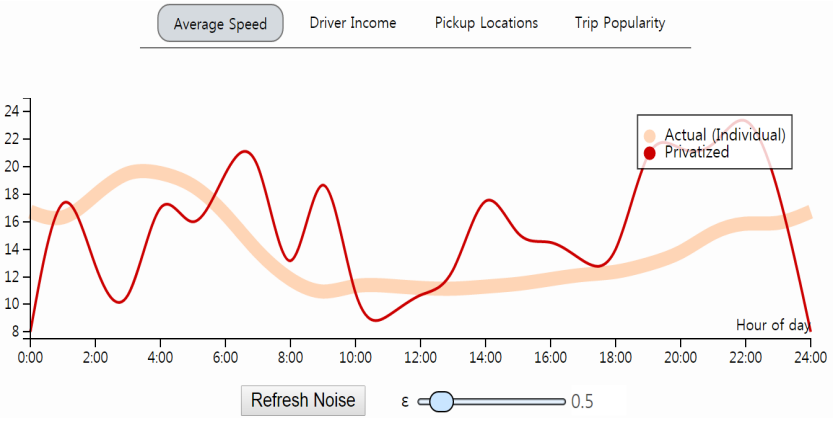
한편, 노출허용 모수 ϵ을 4.02로 하여 약한 정보보호를 적용하면 아래 [그림 4-9]와 같은 결과를 얻는다. 승하차 장소도 진짜 승하차 장소를 추정할 수 있도록 나타나며, 요금은 $10~$15, 팁은 $0~$2로 비교적 참값과 유사하게 나타남을 알 수 있다.



[그림 4-9] 차등정보보호를 적용하여 제공되는 A씨의 이동 경로 관련 정보(ϵ=4.02)

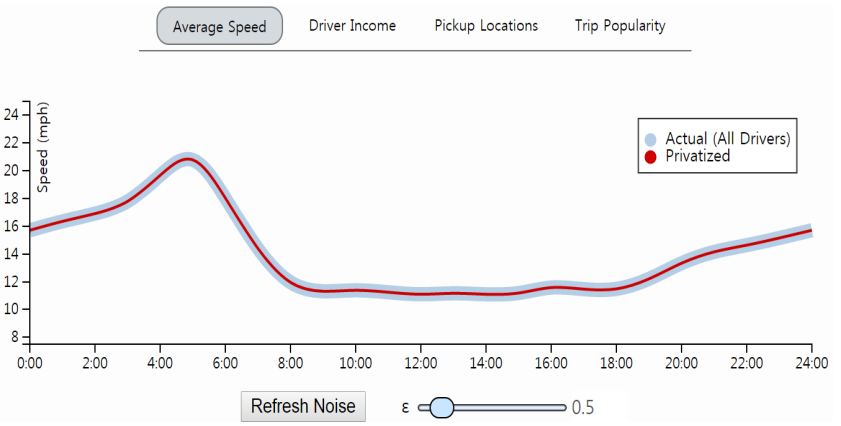
2. 뉴욕시 택시 운전자 데이터베이스 : 운전자의 개별정보보호

다음으로 택시 운전자의 개별 정보를 보호하기 위해 차등정보보호를 적용한 결과를 살펴보자. 택시 운전자에게는 시간대별 운전 속도 및 수입이 개별민감정보가 된다. 먼저 강한 정보보호를 적용한 경우이다. 아래 [그림 4-10]은 한 택시 운전자의 실제 시간대별 평균 운전 속도를 나타낸다. 연한 색 띠로 나타난 것은 원래 자료에서 얻은 이 운전자의 시간대별 평균 운전 속도이다. 새벽 시간대에 빠르게 달리다가, 교통량 증가로 인해 오전에 속도가 감소하고, 밤 시간이 되면 다시 속도가 점점 증가하는 것을 알 수 있다. 가늘고 진한 선은 차등정보보호 처리된 운전자의 평균 운전 속도이다. 원자료의 정보와 전혀 유사하지 않음을 알 수 있다. 이 경우 개별 운전자의 운전속도 정보가 잘 보호되고 있다고 하겠다.



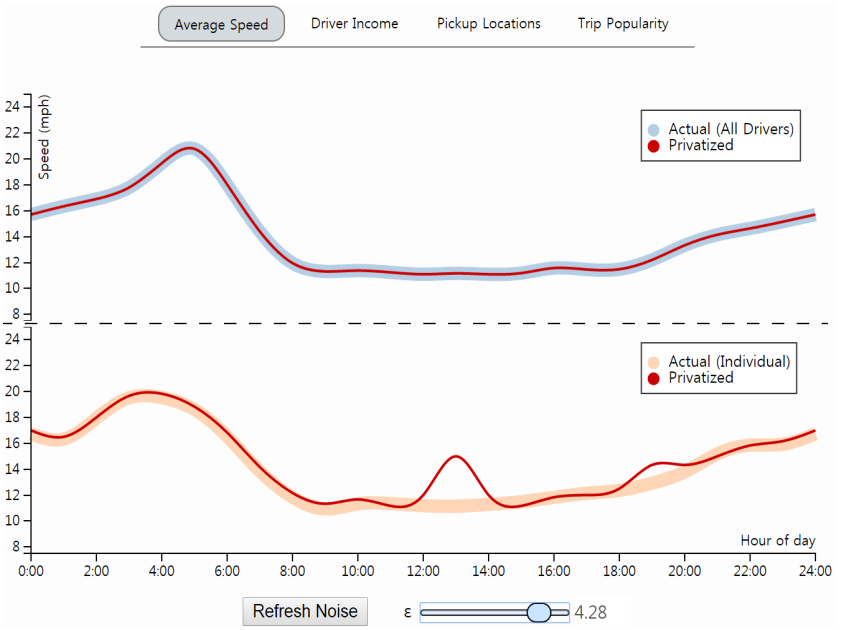
[그림 4-10] 차등정보보호를 적용하여 제공되는 택시 운전자 1인의 평균 속도(ϵ=0.5)

다음 [그림 4-11]은 전체 택시 운전자의 실제 시간대별 평균 운전 속도를 나타낸다. 위의 [그림 4-10]과 마찬가지로 연한 색 띠는 원래 자료에서 얻은 시간대별 평균 운전 속도를 나타낸다. 또, 두꺼운 선은 차등정보보호 처리된 전체 운전자의 평균 운전 속도를 나타낸다. [그림 4-10]과는 다르게 이 두 그래프가 거의 일치하므로, 우리는 개인별 자료를 제공할 때는 차등정보보호 처리로 인한 정보손실이 매우 컸으나, 전체 집단에 대한 자료 분석에서는 차등정보보호 적용 여부가 분석 결과에 별다른 영향을 끼치지 않음을 알 수 있다.



[그림 4-11] 차등정보보호를 적용하여 제공되는 전체 택시 운전자들의 평균 속도(ϵ=0.5)

같은 자료에 약한 정보보호를 할 경우 아래 [그림 4-12]와 같은 결론을 얻을 수 있다. 아래 그림의 개별 운전자 평균 운전속도는 참값에 가깝게 드러나고, 위쪽 그림에서는 전체 운전자의 평균 속도가 강한 정보보호 조치를 한 경우와 큰 차이가 없음을 알 수 있다.



[그림 4-12] 차등정보보호를 적용하여 제공되는 택시 운전자들의 평균 속도(ϵ=4.28)

결론적으로, 차등정보보호를 만족하도록 라플라스 매커니즘을 적용하여 자료를 제공하면, 이 사례에서는 노출허용 모수의 수준에 따라 개별정보는 보호할 수 있으며, 노출허용 모수 값에 상관없이 자료 전체에 대해서 정보손실이 거의 없는 분석 결과를 얻을 수 있음을 알 수 있다.

다음으로 운전자의 연간 수입에 대해 강한 차등정보보호를 적용한 결과를 살펴보자. 아래 [그림 4-13]에서는 10명의 개별 운전자 연간 수입을 막대 형태로 나타내고 있다. 첫 번째 운전자는 연간 수입이 $50,000불이지만 차등정보보호 처리되어 제공되는 결과를 통해서는 $105,000불로 나타난다. 이와 같이 개별 운전자 연간 수입의 참값과 차등정보보호 처리 이후 제공되는 값들의 차이가 큼을 알 수 있다. 반면에 중앙 부분의 띠와 선으로 표현되어 있는 것은 전체 운전자의 연간 수입 평균을 나타낸다. 차등정보보호 처리로 인한 차이가 없음을 알 수 있다.



[그림 4-13] 차등정보보호를 적용하여 제공되는 택시 운전자들의 평균 속도(ϵ=0.5)

약한 정보보호를 적용한 경우는 다음 [그림 4-14]에 나타나 있다. 개별 연간 수입이 참값과 유사하게 드러날 수 있다. 운전자 평균 운전 속도와 마찬가지로 차등정보보호 노출허용 수준에 따라 개별정보에 대한 정보손실은 규모가 달라지며, 전체 자료에 대한 평균값은 노출허용 수준에 상관없이 정보손실이 작음을 볼 수 있다.



[그림 4-14] 차등정보보호를 적용하여 제공되는 택시 운전자들의 평균 속도(ϵ=4.08)

제5절 맺음말

본 연구에서는 통계기관이 통계를 일반에 제공할 때, 개별정보는 보호하고 유용성은 높은 자료를 제공하기 위한 최신 연구 동향에 관하여 '차등정보보호' 위주로 소개하였다. 빅데이터 시대에는 자료 크기 때문에라도 기존의 일방적인(non-interactive) 자료배포 방식의 어려움이 예상되며, 상호적인(interactive) 방식으로의 진화가 예상된다. 이를 대비하여 차등정보보호를 적용한 자료제공에 관해 국내에 쉽게 소개하여, 통계기관의 자료제공 전략 수립에 도움이 되는 것을 목적으로 본 보고서를 작성하였다.

노출제어 처리된 자료제공의 방법은 크게 네 종류로 분류해볼 수 있다. 우선 매스킹 기법들이란 주로 국가 통계기관들에서 사용하고 있는 전통적인 다양한 기법들을 일컫는다. 매스킹 기법들만으로는 노출위험과 정보손실을 동시에 축소시키기는 어려우므로, 이를 극복하기 위한 방안으로 온라인 분석 시스템과 매스킹 활용, 재현자료 작성 및 활용, 차등정보보호 적용이 연구되어 왔다. 2절에서는 매스킹 기법들, 노출위험과 정보손실 측도, 온라인 분석 시스템과 재현자료를 소개하였다. 3절에서는 각 방안들을 비교하고, 차등정보보호의 개념과 이론적인 내용을 최대한 쉽게 설명하였다.

이어 4절에서는 차등정보보호를 적용하여 자료를 제공하는 사례를 살펴보았다. 뉴욕 택시 운전자 데이터베이스 사례에서, 차등정보보호를 적용하면 노출허용 모수 ϵ의 선택에

따라 개별 운전자에 대한 정보 노출을 제어할 수 있었다. 또한, 노출허용 수준에 상관없이 전체 자료에 대한 정보는 정보손실이 거의 없이 제공할 수 있었다. 즉, 데이터베이스를 제공하는 목적, 모집단에 관한 정보 제공에 성공적이면서 개별정보는 보호하는 효과를 보임을 알 수 있다.

차등정보보호의 적용이 모든 경우에서 성공적인 것은 아니며, 현재 학계 및 다양한 기관에서 연구가 진행 중이다. 또한 국가통계 제공에 있어 어떠한 방식으로 활용될 수 있는 방법인지에 대한 연구도 더 필요하다. 향후 통계기관들이 적절한 노출제어 방안을 확보해 가는데 있어 본 보고서가 적절히 활용될 수 있기를 기대한다.

참고문헌

박민정, 권순필, 심규호(2013), 『가계금융·복지조사 마이크로데이터 제공을 위한 매스킹 방안』, 통계개발원.

박민정(2014), 『가계금융·복지조사 매스킹 방안 평가』, 통계개발원.

박민정(2015), 『마이크로데이터 매스킹 기법 실무 활용 안내서』, 통계개발원.

Bethlehem, J. G., Keller, W. J. and Pannekoek, J.(1990), "Disclosure Control of Microdata", Journal of the American Statistical Association, 85(409), pp. 38~45.

X Danker, F. K. and Emam, K. E.(2013), "Practicing Differential Privacy in Health Data: A Review", Transactions on Data Privacy, 5, pp. 35-67.

Duncan, G. T. and Lambert, D.(1989), "The Risk of Disclosure for Microdata", Journal of Business and Economic Statistics, 7, pp. 207-217.

Duncan, G. T., Elliot, M. and Gonzalez J. J. S.(2011), "Statistical Confidentiality: Principles and Practice", Springer.

Dwork, C.(2006), "Differential Privacy", in Automata, Languages and Programming, vol. 4052, Bugliesi, M., Preneel, B., Sassone, V. and Wegener, I., Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1-12.

Krenzke, T., Gentleman, J. F., Li, Jianzhu and Moriarity, C.(2013), "Addressing Disclosure Concerns and Analysis Demands in a Real-Time Online Analytic System", Journal of Official Statistics, 29, pp. 99-124.

Park, M.-J. and Kim, H. J.(2016), "Statistical Disclosure Control for Public Microdata: Present and Future", The Korean Journal of Applied Statistics, 29(6), pp. 1041-1059.

Reiter, J.(2005), "Estimating Risks of Identification Disclosure in Microdata", Journal of the American Statistical Association, 100(472), pp.1103-1112.

Skinner, C. and Shlomo, N.(2008), "Assessing Identification Risk in Survey Microdata Using Log-Linear models", Journal of the American Statistical Association, 103(483), pp. 989-1001.

<부록>

1. 의사결정론 기반 노출위험 측도

키변수 조합의 유일성을 활용하는 것 이외의 노출위험을 측정하는 또 다른 접근은 통계적 의사결정이론(decision theory)을 활용하는 것이다. 외부인이 노출시키려는 목표 개체를 t라고 하자. 외부인이 가진 t에 관한 정보에 대해 적절한 가정을 세우고, 통계기관이 노출제어 처리를 시행한 후에 공표한 마이크로데이터 Z에서 각 개체들이 t일 확률을 단계적으로 조건부로 얻어가는 과정을 통해 노출위험을 측정할 수 있다. 이러한 외부인의 의사결정에 근거한 노출위험은 Duncan과 Lambert(1989)에서 출발한다고 할 수 있고, Reiter(2005)에서는 이 방식을 발전시켜 미국의 경제활동인구조사(Current Population Survey, CPS)에 적용한 결과를 설명하고 있다.

외부인에 대해 의사결정이론을 활용한 노출위험 측정은 여러 장점을 가지고 있지만, 실무에서 활용하기에는 이론적인 난이도가 높은 어려움이 있다. 또한 프로그램으로 구현하기에는 일반화시키기 어려운 단점을 가지고 있기도 하다. 때문에 자체적으로 프로그래밍이 가능한 전공자가 아닌 이상은, 실제 자료의 노출위험을 측정할 때는 이미 R패키지 등에서 구현되어 있는 유일성 기반 노출위험 측도들을 사용해야 한다. 실제 자료에 각 노출위험 측도를 적용한 연구 결과를 아래에서 계속하여 소개하므로, 관심 있는 독자는 참고하기 바란다.

2. 노출위험 측도 활용 사례

여기서는 본문과 '<부록> 1.'에서 언급한 두 종류의 노출위험 측도를 활용한 연구 결과를 소개하도록 하겠다. 모집단 유일성 기반 노출위험 측도를 활용한 경우에 대해 두 가지 분석 결과를, 의사결정이론에 기반한 경우에 대하여 한 가지 분석 결과를 정리하였다.

먼저 모집단 유일성 기반 노출위험 측도의 경우에 대해, 다음 <부표 4-1>은 2012년 통계청의 가계금융·복지조사 마이크로데이터 연구 사례를 보여주고 있다. 여기서는 12개 변수를 대상으로 유일성에 근거하여 노출위험을 측정하였다(박민정 외, 2013). 변수 이름은 각각 시도(V1), 성별(V2), 연령(V3), 교육정도(V4), 종사상지위(V5), 동거여부(V6), 혼인상태(V7), 직업(V8), 가구원수(V9), 주택유형(V10), 주거면적(V11), 주택소유형태(V12)이며, 표에서 각 변수를 포함할 때는 1, 포함하지 않을 때는 0으로 표시하였다.

표본에 포함된 개체수는 19,744이며, U1은 1로 표현된 키변수들을 기준으로 표본에서

유일한 개체수를 의미하고, U2는 해당 키변수 조합에 속하는 개체수가 2개인 경우를 나타낸다. 마지막 열의 risk는 가중값을 고려하여 해당 키조합에 대해 각 개체들이 모집단에서 유일할 확률(%)의 평균을 나타낸다. 즉, 개체별 노출위험의 평균값을 마이크로데이터 파일의 노출위험으로 사용하여 표에 나타내었다.

첫 번째 표에서 S1은 V2~V5, 총 4개 변수를 키변수로 사용하여 노출위험을 측정한 경우를 나타낸다. 전체 19,744개 중에서 유일한 개체수는 80, 개체수가 2개인 경우는 126이며, 노출위험은 0에 가깝다. 두 번째 표의 S1은 V2~V5, 총 4개 변수에 지역(시도) 변수 V1을 추가한 경우이다. 유일한 개체는 1938로 증가하고 노출위험은 0.07의 값을 가지는 것을 볼 수 있다. 즉, 4개 변수만 공표하는 데서 지역 변수를 추가하여 제공하려 할 경우 노출위험이 증가하는 것을 수치로 확인할 수 있다.

〈부표 4-1〉 공표 범위(시도 변수 V1 포함 여부)에 따른 유일한 개체수와 노출위험

(단위: %)

	키변수 포함 여부												노출위험		
	V1	V2	V3	V4	V5	V6	V7	V8	V9	Vl0	Vl1	Vl2	U1	U2	risk
S1	0	1	1	1	1	0	0	0	0	0	0	0	80	126	0
S2	0	1	1	1	1	0	0	0	0	0	0	1	586	578	0.02
S3	0	1	1	1	1	0	0	0	0	0	1	0	722	780	0.03
S4	0	1	1	1	1	0	0	0	0	0	1	1	1962	1472	0.06
S5	0	1	1	1	1	1	1	1	1	1	1	1	11829	2722	0.29

	키변수 포함 여부												노출위험		
	V1	V2	V3	V4	V5	V6	V7	V8	V9	Vl0	Vl1	Vl2	U1	U2	risk
S1	1	1	1	1	1	0	0	0	0	0	0	0	1938	1740	0.07
S2	1	1	1	1	1	0	0	0	0	0	0	1	4854	2826	0.14
S3	1	1	1	1	1	0	0	0	0	0	1	0	5746	3178	0.17
S4	1	1	1	1	1	0	0	0	0	0	1	1	9001	3626	0.25
S5	1	1	1	1	1	1	1	1	1	1	1	1	17138	1647	0.42

두 번째 표의 S5는 12개 변수를 모두 포함할 경우, 전체 19,744개 중에서 유일한 개체는 17,138이고, 파일 수준의 노출위험은 0.42가 됨을 보여준다. 이를 첫 번째표 S5의 시도 변수를 포함하지 않은 경우의 노출위험인 0.29에 비교하면 약 1.5배 증가한 값임을 알 수 있다. 시도변수 포함에 의해 노출위험이 1.5배 가량 증가하기는 하지만, 조사 자료이기 때문에 가중값의 영향으로 노출위험 0.42% 자체가 매우 낮은 수준임을 주시할 필요가 있다.

이 사례에서는 유일성 기반 노출위험은 공표하는 키변수 범위에 따라 달라지며, 표본조사의 경우에는 가중값이 반영되어 비교적 낮은 수준의 노출위험 정도를 보이는 것을 확인할 수 있다. 다만, 유일성 기반 노출위험은 모형에 따라 과소 혹은 과대추정 결과를 보이며, 외부인이 가진 정보를 고려하지 않으므로 보수적으로 활용할 필요가 있다. 참고로, 자료 분석은 음이항 모형을 통해 노출위험을 계산하는 R패키지 sdcMicro를 사용하였다.

지금까지 유일성을 근거로 마이크로데이터의 개체별 노출위험 및 파일 단위의 노출위험을 측정할 수 있음을 설명하였다.[1] 두 번째 예에서는 유일성 기반 노출위험을 활용하여 노출위험을 일정 수준 이하로 유지하는 표본의 크기를 계산할 수도 있음을 설명하겠다. 마이크로데이터를 공개할 때, 노출위험을 기준 값 이하로 유지하기 위해 공표할 수 있는 지역 단위 크기를 결정할 수 있다는 것은 실무적으로 매우 필요한 부분이기도 하다.

주어진 기준 값 이하로 노출위험을 유지하면서 마이크로데이터 파일을 제공하기 위해, 필요한 최소 표본의 개수를 계산하는 연구 결과는 Bethlehem 외(1990)에 제시되어 있다. 사용된 마이크로데이터는 4개의 변수, H(가구구성), A(나이), M(혼인), S(성별)와 n =8,399개의 레코드로 이루어져 있고, 각 변수 H, A, M, S는 각각 24, 14, 2, 2개의 범주를 가지고 있다. 노출위험 기준 값은 유일성 비율 0.1%로 하였으며, 키변수 종류에 따라 필요한 모집단 최소 크기를 계산한 결과는 다음 <부표 4-2>와 같다. 변수를 H, 하나만 공표할 경우 필요한 모집단의 최소 크기는 1,411이다. 이는 최소 지역 단위에 포함된 개체가 1,411 이상이어야 유일성에 근거한 노출위험 0.1%를 유지할 수 있다는 의미이다. 또한 4개 변수를 안전하게 공표하기 위해서는 46,228보다 작은 인구수를 가지는 지역 단위를 포함해서는 안 된다는 결론을 내릴 수도 있다. 다만, 추정 결과가 관측 결과에 비하여 과소 추정되어 있으므로, 실제로 필요한 모집단 최소 크기는 <부표 4-2>의 결과보다 더 커야할 것으로 짐작된다.

1) 파일 단위의 노출위험은 개체별 노출위험의 평균이나 합을 활용해 계산한다.

〈부표 4-2〉 노출위험 0.1% 이하 유지를 위한 모집단의 최소크기(Bethlehem 외, 1990)

키변수 공표범위	키변수 조합 수 (실제 셀의 수)	표본의 유일성 개수		모집단 최소 크기
		관측 결과	추정 결과	
H	24	0	.08	1,411
H×A	288	23	21.9	14,116
H×A×M	554	50	37.7	29,252
H×A×M×S	1,108	108	80.1	46,228

마지막으로 외부인의 의사결정에 근거한 노출위험 측정 연구 사례로 Reiter(2005)의 CPS 자료 노출위험을 계산한 결과를 살펴보자. 사용된 변수는 총 9개이고, 모집단의 수는 104,781,947로 추정되며, 조사된 자료의 크기는 51,016이다. 노출제어를 위해 통계기관에서 매스킹 처리는 나이(G)의 경우 재코딩, 인종(R)과 혼인상태(M)에 대해서는 자료 교환, 민감변수인 부동산 보유세(P)에 대해서는 잡음추가를 실시하였다.

외부인은 목표 개체인 표적에 대해 성별, 나이, 인종, 혼인상태, 총 4개 변수의 정보를 가지고 있다고 하자. 표적의 종류를 평범한 개체(everyman), 유일한 개체(unique), 고소득자(I), 고액 납세자(P)로 다양하게 가정해 본다. 다음 <부표 4-3>에서는 변수에 대한 설명과 표적 유형별 자료 값을 제시하였다. 마지막 표에서는 통계기관이 아무 조치도 취하지 않은 경우(-), 재코딩만 한 경우(G), 자료교환만 한 경우(R, M), 재코딩 및 자료교환을 모두 한 경우(G, R, M)에 대하여 의사결정이론을 사용하여 노출위험을 측정한 결과를 정리하였다. 괄호 안의 숫자는 표본에서 표적과 같은 키조합을 가지는 개체들의 개수가 나타나 있다. 표적 유형마다 매스킹 처리를 할수록 노출위험이 줄어들고, 자료교환보다는 재코딩의 효과가 큰 경우가 대부분임을 확인할 수 있다.

<부록>을 통해서라도 이렇게 노출위험의 이론적 측면과 실무 활용성을 자세히 설명하는 이유는 차등정보보호라는 개념의 중요성이나 의의를 상대적으로 부각시키기 위해서이다. 기존의 매스킹 체계에서는 노출위험을 측정하는 것이 이처럼 복잡한 문제이며, 범용적인 측도의 부재가 노출제어 방법론 발달을 저해한다는 것이 학계의 인식이었다고 볼 수 있다. 이에 비해 차등정보보호는 노출위험 측도 개념 자체를 완전히 바꾸고, 보다 범용적이고 간단한 기준을 제시하고 있다고 할 수 있다. 다만, 다른 노출제어 접근들과 마찬가지로, 차등정보보호 역시 정보보호를 강화하기 위해서는, 노출제어 과정을 통해 일어나는 정보손실을 증가시킬 수밖에 없다는 공통적인 한계를 가진다.

〈부표 4-3〉 Reiter(2015)의 CPS 자료 설명, 표적의 유형, 노출위험 측정 결과

변수명	범주수 또는 범위
X (sex)	2 (male, female)
R (race)	4 (white, black, american indian, asian)
M (marital status)	7
E (highest attained education level)	16
G (age, years)	15 ~ 90
C (child support payments, $)	0, 1 ~ 23,917
S (social security payments, $)	0, 1 ~ 50,000
P (household property taxes, $)	0, 1 ~ 99,997
I (household income, $)	-21,011 ~ 768,742

		X	G	R	M	P	I
표적 유형 1 (Everyman)	원래 값	M	43	1	1	635	40000
	변환 값	M	40-44	1	1	596	40000
표적 유형 2 (Unique)	원래 값	F	39	3	3	0	12700
	변환 값	F	35-39	3	1	0	12700
표적 유형 3 (Big I)	원래 값	M	57	1	1	1100	768742
	변환 값	M	55-59	1	1	1210	768742
표적 유형 4 (Big P)	원래 값	F	79	1	4	99997	94552
	변환 값	F	75-79	1	1	100033	94552

노출위험	-	G (재코딩)	R,M (자료교환)	G (재코딩) R,M (자료교환)
표적 유형 1 (Everyman)	.0022 (455)	.00045 (2230)	.0023 (410)	.00045 (2026)
표적 유형 2 (Unique)	1 (1)	1 (1)	.022 (10)	.0047 (49)
표적 유형 3 (Big I)	.0029 (345)	.00067 (1498)	.0031 (300)	.00069 (1355)
표적 유형 4 (Big P)	.0060 (166)	.0013 (776)	.0046 (169)	.00097 (836)

3. 매스킹 기법 요약[2)]

가. 전반적 재코딩(global recoding)[3)]

전반적 재코딩이란 특정 속성이 가지는 범주를 더 상위 범주(less-specific categories)로 묶는 것을 말한다. 예를 들면 직업이 의사와 간호사일 경우 모두 의료계 종사자로 묶는 것을 생각할 수 있다. 연령 변수도 마찬가지로 각 세 연령을 5세나 10세 단위로 묶어 전체 마이크로데이터의 익명성을 확보하는데 도움을 줄 수 있다. 참고로 전반적 재코딩 중에서 특정 값 이상/이하를 묶는 것을 top/bottom 코딩이라고 하며, 그 그룹에 속한 값들을 해당 그룹의 평균이나 중앙값으로 대체하여 표시하곤 한다. 보통은 각 키변수에 대하여 빈도수가 너무 적은 그룹이 생기지 않도록 재코딩 작업을 하는 것이 바람직하다고 할 수 있다.

나. 국소적 감추기(local suppression)

마이크로데이터를 행은 개체(레코드)들, 열은 속성 변수들로 구성된 행렬처럼 생각할 수 있다. 감추기(suppression)는 보통 레코드나 속성 차원에서 이루어진다. 예를 들면, 특정 금액 이상의 고소득자의 자료를 모두 감추거나(레코드 감추기), 레코드의 식별을 간접적으로 도울 수 있는 지역 변수를 모든 레코드에 대하여 감추는 것(속성 감추기) 등을 생각할 수 있다. 즉, 행렬에서 한 행 혹은 한 열을 통째로 감추는 것이다. 이러한 감추기는 비밀보호는 확실하나 너무 많은 정보의 손실을 야기한다.

국소적 감추기(local suppression)는 노출위험을 높이는 속성변수 조합에서 몇 개의 값만을 감추는 것을 말한다. 즉, 마이크로데이터 행렬에서 열이나 행 대신에 셀을 감추는 것을 국소적 감추기로 생각할 수 있다. 예를 들면, 변수 조합이 지역·연령·성별·직업이면서 노출위험이 높은 어떤 레코드의 해당 변수들의 값이 A지역·50대·여성·군인일 경우, 성별에서 여성이라는 값이나 직업에서 군인이라는 값을 감추는 것을 생각할 수 있다. 만약 자료 이용자가 직업에 관한 정보를 더 필요로 할 경우에는 직업 대신에 성별을 감추는 것이 유익할 것이다.

국소적 감추기를 사용할 경우에는 레코드별 노출위험을 측정하여 노출위험이 높은 레코드를 선별하는 과정과 해당 레코드의 값들 중 어느 변수의 값을 감출지 선택하는

2) 마이크로데이터 매스킹 기법을 설명하는 내용은 대부분 박민정(2015)에서 발췌하였다. 더 자세한 내용을 알고 싶거나 프로그램을 이용해 매스킹 처리를 해보고 싶은 독자들은 해당 참고문헌을 살펴보기 바란다.

3) 전반적 재코딩, 국소적 감추기, 국소통합, 잡음추가, 자료순위 교환, 자료섞기 등은 우리말로 번역된 적절한 용어가 없어서 통계개발원에서 2013년 연구 과제를 수행할 때 번역했던 용어를 사용하였다.

과정이 필요하다. 먼저 공표하는 변수 조합들에 대하여 유일한 레코드가 발생했을 때 국소적 감추기를 통해 익명성[4]을 2 이상 확보하도록 할 수 있다. 다음으로 최대한 감추어지지 않기를 원하는 변수를 선택하는 알고리즘을 사용할 수 있다. 국소적 감추기는 수행할 때마다 정보 손실을 초래하기 때문에 통계개발원에서 수행하였던 기존 연구에서는 마이크로데이터에 국소적 감추기 외의 매스킹 기법들을 적용한 후에, 마지막으로 필수적인 수준의 익명성 획득에 활용하고자 하였다.

다. 국소통합(microaggregation)

국소통합이란 자료를 최소 k개(보통 3개) 이상의 레코드를 가지는 그룹으로 묶고, 각 그룹의 자료 값들을 그 그룹의 평균이나 중앙값 등의 동일한 한 값으로 대체하는 기법을 말한다. 레코드를 묶어 그룹을 만들 때, 단일 변수를 기준으로 할 수도, 여러 변수를 모두 고려하여 그룹을 정할 수도 있다. 여러 변수를 한꺼번에 고려하기 위해서는 주성분을 이용하거나, 네델란드 통계청에서 제안했던 자료 간 거리가 먼 것부터 묶어가는 MDAV[5] 알고리즘을 이용할 수도 있다. 그 외에도 다양한 종류의 그룹화 기법이나 그룹 내 자료의 수 k의 결정에 관해 비교한 연구들도 있다.

라. 잡음추가(noise addition)

잡음추가 기법은 연속형 변수의 자료에 주로 적용하며, 매스킹 처리로 인해 자료의 값이 변하게 되는 자료 변조적(perturbative) 기법 중 하나이다. 잡음추가 기법은 통계적인 잡음을 원래 자료에 더하여 특정 변수의 정보가 공유된 다른 마이크로데이터와의 정확 매칭(excat matching)을 피하기 위해 사용된다.

세부 알고리즘으로는 먼저 각 변수별로 원래 자료의 평균과 분산을 가지는 정규 분포를 찾아 그 정규 분포에서 잡음을 발생시켜 더하는 것이 있으며, 이를 비상관(uncorrelated) 잡음추가라고 한다.

다음으로 원래 자료에서 매스킹 처리를 하고자 하는 모든 변수들의 공분산 행렬에 비례하는 공분산을 가지는 분포에서 잡음을 발생시켜 원래 자료에 더할 수 있는데 이를 상관(correlated) 잡음 기법이라 한다.

비상관 잡음추가보다는 상관 잡음추가가 원래 자료의 구조를 더욱 잘 보존하는 것이 당연하나, 상관 잡음추가 기법을 사용하여도 원래 자료에서 정규성 가정이 성립하지

4) 익명성(k-anonymity)이란, 특정 변수 조합에 대해 같은 값을 가지는 레코드의 수를 말한다.
5) Multivariate Microaggregation based on Maximum Distance to Average Vector

않는다면 매스킹 후에 자료의 구조가 심각하게 왜곡될 수도 있다. 이에 따라 상관 잡음 기법의 로버스트한 알고리즘이 제시되었고, 이 알고리즘 역시 sdcMicro 패키지에 구현되어 있다.

그 외에 자료의 크기를 고려하거나 자료 중에서 이상값들을 골라서 잡음을 더하는 알고리즘들도 존재한다.

마. 자료순위 교환(rank swapping)

자료순위 교환(rank swapping) 기법은 범주형 변수들에 주로 사용되는 자료교환(swapping)[6] 기법의 확장으로 생각할 수 있으며, 연속형 변수에서 자료를 서로 교환하는 것으로 간주할 수 있다. 즉, 자료순위 교환이란 자료를 정렬하고 제한된 범위 내에서 자료를 서로 교환하는 것을 말한다.

바. 자료섞기(shuffling)

자료섞기(shuffling)[7]는 매스킹된 자료를 생성하기 위해 조건부 분포를 이용한다. 조건부 분포를 이용한다는 것은 표본으로부터 모집단의 분포를 모수적으로 추정하여 새로운 자료 세트를 생성한다는 것을 말한다. 기본적으로 자료가 정규분포를 따른다고 가정하여 자료섞기 작업을 수행할 수 있으나, 가정을 확장하여 자료가 치우친 t분포를 따를 경우에 대한 자료섞기 알고리즘도 연구되어 있다. 세부 알고리즘으로는 일반적 가법 자료 변조(GADP[8])를 이용하는 경우 및 정규 코풀라(normal copula)를 이용할 경우를 고려할 수 있다.

사. 빈도(frequency)표 매스킹

일단 빈도표를 작성하면 정해진 기준보다 값이 작은 셀들은 위험하다고 판단하여 감추기 처리한다. 이때 주변합에서 안전한 셀들의 값을 빼서 위험한 작은 셀의 값을 거꾸로 계산해낼 수 있으므로, 안전한 셀들을 대상으로 추가적인 셀 감추기(secondary cell suppression)를 실시할 것이 권장된다.

---

6) sdcMicro 패키지의 자료교환(swapping) 명령어를 이용할 경우에는 얻어진 결과에 오류가 있는 것으로 판단된다.

7) 잡음추가나 자료교환 기법은 자료 변조적(perturbative) 기법으로 주로 언급되나, 자료섞기로 번역한 shuffling은 자료의 분포를 이용하여 자료를 완전히 바꾸므로 자료 변조(perturbation) 기법으로 구분되기도 한다.

8) General Additive Data Perturbation

다른 방법으로 주어진 값을 기준으로 빈도값들을 반올림(rounding) 처리하는 방법이 있다 예를 들어 기준값이 5이면 모든 빈도를 0, 5, 10, ... 등으로 바꾸는 것이다. 한편, 반올림할 때 확률적으로 값을 결정할 수도 있으며 이를 랜덤 라운딩(random rounding)이라고 한다.

빈도표 매스킹의 어려운 점은 센서스 빈도표와 같이 하나의 마이크로데이터에서 매우 많은 개수의 빈도표를 생성하는 경우에 발생한다. 작은 셀들의 노출을 방지하면서 빈도표들 사이의 일치성을 확보하는 것이 어렵기 때문이다. 이에 관한 최근 연구로는 Shlomo(2007)[9]가 있다.

차. 총계(magnitude)[10]표 매스킹

특정 셀에 속하는 특정 응답자의 값을 $p\%$ 이내의 상하한 구간 내로 추정할 수 있을 때 해당 셀을 민감하다고 판명하고 공개하지 않을 수 있다. 이를 $p\%$-규칙이라고 부른다.

한편, n개의 큰 값을 가지는 개체들의 총계가 셀 값의 $k\%$를 넘게 차지하는 경우 민감한 셀로 규정하고 해당 셀을 공개하지 않을 수 있다. 이는 적은 수(n)의 응답자가 전체 셀 값에 대해 큰 비중($k\%$)을 차지하면 민감한 셀로 판단한다는 의미이다. 이를 (n, k)-규칙이라고 부른다.

9) Shlomo, N. (2007), Statistical disclosure control methods for Census frequency tables, International Statistical Review, 75, 2, pp. 199-217.

10) 본 보고서에서는 magnitude를 총계라 번역하였다. 기타 국내 문헌에서 크기 등으로 번역되기도 하였다.

제5장

차등정보보호에 관한 연구

박민정 · 이용희 · 권성훈

| 요 약 |

4차 산업혁명시대가 도래함에 따라 자료제공과 개인정보보호의 중요성이 크게 부각되고 있다. 이를 동시에 해결하기 위한 최근의 기술적인 대안으로서 차등정보보호를 설명한다.

서론에서는 통계적 비밀보호 기법들을 이해하고 비교할 때 노출위험과 정보손실을 종합적으로 고려해야 함을 이해한다. 2절은 차등정보보호의 주요 개념을 다양한 예제를 통해 설명한다. 이를 통해 차등정보보호의 개념적 이해를 제시하고, 고려할 사항을 정리한다. 전문 연구자들은 3절을 통해 차등정보보호를 이론적으로 접근할 수 있다. 활용 측면에서 차등정보보호 발전 상황은 4절에 제시한다. 결론에서는 향후 연구 방향을 논의한다.

주요 용어 : 차등정보보호, 재현자료, 위험-유용성 지도

✍ 이 연구는 2018년 하반기 통계개발원 연구보고서에 수록된 자료입니다.

제1절 서 론

4차 산업혁명 시대의 데이터는 과거의 석유처럼 중요한 자원이 될 수 있다. 데이터를 활발히 공유하는 것이 중요한 시대가 되었다. 그러나 아무리 데이터가 쏟아져 나오는 시대라고 해도, 개인정보보호 역시 절대 놓쳐서는 안 될 중요한 가치이다.

개인정보보호에 대한 인식이 강화되는 상황에서, 개인의 자발적 동의[1]없이 데이터를 유통시키기 어렵다. 반면 분석 가치가 높은 데이터일수록 개인 수준의 정보가 자세히 수록되어 있게 마련이다. 즉, 이 시대에 필수적인 두 요소인 데이터의 활용과 개인정보보호는 상충 관계에 있다. 그럼에도 데이터의 활용과 개인정보보호가 반드시 동시에 이루어져야 한다는 것은 시대적인 요구이다.

이런 상황에서 각국의 통계작성기관은 자료를 적극적으로 제공하고, 개인정보 노출을 일으키지 않기 위해 기술적으로 많은 노력을 해왔다. 최근에는 데이터를 기반으로 사업을 하는 구글이나 애플 등의 IT기업들도 최신 비밀보호 방법론을 연구·개발하고, 나아가 새로운 기법을 실제로 적용하기 위해 다양한 노력을 하고 있다. 기존의 매스킹(masking) 처리만으로는 데이터의 활용과 개인정보보호라는 두 목적을 달성할 수 없었지만, 이와 같은 다양한 노력의 결과로 재현자료(synthetic data)와 차등정보보호(differential privacy)라는 새로운 방향들이 제시되었다. 본 연구의 서론에서는 노출위험을 일정 수준 이하로 제어하는 차등정보보호를 연구해야 할 필요성을 설명한다. 이를 위해 먼저 매스킹, 재현자료, 차등정보보호를 위험-유용성 관점(risk-utility framework)에서 정리한다.

매스킹(masking) 처리는 전통적 방식으로 자료를 적절히 변환함으로써 자료를 구성하는 각 개체들이 식별되거나 특정 속성이 노출되지 않도록 하는 것을 일컫는다. 다음 <표 5-1>의 간단한 데이터 예제를 통해 노출위험과 매스킹 처리에 대해 생각해보자. 여기서 키변수란 각 개별 응답자의 특성을 나타내는 변수이다. 키변수는 누구나 쉽게 정보를 얻을 수 있는 경우가 많다. 민감변수는 자료 구축에 참여한 각 개체가 노출을 원하지 않는 변수이다. 한편 <표 5-1>의 노출위험 측도 중, k-익명성은 특정 변수 조합에 대해 같은 값을 가지는 레코드의 수가 k개 이상임을 말한다. 즉, 주요 변수의 각 구별 패턴이 자료 내에서 적어도 k개 이상임을 의미한다. l-다양성(포괄성)은 주요 변수들에서 같은 패턴을 가지는 개체들이 민감변수에 대해 적어도 l개의 자주 등장하는 값을 가진다는 것을 말한다.

1) 동의란 무엇이며 어떤 방법으로 얻어야 하는가는 또 다른 연구주제이기도 하다.

〈표 5-1〉 간단한 데이터 예제에서 노출위험 측정

	성별 (키변수 1)	연령 (키변수 2)	질병번호 (민감변수)	노출위험 측도	
				k-익명성	l-다양성
1	1	1	50	3	2
2	1	1	50	3	2
3	1	1	42	3	2
4	1	2	42	1	1
5	2	2	62	2	1
6	2	2	62	2	1

<표 5-1>에서 처음 세 개체는 동일한 특징, 즉 동일한 키변수 값들을 가지므로 익명성 3을 확보한다. 반면 네 번째 개체는 키변수 조합에서 유일한 특징을 가지므로, k-익명성 값이 1이며 데이터에서 이 개체가 식별된다고 판단한다. 마지막 두 개체는 익명성 2를 가진다. 따라서 키변수만을 고려하고 k-익명성을 기준으로 노출위험을 정의하면 네 번째 개체가 가장 노출위험이 높다.

한편 노출위험을 정의하는 기준을 l-다양성으로 바꾸면 판단 결과가 달라진다. l-다양성을 기준으로는 마지막 세 개체의 노출위험이 동일하기 때문이다. 키변수 조합 (2,2)를 가지는 마지막 두 개체는 동일한 민감정보를 가진다. 따라서 이 두 개체가 62번 질병에 걸렸다는 정보가 노출되어 세 번째 개체와 마찬가지로 l-다양성이 1이 되므로 노출위험이 같아진다. 이렇게 민감변수의 다양성을 고려해서 정보보호를 측정하는 경우, 키변수 조합별 익명성을 근거로 노출위험을 측정할 때에 비해 개인정보 침해에 대한 판단이 달라짐을 알 수 있다. 따라서 자료제공 여부를 결정할 때, 어느 노출위험 측도를 기준으로 판단할 것인가 하는 문제를 고려해야 한다.

이제 매스킹 처리를 통해 주어진 노출위험 측도를 낮추는 것을 생각해보자. 여러 가지 방법이 있을 수 있겠지만, 마지막 세 개체에 대해서 키변수 2에 해당하는 값들을 NA 처리하는 방법(국소 감추기, local suppression)을 선택한다고 하자. 그러면 다음 <표 5-2>에 보는 것처럼 3개 이상의 매스킹 처리 방안이 존재할 수 있다. 이들 중 어느 방안이 가장 타당한 것일까? 세 방안은 노출위험을 줄이는 대신 정보손실을 늘리거나, 또는 그 반대의 효과를 가지기 때문에 딱히 무엇이 왜 타당한지 말하기가 어렵다. 이처럼 매스킹 처리를 통해서는 노출위험을 줄일 수 있는 다수의 시나리오가 존재한다. 따라서 여러 매스킹 방안들 중에서 어느 시나리오를 적용해서 공표할 것인가를 선택해야 한다.

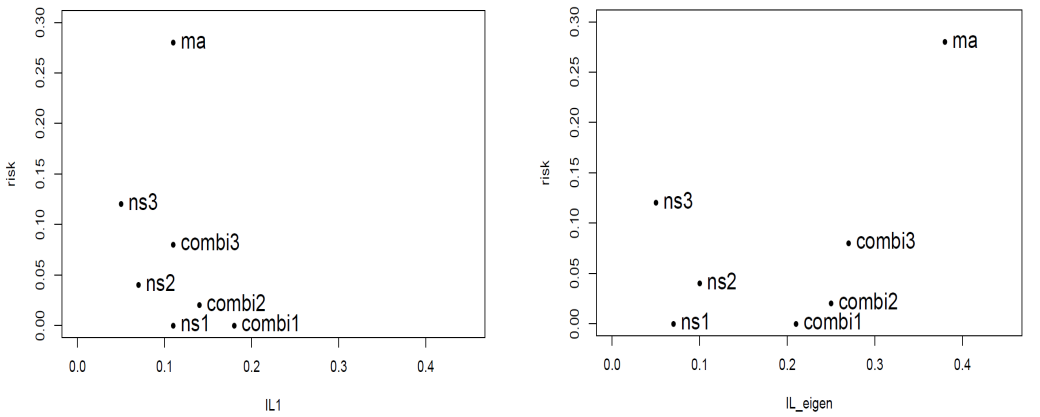
〈표 5-2〉 국소 감추기를 통한 매스킹 처리 예제

	성별 (키변수 1)	연령 (키변수 2)			질병번호 (민감변수)	빈도 (k-익명성)	l-다양성
1	1	1			50	3	2
2	1	1			50	3	2
3	1	1			42	3	2
4	1	NA	NA	NA	42	1 → ?	1 → ?
5	2	2	2	NA	62	2 → ?	1 → ?
6	2	2	NA	NA	62	2 → ?	1 → ?

기존의 매스킹 처리 및 평가 방안에 관한 문헌들을 살펴보면, 위험-유용성 지도(risk-utility map) 위에 각 매스킹 방안들을 표시하고 선택하는 방안이 제시되어 있다. 예를 들면 다음 [그림 5-1]과 같이 여러 매스킹 시나리오[2]에 대해 노출위험(세로축)과 정보손실(가로축)의 값을 계산해서 표시하고 나서, 노출위험과 정보손실이 모두 적은, 즉 원점에 가까운 방안을 선택할 수 있다(박민정, 2014). 다만 문제는 정보손실과 노출위험은 서로 상충관계에 있으므로, 원점과의 거리가 동일하다면 도대체 어느 방법이 좋은지 선택하기 어렵다는 것이다. 예를 들면 위험-유용성을 종합적으로 고려한다면 [그림 5-1]에서 시나리오 “ns1”과 “ns2”가 동일한 품질의 자료일 수 있다. 이처럼 매스킹 방안들 사이에 적절한 평가 기준을 세우는 것은 잘 해결되지 않는 문제였다.

또한 노출위험과 정보손실을 계산하는 측도들이 다양하다는 문제점도 있다. [그림 5-1]의 왼쪽은 정보손실 측도를 계산할 때 매스킹 전후 자료값의 거리(IL1)를 이용했고, 오른쪽은 고유값 사이의 거리(IL_eigen)를 이용했다. 시나리오 “ns1”을 예로 들면 정보손실 측도가 바뀌면서 원점까지의 상대적인 거리가 줄어들므로 다른 시나리오와 비교 평가 결과도 달라진다.

2) ns1, ns2, ns3은 잡음을 추가하는 모형들을 적용해서 변환된 마이크로데이터의 노출위험과 정보손실 값을 표시하고, ma는 군집화(microaggregation)를 적용한 마이크로데이터, combi1, combi2, combi3은 국소통합과 잡음추가를 동시에 적용해서 변환된 마이크로데이터의 노출위험과 정보손실 값을 나타낸다.



[그림 5-1] 위험-유용성 지도 위에 표현된 다양한 매스킹 시나리오

결론적으로 매스킹 기법을 사용할 때는, 자료를 변환시키는 수많은 방법들 중에 무엇을 선택하는 것이 타당한지 판단하기 어렵고, 적절한 노출위험 및 정보손실 측도를 규정하는 문제도 명확하게 설명하기 쉽지 않다. 노출위험 측도의 경우, 초창기에는 k-익명성(anonymity)[3]이 제안되었었고, 그 부족함을 만회하기 위해 l-다양성(diversity)[4]이 제시되고, 이 두 개념의 문제점을 비판하며 t-근접성(closeness)[5]이 나오는 등, 어찌 보면 다람쥐 쳇바퀴 도는 것처럼 새로운 측도가 계속해서 제시되지만, 궁극적인 해결책에 다가가지 못하는 측면이 있었다.

그러나 정책 결정자들이 데이터 활용과 개인정보보호를 모두 해결하는 의사결정을 할 수 있기 위해서는 변환된 여러 데이터 중 무엇을 제공할지 판단이 가능해야 한다. 이를 위해 위험-유용성 관점(risk-utility framework)을 적용할 수 있다. 위험-유용성 관점에서 매스킹 방안을 적절히 평가하기 위해서는 노출위험과 정보손실을 측정하는 측도(measure)가 대단히 중요하다. 나아가 바람직한 방향으로 비밀보호[6] 방법론이 발전해 나가는데 있어서도 노출위험과 정보손실 측도의 중요성은 더할 나위 없이 크다고 하겠다. 적절한 노출위험과 정보손실 측도가 제시된다면 위험-유용성 지도에서 최적의 방안을 찾아가는 방향으로

3) Sweeney, L.(2002), k-Anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10, pp. 557-579.

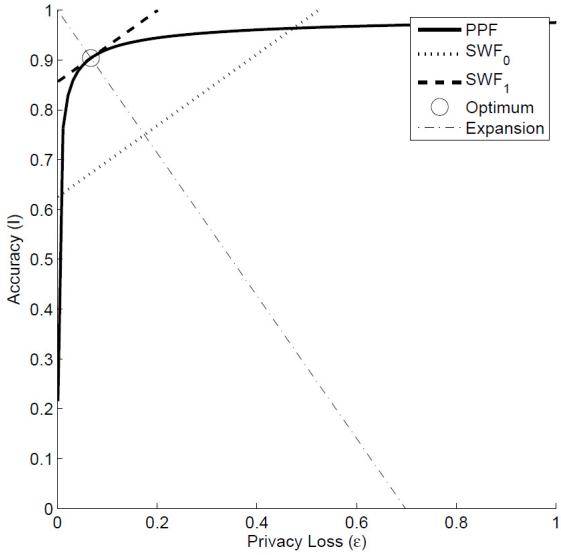
4) Machanavajjhala, A., Gehrke, J., Kifer, D. and Venkitasubramaniam, M.(2006), l-Diversity: Privacy Beyond Kappa-Anonymity. in Proceedings of the 22nd International Conference on Data Engineering (ICDE '06).

5) Li, N., Li, T. and Venkatasubramanian, S.(2007), t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. 2007 IEEE 23rd International Conference on Data Engineering.

6) 통계적 노출위험 제어/제한(statistical disclosure control/limitation), 통계적 비밀보호(statistical data confidentiality), 데이터 프라이버시(data privacy) 등의 용어는 같은 의미를 가지는 것으로 간주한다.

구체적인 알고리즘 연구가 진행될 수 있기 때문이다. 따라서 노출위험과 정보손실 측정에 관한 활발한 연구가 필요하다.

참고로 다음 [그림 5-2]는 최근 제안된 위험-유용성 지도를 보여준다. [그림 5-1]과의 차이점은 노출위험 기준을 가로축에, 정보손실과 반대로 움직이는 유용성 기준을 세로축에 두었다는 것이다. 따라서 마치 ROC 곡선처럼 어떤 방안에 대한 평가 결과가 그림의 왼쪽 상단에 나타날수록 바람직한 방법임을 나타내 준다. 한편 위험-유용성 지도 위에 표현된 여러 방법 중 무엇이 적절한가에 대해 평가하기 위한 논의도 시작되고 있다. 이렇게 위험-유용성 관점에서 비밀보호 방식을 평가하는 움직임이 최근 "formal privacy"라는 이름으로 나타나고 있으며(Vilhuber와 Schmutte, 2017), 앞으로도 이와 관련된 연구가 지속적으로 이루어질 것으로 기대된다.



출처: Abowd and Schmutte(2017)

[그림 5-2] 최근 제안된 위험-유용성 지도 사례

정리하면 매스킹의 기법들은 위험-유용성 관점에서 명확한 대안을 제시하지 못한다. 여러 방식으로 자료를 변환하는 것이 가능하지만 타당성을 입증하기 어렵기 때문이다. 원래 자료와 매스킹 기법을 적용해서 변형된 자료 사이의 거리가 멀면 노출위험이 적고, 정보손실이 크다. 그 거리가 가까우면 반대의 현상이 일어난다. 즉, 매스킹 처리라는 범위 내에서는 서로 상충관계에 있는 노출위험과 정보손실 두 측도를 동시에 감소시키는 방법을 찾기가 어렵다고 볼 수 있다. 따라서 근본적으로 새로운 접근이 필요하다.

위험-유용성 관점에서 많은 한계를 보이는 매스킹 기법을 극복하기 위해, 최근 출현한 새로운 두 종류의 비밀보호 방법은 재현자료와 차등정보보호라고 할 수 있다. 먼저 재현자료는 원래 데이터와 동일한 분포를 따르도록 생성한 자료를 말한다. 개념적으로 완전재현자료(fully synthetic data)는 원래의 데이터를 전혀 포함하지 않으므로[7] 노출위험이 없다. 또한 원래 데이터와 동일한 분포를 알아낼 수 있다면 정보손실을 획기적으로 줄일 수 있어 궁극적인 비밀보호 대안이 될 수 있어 보인다. 위험-유용성 관점에서 재현자료 관련 연구는, 노출위험은 없는 것으로 간주하고 정보손실을 줄여 자료를 생성하는 방법에 집중해 왔다고 할 수 있다. 따라서 매스킹에서 복잡했던 의사결정 문제가 단순화된다는 장점이 있다.

재현자료 연구에 있어서 단점은 다음과 같다. 우선 원래 데이터가 가지는 분포를 추정하고, 그 분포에서 원래 자료와 유사한 자료를 생성하는 것이 쉬운 일이 아니다. 따라서 정보손실을 줄이기 쉽지 않으므로, 원자료를 대신할 만족스러운 수준의 재현자료 생성 결과를 얻기가 쉽지 않았었다. 또한 최근에는 재현자료의 노출위험이 정말 없다고 할 수 있는지(Ruiz 외, 2018), 혹은 차등정보보호를 만족시킬 수 있는지(Machanavajjhala 외, 2008) 등의 논의도 이루어지고 있다. 차등정보보호를 만족하는 재현자료에 대해서는 4절에서 더 논의하도록 한다. 재현자료에 대한 자세한 설명은 기존의 보고서(박민정과 김정연, 2017) 및 논문들(박민정과 김항준, 2016; 박민정과 김정연, 2019)과 함께 수록된 참고문헌을 참고하도록 한다.

마지막으로 차등정보보호는 노출위험에 대한 새로운 정의로 이해할 수 있다. 차등정보보호를 제안한 Dwork(2006)은 노출에 대한 개념을 기존과 다른 방식으로 새롭게 제안하고, 나아가 수학적으로 엄격한 정의를 제시했다. 노출위험에 대한 수학적인 기준이 제시된다는 측면에서 차등정보보호는 “formal privacy”에 부합한다고 여겨진다. 위험-유용성 관점에서 차등정보보호를 만족시키며 통계를 생산하는 것은 주어진 수준 이하로 노출위험을 제어한다는 것을 의미한다. 즉, 차등정보보호를 만족하도록 생성된 자료는 주어진 기준에서 안전하다고 판단할 수 있다. 따라서 대다수의 차등정보보호 관련 연구 역시 정보손실을 줄이기 위해 적절한 알고리즘을 제안하는데 집중되어 왔다고 볼 수 있다.

한편 차등정보보호는 대용량 데이터를 사용하면서 개인정보보호를 달성하기 위해 구글, 애플 등에서 이미 활용되고 있기도 하다. 비록 국내에서는 아직 생소한 개념이라고 하더라도, 국가통계기관이 차등정보보호를 바르게 이해하려는 노력을 하는 것은 현 시점에서 적절하다고 생각된다. 본 연구에서는 새로운 비밀보호 방법으로서 차등정보보호를 주제로 다루도록 한다.

---

7) 이에 관한 최근의 논의는 Drechsler(2018)을 참고할 수 있다.

차등정보보호를 바르게 이해하기 위해서 '18년 총 5회에 걸쳐 각 분야의 통계학 전문가들이 모여 세미나를 개최하고 주요 문헌들을 정리했다. 그러한 과정을 통해 일반 학생에서부터 각 분야 전문가에 이르기까지 차등정보보호에 대한 이해를 돕는 것을 목표로 본 보고서가 작성되었다. 데이터 활용과 개인정보보호, 상충하는 두 목적을 함께 달성하기 위해 기술적인 대안에 관한 연구가 국내 각 분야에서 더욱 활발히 진행되기를 희망한다.

참고로 본 연구는 통계적 비밀보호 또는 데이터 프라이버시 분야에 속한다. 주의할 것은 비밀보호는 보안(data protection, security)과는 다른 개념을 가진다는 것이다. 보안은 자료가 유출되는 것을 방지하기 위한 다양한 IT분야의 기술을 다루는 것다. 하지만 비밀보호는 이미 공표 또는 유통되는 자료를 다루며, 공표하는 자료에서 개별적인 수준의 민감한 정보가 노출되는 것을 제한하는 것이 목적이다. 비밀보호와 보안을 분리해 이해할 필요가 있다.

본 보고서는 다음과 같이 구성되어 있다. 1절 서론에서는 노출위험-유용성 관점을 소개하고, 매스킹 처리, 재현자료, 차등정보보호의 의의를 논했다. 2절에서는 다양한 예를 통해 차등정보보호의 개념을 이해할 수 있도록 설명하고, 수학적 정의와 성질에 대해서도 논의한다. 이는 통계학이나 전산학 등의 전공자가 아니더라도 차등정보보호를 이해할 수 있도록 돕기 위한 의도가 있다. 3절은 통계학적 관점에서 차등정보보호를 엄밀하게 이해할 수 있도록 구성되었다. 차등정보보호 관련 주요 통계학 논문에 근거해서 차등정보보호를 만족하며 도달할 수 있는 정보손실에 관해서도 논의한다. 이는 차등정보보호를 만족하면서 정보손실을 줄이는 새로운 알고리즘 연구를 돕기 위한 목적이 있다. 관련 전공자가 아닌 일반 독자들은 3절을 건너뛰어도 된다.

4절에서는 차등정보보호를 활용하는 사례들을 소개한다. 단순히 라플라스 잡음을 추가하는 것은 자료 유용성 측면에서 효율적이지 않은 경우가 많다는 것을 설명하고, 통계적 모형을 활용해서 차등정보보호를 만족시키는 사례들을 소개한다. 한편 국가통계(official statistics) 관련 데이터베이스에 대해 실무적으로 사용할 수 있을 정도로 정보손실을 획기적으로 줄이는 차등정보보호 적용 사례를 찾는 것은 쉽지 않았다. IT기업들을 제외하면 미국 OnTheMap[8]의 사례가 차등정보보호 활용에 있어 가장 성공적인 것으로 평가되고 있다. OnTheMap 관련 논문은 차등정보보호를 만족하는 재현자료 생성과 확률적 차등정보보호의 제안을 논의하며, 4절에서는 그러한 내용을 정리해서 소개하도록 한다. 참고로 본 보고서의 2절, 3절, 4절은 각기 다른 저자가 기술했다. 때문에 중복되는 내용이 있을 수도 있겠지만, 차등정보보호가 워낙 이해하기 어려우므로 다양한 목소리로 들어 보는 것이 바람직한 이해에 도움이 될 수 있다. 보고서는 5절의 맺음말로 마친다.

8) https://onthemap.ces.census.gov

제2절 개념적 이해

1. 빅데이터 시대와 개인정보보호

가. 새로운 환경: 빅데이터와 실시간 정보

현대를 사는 사람들은 자신의 많은 정보들이 알게 모르게 노출되고 저장되며 이용되고 있다는 사실을 때때로 인식한다. 그럴 때마다 두려움이 슬쩍 피어오르지만 이러한 두려움이 일상이 되면 금세 무감각해진다. 거리의 수많은 폐쇄회로 카메라들이 우리를 촬영하고, 대중교통을 이용할 때마다 경로와 목적지가 저장되며, 휴대폰을 사용할 때마다 개인의 다양하고 방대한 정보가 저장되고 있다는 사실은 이제 비밀이 아니다. 이러한 개인정보의 유출은 두려움도 주지만 현대인들에게 엄청난 편의를 제공하고 있는 것도 사실이다. 폐쇄회로 카메라 때문에 범죄를 예방하고 더 나아가 테러범을 체포하는 경우도 있다. 또한 대중교통을 이용하는 사람들의 행태를 분석하여 최적의 이동 경로를 제공하고 교통 환경을 개선할 수 있으며, 휴대폰과 인터넷을 사용하는 행동을 분석하여 이용자에게 놀랍도록 최적화된 상품을 추천해주기도 한다.

20세기 중반에 컴퓨터의 보급으로 시작된 IT혁명은 사람들의 생활에 큰 변화를 주기 시작하였다. 빠른 계산과 더 많은 자료를 저장할 수 있는 기술, 인터넷과 모바일 기구로 정보를 교환하는 환경은 우리의 생활양식을 완전히 바꾸어 놓았다. 최근 빅데이터(big data)와 인공지능의 시대가 열리면서 몇 년 전까지만 해도 상상할 수 없었던 방대한 양의 정보를 저장하고 처리하는 것이 가능해졌다. 우리의 삶이 편리해졌지만 사생활에 대한 두려움도 유례없이 커지고 있으며, 종종 심각한 사생활 침해가 일어나기도 한다.

전통적으로 개인의 민감한 자료는 정부기관, 병원, 은행과 같은 공신력이 높고 강한 규제를 받는 기관만 보유할 수 있었다. 하지만 이제는 구글, 아마존, 애플, 네이버 등과 같은 거대한 IT기업들이 더욱 민감하고 방대한 개인의 자료를 수집하고 있다. 더 나아가 규모가 작은 기업이나 개인들도 다양하고 많은 개인정보를 쉽게 수집하고 이용할 수 있다. 더욱 중요한 것은 이렇게 수집된 다양하고 방대한 정보들이 서로 결합되어 전에는 상상할 수 없었던 독창적이고 새로운 가치와 편의를 창출하는 시대가 열렸다는 사실이다. 하지만 이러한 환경의 변화는 민감한 개인정보의 유출을 억제할 수 있는 과거의 방법들이 무력화될 수 있는 새로운 위험도 동시에 창출하였다. 좋은 점이 생기면 반대로 나쁜 점도 생기는 세상의 이치처럼, 빅데이터 시대는 우리에게 전에 없던 편리성과 신속성을

주었지만 개인의 사생활이 더욱더 침해될 수 있는 환경을 만들었다. 이제 개인정보를 어떻게 보호할 수 있는지에 대한 새로운 개념과 방법을 심각하게 고민해야 하는 과제가 주어졌다.

개인정보보호에 대한 본격적인 논의를 위해 빅데이터 시대의 변화된 환경을 몇 가지 정리해보면 다음과 같다.

- 첫째로, 자료에서 레코드 개수와 항목(변수) 개수가 과거에 비해 급격히 증가하였다. 이는 데이터 증가 추세에 따른 당연한 변화이다.

- 두 번째로, 공개하는 자료의 정확성이 증가하고 있으며 정보의 교환이 실시간으로 이루어진다. 자료를 수집하는 기법의 개발, 다양하고 정확한 센서의 출현, 저장 기술의 발전으로 자료의 정확성이 크게 향상되었다. 예를 들어 최근에는 장소에 대한 자료로서 행정구역이나 주소보다 더 세밀한 GPS 정보가 제공된다.

- 마지막으로, 사람들이 사용하기를 원하는 자료의 종류가 다양해지고 그 개수가 늘어나고 있다.

이러한 변화로 인해 서로 다른 자료를 연계하여 개인정보를 유추하거나 알아낼 수 있는 기회가 급격하게 증가하고 있으며, 노출된 정보의 정확성도 높아짐으로써 개인의 민감한 정보가 유출될 가능성은 점점 커지고 있다.

나. 노출에 대한 새로운 위험들

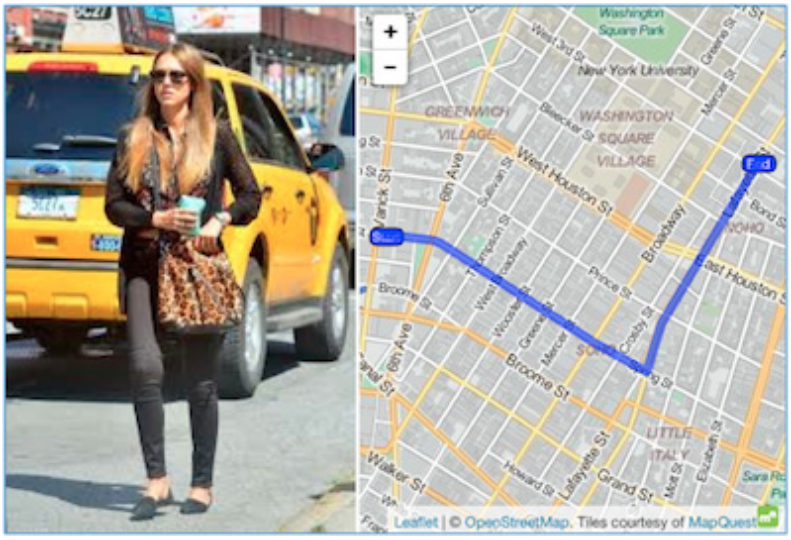
이제 몇 가지 간단한 사례를 들어서 우리가 처한 개인정보 노출에 대한 새로운 위험에 대하여 생각해보자.

아래 제시된 예는 1997년 실제로 일어난 사건이다. 개인정보 노출의 위험성에 대한 경종을 울리기 위하여 당시 MIT 대학원생이던 Sweeny는 개인(미국 매사추세스 주지사)의 민감한 정보인 의료보험 기록을 유출시킬 수 있음을 보였다(Sweeny, 1997). 이 사건은 자료를 대중에게 공개할 때 개인을 식별할 수 있는 정보를 철저하게 제한하여 공개하더라도, 외부자료와의 연계를 통하여 민감한 정보가 유출될 수 있다는 사실을 보였다. 이는 상당한 반향을 일으켰다. 또한 이를 통해 개인정보를 보호하도록 고안된 전통적인 방법들이 무력화될 수 있다는 사실도 알려지게 되었다.

1990년대 후반, 미국 매사추세츠주 공무원에게 건강보험을 제공하는 보험공단은 공무원들의 병원 방문에 대한 의료보험 데이터를 익명화한 후, 연구자들에게 이용할 수 있도록 허용하였다. 공개된 자료를 분석하고 그 결과를 이용한다면, 의료서비스를 개선할 뿐만 아니라 더 나아가 의료비용을 줄일 수 있는 새로운 정책의 개발로 이어질 것을 기대했기 때문이다. 물론 보험공단은 보험기록 자료를 대중에게 제공할 때, *개인정보를 보호하기 위하여 이름, 주소, 사회보장번호 등, 개인을 직접 식별하는 데 사용할 수 있는 모든 정보를 삭제하였다.*

MIT의 대학원생 Latanya Sweeney는 이렇게 공개된 익명화(비식별화) 자료에서 주지사 개인의 보험기록을 알아내 보기로 했다. 먼저 Sweeny는 일반인이 사용할 수 있도록 공개하는 유권자 등록기록을 소액의 비용으로 입수하였다. 그리고 삭제 대상이 아닌 변수들인 우편번호, 생년월일 및 성별을 이용하여 유권자 등록기록과 익명화된 보험 자료와 서로 대조, 연계했다. 마침내 주지사의 성별, 생년월일 및 우편번호와 일치하는 단 하나의 보험 기록을 찾을 수 있었다.

좀 더 최근의 사례로 2014년 미국의 보안회사인 Neustar가 미국 뉴욕시의 택시운행기록 자료를 분석한 결과를 살펴보자. 이 사례는 뉴욕시가 공개한 자료를 이용하여 특정 연예인들의 택시 승하차 장소와 지불한 요금에 대한 정확한 정보를 알아낼 수 있다는 것을 보여준다. 물론 뉴욕시의 자료 공개는 이러한 빅데이터 분석으로 운전기사들의 수입 추정, 최적경로 추천 등 새롭고 유용한 정보를 창출할 수 있다는 빅데이터 활용의 좋은 점도 보여준다. 하지만 민감한 정보가 유출될 수 있다는 부정적인 측면도 다음과 같이 드러난다.



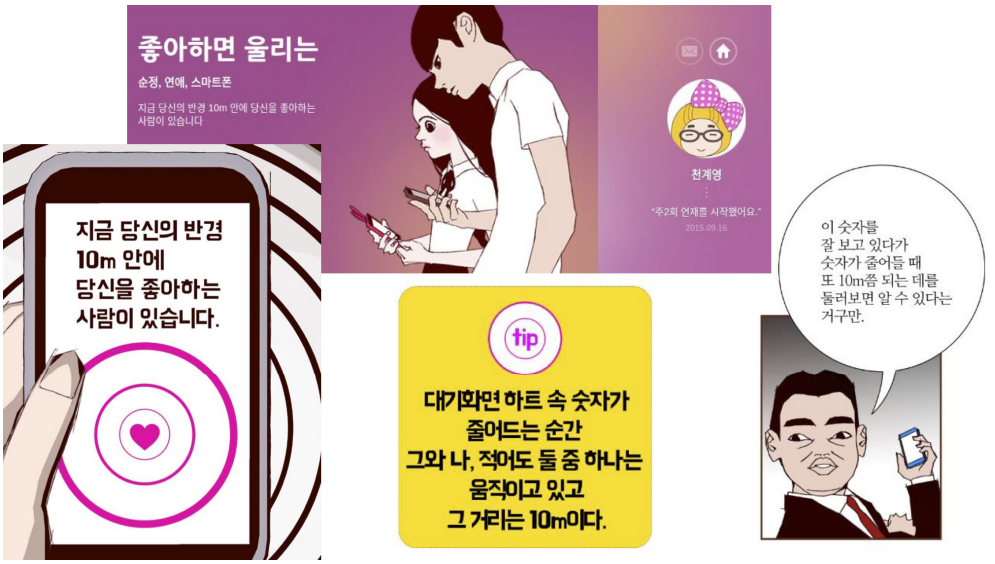
2013년 뉴욕시는 정보공개법에 의거하여 택시들이 운행한 기록을 공개하였다. 공개된 자료는 뉴욕시의 모든 택시들의 운행기록을 포함하며, 출발 장소와 시간, 도착 장소와 시간, 요금과 팁으로 구성되어 있다. *개인정보보호를 위해 택시의 차량 번호는 다른 번호로 가명화 처리하여 공개하였다.*

보안회사 Neustar의 한 연구자는 공개된 택시 운행기록 자료를 얻고, 어느 가십 기사에서 두 명의 영화배우가 뉴욕시에서 택시를 잡는 사진 두 장을 찾아내었다. 연구자는 사진이 찍힌 시간과 장소를 운행기록과 연계하여 배우들이 택시를 타고 도착한 목적지와 지불한 요금 그리고 팁의 정확한 금액을 쉽게 알아 낼 수 있었다.

또한 민감한 장소를 자주 출입하는 사람들이 택시를 타고 출발한 지점을 정확하게 알아내고, 누가 그 건물에 거주하는지도 알 수 있다는 사실을 보여주기도 하였다. (사진, 자료 출처: Neustar, 2014, Company Blog)

위에서 제시한 두 가지 사례는 전통적으로 이용되어 온 개인정보 노출제한 방법들이 이제는 더 이상 유효한 효과를 내지 못한다는 것을 단적으로 보여준다. 개인의 신분을 파악할 수 있는 직접 식별변수를 공개하지 않거나, 특정 변수를 가명 처리하는 등의 전통적인 방법은 한계가 있음을 보여준 좋은 예라고 할 수 있다.

이제 마지막으로 현재 실제로 일어나고 있는 일은 아니지만 머지않아 사람들이 직면하게 될 새로운 환경에 대한 암시를 주는 예를 살펴보자.



위의 그림은 천계영 작가의 인기 웹툰 "좋아하면 울리는"에서 몇 장면을 가져온 것이다. 이 웹툰은 10미터 반경 안에 나를 좋아하는 사람의 숫자를 알려주는 스마트폰 어플이 있다는 설정 하에서 벌어지는 학교드라마이다. 나를 좋아하는 새로운 사람이 반경 안에 들어오면 알람을 울려서 알려주지만 누구인지 정확한 정보를 알려주지 않는다. 때문에 나를 좋아하는 사람이 누구인지 알려고 여러 가지 아이디어를 내고 노력하는 일이 벌어지게 된다. 위의 그림은 나를 좋아하는 사람의 정체를 알아내는 하나의 방법을 보여준다. 이는 화면을 보고 있다가 숫자가 하나 줄어들면 반경 10미터 부분을 관찰하여 나를 좋아하는 사람을 찾을 수 있다는 것을 보여주는 장면이다. (사진 출처: 천계영, 2015, "좋아하면 울리는")

이 웹툰의 배경과 이야기는 우리가 가까운 미래에 만나게 될 정보의 이용환경과 개인정보 누출에 대한 큰 암시를 준다.

첫 번째로 자료를 이용하는 환경이 실시간이라는 것이다. 과거에는 공신력 있는 기관이 노출제어 기법을 적용한 자료를 공개하고, 이용자들이 분석하는 자료는 동일한, 고정된 정보로 구성되어 있는 상황이었다. 그러나 이제는 실시간으로 자료가 수집되고, 수집된 자료를 분석하여 이용자가 연속적으로 갱신되는 정보를 얻을 수 있는 경우가 많으며, 그러한 환경이 점점 확대되고 있다. 현재도 데이터베이스에 접속해서 그 안의 자료를 여러 가지 방법으로 분석하고 다양한 결과를 받을 수 있는 대화형 자료 분석

서버(interactive data analysis server)가 존재한다. 이러한 대화형 서버가 빅데이터와 함께 발전하면 웹툰에 나오는 장면과 유사한 서비스들이 많이 등장할 것이다. (지금도 구글 지도는 이용자가 이동하면 동선을 파악하여 식당을 추천해 준다!)

이 웹툰의 두 번째 시사점은 개인의 정보가 누출될 수 있는 새로운 가능성을 알려준다는 것이다. 웹툰의 내용은 10미터 반경에 있는 사람들 중 한 명이 그 범위를 벗어났을 경우, 이용자가 제공받는 정보의 변화를 감지하여 찾고 있는 사람의 정체를 알 수 있다는 것이다. 즉 갱신된 정보가 지속적으로 교환되는 상황에서, 정보가 생성되는 과정을 이용자가 잘 이해하고 있다면 일련의 사소한 방법을 적절히 적용하여 중요한 개인정보를 쉽게 알아내는 것이 가능하다는 사실을 알려준다.

이 웹툰에서 나오는 개인정보 유출(나를 좋아하는 사람의 정체가 드러나는 사건)을 방지하려면 어떠한 정보보호의 개념이 정의되어야 할까? 또 그에 적합한 노출제어 방법은 어떤 것일까? 그에 대한 해답은 차등정보보호(differential privacy)가 제공하고 있다. 예를 들어 실제로 나를 좋아하는 사람이 10미터 반경을 벗어난 순간 이용자에게 보내지는 정보를 약간 바꾸거나(예를 들어 두 명이 동시에 벗어났다고 거짓 정보를 주거나), 반경의 크기를 수시로 조금씩 바꾸어(예를 들어 반경을 9, 10, 11미터 중 하나를 임의로 선택하여) 정보를 주면 개인정보 누출의 가능성은 크게 줄어들 것이다. 이러한 방법이 바로 차등정보보호가 제시하는 정보보호의 정의를 구현하는 것이다.

마지막으로 웹툰에서 다음의 상황을 상상해보자. 내 주위 10미터 반경 안에 현재 10명이 있는데 나를 좋아하는 사람이 한 명이라고 알람이 울렸다. 만약 10명 중 9명이 모두 내가 아는 사람들이고 나를 좋아할 가능성이 전혀 없는 사람들이라면, 마지막 한 명이 나를 좋아하는 사람이라는 사실을 알게 된다. 차등정보보호는 이러한 경우에 이용자가 본인을 좋아하는 사람이 누구인지 파악하기 어렵게 하는 것을 정보보호(privacy protection)로 정의하는 통계적 노출제어 기준이다.

2. 차등정보보호의 기준과 성질

가. 새로운 시각 : 정보의 생성과 정보보호 기준

빅데이터 시대의 환경 변화 및 그에 따른 새로운 위험에 대처하기 위하여, 정보보호에 대한 새로운 개념을 제시해야 하는 필요성이 최근 연구자들을 자극하였다. 전통적으로 데이터베이스를 구축하고 정보를 제공하는 IT분야에서는, 개인정보가 보호되는 데이터베이스(privacy protected database)를 구축하는 여러 가지 방법들, 즉 암호화, 익명화, 코딩 등에 대한 방법들이 개발되어 왔다. 자료를 분석하는 통계학 분야에서는, 자료를 공개할

때 공개된 자료에서 정보 누출의 위험을 억제하는 잡음첨가, 바꾸기, 교환, 재코딩 등에 대한 방법이 제시되어 왔다. 대표적으로 k-익명화(k-anonymization)는 공개하는 자료에서 식별 변수들(key variables)의 각 조합에 해당하는 레코드 수를 제어함으로써 유일한 개체의 정보가 연계를 통하여 누출될 가능성을 줄이는 방법이다(Sweeny, 2002; Li et al., 2007; Duncan et al., 2011; Lee, 2013; 박민정, 2015).

위에서 언급한 전통적인 방법들은 노출제한 방법을 적용하여 자료를 공개하더라도 여전히 자료에 남아있는 개인정보 유출 위험성을 정보보호의 기준으로 고려한 것이다. 즉, 대중에게 공개되는 최종적인 자료가 얼마나 개인정보 노출의 위험성을 가지는지에 따라서 정보의 보호 정도가 정해진다. 예를 들어 k-익명화 방법은 k의 값이 커질수록 정보보호가 강화된 자료이며, 식별변수가 많이 공개될수록 정보유출의 가능성이 높다고 판단한다.

앞에서 제시된 최근 사례들은 전통적인 노출제어 방법이 적용된 비교적 안전한 자료라고 판단하여 자료를 공개했는데도 불구하고, 공격자의 교묘하고 지속적인 노력에 의하여 예상하지 못한 개인의 민감한 정보가 누출될 수 있다는 것을 보여준다. 매사추세스 주의 보험기록 누출 사례는 예상하기 힘든 두 자료의 결합으로 발생하였으며, 뉴욕시의 택시 운행기록에 대한 사례에서는 공개한 자료에 아예 유명 배우에 대한 정보는 존재하지 않는다. 웹툰의 사례는 아직 일어난 사건은 아니지만 우리가 곧 직면할 환경에 대한 암시를 준다. 이렇게 상상할 수 없을 만큼의 다양하고 복잡한 경로를 통하여 개인정보가 유출되기 때문에, 단순한 유출의 경우를 예상하고 고안된 전통적인 방법들은 그 목적을 달성할 가능성이 매우 작아진다.

이러한 상황의 변화에 대응하여 개인정보보호에 대한 새로운 시각이 최근 제시되고 있다. 이제 개인의 정보보호는 최종적으로 공개되는 정보(공개자료, 분석결과; output)가 가진 유출 위험성만을 고려해서는 부족하다. 대신에 자료가 만들어지거나 정보가 생성되는 과정(process, analysis, computation)을 고려하는 정보보호의 기준을 만들고, 그에 적합한 정보보호 방법이 적용되어야 한다는 새로운 관점이 제시되고 있다(Nissim et al., 2017; Narayanan et al., 2015; Kenthapadi et al., 2013).

전통적 방식의 자료의 공개는 단순한 일회성 절차에 기반을 둔 것들이 많다. 공신력 높은 기관이 보유하고 있는 자료(input)에 여러 가지 보안 절차를 적용하여 최종적으로 공개할 자료(output)를 만들고, 대중은 이렇게 공개된 자료를 이용하는 과정을 따르기 때문이다. 많은 경우, 자료가 생성되는 과정이 한 번으로 끝나고(예를 들어 조사통계) 대중은 공개된 자료만 볼 수 있다. 때문에 정보유출의 위험성과 정보보호의 기준은 공개된 자료(output)에서 정의되었다. 그 대표적인 예가 k-익명성이다. k-익명성의 기준은 공개할 자료에만 적용되는 것이고, 어떻게 자료가 만들어졌는지에 대한 고려는 하지 않는다.

그러나 이제 사람들은 실시간으로 분석된 정보를 지속적으로 얻을 수 있는 환경에 살고 있다. 주식의 등락, 미세먼지의 농도 등 사람들이 관심 있어 하는 정보는 실시간으로 갱신되며 쉽게 확인할 수 있다. 더 나아가 빅데이터와 클라우드 환경이 발달하면서 데이터베이스에 쉽게 접근할 뿐만 아니라, 대화형 자료 분석 서버 등을 이용하여 필요한 정보를 맞춤형으로 얻을 수도 있다. 이러한 환경에서는 악의를 가진 공격자가 정보의 생성 과정과 도구들을 활용하여 민감한 정보를 유출할 수 있는 가능성이 매우 커진다.

다음은 Nissim et al.(2017)에서 제시된 두 가지 예제이다. 이 간단한 예제를 통해서 정보보호의 기준을 정하기 위해서는 공개되는 최종 자료뿐 아니라 정보를 생산하는 과정, 즉 입력과 출력 간의 상호 관계를 고려해야 한다는 것을 알 수 있다.

> A와 B는 대학교의 교직원으로서 학생들의 학업 성취도에 관한 요약 통계를 발표 자료에 포함시키고자 한다. 이를 위해 대학교를 대표하는 한 명의 학생과 성적에 대한 정보를 발표하려고 한다. 개인정보보호법은 교육 기관이 보유한 특정 학생의 정보를 보호하도록 정하였기 때문에, 특정 학생의 이름이나 성별 등 학생의 신상 정보를 공개할 수는 없다.

> A는 학생들을 잘 대표할 수 있는 사람이 현재 학생회장이라고 판단하여 학생회장의 이름, 성별, 신분을 모두 숨기고 학교를 대표하는 학생의 평균학점이 3.2라고 발표한다.

> B는 학교 학생들의 정보를 보고 가장 흔한 성이 “김”씨, 가장 흔한 이름이 “철수”, 전체 학생들 평균학점의 중앙값이 3.1임을 파악하고, 학교를 대표하는 김철수 학생의 평균학점이 3.1이라고 발표한다.

첫 번째 A의 예는 겉으로 보기에는 이름, 성별 등 학생의 신분에 대한 어떠한 정보도 공개하지 않았기 때문에 안전할 것 같다. 하지만 자료의 생성과정을 알고 있다면, 즉 선택된 학생이 학생회장이라는 사실을 안다면, 쉽게 개인정보가 노출될 수 있다. 두 번째 B의 예는 겉으로 보기에는 이름을 공개하여 매우 노출 위험이 높아 보이지만, 자료가 생성되는 과정을 알면 실제 위험은 첫 번째 예제보다 낮다는 사실을 보여준다.

나. 위험을 대하는 새로운 기준: 한 사람의 차이

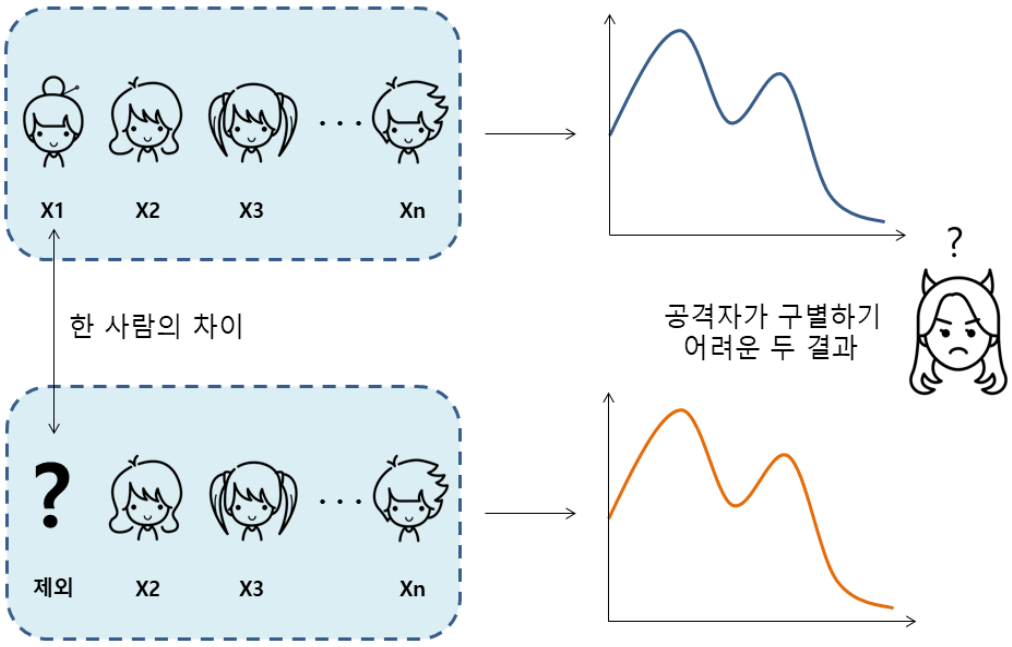
지금까지 빅데이터의 등장, 실시간 자료의 상호 교환, 데이터베이스에 접속하여 지속적으로 정보를 얻을 수 있는 자료 분석 서버 등, 최근에 우리가 처한 상황의 변화를 살펴보았다. 나아가 이러한 환경에서는 정보보호 여부를 공개된 자료를 기준으로 정하는 것이 아니라, 정보를 생산하는 절차를 고려해야 한다는 새로운 시각을 알아보았다. 이제 이러한 변화와 새로운 시각에서는 어떤 기준과 방법을 통하여 개인정보 노출의 위험을 막을 수 있는지를 생각해보자.

최근에는 온라인 조사, 핸드폰의 다양한 어플리케이션 등을 통하여 개인의 정보를 제공하게 된다. 조사통계에 응답자로서 참가하는 경우뿐만이 아니라, 이렇듯 새로운 방식으로도 개인의 정보를 제공하게 되면서 개인정보 노출의 위험을 좀 더 진지하게 다루어 볼 필요가 있다. 다음과 같은 단순한 가상의 예제를 생각해 보자.

> 철수는 개인의 소득과 자산의 현황을 조사하는 전국적인 조사통계의 응답자로 선정되었다. 선정된 사람은 자신의 직장, 소득, 부채 및 자산에 대한 정보를 제공하고 소정의 보상금을 받는다.
>
> 철수는 노출제한이 적용된 자료가 공개된 경우에도 개인의 정보가 누출될 위험성이 있다는 사실을 알고 있다. 철수가 실제로 우려하는 것은 자기가 조사에 참가하여 정보를 제공한 후에 자신의 자산에 대한 정보가 유출되어 신용등급 등에 나쁜 영향을 주게 되는 사건이 일어날 수 있다는 것이다.

전통적인 조사통계 시절에 비하여, 최근의 정보를 수집하고 공개하는 새로운 환경에서, 위의 예와 같은 노출에 대한 위험은 사람들이 더욱 걱정하는 부분이 되었다. 또한 사람들이 우려하는 개인정보 노출이 실제로 일어나는 경우도 많아지고 있다.

이러한 개인정보의 노출위험과 보호를 새롭게 이해하려면, 어떤 환경에서 개인정보가 완벽하게 보호될 수 있는지 생각해볼 필요가 있다. 위의 철수의 예에서 철수가 개인정보를 지키기 위해 취할 수 있는 가장 자연스러운 방법은 조사에 참가하지 않는 것이다. 철수가 조사에 참가하지 않으면 공개된 자료에는 자신의 정보가 전혀 없으므로 자신의 정보가 노출될 가능성은 없다. 이러한 시나리오는 [그림 5-3]으로 설명할 수 있다. [그림 5-3]은 철수가 조사에 참가하여 공개되는 자료를 구성하는 경우와 철수가 조사에 참가하지 않는 경우라는 두 가지 시나리오를 나타낸다.



[그림 5-3] 한 개인의 참여여부에 따른 두 개의 시나리오와 차등정보보호

차등정보보호는 위에서 본 것처럼 한 명의 개인이 실제로 자료의 분석에 포함되더라도 마치 포함되지 않는 경우와 같은 안전한 보안 환경을 목표로 한다. 더 나아가 차등정보보호는 이러한 안전한 환경이 철수 한 명뿐 아니라 조사에 참가한 모든 사람에게 제공되는 것을 목표로 한다.

만약 철수가 조사에 참가하여 데이터베이스에 그의 정보가 포함되었다고 하자. 차등정보보호 기준으로 보호된 환경이라면, 철수가 포함되어 제공되는 자료에서 얻을 수 있는 철수에 대한 정보는 철수가 제외된 자료에서 얻을 수 있는 철수에 대한 정보보다 크지 않다. 즉, 철수가 데이터베이스에 있는 경우에 생산된 정보와 없는 경우에 생산된 정보의 차이가 없어야 정보보호가 실현되었다고 보는 것이 차등정보보호의 의미이다.

다. 개인정보를 보호하는 방법

위에서 살펴본 차등정보보호의 정의와 기준을 만족하는 환경을 구현하려면 구체적으로 어떤 방법을 사용해야 할까? 이제 Nissim et al.(2017)에서 소개된 간단한 사례를 들어서 차등정보보호를 구현하는 대표적인 방법을 제시하려고 한다. 먼저 노출을 제한하는 조치를 취하지 않아서 개인의 민감한 정보가 노출되는 예를 생각해 보자.

A와 B는 사립 대학교의 교수이다. 두 교수 모두 학생이 받는 장학금 지원과 관련된 정보와 기타 학생의 개인정보가 들어있는 데이터베이스를 이용하려고 한다. 학교가 보유한 학생 데이터베이스는 개인정보가 포함되어 있기 때문에 데이터베이스에 대한 이용이 철저히 제한된다.

이용허가를 얻기 위해 교수 A와 B는 기밀 유지 교육을 받고 데이터베이스에서 얻은 개인정보의 사용 및 공개를 금지하는 데이터 사용 계약서에 서명함으로써 개인자료 처리를 위한 대학의 규정에 따르겠다고 약속한다.

3월에 교수 A는 이 데이터베이스의 정보를 기반으로 한 연구 결과를 신문 기사에 발표하였다. 이 기사에는 다음과 같은 분석결과가 실렸다. "대학의 현재 재학생은 3,000명이며 이중 200명은 연간 2억 원 이상의 수입을 가지는 가구의 구성원입니다." 교수 A는 3,000명과 200명과 같은 집계된 통계만을 발표했으므로 개인정보는 공개되지 않다고 판단한다.

다음 달 4월 교수 B는 다음 통계를 포함하는 기사를 발표한다. "대학교 현재 재학생 2,999명 중 199가구가 연간 2억 원을 초과하는 가구 소득을 가지고 있습니다."

교수 A 와 B는 같은 데이터베이스를 이용하여 둘 다 비슷한 정보를 발표하지만 서로 상대방의 연구결과는 모른다.

다시 아래에서 제시된 가상의 사례는 서로 전혀 알지 못하는 독립적인 주체에 의하여 생산된 자료 이용 결과가 다른 외부의 정보와 결합되면 개인정보 노출이 일어날 수 있다는 것을 보여준다.

학생 C는 이 두 기사를 읽고 3월과 4월 사이에 한 명의 학생이 대학을 그만두었으며 그 학생의 학부모는 연간 2억 원 이상의 수입을 얻는다고 결론을 내린다. 한편 학생 C는 주위의 친구들에게 물어서 학생 D가 3월 말쯤에 중퇴했다는 사실을 알게 된다. 이제 학생 C는 친구들에게 학생 D의 부모님이 아마도 1년에 2억 원 이상의 소득을 올린다고 알려준다.

학생 D는 이것에 대해 듣고 친구가 그의 부모의 재정 상태에 대해 알게 된 것에 당황한다. 그는 대학에 이러한 개인정보 누출에 대하여 항의하였고 대학은 교수 A와 B에게 이러한 개인정보 유출이 일어난 것에 대한 해명을 요구한다. 하지만 교수 A와 B은 모두 개인정보의 보호를 위하여 집계된 정보만을 발표했으며 발표된 정보만으로는 어떤 개인도 식별하지 못한다고 주장한다.

학생 C도 교수들의 연구를 보기 전까지 학생 D의 정보에 관심을 가진 것이 아니다. 이러한 사례와 같이 몇 개의 관계가 없을 것 같이 보이는 정보들이 결합되어 새로운 정보를 생성하여 결과적으로 개인의 정보가 노출되는 사건을 미리 예측하기는 매우 어렵다. 또한 정보의 생산자들(예를 들어 교수 A와 B)이 다양한 정보들의 결합으로 개인의 신분이 노출되는 경우를 모두 상상 또는 예상하여 이를 방지하는 노출 제한 기법을 적용하는 것은 실제로 불가능하다.

이제 위의 사례에 대하여 차등정보보호를 구현한 방법이 적용된 환경을 생각해보자.

사립 대학교에서는 연구 목적으로 학생 데이터베이스를 이용하려는 연구자들에게 자료 분석 서버를 통하여 필요한 통계에 대한 질의어(쿼리, query)를 접수하도록 한다. 자료 분석 서버는 연구자의 질의어에 대응해서 통계를 생산할 때 적당한 잡음을 첨가하여 연구자에게 보낸다. 즉, 연구자는 자신이 접수한 질의에 대한 정확한 답을 얻는 것이 아니라 근사적인 값을 얻는 것이다.

이제 교수 A 와 B는 위와 같은 절차를 따라서 데이터베이스를 분석하고, 얻어낸 근사적인 결과를 이용해 다음과 같은 연구결과를 발표한다.

3월에 교수 A는 이 데이터베이스의 정보를 기반으로 한 기사를 발표한다. 이 기사에는 다음과 같은 분석결과가 실렸다. “대학의 현재 재학생은 3,002명이며 이중 202명은 연간 2억 원 이상의 수입을 가지는 가정의 구성원입니다.”

다음 달 4월 교수 B는 다음 통계를 포함하는 기사를 발표한다. “대학의 현재 재학생 2,998명 중 197가구가 연간 2억 원을 초과하는 가구 소득을 가지고 있습니다.”

이제 학생 C는 이 두 기사를 읽고 더 이상 3월과 4월 사이에 한 명의 학생이 대학을 그만두었다는 사실을 추론할 수 없다. 따라서 학생 D가 3월 말에 중퇴했다는 사실을 알게 되더라도, 학생 D의 부모님이 1년에 2억 원 이상의 소득을 올린다는 개인정보는 유출되지 않는다.

위에서 교수 A와 B가 얻은 차등정보보호가 적용된 분석결과(3,002, 202, 2,998, 197)는 질의어에 대한 정확한 답(3,000, 200, 2,999, 199)이 아니라 정교하게 설계된 잡음첨가 방법이 적용된 근사적인 답이다. 이렇게 데이터베이스에 요구된 질의에 대하여 잡음이 첨가된 응답을 생산하는 절차를 차등적으로 보호된 계산(differentially private computation)이라고 한다.

차등(differential)이라는 사전적인 의미는 차이, 격차, 차별이라는 뜻이다. 차등정보보호(differential privacy)는 입력에서 “작은 차이”가 출력에서 “중대한 차이”를 줄 수 있는 자료생성과정을 제어하여 정보보호를 구현하려는 개념이다. 입력에서의 “작은 차이”는 동일한 데이터베이스가 특정한 하나의 레코드를 포함하는지 하지 않는지에 대한 차이이다. “중대한 차이”는 미세하게 다른 두 개의 자료에서 얻은 출력의 차이를 보고 역으로 입력의 작은 차이(측정 레코드의 포함 여부)를 유추할 수 있는 것을 의미한다. 즉, 차등정보보호는 미세하게 다른 데이터베이스의 분석 결과들이 보이는 차이를 간파하여 공격자가 개인의 정보를 유출할 수 있는 위험을 제거하자는 개념이다(Dwork 2006; Dwork et al., 2006).

개인정보 노출 문제를 다루는데 있어 과거의 전통적인 정보보호 방법들은 하나의 레코드가 유일한 것(uniqueness)이 되는 위험을 낮추는 것을 목표로 하였다(Franconi와 Polettini, 2004). 따라서 공개된 자료에서 유일한 정보를 가진 레코드의 수를 제한하는 방식의 방법들이 주로 고안되었으며, 결과적으로 공개할 수 있는 정보의 다양성이 크게 제한되었다. 이러한 현상은 앞에서 언급한 것과 같이 최종적으로 공개하는 자료의 위험성만을 고려하기 때문에 발생한다. 빅데이터 시대에서 요구되는 분석환경, 즉 자료를 이용하려는 시도가 많아지고 분석의 목적이 다양한 환경에서는 유일성의 제한에 관한 개념만으로 유용성과 정보보호 모두를 획득하기가 어렵다.

반면, 자료를 이용하려는 시도가 많아지고 분석의 목적이 다양하게 변화된 환경에서 차등정보보호는 자료에서 정보를 생성하는 과정을 염두에 둔다. 차등정보보호는 “두 개의 자료에 존재하는 작은 차이”에 의한 정보유출의 위험성을 고려하는 개념으로 이해할 수 있으며, 차등(differential)의 의미도 이러한 관점에서 해석할 수 있다.

또한 차등정보보호는 오래전부터 있었던 “개인정보가 보호된 통계적 데이터베이스(privacy-preserving statistical database)”를 구현하기 위한 정보보호 기준이며, 동시에 새로운 정의라고 할 수 있다(Agrawal과 Srikant, 2000; Bertino et al., 2005). 오랫동안 연구자들은 데이터베이스가 모집단에 대한 유용한 정보를 제공할 때, 개인의 민감한 정보가 유출될 가능성을 낮추는 다양한 방법들을 개발해 왔다. 그러나 대부분 특정 분석 방법에 대한 개별적인 방법들이 주로 제시되었고, 넓은 의미에서 정보보호의 통합적인 정의에 대한 연구는 미진하였다. 이에 비해, 차등정보보호는 개인정보를 보호할 수 있는 데이터베이스를 구현할 때 사용될 수 있는 일반적인 기준에 대한 정의이다.

마지막으로 차등정보보호는 예상할 수 없는 정보들의 결합이 증가하는 경우에 있어서도, 이에 의한 노출의 위험성이 급격하게 증가하지 않도록 설계된 정의이다. 앞의 예에서 보았듯이 데이터베이스를 분석한 독립적인 연구자들과 우연한 단 한 명의 공격자에 의해서도 예상할 수 없는 정보의 결합이 이루어지고, 나아가 그로 인한 개인정보의

노출이 발생할 수 있다. 그러나 차등정보보호의 기준이 구현된다면, 각각 독립적으로 생산된 정보가 서로 결합되어도 개인에 대한 정보를 유추할 수 있는 노출 위험이 급격하게 증가하지는 않는다. 앞의 예에서 차등정보보호를 기준으로 주어진 잡음이 첨가된 결과가 각 연구자들에게 제공되면, 그것들이 서로 결합되어도 개인의 정보가 유출될 새로운 위험성을 크게 증가시키지는 않는다. 다만 여기서 주의할 사항은 지속적으로 정보가 생산되는 경우에는 비록 잡음이 첨가되었더라도 노출 위험성이 증가하기는 한다는 것이다. 차등정보보호가 아무리 정교하게 구현되었더라도 공개되는 정보의 양이 증가하면 개인정보 유출의 위험성을 일정한 상태로 유지하거나 낮추는 것은 불가능하다. 다만 차등정보보호를 적용하는 것으로써 위험이 증가하는 속도를 제어할 수 있는 것이다.

예를 들어 살펴보자. 앞에서 제시한 사례에서 자료 분석 서버가 교수 A와 B가 질의한 통계분석의 결과를 제공할 때, 실제 결과와 잡음이 첨가된 근사 결과의 차이는 최대로 2명이다. 앞의 사례보다 더 작은 잡음을 첨가하면(예를 들어 최대 차이를 1명으로 하면) 연속적인 분석 결과의 결합에 따른 노출의 위험성이 더 빠르게 높아질 것이다. 더 큰 잡음을 첨가하면(예를 들어 최대 차이를 5명으로 하면) 노출의 위험성은 더 느리게 증가할 것이다. 또한 전통적인 노출제어 방법과 유사하게, 잡음의 크기가 커지면 분석 결과의 유용성(utility)은 감소한다.

차등정보보호가 구현된 결과라도, 비록 적은 양이라 할지라도, 개인의 정보는 유출되는 것이고 일련의 분석이 차례로 진행되면 개인정보의 유출 위험성은 계속 증가하게 된다. 이러한 점을 고려하여 차등정보보호에서 증가되는 노출의 위험성을 어떻게 어떤 속도로 제어할 수 있는지에 대한 방법을 생각해야 한다.

라. 차등정보보호에 의한 노출위험 제어

노출제한 방법의 적용 여부에 관계없이 자료가 공개되면 자료에 포함된 개인에 대해서는 정보 노출의 위험이 발생한다. 노출제한 방법을 적용하는 이유는 이를 적용하지 않는 경우보다 노출의 위험을 낮추기 때문이며, 노출의 위험이 얼마나 줄어들었는지 판단하는 것은 현실적으로 매우 중요한 일이다.

이 절에서는 차등정보보호가 어떤 위험을 제어할 수 있고 얼마나 제어할 수 있는지 도식적인 예제를 통하여 소개하려고 한다. 아래의 예제는 보험회사, 통신회사와 가상적인 고객의 관계를 통하여 차등정보보호로 보호할 수 있는 개인정보, 보호할 수 없는 영역, 노출 위험의 상대적 크기 등을 설명하고자 인위적으로 단순하게 만든 것이다.

1) 사고의 위험과 보험료

자동차 사고보험을 다루는 보험회사와 그 회사를 이용하고 있는 한 명의 고객에 대한 사례가 다음과 같다.

> 자동차 보험회사는 차량에 의한 사고가 날 때마다 고객들로부터 받은 보험료에서 보험금을 지급한다. 고객들로부터 보험료를 받을 때, 고객이 모두 똑같은 금액을 지불하지는 않는다. 비싼 차를 소유할수록, 사고를 낼 확률이 많은 사람일수록 더 비싼 보험료를 회사에 지불한다. 회사는 어떻게 보험료를 산정할까? 자동차 보험회사는 과거의 차량 사고 기록을 분석하여 개인별 보험료를 책정한다. 20대인 사람이 50대인 사람보다 보험료가 많고, 최근 사고를 자주 낸 사람이 오랫동안 무사고인 사람보다 보험료가 높다. 보험료가 상대적으로 높은 이유는 사고가 날 확률이 높기 때문이다. 보험회사는 사고 정보와 다양한 연구결과를 분석하여 사람들과 차량의 특성에 따른 사고 확률을 계산하여 보험료를 정한다.
>
> 이제 배달업에 종사하는 40세 김철수는 차량 1대를 소유하고 있으며, 보험회사에 2018년에 연 10만 원을 보험료로 내고 있다고 하자. 보험회사는 차가 완전히 파손되었을 경우 100만 원의 손해보험금을 지급해야 하고 김철수의 나이, 직업, 차종에 의하여 사고가 날 확률이 10%라고 계산하여 10만 원의 손해 보험료를 결정하였다.

2) 피할 수 없는 노출 위험과 노출위험의 최소 기준

이제 다음과 같은 경우를 생각해 보자.

> 어떤 연구자가 자동차 사고 정보를 심층적으로 분석하였다. 그리고 후진 시에 차량 후방에 대한 영상을 보여주지 못하는 차량의 사고의 위험이 평균적으로 $2\%p$ 더 높다는 연구결과를 발표하였다. 보험회사는 후방 카메라가 없는 차량에 대하여 2019년에 보험료를 추가로 올리는 결정을 내렸다.
>
> 김철수의 차는 후방 카메라가 없는 모형이기 때문에 보험사는 2019년에 김철수의 보험료를 10만 원에서 12만 원으로 올린다고 통보하였다. 이러한 결정은 후방카메라가 없는 김철수의 차가 사고 날 확률이 12%라는 추정에 의거한 것이다.

위의 사례와 유사한 보험료의 변화는 흔히 일어나는 일이다. 전체 모집단의 특성을 분석하여 그 결과를 보험료 산정에 이용하는 일은 자주 일어난다. 예를 들어 담배를 피우는 사람이 안 피우는 사람보다 심장질환이나 특정 암의 발생확률이 높다는 것은 잘 알려진 결과이며, 따라서 흡연의 유무에 따라 보험료가 다르다는 것도 잘 알려진 사실이다.

위의 사례는 일종의 추정노출(inferential disclosure)의 결과라고도 할 수 있다. 즉 연구에 자신이 참가했건 참가하지 않았건, 일반적인 모집단의 특성에 대한 추정의 결과로 인하여 자신의 정보가 노출되는 것이다. 노출된 정보는 개인에게 나쁘게 작용할 수 있다. 위의 예에서 김철수의 차에 후방카메라가 없기 때문에 있는 차보다 사고 확률이 $2\%p$ 높을 것이라는 추정이 있었다. 이는 개인에게 불리하게 작용하여 보험료가 2만원 증가하였다. 담배를 피우는 사람이 안 피우는 사람보다 폐암의 가능성이 2배 높다는 연구결과가 발표되었다고 하자. 생명보험 회사가 김철수가 담배를 피우는 사실을 알고 있다면 그에게 폐암이 발생할 확률이 2배 높다는 사실을 추론할 수 있고, 결과적으로 보험료가 올라갈 수도 있다.

이러한 유형의 노출은 대부분의 경우 크게 심각하지 않을 것 같지만, 민감한 정보가 노출되는 경우가 있을 수도 있다. 만약에 배달 업종에 근무하는 사람들의 평균 연봉이 5000만 원이라는 정보가 발표되었다면, 근무 기간이 오래된 김철수의 연봉이 최소한 5000만 원일 것이라고 추론하게 된다. 연봉이 민감한 정보라면 이러한 노출로 생기는 개인정보 유출도 사소한 것이 아닐 수 있다.

이제 개인정보 유출에 대한 가능성과 그로 인한 결과를 확률과 연관하여 생각해보자. 김철수의 자동차 보험료는 그가 사고를 낼 확률에 비례하여 책정되었다. 앞의 예제를 보면 후방카메라가 없는 차가 있는 차보다 사고의 확률이 $2\%p$ 높으므로 이 확률에 의하여 계산된 보험료의 증가율도 $2\%p$로 정해진다. 처음에는 차의 가치가 100만 원이고 사고가 발생할 확률이 10%라는 사실로 보험료가 10만 원으로 책정되었다. 이 후, 후방카메라가 없는 차가 있는 차보다 사고의 가능성이 $2\%p$ 높으므로 보험료를 2만 원 올린 것이다. 이때 김철수에게 자신의 정보가 이용되지도 않은 연구의 결과로 인한 $2\%p$ 보험료 인상이 일어났으며 이는 일종의 추정노출로 인한 것이다. 이러한 보험료 인상은 피할 수 없는 노출의 결과이다.

차등정보보호를 포함한 일반적인 정보보호 방법으로 이러한 노출은 막을 수는 없다. 이러한 유형의 노출은 최근 더욱 증가하고 있으며 앞으로도 증가할 것이지만, 이제 이러한 피할 수 없는 노출을 정보보호의 범주로 고려하지는 않는다.

다만 개인정보 노출의 위험을 평가하거나 정보보호의 기준을 고려할 때, 이러한 유형의 노출 위험성을 가장 낮은 수준의 위험, 즉 가장 낮은 수준의 노출위험으로 생각할 수 있다. 어떤 정보보호를 위한 방법으로 개인 노출의 위험을 얼마나 낮출 수 있는지 판단할 때 이러한 유형의 노출 위험성을 기준(baseline)으로 고려할 수 있다.

다음에 나오는 사례들에서는 위의 사례에서 나타난 피할 수 없는 노출로 인한 개인정보의 유출과 그로 인한 결과(후방 카메라가 없으면 사고의 위험이 $2\%p$ 증가하기 때문에 보험료가 증가한다)를 기준으로 삼아서 노출의 위험을 비교하려고 한다. 이러한 논리는 차등정보보호의 기준을 만드는데 사용되는 논리이기도 하다.

3) 정보의 제공과 노출의 위험

이제 개인정보를 제공할 때 유용성과 위험성에 대한 사례를 살펴보자.

> 김철수가 이용하는 통신회사는 네비게이션을 무료로 사용하는 대신 차량의 운행기록 데이터(운행 경로와 속도 등)를 회사의 서버에 저장하고 분석할 수 있게 해달라는 제안을 하였다. 회사는 운행기록을 분석하여 더 좋은 네비게이션 서비스를 하는 것이 자료 저장의 목적이라고 설명하였다.
>
> 김철수는 통신회사의 제안을 받아들여서 무료로 서비스를 이용하고 싶다. 또한 운행 자료도 제공하여 향후 더 빠른 최적의 경로를 찾는 개선된 서비스를 받고 싶다. 하지만 자신의 정보가 유출되는 것을 우려하고 있다.
>
> 정보제공을 우려하는 가장 큰 이유는 통신회사와 보험회사가 고객 정보를 공유하고 있으며, 특히 보험회사는 운전자의 운전 행태를 분석하여 보험료를 차등적으로 부가하고자 한다는 소문을 들었기 때문이다.

위와 같은 상황은 요즘에 흔히 일어날 수 있는 상황이다. 빅데이터의 시대에 들어서면서 사람들이 다양하고 편리한 새로운 서비스를 이용할 기회도 생기지만, 개인정보 유출에 의한 피해도 함께 증가하게 된다.

김철수는 자신의 정보(운행 기록)가 보험회사에 제공되면 보험회사가 그 정보를 포함해 자료를 분석하고 그 분석의 결과를 기반으로 보험료가 오를 수 있다는 우려를 하게 된다. 즉, 개인의 중요한 정보가 노출되어 그 결과가 자신에게 불리한 상황이 생길 수 있다는 생각을 하게 된다. 이는 충분히 있을 수 있는 일이다. 또 한편으로는 자신의 정보를 제공하면 네비게이션을 무료로 이용할 수 있고 더 좋은 서비스를 받을 수 있다는

이점도 있다. 이렇게 고객은 개인정보 노출의 "위험"과 경제적 혜택 및 편리성 증진이라는 "이익" 사이에서 갈등하게 된다.

반면 통신회사는 고객의 운행기록을 수집하여 더 좋은 서비스를 개발하고 싶고, 나아가 보험회사와 고객의 정보를 공유하여 새롭고 가치 있는 정보를 얻고 싶을 것이다. 보험회사 또한 다른 외부의 유용한 자료를 자신이 가지고 있는 자료에 더하여 보험료 계산을 더 정교하고 정확하게 하고 싶어 한다. 이러한 경우 통신회사와 보험회사는 실제로 어떤 방법으로 고객의 개인정보를 보호하고 어떻게 고객을 설득하여 운행기록을 얻을 수 있을까?

정보를 수집하는 통신회사는 개인정보를 보호하는 적절한 방법을 적용해야 하고, 적용된 방법이 노출의 위험성을 얼마나 낮출 수 있는지 알 수 있어야 한다. 차등정보보호는 피할 수 없는 노출에 의한 위험을 기준으로 했을 때, 자신이 포함된 자료에서 개인의 정보가 유출되어 발생하는 위험의 상대적 크기를 확률적으로 제한하는 기준이다.

4) 개인정보 제공 후 노출에 의한 결과

다음과 같이 개인의 정보가 제공된 후, 노출이 일어나고 개인에게 손해되는 결과가 발생하는 사례를 생각해 보자.

> 만약 김철수가 통신회사의 제안을 수락하여 운행기록을 제공하고, 보험회사는 통신사가 제공한 고객의 차량 운행 자료를 이용한 분석을 실시했다고 가정하자.
>
> 김철수의 운전 기록과 행태를 분석한 결과, 보험회사는 김철수가 주로 이용하는 경로들이 다른 경로보다 사고율이 높은 사실을 알아냈다. 또한 김철수가 사고 낼 확률이 10%가 아닌 30%라고 결론을 내렸다.
>
> 이러한 사고확률 20%p 증가라는 결론을 보험료에 반영하면 김철수의 보험료는 추가로 20만 원이 올라야 한다. 즉, 총 보험료가 10만 원에서 32만 원으로 올라야 한다. 2만 원은 후방카메라의 부재에 의한 사고위험 증가를 고려한 추가 보험료이다.

자신의 정보를 제공할 때 위와 같이 개인정보의 노출로 인하여 손해를 보게 될 가능성이 있다면, 많은 사람이 자신의 정보를 공개하지 않을 것이다. 따라서 빅데이터 시대에 새로운 가치를 창조하려는 기업이나 기관은 개인의 정보를 전혀 수집할 수 없게 된다.

개인의 자료를 보유하고 있는 통신회사가 자료를 제공할 때 개인정보보호를 위하여 개인의 신분을 그대로 노출시킬 수 있는 직접 식별변수(이름, 성별, 주소 등)를 제거하였더라도, 보험회사가 자신이 보유한 보험 자료를 추가로 활용해서 개인을 식별할 수 있는 가능성은 존재한다. 예를 들어서 김철수가 과거에 낸 사고 기록을 검토하여 제공된 자료의 경로와 연계하여 개인을 식별할 수 있다.

따라서 위의 예에서 개인의 정보를 수집하는 통신회사가 보험회사에 정보를 제공하려면 정보의 유출을 우려하는 고객의 불안을 해소할 수 있는 적절한 개인정보보호 방법을 적용해야 한다. 이러한 방법을 적용할 때 고객에게 자신의 정보가 얼마나 잘 보호되고 있는지 구체적이고 양적인 기준을 제시할 수 있다면 매우 유용할 것이다.

5) 차등정보보호에서 개인정보 노출의 위험

이제 차등정보보호가 적용된 자료 분석 서버를 통해 보험회사가 자료를 이용할 수 있게 하며, 그 결과를 제공하는 경우를 생각해 보자

> 김철수가 통신회사의 제안을 수락하여 운행기록을 제공한다고 하자. 또한 통신사는 개인의 운행기록을 그대로 제공하는 대신, 차등정보보호 방법[9]이 적용된 분석의 결과만 볼 수 있게 한다고 하자. 즉, 통신회사는 보험회사에게 차등정보보호가 적용된 자료 분석 서버를 이용할 수 있도록 제공한다. 차등정보보호를 적용한 자료 분석 서버는 개인정보의 보호기준으로서 $\epsilon = 0.01$을 이용한다.
>
> 이제 보험회사가 차등정보보호를 적용한 자료 분석 서버에서 김철수의 운행기록을 포함한 자료를 이용하여 다양한 분석을 수행한다. 이러한 차등정보보호의 기준이 적용된 분석 결과를 이용하여 인상시킬 수 있는 김철수 보험료의 최대 증가분은 $2\times(1+\epsilon)$만 원이다. $\epsilon = 0.01$인 경우 가능한 보험료의 최대 증가분은 2×1.01만 원, 즉 2만 2백 원이다.

만약에 통신회사가 차등정보보호를 적용한 자료 분석 서버를 이용하게 한다면, 보험회사는 김철수라는 특정한 사람에 대한 정보만을 알아내기 매우 힘들다. 차등정보보호가 적용된 분석에서는 “김철수가 포함된 자료를 분석한 결과”와 “김철수가 제외된 자료를 분석한 결과”의 차이가 크지 않아야 하기 때문이다. 즉, 김철수가 포함된 자료에서 김철수에

9) 차등정보보호를 만족하도록 자료를 변환 또는 생성하는 알고리즘, 메커니즘을 일컫는다.

대하여 얻을 수 있는 정보의 크기는 김철수가 제외된 자료에서 얻을 수 있는 그에 대한 정보의 크기와 비교해서 큰 차이가 나서는 안 된다.

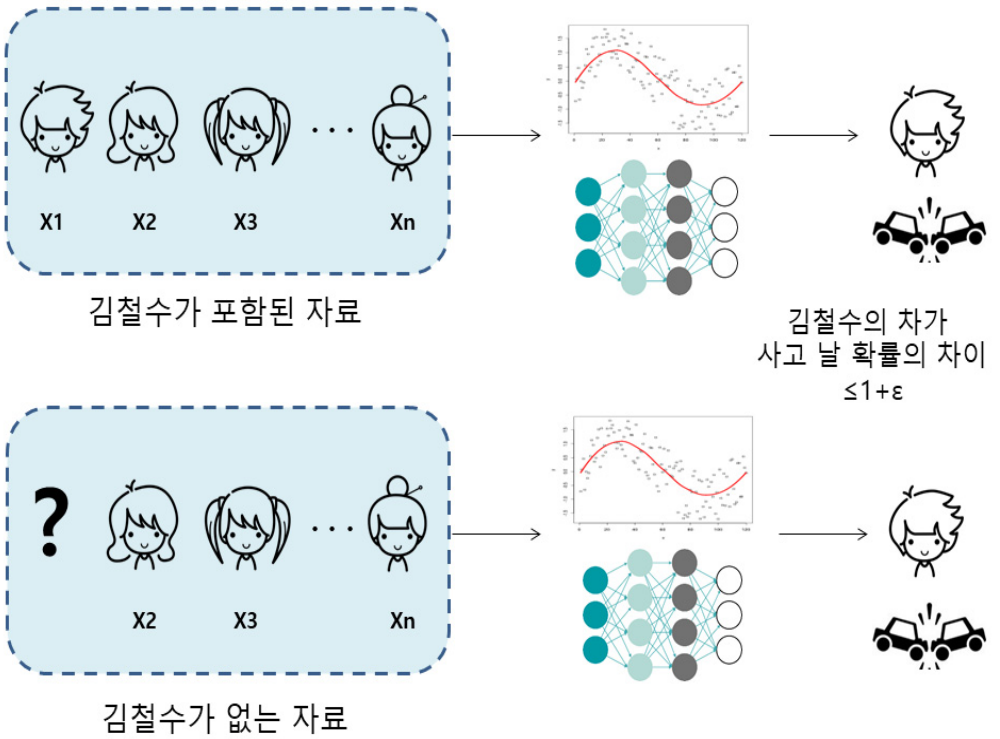
차등정보보호를 적용하면, 자료를 분석할 때 특정 개인의 정보만을 알아내기 힘들다. 이때 보호기준으로서 $\epsilon = 0.01$의 의미는 다음과 같다. 특정 개인에 대한 정보를 알아내는 사건을 최악의 경우라 하자. 보호기준 ϵ은 이런 경우의 노출위험도 피할 수 없는 노출의 위험(기준이 되는 최소 위험)보다 $(1+\epsilon)$배 크다는 것을 의미한다. 이러한 차이의 상대적 최대 한계를 ϵ이라고 하고, 이를 개인정보 예산(privacy budget)이라고 부른다. 개인정보 예산 ϵ이 커지면 개인정보의 유출 위험성이 커지며 ϵ이 작아지면 위험성이 줄어든다. 만약 개인정보 예산이 0이라면($\epsilon = 0$) 이는 개인의 정보가 완벽하게 보호되었다는 의미이며, 동시에 분석의 유용성(utility)은 거의 사라지게 된다. 이러한 의미에서 개인정보 예산 ϵ은 한 번의 분석을 수행할 때 사용하는 "개인정보의 총 크기"로 생각할 수 있다.

이제 개인정보 예산 ϵ의 의미를 좀 더 구체적으로 설명해보자.

6) 개인정보 유출의 최소 기준과 그에 따른 위험

후방카메라 부재에 의한 사고 가능성의 증가로 보험료가 2만 원 인상된 것은 차등정보보호의 적용여부에 관계없이 피할 수 없는 노출로 일어날 수 있는 결과이다. 즉 차등정보보호가 적용된 환경이든 아닌 환경이든, 김철수가 분석에 포함되었던 아니던 간에, 모집단에 대해 후방카메라가 없는 차량에서 사고 날 확률이 평균적으로 $2\%p$ 높다는 정보를 추론할 수 있다. 추론에 의한 노출이 일어나서 후방카메라가 없는 김철수의 보험료가 2만 원 증가하는 것은 어떠한 정보보호 기준으로도 막을 수 없다. 따라서 "후방카메라 부재로 인하여 사고발생 확률이 증가한다는 추론"에 의해 발생한 "김철수의 보험료 인상이라는 사건"을 "개인정보 유출에 의한 가장 낮은 수준의 기준 위험"으로 생각할 수 있다.

이제 "김철수가 제외된 자료로 얻을 수 있는 김철수의 사고 확률"을 생각해 보자([그림 5-4]). 이는 김철수를 제외한 전체 집단을 분석하여 추정된 결과를 활용하여 얻을 수 있는 정보이다. 이 데이터베이스로는 아무리 잘 해도 김철수의 개인의 정보를 알 수 없다. 즉 김철수가 없는 데이터베이스가 주어진 상황에서 김철수의 사고에 대한 확률을 추정하려면, 전체 모집단(물론 김철수 1명은 빠져있다)에 대한 추론에서 얻은 정보를 이용해서 김철수에 대한 추론을 할 수밖에 없다. 이런 상황은 앞에서 언급한 피할 수 없는 노출에 의한 개인정보 유출과 그에 따른 확률 계산에 국한되는 것이다.



[그림 5-4] 차등정보보호에 의한 노출 위험 제어

따라서 김철수가 제외된 자료를 분석한다면 위의 예제에서 제시한 일반적인 연구의 결과처럼 후방카메라가 없는 차는 사고확률이 $2\%p$ 증가한다는 것만을 알 수 있을 것이다. 이로 유추할 수 있는 것은 오직 "후방카메라가 없는 차를 소유한 김철수의 사고 확률이 $2\%p$ 증가한다."는 것이다. 이는 김철수의 개인정보를 이용하지 않고 김철수에 대해 얻을 수 있는 최대한의 정보이며, 김철수를 제외한 자료를 분석하는 경우 김철수 개인에 대해서는 이보다 더 좋은 정보를 얻을 수 없다.

7) 개인정보 노출위험의 최대 한계

차등정보보호가 적용된 자료 분석 시스템에서는 하나의 개인이 포함된 자료를 분석할 때 그 개인의 정보가 유출될 확률과 그에 따른 위험을 생각한다. 앞 예제의 경우, "김철수가 포함된 자료로 얻을 수 있는 김철수 노출 사고 확률"과 "그 확률에 기반한 보험료의 인상"으로 생각하면 된다.

우선 다시 한 번 차등정보보호의 개념을 생각해보자. 차등정보보호는 앞에서 정의한 “하나의 개인이 제외된 자료에서 얻을 수 있는 그 개인의 정보”를 기준으로 “개인이 포함된 자료에서 얻을 수 있는 개인의 정보”를 비교하고, 그 상대적인 최대 차이가 주어진 개인정보 예산 ϵ을 넘을 수 없다는 기준을 제시하는 것이다. 노출 위험의 상대적 차이는 피할 수 없는 최소의 위험에 대한 상대적인 위험의 크기이며, 이 위험의 크기는 확률의 단위로 계산한다.

차등정보보호에 의한 분석을 통해서 얻을 수 있는 두 확률의 상대적인 차이를 다음과 같이 나타낼 수 있다. 이것이 바로 차등정보보호의 기준이다.

$$\frac{\text{김철수가 포함된 자료로 얻은 김철수의 사고 확률}}{\text{김철수가 제외된 자료로 얻은 김철수의 사고 확률}} \leq 1+\epsilon$$

결론적으로 차등정보보호가 적용된 분석에서는, 김철수가 포함된 자료를 분석한 결과와 김철수가 없는 자료를 분석한 결과가 크게 다르지 않다. 또한 그 최대 한계는 $1+\epsilon$이다. 즉 자료 안에 특정 개인의 포함여부에 따른 개별 정보 노출 차이의 크기는 아주 작으며, 그 상대적인 노출 차이의 양을 확률의 비율로 나타내면 $\epsilon = 0.01$이라는 것이다.

앞에서 제시한 예에서는 보험료가 확률에 비례하여 책정되므로 차등정보보호의 기준을 보험료의 단위로 나타낼 수 있다. 보험료 인상의 단위를 이용해 두 확률 간 비율의 최대 한계 기준을 생각해보자.

$$\frac{\text{김철수가 포함된 자료로 얻은 보험료 인상치}}{\text{김철수가 제외된 자료로 얻은 보험료 인상치}} \leq 1+\epsilon$$

이제 차등정보보호의 기준을 만족하는 분석시스템을 통해 추론할 수 있는 김철수의 내년 보험료 최대 인상치는 $2\times(1+\epsilon)$만 원이 된다. 위의 식에서 “김철수가 제외된 자료로 얻은 보험료의 인상”은 앞에서 설명한 것과 같이 “피할 수 없는 노출에 의한 위험에 의한 최소 기준”으로 생각할 수 있어 다음과 같은 식을 유도 할 수 있기 때문이다.

$$\begin{aligned}
&\text{김철수가 포함된 자료로 얻은 보험료 인상치} \\
&\quad \leq (\text{김철수가 제외된 자료로 얻은 보험료 인상치}) \times (1+\epsilon) \\
&\quad \leq (\text{피할 수 없는 노출에 의한 보험료 인상치}) \times (1+\epsilon) \\
&\quad = 2\times(1+\epsilon)
\end{aligned}$$

이러한 결과는 통신회사가 고객들이 개인의 자료를 수집하려고 할 때, 통신회사가 추구하는 개인정보보호의 기준으로 고객에게 설명할 수 있다. 즉 차등정보보호라는 개인정보보호 기준을 설명할 때, 개인정보 유출에 의해 발생할 수 있는 손해를 확률로 구해지는 보험료의 증가라는 양적인 기준을 이용해 설명할 수 있다. $\epsilon = 0.01$의 경우, 아래와 같다.

> 고객이 운행기록을 제공해도, 통신회사는 차등정보보호가 적용된 분석결과만 보험회사가 이용하도록 제공할 것이며, 이러한 경우 개인정보는 매우 안전하다고 설명할 수 있다. 안전한 정도는 고객이 자신의 정보를 제공하지 않아도 발생할 수 있는 보험료 인상치 2만 원을 기준으로 할 때, 개인정보 이용(또는 노출)으로 발생할 수 있는 추가 보험료의 최대 인상치는 그것의 1%인 200원이라고 말할 수 있다.

마. 차등정보보호에서 고려할 사항

1) 차등정보보호는 정의이며 방법이 아니다.

차등정보보호는 정보보호의 기준으로 사용할 수 있는 정의(definition)의 개념이며 방법(method)은 아니다. 앞에서 살펴본 예제 중에서 교수 A와 B가 연구결과를 발표할 때 재학생의 수 및 2억 원을 초과하는 가구의 수에 잡음을 첨가하는 방법은 차등정보보호의 기준을 만족시킬 수 있는 방법이며, 기준이 아니다. 통신회사와 보험회사의 예제에서는 통신회사가 채택한 차등정보보호의 기준을 제시한 것이지 실제로 통신회사가 사용한 정보보호 방법은 구체적으로 명시하지 않았다.

차등정보보호의 기준은 정보보호 예산 ϵ으로 표현되며, 이를 만족할 수 있는 정보보호의 방법을 선택하여 적용해야 한다. 교수 A와 B의 연구발표 예제에서 참가하는 잡음의 크기는 정해진 정보보호 예산 ϵ에 반비례할 것이다.

통신회사와 보험회사의 예제에서 확률을 계산할 때 로지스틱 회귀분석을 이용한다면, 다음에 제시된 회귀분석에 사용될 수 있는 방법들을 사용할 수 있다.

2) 차등정보보호를 구현하는 방법은 주어지는 분석의 결과에 따라 다르다.

차등정보보호는 자료 분석 서버가 이용자에게 제공하는 통계적 분석의 결과에 대하여 정의된다. 따라서 차등정보보호를 구현하는 구체적인 방법은 결과의 형태에 따라 다를 수 있다. 다음은 대표적인 분석 결과 유형별로 차등정보보호를 위한 방법들의 예이다.

- 도수(count, frequency)

도수는 가장 흔히 사용되는 자료 분석의 결과로서 제시된 조건을 만족하는 자료 단위들의 개수이다. 차등정보보호를 구현하는 대표적인 방법은 교수 A와 B의 연구에 대한 예제처럼 도수에 잡음(noise)을 첨가하는 것이다.

- 히스토그램(histogram)

히스토그램은 주어진 자료에서 관심의 대상이 되는 값이 취할 수 있는 범위를 여러 개의 구간으로 나누고 그 구간에 해당하는 단위들의 개수들을 표시하는 것이다. 집계표도 하나의 히스토그램 예이다. 차등정보보호를 구현하는 대표적인 방법으로는 각 구간마다 얻어진 도수들에 잡음을 첨가하거나 구간의 길이를 약간씩 변화시키는 방법이 있다.

- 회귀분석(linear regression)

회귀분석은 설명변수와 반응변수의 관계를 함수식으로 표시하여 추정하는 방법이며, 회귀분석의 결과는 함수식에 포함된 회귀계수(regression coefficient)의 추정량이다. 그러므로 차등정보보호를 적용하려면 추정된 회귀계수에 잡음을 첨가하거나, 아예 반응변수 등에 잡음을 첨가할 수 있다.

- 기계학습(machine learning)

AI 발달을 가능하게 한 다양한 기계학습 방법들(예를 들어 분류, 클러스터링)에 대해서도 차등정보보호 기준을 만족할 수 있도록 하는 다양한 방법들이 최근 활발하게 제안되어 연구되고 있다. 이러한 연구들에서는 새로운 방법의 개발이 시도되고 있을 뿐만 아니라, 통계적 데이터베이스에서 개인정보보호를 위해 제시되어 온 전통적인 방법들이 과연 차등정보보호의 기준을 만족하는지에 대한 고찰도 이루어지고 있다(Zhu et al., 2017).

3) 차등정보보호 기준은 하나의 제시어(one query)에 대한 기준이다.

차등정보보호의 기준은 기본적으로 이용자가 데이터베이스에 하나의 제시어(one query)를 요구할 때 그 결과에 대한 기준이다. 만약 공격자가 여러 개의 제시어를 연속적으로 사용하면 노출의 위험성은 커지게 된다. 어떤 정보보호 방법도 이러한 현상을 피할 수는 없다. 무한하게 많은 쿼리를 데이터베이스에 제시하게 되면 결국 개인의 정보는 노출될 수밖에 없다.

하지만 차등정보보호가 가지는 가장 중요한 성질 중의 하나는 이러한 정보의 요구가 연속으로 일어나더라도 그로 인해 발생하는 위험의 누적 속도를 구체적으로 알 수 있다는 것이다. 만약에 같은 자료에 대하여 차등정보보호가 적용된 두 개의 쿼리에 대한 결과가

각각 개인정보 예산 ϵ_1과 ϵ_2로 생성되었다고 하자. 예를 들어서 첫 번째 쿼리는 "소득이 2억 원 이상인 가구의 수"이고 두 번째 쿼리는 "소득이 3억 원 이상인 가구의 수"라고 하면 출력되는 결과는 두 개의 도수(count)가 될 것이다. 이제 주어진 각각의 개인정보 예산에 맞추어 잡음을 첨가한다. 그러면 두 개 쿼리 결과를 조합(composition)해서 생길 수 있는 위험은 두 쿼리의 위험의 합 $\epsilon_1 + \epsilon_2$를 넘지 않는다.

이러한 결과를 통해서 알 수 있는 것은 정보 공개가 증가하면 유출의 가능성이 증가한다는 필연적인 사실을 피할 수는 없지만, 차등정보보호는 정보의 연속적인 공개에 따른 노출위험의 증가속도를 단순하게 누적되는 것으로 제어할 수 있다는 사실이다. 별다른 기준이 없이 어떤 정보보호 방법이 적용되는 경우, 정보의 양이 증가하면 노출 위험성은 급격하게 증가하는 경우가 대부분이다. 또한 이렇게 정보의 노출이 증가하는 경우 위험의 누적을 정확하게 수량화할 수 있는 정보보호 기준은 차등정보보호가 유일하다.

4) 차등정보보호는 외부에 추가정보가 있는 경우 그 기준이 만족되지 않을 수 있다.

어떤 자료에 대해서 데이터베이스와 별개로 존재하는 외부정보(auxiliary information)가 존재하면, 개인의 신분 노출은 쉽게 일어날 수 있다. Dwork(2006)은 Dalenius(1977)의 명제가 불가능하다는 사실을 증명하면서 다음과 같은 예를 들어 이를 설명하였다.

> 만약에 키(height)가 민감한 정보이며 주어진 데이터베이스에서 자료의 평균키만을 출력한다고 하자. 이용자가 데이터베이스에서 "한국 여자의 평균키는 162.3cm이다"라는 정보를 얻는다. 이 이용자는 다른 경로를 통하여 "김순이의 키는 한국여자 평균키보다 2cm 더 크다"라는 외부정보를 알게 된다. 이제 이를 조합하여 "김순이의 키는 164.3cm이다."라는 민감한 개인정보를 얻게 된다.

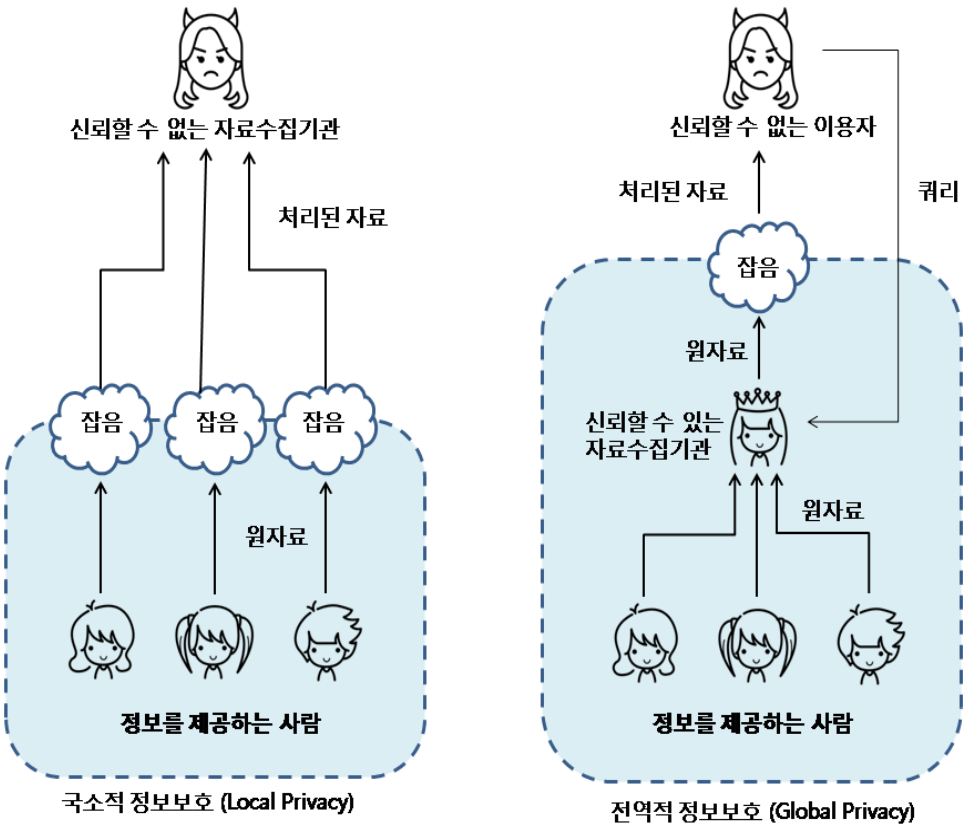
Dwork(2006)에서 위와 같은 특별한 예제를 확장하여 외부정보가 존재하면 일반적인 경우 개인의 정보는 궁극적으로 보호될 수 없다는 것을 증명하였다.

차등정보보호는 외부정보가 있는 경우를 고려하지 않는 정의이다. 차등정보보호를 만족하도록 적용된 정보보호 방법이 외부정보가 있는 경우에도 제시된 기준을 만족하는지 알 수 없다. 만약에 공격자가 차등정보보호가 적용된 자료와 외부자료를 결합할 수 있다면 개인정보의 유출 위험성은 제시된 기준보다 커질 것이다.

바. 차등정보보호의 확장과 한계

1) 국소차등정보보호

최근 일어난 일련의 개인정보 누출에 관련된 사건들로 인하여 사람들은 “자료를 수집하는 신뢰할 수 있는 기관”에 대한 개념에 큰 의심을 품게 되었다. 대표적인 예로 최근에 일어난 페이스북(facebook)의 개인정보 유출 사건을 들 수 있다. 전통적으로 개인은 신뢰할 수 있는 기관에게 개인의 정보를 제공하고, 자료를 수집한 기관은 개인정보를 보호할 수 있는 정보보호 방법을 적용한다고 예상한다. 또한 일반 이용자에게는 이렇게 가공된 정보가 제공된다고 생각하므로 정보의 교환이 이루어질 수 있었다. 이러한 과정에서 자료를 수집하는 기관에 대한 신뢰가 개인정보보호의 필수적인 요인으로 여겨졌다. 그러나 최근에 이러한 신뢰받는 기관의 부주의나 실수로 인하여 개인의 정보가 유출되는 사건들이 발생해서 자료수집 기관에 대한 신뢰 수준이 낮아지고 있다. 이러한



[그림 5-5] 국소적 차등정보보호와 광역적/전역적 차등정보보호

환경으로 개인이 기관에 정보를 제공할 때부터 정보보호 방법을 적용하여 제공해야 한다는 필요성이 제기되었다. 그런 정보보호 환경을 앞에서 살펴본 기관이 수집한 정보를 제공할 때 적용하는 기준(global privacy)과 구별하여 국소적인 정보보호 기준(local privacy)이라고 부른다(Erlingsson et al., 2014;. Duchi et al., 2013).

차등정보보호가 국소적인 기준으로 적용되는 환경, 즉 개인이 자료를 수집할 기관에게 정보를 제공할 때 차등정보보호의 기준에 맞추어 제공하는 것을 국소차등정보보호(local differential privacy)라고 부른다. 반대로 앞에서 살펴본 기관이 수집한 자료를 일반 이용자에게 이용할 수 있게 공개하는 기준을 광역/전역차등정보보호(global diffrential privacy)라고 한다. 두 개의 차등정보보호의 차이점과 절차를 다음과 같은 [그림 5-5]로 설명할 수 있다.

국소차등정보보호에서는 개인이 정보를 수집하는 기관에게 자료를 제공할 때 자료에 정보보호 방법을 적용하여 실제 자료와는 차이가 있는 자료를 제공하는 것이다. 애플(apple)사가 이용자의 이모티콘 사용 현황을 파악하기 위해 개인의 정보를 받을 때 이모티콘 종류의 일부를 바꾸어 저장하는 방법이 대표적인 예이다.

2) 레코드의 연관성

차등정보보호는 그 개념을 정의할 때 자료가 서로 관련 없이 독립적인 관계에 있다고 가정한다. 차등정보보호에서 필수적으로 나오는 개념은 하나의 개인 또는 개체가 포함된 자료와 제외된 자료의 차이이다. 대부분의 정형화된 자료는 전체 자료에서 하나의 레코드를 제외하여도 다른 레코드의 값에 영향을 주지 않는다. 앞에서 살펴본 예제에서처럼 학생들의 개인 기록이나 보험회사의 고객정보들은 하나의 레코드를 제외하여도 다른 개체의 정보는 변하지 않는다.

하지만 특정 비정형 자료에서는 하나의 레코드를 제외하면 다른 자료의 정보들이 변하는 경우가 있다. 예를 들어서 사회망(social network)을 설명하는 자료는 개체(node)와 서로 다른 개체들을 이어주는 관계(edge)로 구성된다. 이러한 자료에서 하나의 개체를 제외하면 그 개체뿐만 아니라 다른 개체와의 관계도 제거된다. 따라서 다른 개체에 대한 정보도 변하게 된다. 이러한 예들은 구성된 자료들이 서로 독립이 아니고 특정한 관계를 가지는 경우이며, 차등정보보호가 적용되더라도 적용된 기준을 만족하지 않는 현상이 생길 수 있다. 다른 예로서 국가통계기관이 집계표(statistical table)는 정확하게 발표하고 이를 생산한 마이크로자료(microdata)를 공개할 때도 집계표와 마이크로자료의 연관성에 의하여 차등정보보호가 그 기능을 제대로 만족하지 않는 경우가 생긴다. 이렇게 구성된 자료가 독립이 아니고 서로 연관이 있는 경우 차등정보보호의 기준이 구현되지 않는 경우도 있다(Kiefer와 Machanavajjhala, 2011).

3. 차등정보보호의 확률적 정의

가. 차등정보보호의 정의

차등정보보호의 개념은 한 개의 개체 또는 단위가 자료에 추가로 포함될 때 증가하는 위험을 측정하는 것이다. 따라서 차등정보보호는 자료의 전체를 보호하는 의미에서 위험을 측정하기보다는, 자료의 개수가 미세하게 변화함에 따라 증가하는 위험을 상대적으로 측정하고 이를 제어하는 기준이라고 할 수 있다. 앞에서 설명할 때 차등이라는 의미를 "하나의 차이가 날 때"라고 설명하였는데, 이를 좀 더 현실적으로 말하면 "미세하게 하나의 차이가 날 때"라고 말할 수 있을 것이다. 차등정보보호에 대한 확률적 정의에 필요한 가정은 다음과 같다.

- 데이터베이스 D_1과 D_2를 고려하고, 각각의 데이터베이스에 포함된 개체 수의 차이가 한 개라고 하자.

- 임의의 노출제한 방법을 확률함수 K라고 표현하자. 노출제한 방법으로 잡음첨가, 바꾸기, 재코딩, 표본추출 등 개인의 노출을 제한하는 모든 방법을 고려한다. 잡음첨가나 표본추출 등은 확률적인 임의방법(randomization)에 의한 것이고 재코딩, 바꾸기 등은 결정적인(deterministic) 방법이다. 이러한 모든 방법을 확률함수로 보고, $K(D)$는 노출제한 방법이 적용된 공개용 데이터베이스라 한다.

이러한 가정 하에서 ϵ-차등정보보호(ϵ-differential privacy)란 다음과 같은 조건을 만족하는 것이다.

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

위의 식에서 집합 S는 데이터베이스의의 일부분으로 노출제한 방법이 적용된 모든 가능한 자료의 집합을 나타낸다. 따라서 위의 기준은 데이터베이스에서 특별한 하나의 부분 집단에만 적용되는 것이 아니라, 데이터베이스 전체를 포함하여 추출할 수 있는 모든 가능한 부분집단 S에 적용된다.

위의 정의에서 사용된 확률 P는 노출제한 방법 K를 확률적 함수로 고려했기 때문에 생기는 임의성(randomness)에 대한 확률이다. 만약 쿼리가 특정한 조건을 만족하는 개체들의 도수(count)이고, 노출제한 방법이 잡음을 첨가하는 방법이라 하자. 그러면 쿼리의 결과는

"잡음이 첨가된 도수" $K(D)$이고, 확률 P는 잡음이 첨가된 도수의 분포, $P[K(D) \subseteq S]$를 의미한다. 위에서 정의된 차등정보보호는 하나의 개인 또는 개체가 제외되어도, 공개된 자료, "잡음이 첨가된 도수"로부터 얻어지는 정보가 유의하게 변하지 않는다는 것을 의미한다. 차등정보보호에 대한 자세한 이론과 응용은 Dwork et al.(2006), Dwork(2006, 2008) 및 Zhu et al.(2017)에서 찾을 수 있다.

차등정보보호의 기준은 다음과 같이 통계학에서 자주 사용되는 가능도비(likelihood ratio) 형태로도 표현할 수 있다.

$$\log \frac{P[K(D_1) \in S]}{P[K(D_2) \in S]} \leq \epsilon \quad for\ all\ S \subseteq range(K)$$

차등정보보호의 정의에서 개인정보 예산 ϵ이 아주 작은 수일 경우(예를 들어 $\epsilon = 0.01$) 근사식 $e^{\epsilon} \approx 1 + \epsilon$ 을 이용하면 그 기준을 다음과 같이 나타낼 수 있다.

$$\frac{P[K(D_1) \in S]}{P[K(D_2) \in S]} \leq e^{\epsilon} \approx 1 + \epsilon \quad for\ all\ S \subseteq range(K)$$

위의 식은 앞에서 자동차 사고의 확률을 이용하여 차등정보보호를 설명할 때 사용한 식과 동일한 형태이다. 위의 식을 예제의 식과 연관시키면 $P[K(D_1) \in S]$는 "김철수가 포함된 자료(D_1)를 이용하여 구한 김철수가 사고를 낼 확률"며 $P[K(D_2) \in S]$는 "김철수가 제외된 자료(D_2)를 이용하여 구한 김철수가 사고를 낼 확률"이라고 해석할 수 있다.

차등정보보호 두 확률의 비율로서 상대적인 관계의 상한(upper bound)을 제어하는 기준이기 때문에 매우 강한 정도의 정보 보호 기준을 의미한다. 예를 들어서 위의 식에서 분모의 확률이 매우 작은 경우에는 분자의 확률이 분모의 크기에 비해 상대적으로 작아야 한다. 중요한 것은 이러한 기준이 데이터베이스의 모든 가능한 부분 집단에 성립해야 한다는 것이다. 이렇게 강력한 차등정보보호는 현실적으로 구현이 불가능한 경우가 있으므로, 원래의 기준을 조금 완화해 (ϵ, δ)-차등정보보호((ϵ, δ)-differential privacy)가 제시된 바 있다. (ϵ, δ)-차등정보보호의 기준은 아래와 같다.

$$P[K(D_1) \in S] \leq e^{\epsilon} P[K(D_2) \in S] + \delta \quad for all\ S \subseteq range(K)$$

주어진 δ로 두 확률의 차이를 상대적이 아닌 절대적인 차이(difference) 이내에 있게 하는 조건을 주어 ϵ-차등정보보호의 기준보다 조금 완화된 기준을 제시한다.

나. 통계량의 민감도

차등정보보호는 자료 분석 서버가 주어진 쿼리에 대한 결과를 생성할 때 고려하는 기준이므로 쿼리의 성질에 의존한다. 따라서 쿼리의 결과, 즉 통계량(estimator, statistic)의 성질에 의존한다고 할 수 있다. 예를 들어 쿼리가 특정한 조건을 만족하는 도수(count)라면 통계량은 한 개의 정수로 주어지며, 쿼리가 자료의 히스토그램이라면 결과는 주어진 구간에 대한 도수들의 집합이 된다.

Dwork(2006)은 차등정보보호의 구현을 위하여 데이타베이스의 쿼리에 대한 민감도(sensitivity)를 정의하였다. 데이터베이스 D에 대한 쿼리를 함수 f로 표현한다면 $f(D)$는 쿼리에 대한 결과(통계량)이다.

$$f : D \rightarrow R^k$$

만약 쿼리 f가 히스토그램이라고 한다면 $f(D)$는 데이터베이스 D에 대하여 주어진 m개의 구간과 각 구간에 해당하는 도수(frequency),

$$f(D) = (n_1, n_2, ..., n_m)$$

이다.

이러한 가정 하에서 쿼리 $f(D)$의 민감도 Δf 는 다음과 같이 두 쿼리 결과물의 차이로 정의한다.

$$\Delta f = max \| f(D_1) - f(D_2) \|_1$$

여기서 $\|x\|_1$ 은 주어진 개체 x의 성분 중에 최대값을 의미한다. 여기서 두 데이터베이스 D_1 과 D_2 는 개체의 수가 하나가 차이가 나는 모든 경우를 고려한 것이다. 예를 들어 히스토그램의 민감도는 $\Delta f = 1$ 이다. 왜냐하면 한 개의 개체가 새롭게 자료에 포함되는 경우, 아래와 같이 히스토그램은 첨가된 개체가 포함된 k번째 구간의 도수가 하나 증가하고 나머지 도수는 변하지 않기 때문이다.

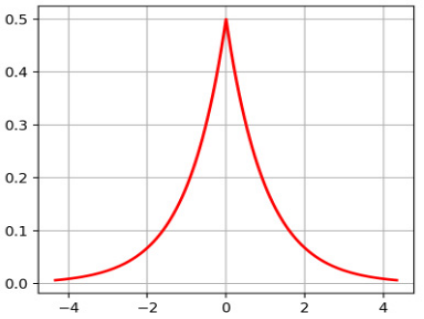
$$f(D_1) = (n_1, n_2, ..., n_k, ..., n_m), \quad f(D_2) = (n_1, n_2, ..., n_k + 1, ..., n_m)$$

많은 경우에 민감도의 값은 크지 않지만 쿼리의 종류에 따라 그 크기가 상한(upper bound)을 가지는 경우도 있다.

다. 차등정보보호의 구현

쿼리의 민감도가 계산되었다면, ϵ-차등정보보호를 구현하는 대표적인 방법은 쿼리의 결과에 표준편차의 크기 $\sigma = \Delta f/\epsilon$을 가지는 이중지수분포(double exponential distribution; Laplace distribution)에서 생성된 잡음을 첨가하는 것이다. 이중지수분포의 모양은 다음 [그림 5-6]과 같이 뾰족한 모양을 한 대칭분포이며, 표준편차 σ의 크기에 따라 퍼진 정도가 변하는 분포이다.

$$f(x;\sigma) = \frac{1}{2\sigma} exp\left(-\frac{|x|}{\sigma}\right).$$



[그림 5-6] 이중지수분포의 확률밀도함수

예를 들어 도수(count, frequency)에 이중지수분포에서 독립적으로 발생된 잡음들을 첨가한다. 잡음을 첨가하는 방법 K는 다음과 같이 나타낼 수 있다.

$$K(D) = f(D) + Lap(\sigma).$$

위의 식에서 $Lap(\sigma)$는 표준편차 σ를 가지는 이중지수분포를 따르는 확률변수이다. 잡음이 표준편차가 $\sigma = \Delta f/\epsilon$인 이중지수분포를 따른다고 가정하면, ϵ-차등정보보호의 기준을 다음과 같이 유도할 수 있다.

$$\frac{P[K(D_1) = y]}{P[K(D_2) = y]} = \frac{P[f(D_1) + Lap(\Delta f/\epsilon) = y]}{P[f(D_2) + Lap(\Delta f/\epsilon) = y]}$$

$$= \frac{P[Lap(\Delta f/\epsilon) = y - f(D_1)]}{P[Lap(\Delta f/\epsilon) = y - f(D_2)]}$$

$$= \frac{exp[-|y - f(D_1)|\epsilon/\Delta f]}{exp[-|y - f(D_2)|\epsilon/\Delta f]}$$

도수는 하나의 수를 결과로 가지는 쿼리이며 두 쿼리의 차이 $|f(D_1)-f(D_2)|$의 최대값은 민감도 Δf보다 작거나 같다. 이러한 성질을 이용하면 다음과 같은 상한을 유도할 수 있다.

$$\frac{exp[-|y-f(D_1)|\epsilon/\Delta f]}{exp[-|y-f(D_2)|\epsilon/\Delta f]} \leq exp[\epsilon|f(D_1)-f(D_2)|/\Delta f]$$

$$\leq exp(\epsilon)$$

위의 잡음첨가 방법을 보면 개인정보 예산 ϵ의 크기에 반비례하여 잡음의 크기(잡음 분포의 표준편차)가 커지는 것을 알 수 있다. 개인정보 예산 ϵ이 작을수록 정보보호가 강해지는 것이므로 첨가하는 잡음 분포의 분산은 커지게 된다.

이러한 잡음첨가 방법은 모든 쿼리에 적용될 수 있는 것은 아니다. 예를 들어 쿼리의 결과가 문자열, 그림 등이라면 잡음첨가를 할 수 없으며 다른 방법을 사용해야 한다(Zhu et al., 2017).

제3절 이론적 이해

개인정보보호에 대한 연구가 고도화되면서 단순히 비식별 데이터(sanitized data)[10]를 구성하고 공개하는 수준에서 다뤄지던 정보보호 과정이 좀 더 복잡하고 특수한 상황에 적용할 수 있는 체계적인 방법으로 발전하고 있다. 특히 이러한 고도화 과정에서 가장

10) 이 절에서 사용하고 있는 “sanitized data”는 “원래 자료와 다르지만, 어떤 알고리즘을 통해 변형되어 안전하게 처리된 이후에 제공되는 자료”라는 의미를 가진다. 현재 국내에 이런 의미로 통용되고 있는 용어는 “비식별 데이터”라고 할 수 있다. 비식별 데이터는 2016년 관계부처 합동으로 발간된 “개인정보 비식별 조치 가이드라인”에서 사용하고 있는 용어이기도 하다. 통계적 비밀보호라는 개념이 국내에 아직 생소한 중에, 실무적으로 가장 널리 사용되고 있는 용어이다.

다만 “비식별 데이터”라는 용어를 사용할 때, 자료를 안전하게 처리하기 위해 사용하는 방법은 매스킹 처리에 국한되는 경향이 있다. 매스킹 처리 중에서도 키변수를 변형시켜 식별 가능성을 감소시키도록 변형된 자료를 지칭하는 경우가 대부분이다. 반면 “sanitized data”라는 표현은 방법에 상관없이 안전하게 변형했다는 의미를 가진다고 볼 수 있다. 따라서 “sanitized data”는 “비식별 데이터”보다 상위 개념이다.

사실 “sanitized data”를 “재현자료(sanitized data)”라고 번역하는 것이 더 적절하겠지만, 차등정보보호와 함께 통계적 비밀보호 분야의 미래 지향적 대안으로서 연구되고 있는 “재현자료(synthetic data)”와 번역이 겹치는 문제가 있다. 때문에 “비식별 데이터”가 정확한 표현은 아니지만, 본 보고서에서는 이를 “sanitized data”를 번역한 용어로 사용하도록 한다. 본 보고서에서는 이 용어를 “어떤 알고리즘을 통해 변형되어 안전하게 처리된 이후에 제공되는 자료”라는 대단히 일반적인, 광의의 의미로 사용함을 주지하기 바란다.

중요한 역할을 하고 있는 연구가 차등정보보호(differential privacy)에 대한 연구이다. 차등정보보호의 장점은 개인정보가 얼마나 보호되고 있는가를 계량적으로 측도할 수 있도록 정보보호의 강도에 대한 엄밀한 정의를 제공한다는 점이다. 이러한 엄밀한 정의는 새로운 정보보호 방법을 개발 혹은 평가하거나 서로 다른 정보보호 방법을 비교하는 등, 개인정보보호와 관련해 발생 가능한 다양한 상황에서 매우 간결하고 유용하게 사용될 수 있다.

이 절에서는 차등정보보호와 국소차등정보보호(local differential privacy)를 비교하여 소개하고 각각의 주제에 대한 최근의 연구들을 이론적인 측면에서 이해할 수 있도록 필요한 개념들을 정리하였다. 차등정보보호와 국소차등정보보호의 차이는 수학적 정의를 통해 정확하게 파악할 수 있겠지만, 좀 더 직관적으로 자료를 제공하는 주체가 자료를 수집, 관리, 분석하는 주체를 신뢰하는가 그렇지 않은가를 기준으로 설명할 수도 있다. 자료를 제공하는 주체가 자료를 수집하는 주체를 신뢰한다면 원래의 자료가 그대로 수집되어 저장되지만, 만약 그렇지 않다면 수집 단계에서 직접 정보보호 처리를 진행하고 원래의 자료는 저장하지 않는다. 따라서 국소차등정보보가 차등정보보호보다 좀 더 강력한 의미의 정보보호를 요구한다는 점을 알 수 있다.

이 절은 Wasserman과 Zhou(2010)와 Duchi et al.(2018)에서 연구한 내용을 위주로 작성하였으며, 가독성을 높이기 위하여 문맥상 정확히 파악할 수 있는 경우에는 사용된 기호에 대한 설명을 생략하였다. 이 절의 내용을 정확히 이해하기 위해서는 통계학 석사과정 정도의 지식이 필요하지만, 전반적인 개념을 파악하고 활용 방법 등을 이해하는 과정은 특별히 어렵지 않다고 판단된다. 하지만 소개된 방법론을 이론적인 측면에서 평가하는 내용은 일반적인 독자에겐 좀 어려운 부분이며, 이 절의 마지막 부분에 정의된 차등정보보호 위험함수와 최소최대위험을 먼저 숙지하는 것이 이해에 도움이 된다.

1. 차등정보보호

개인정보보호의 기본 원칙은 수집된 원자료(raw data)를 외부에 직접 노출하지 않는 것이다. 하지만 원자료가 가지는 유용한 정보를 활용해야 하는 경우는 언제든 발생할 수 있다. 따라서 원자료가 가지는 정보의 손실은 최소화하고 동시에 개인정보보호는 최대화하는 것이 가장 이상적인 방법이다. 차등정보보호는 이러한 상황을 해결하기 위하여 원자료 대신 일종의 비식별 데이터를 구성하여 제공하는 방법이다. 이를 위한 과정에는 자료를 수집하고 처리하는 방법은 물론 분석을 위한 확률적 개념 등이 요구된다. 따라서 차등정보보호에 대한 이론적인 연구는 전산학(computer science), 통계학(statistics) 그리고 결정이론(decision theory) 등의 분야에서 광범위한 주제로 진행되고 있다.

가. 차등정보보호의 정의 및 성질

차등정보보호는 원자료로부터 비식별 데이터를 얻는 과정에서 사용되는 조건부 분포의 성질과 밀접한 관련이 있으며, 조건부 밀도함수의 값의 비율에 대한 상한(upper limit)의 크기와 존재성을 포함한다. 다음의 정의는 Dwork et al.(2006) 등의 연구에 기술된 내용을 정리한 것이다. 이 정의의 핵심은 원자료에서 변형된(sanitized) 자료가 개인정보를 얼마나 보호하고 있는가를 양적으로 측도할 수 있다는 것이다.

[정의 1] 원자료 $X=(X_1,...,X_n)$가 주어져 있을 때 비식별 데이터 $Z=(Z_1,...,Z_k)$의 조건부 밀도함수(conditional density function)를 $f_{Z|X}$라고 하자. 만약 $f_{Z|X}$가 적당한 $\alpha > 0$[11]에 대하여 아래의 부등식을 만족하면 $f_{Z|X}$를 X에 대한 α-차등정보보호 방법(α-differential privacy mechanism)이라고 하고 Z를 X에 대한 α-차등정보보호 자료(α-differential privacy 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. ■

위 정의에서 조건부 밀도함수가 존재하지 않는 경우 Wasserman과 Zhou(2010)가 제안한 것처럼 조건부 분포함수를 사용할 수 있다. α-차등정보보호의 정의는 $X=u$인 경우 조건부 밀도함수와 $X=v$인 경우 조건부 밀도함수의 값의 비율이 어느 정도 크기인가에 대하여 기술하고 있다. 다시 말하면 원자료의 값이 u인 경우의 비식별 데이터와 원자료의 값이 $v(\neq u)$인 경우의 비식별 데이터가 분포의 측면에서 서로 얼마나 다를 수 있는가를 측도하는 것이다. 예를 들어 두 비식별 데이터의 분포가 충분히 다를 수 있다면, 즉 α가 매우 크다면, 비식별 데이터의 분포가 원자료의 차이에 심하게 의존한다는 뜻이 된다. 이는 개인정보의 유출 가능성이 매우 크다는 의미가 된다. 역으로 두 비식별 데이터의 분포가 거의 동일하다면, 즉 α가 0에 매우 가깝다면, 비식별 데이터의 분포가 원자료의 차이에 영향을 받지 않는다는 뜻이 된다. 따라서 이는 개인정보의 유출 가능성이 매우 작다는 것을 의미한다.

11) 앞 절과 달리, 이 절에서는 ϵ 대신에 α 을 사용해 차등정보보호를 설명한다. 이는 주요 참고문헌들의 표현을 따른 것이다. 주요 참고문헌에서는 ϵ 은 주로 잡음을 표기하기 위해 사용하고, 차등정보보호 모수는 ϵ 대신에 α 를 사용해 표기했다.

위 정의에서 에서 주목해야 할 부분은 $\|u-v\|_0 = 1$인 조건이다. 이 조건은 u와 v가 오직 하나의 원소만 다르다는 것을 의미하며, 따라서 이 정의는 오직 하나의 자료만 다른 두 원자료로부터 생성한 비식별 데이터의 분포가 서로 얼마나 차이가 나는가를 고려하고 있는 것이다. 이러한 특별한 상황을 고려하는 이유는 이러한 상황이 개인정보를 침해하려는 외부의 공격자 입장에서 가장 유리한 상황이라고 판단하기 때문이다. 즉, 하나의 자료만 다른 상황에서 정보보호가 가능하다면 다른 어떠한 상황에서도 정보보호가 가능하다는 의미를 내포하고 있다.

[정의 1]은 비전문가의 입장에서 직관적으로 이해하기에는 어느 정도 난해한 측면이 있다. Wasserman and Zhou(2010)는 이 정의를 바탕으로 α-차등정보보호를 일종의 통계적 검정(statistical test) 가능성의 여부로 해석하였다. 그리고 α-차등정보보호의 개념을 좀 더 이해하기 쉽도록 다음과 같은 정리를 소개하고 있다.

[정리 1] 원자료를 $X=(X_1,...,X_n)$이라 하고 아래와 같은 가설을 검정하기 위하여 비식별 데이터 $Z=(Z_1,...,Z_k)$로부터 신뢰수준(significance level)이 γ인 검정법 T_i를 구성한다고 하자.

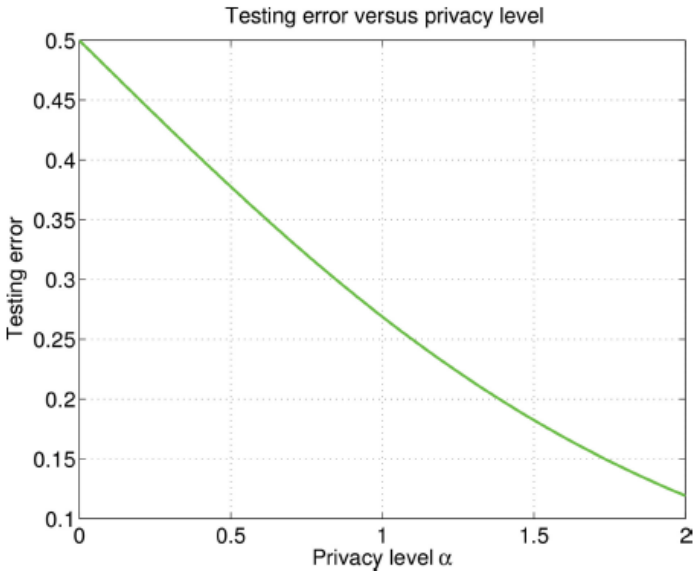
$$H_0 : X_i = u \ \text{ vs } \ H_1 : X_i = v, \quad u \neq v.$$

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

[정리 1]은 비식별 데이터 Z를 사용하여 원자료 X의 값 중에서 특정 자료 X_i의 값을 식별(identification)할 수 있는가를 설명하고 있다. 이 정리의 결과에 의하면 $\alpha \to 0$인 경우 $\gamma \exp(\alpha) \to \gamma$이므로 검정법의 신뢰수준과 검정력이 같아짐을 확인할 수 있다. 일반적으로 신뢰수준은 $\gamma < 0.1$ 정도이고, 따라서 자료를 변형하는 과정에서 강력한 α-차등정보보호 방법을 적용한다면 특정 원자료의 값을 정확하게 식별할 수 없다는 결론을 얻는다. Duchi et al.(2018)은 Wasserman과 Zhou(2010)의 아이디어를 바탕으로 검정오차(testing error)를 다음과 같이 정의하였다.

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

다음 [그림 5-7]을 참고하면 α가 0에 가까워지면서 검정오차가 1/2에 가까워지므로 임의의 어떠한 검정법도 무작위 추측(random guess)에 불과하다는 것을 알 수 있다.



출처 : Duchi et al.(2018)

[그림 5-7] 정보보호 수준에 따른 검정 오차

이 정리의 내용 중 주목할 만한 사실은 외부의 공격자가 차등정보보호 자료를 생성한 조건부분포 $F_{Z|X}$를 알고 있다고 해도 결과가 동일하다는 점이다. 따라서 비식별 데이터를 구성하는 과정 일체를 외부에 공개하여 자료를 제공하는 주체에 대한 신뢰를 확보할 수 있다는 점이 α-차등정보보호의 장점 중 하나라고 할 수 있다.

다음 정리는 Wasserman과 Zhou(2010)에서 소개되었으며 α-차등정보보호 자료의 성질에 대하여 기술하고 있다.

[정리 2] $e=(e_1,\ldots,e_m)$를 밀도함수 f_ϵ에서 추출한 임의표본(independent and identically distributed random variables)이라고 하자. 만약 e가 원자료 $X=(X_1,\ldots,X_n)$와 독립이면 e와 X로부터 얻은 확률벡터 $Z(X,e)$는 다음과 같은 성질을 가진다.

① $Z(X,e)$가 X에 대한 α-차등정보보호 자료이면 임의의 함수[12] 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. ■

12) 측도가능한 함수(measurable function)

α-차등정보보호 자료를 만드는 방법은 매우 다양하지만 가장 일반적인 방법은 원자료와 독립인 임의표본을 사용하는 것이다. 예를 들어 $Z(X,e)=X+e$와 같이 X에 오차 e를 직접 더하여 간단하게 얻을 수 있다. [정리 2]의 ①은 얻어진 α-차등정보보호 자료에 절대값을 취하여 새로운 α-차등정보보호 자료 $W=|X+e|$를 계산해도 상관없다는 것을 의미하며 따라서 생성된 α-차등정보보호 자료를 활용 목적에 맞도록 가공하거나 수정할 수 있게 된다. [정리 2]의 ②는 α차등정보보호 자료의 밀도함수를 알고 있다면 밀도함수로부터 원자료와 전혀 다른 임의표본을 직접 추출할 수도 있다는 것을 의미한다. 따라서 원자료보다 더 많은 수의 차등정보보호 자료를 구성하는 것도 가능하다.

나. 차등정보보호 자료의 생성

이상에서 차등정보보호의 정의와 성질에 대하여 간략히 소개하였다. 다음으로 차등정보보호 자료를 생성하는 방법 중에서 가장 대표적인 방법 세 가지를 소개하도록 한다.

1) 오차를 활용하여 교란하는 방법

원자료로부터 비식별 데이터를 얻는 방법 중 가장 단순한 방법은 원자료에 적당한 오차를 더하여 교란(perturbation)하는 방법이다. 하지만 오차를 더하는 방식에 따라 α-차등정보보호 자료가 될 수도 있고 되지 않을 수도 있다. 이러한 사실을 확인하기 위해 두 가지 간단한 예제를 살펴본다.

[예제 1] 원자료를 $X=(X_1,\ldots,X_n)$이라 하고 오차 $\epsilon=(\epsilon_1,\ldots,\epsilon_n)$이 표준정규분포를 따르는 임의표본이라고 하자. X와 e를 사용하여 다음과 같이 비식별 데이터를 구성할 수 있다.

$$Z_i=X_i+\epsilon_i,\ i\in\{1,\ldots,n\}.$$

이 경우 e와 X가 독립이면 간단한 계산 과정을 거쳐 다음과 같은 등식을 얻을 수 있다.

$$\begin{aligned} sup_{\|u-v\|_0=1} sup_z \frac{f_{Z|X}(z|u)}{F_{Z|X}(z|v)} &= sup_{a\neq b}\ sup_z \frac{f_e(z-a)}{f_e(z-b)} \\ &= sup_{a\neq b}\ sup_z\ exp\left(\frac{(z-b)^2-(z-a)^2}{2}\right). \end{aligned}$$

단, f_e은 표준정규분포의 밀도함수이다. ■

위 [예제 1]에서 a, b와 z는 임의의 실수이므로 조건부 밀도함수의 비율의 상한이 존재하지 않는다. 따라서 단순히 원자료에 오차를 더하는 방법은 직관적으로 이해하기 쉽고 생성 과정이 간단하지만 강력한 정보보호를 기대할 수 없다.

다음의 예제는 Dwork et al.(2006)에서 제안하였으며 자료의 합 혹은 평균에 대하여 α-차등정보보호를 구현하는 방법을 기술하고 있다.

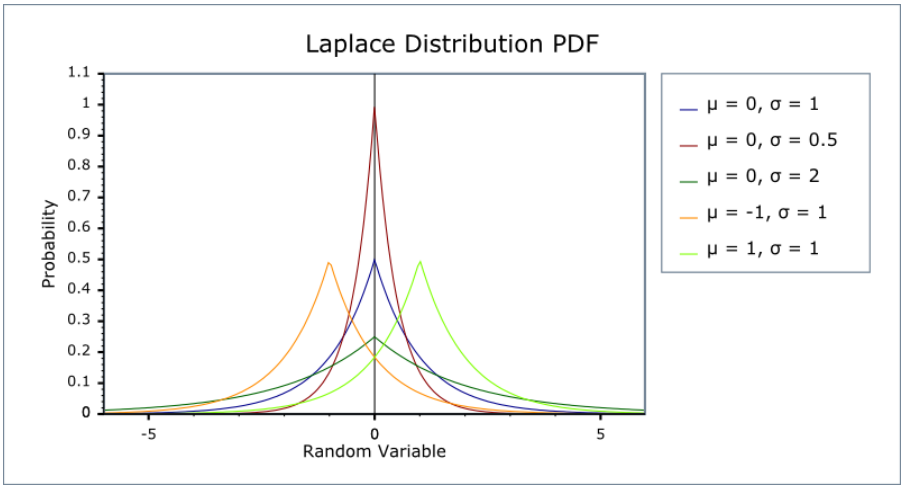
[예제 2] 원자료를 $X=(X_1,...,X_n)$, $X_i \in \{0,1\}$, $i \in \{1,...,n\}$이라 하자. 오차 ϵ의 밀도함수가 $f_\epsilon(r) \propto exp(-|x-\mu|/\sigma)$[13]이고 X와 독립이라고 하자. X와 e를 사용하여 다음과 같이 원자료의 합에 대한 비식별 데이터를 구성할 수 있다.

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

위 비식별 데이터가 X에 대한 α-차등정보보호 자료인가를 확인하기 위하여 간단한 계산 과정을 진행하면 $\sigma = 1/\alpha$인 경우 다음과 같은 부등식을 얻을 수 있다.

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

[예제 2]는 원자료에 독립인 오차를 더하여 원자료의 평균에 대한 α-차등정보보호 자료를 생성할 수 있다는 것을 보여준다. [예제 1]과 [예제 2]의 결과에서 차이가 발생하는 이유 중 하나는 원자료 값의 범위 때문이다. 전자의 경우 원자료 값의 범위에 아무런



출처 https://en.wikipedia.org/wiki/Laplace_distribution

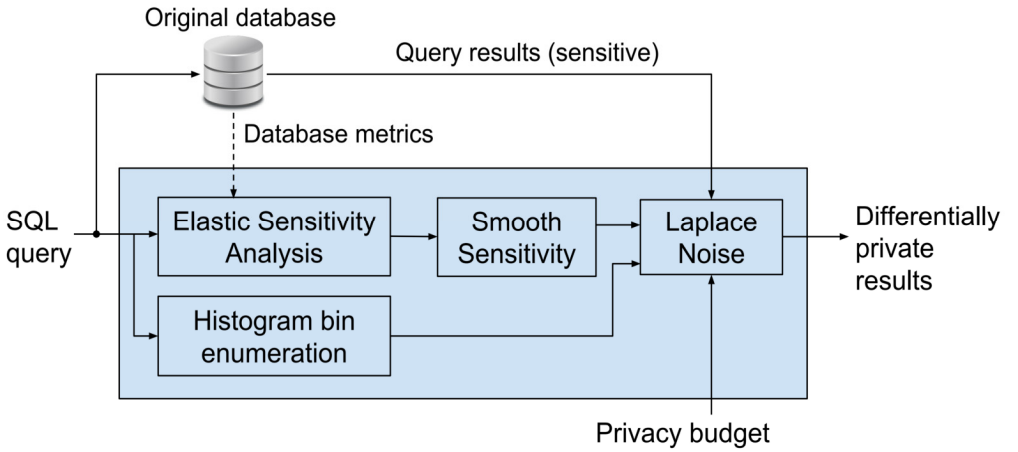
[그림 5-8] 이중지수(Double exponential or Laplace)분포의 밀도함수

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

제한이 없어 상한이 존재하지 않지만 후자의 경우 원자료가 0 또는 1의 값만 가질 수 있기 때문에 상한이 존재한다. 특히 [예제 2]의 경우 오차의 분포가 이중지수분포이므로 상한을 구하는 과정이 매우 쉬워 다음에 소개할 이중지수분포 방법의 기본이 되고 있다. 참고로 [그림 5-8]은 이중지수분포에서 확률밀도함수의 예를 보여준다.

2) 이중지수분포를 활용하여 교란하는 방법

α-차등정보보호 자료를 얻기 위한 가장 단순한 방법은 Dwork et al.(2006)이 제안한 이중지수분포 방법(double exponential mechanism)을 이용하는 것이다. 이중지수분포 방법은 이중지수분포를 따르는 임의표본을 생성하여 원자료에 추가하는 방법이다. 이중지수분포 방법은 원자료의 평균과 같은 단순한 통계량뿐만 아니라 원자료의 히스토그램과 같은 복잡한 통계량에 대하여 α-차등정보보호 자료를 생성할 수 있는 매우 유용한 방법이다. 참고로 다음 [그림 5-9]는 이중지수분포 중에서 라플라스 분포를 따르는 임의표본을 생성하여 원자료에 추가하는 과정을 보여준다.



출처: https://medium.com/uber-security-privacy/differential-privacy-open-source-7892c82c42b6

[그림 5-9] 라플라스 잡음을 추가해 차등정보보호 자료를 생성하는 과정

[정리 3] $e=(e_1,\ldots,e_k)$가 아래에 정의된 밀도함수 f_e에서 추출한 임의표본이라고 하자. e가 원자료 $X=(X_1,\ldots,X_n)$과 독립이면 $Z(X,e)=\xi(X)+e$은 X에 대한 α-차등정보보호 자료이다.

$$f_e(r) \propto exp\left(-\frac{\alpha}{\Delta_{n,k}^{\xi}}|r|\right).$$

단, $\xi : R^n \rightarrow R^k$, $\Delta_{n,k}^{\xi} = sup_{\|u-v\|_0=1}\|\xi(u)-\xi(v)\|_1$이고
$\|\xi(u)-\xi(v)\|_1 = \sum_{j=1}^{k}|\xi_j(u)-\xi_j(v)|$이다. ■

[정리 3]에 기술된 이중지수분포 방법은 거의 모든 정보보호 과정에서 사용 가능한 광범위한 방법이며 α-차등정보보호 자료를 생성하는 절차를 좀 더 자세히 소개하면 다음과 같다.

1) 적절한 함수 ξ[14]를 사용하여 원자료 X를 $\xi(X)$로 변환한다.
2) 함수 ξ의 변동의 상한 $\Delta_{n,k}^{\xi}$을 계산한다.
3) 산포모수(scale parameter)가 $\nabla_{n,k}^{\xi}/\alpha$인 이중지수분포의 밀도함수 f_e에서 임의표본 e을 추출한다.
4) $\xi(X)$에 임의표본 e를 더하여 α-차등정보보호 자료 $\xi(X)+e$를 얻는다.

이중지수분포 방법의 특징은 원자료의 크기를 함수 ξ를 사용하여 제한한다는 점이다. 이러한 제한이 필요한 이유는 [예제 1]과 [예제 2]에서 살펴본 바와 같이 α-차등정보보호 자료를 얻기 위한 교란 과정에서 밀도함수의 비율을 강제로 조정할 필요가 있기 때문이다. [정리 3]은 차등정보보호 자료 생성과정에서 필요한 함수 ξ를 어떠한 방식으로 찾을 수 있는가에 대한 정보는 제공하지 않는다. 하지만 이중지수분포의 산포모수가 원자료의 크기 n과 함수 ξ에 의존하기 때문에 적절한 함수 ξ를 찾는 것이 이중지수분포 방법의 핵심이 된다. 이러한 의존성은 α-차등정보보호 방법의 성능을 결정하며 자세한 논의는 Wasserman과 Zhou(2010)를 참고하기 바란다.

---

14) ξ는 자료 생성 목적에 따라 결정된다. 예를 들어 [예제 2]에서 $\xi(u) = \sum_{i=1}^{n} u_i$, $u_i \in \{0,1\}$이고 $\Delta_{n,k}^{\xi} = sup_{u \neq v}|u-v| = 1$이다.

3) 일반화 이중지수분포를 활용하는 방법

[정리 3]에서 소개한 이중지수분포 방법은 임의표본 오차를 추출하는 밀도함수를 이중지수분포로 제한하고 있다. 이중지수분포의 특징은 밀도함수가 절대값의 형태로 표현된다는 점이며 McSherry과 Talwar(2007)는 이러한 특징을 오차의 강건성(robustness)이라고 표현하고 있다. McSherry과 Talwar(2007)는 이중지수분포의 강건성의 장단점에 대하여 논의하고, 좀 더 일반적인 형태의 이중지수분포로 확장하여 다음과 같은 일반화 이중지수분포 방법을 제안하였다.

[정리 4] 비식별 데이터 $Z=(Z_1,...,Z_k)$가 아래에 정의된 조건부 밀도함수 $f_{Z|X}$에서 추출된 임의표본이면 Z는 원자료 $X=(X_1,...,X_n)$에 대한 α-차등정보보호 자료이다.

$$f_{Z|X}(z|x) \propto \exp\left(-\frac{\alpha\xi(x,z)}{\Delta^{\xi}_{n,k}}\right).$$

단, $\xi: R^n \times R^k \rightarrow R$이고 $\Delta^{\xi}_{n,k} = sup_{\|u-v\|_0=1} \; sup_z |\xi(u,z)-\xi(v,z)|$이다. ■

일반화 이중지수분포 방법이 이중지수분포 방법보다 확장된 방법인 또 다른 이유는 오차를 더하는 과정, 즉 교란 과정을 거치지 않고도 α-차등정보보호 자료를 얻을 수 있다는 점이다. 이러한 점을 명확하게 구분하기 위하여 일반화 이중지수분포 방법에 따라 α-차등정보보호 자료를 생성하는 과정을 기술하면 다음과 같다.

1) 적절한 양의 함수 ξ에 대하여 변동의 상한 $\Delta^{\xi}_{n,k}$을 계산한다.
2) [정리 4]에서 기술한 밀도함수에서 임의표본을 추출하여 α-차등정보보호 자료 Z를 얻는다.

다. 차등정보보호 방법을 활용한 분포 추정과 평가

[예제 2]에서 원자료의 표본평균 대신 정보보호 처리된 차등정보보호 자료를 사용할 수 있음을 살펴보았다. 이를 확장하여 차등정보보호 방법을 활용하여 히스토그램 혹은 밀도함수와 같이 분포를 직접 추정하는 방법을 소개하고, 원자료를 사용하여 추정한 경우와 차등정보보호 자료를 사용하여 추정한 경우를 비교 평가한다.

1) 히스토그램 작성

히스토그램은 가장 기본적인 분포추정 방법이며 자료의 분포적 특징을 확인하기 위한 기본 과정이다. 차등정보보호 방법을 적용하면 원자료의 히스토그램을 차등정보보호 자료로 작성한 히스토그램으로 대체할 수 있다. 원자료를 $X=(X_1,...\ X_n)$이라고 하면 X의 히스토그램은 다음과 같은 형태의 밀도함수로 정의된다.

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

단, h, B_j, C_j는 각각 히스토그램의 평활모수(bandwidth), 간격(width), 간격 B_j에서의 자료의 빈도(frequency)이며 m은 간격의 개수이다. 다음 정리는 이중지수분포 방법을 활용하여 원자료의 히스토그램을 대체하는 방법을 소개하고 있다.

[정리 5] $V=(V_1,...,V_m)$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.

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

위 정리는 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. Wasserman과 Zhou(2010)는 이러한 교란 방법을 사용하는 대신 직접 α-차등정보보호 자료를 생성하여 히스토그램을 작성하는 방법을 제안하였다.

[정리 6] 원자료 $X=(X_1,...,X_n)$이 $X_i\in[0,1],\ i\in\{1,...,n\}$이라고 하자. 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, $\tilde{D}_j = max\{D_j, 0\}$이고 $\hat{q}_j = \tilde{D}_j / \sum_{s=1}^{m} \tilde{D}_s$이다. ■

[정리 6]에서 소개한 차등정보보호 방법의 특징은 차등정보보호 자료의 개수 k를 임의로 조정할 수 있다는 점이다. 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 최소최대추정량(minimax risk estimator)이 된다는 것을 증명하였다. 또한 교란 방법이 평활 히스토그램(smoothed histogram)에서 비식별 데이터를 구성하는 방법보다 성능이 뛰어나다는 것을 추가로 증명하였다. 위에서 소개한 방법 이외에도 다양한 방법으로 히스토그램을 작성할 수 있으며 Machanavajjhala et al.(2008) 등을 참고하기 바란다.

2) 밀도함수의 추정

원자료 $X = (X_1, ..., X_n)$가 $X_i \in [0,1]$, $i \in \{1, ..., n\}$이고 X의 밀도함수 f_X가 $L_2[0,1]$ 공간의 원소라고 하자. 만약 $L_2[0,1]$ 공간의 직교기저(orthogonal basis)가 $\{1, \psi_1, \psi_2, ...\}$이면 f_X는 다음과 같이 표현된다.

$$f_X(x) = 1 + \sum_{j=0}^{\infty} \beta_j \psi_j(x), \ \beta_j = \int_0^1 \psi_j(x) f(x) dx.$$

이 경우 원자료 X로 추정한 밀도함수는 다음과 같이 주어진다.

$$\hat{f}_n(x) = 1 + \sum_{j=0}^{m_n} \hat{\beta}_{n,j} \psi_j(x).$$

단, $m_n = n^{1/(2\gamma+1)}$, $\hat{\beta}_{n,j} = \sum_{i=1}^{n} \frac{\psi_j(X_i)}{n}$이고 $\gamma > 1/2$.

Wasserman과 Zhou(2010)는 다음과 같이 일반화 이중지수분포 방법을 적용하여 밀도함수를 추정하는 방법을 제안하였다.

[정리 7] 비식별 데이터 $Z=(Z_1,...,Z_k)$가 다음과 같은 함수 ξ를 사용하여 [정리 4]에 정의된 일반화 이중지수분포에서 추출한 임의표본이라고 하자.

$$\xi(x,z)=\|\hat{f}_n-\hat{f}_k\|_{L_2}=\left\{\int(\hat{f}_n(x)-\hat{f}_k(x))^2dx\right\}^{1/2}.$$

단, $\hat{f}_k(x)=\sum_{j=0}^{m_k}\hat{\beta}_{k,j}\psi_j(x)$, $\hat{\beta}_{k,j}=\sum_{i=1}^{k}\psi_j(z_i)$이고 $m_k=k^{1/(2\gamma+1)}$이다.

적당한 양수 c_0에 대하여 $\Delta_{n,k}^{\xi}\le 2c_0^2\frac{m_n}{n}$이면 비식별 데이터 $Z=(Z_1,...,Z_k)$는 X에 대한 α-차등정보보호 자료이다. ■

[정리 7]은 [정리 4]에서 소개한 일반화 이중지수분포 방법에서 함수 ξ를 어떠한 방식으로 구성할 수 있는지 소개하고 있다. Wasserman과 Zhou(2010)는 추가로 이중지수분포를 사용하여 교란하는 방법을 제안하고, 각각의 방법에서 차등정보보호 자료 Z를 사용하여 추정한 밀도함수 추정량 $\hat{f}_Z$의 최대위험(maximum risk)을 비교하였다. 적당한 정칙조건 하에서 일반화 이중지수분포를 사용하는 경우

$$E_{L_2}(f,\widehat{f_Z})=O(n^{-\gamma/(2\gamma+1)}).$$

이고, 이중지수분포와 교란 방법을 사용하는 경우

$$E_{L_2}(f,\widehat{f_Z})=O(n^{-2\gamma/(2\gamma+1)}).$$

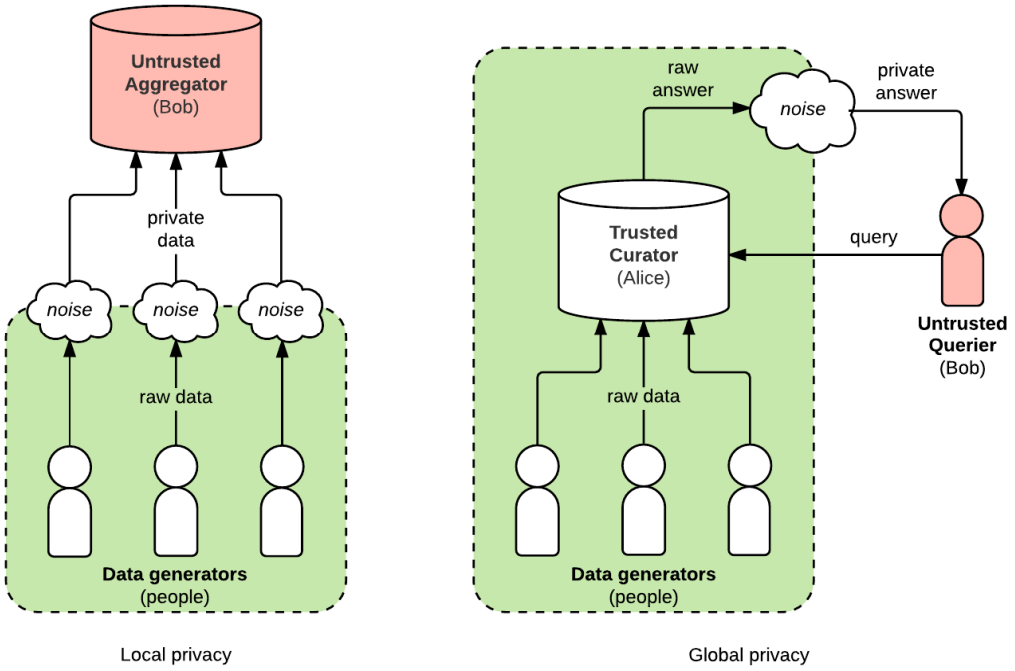
이므로 교란 방법이 위 정리에서 제시한 방법보다 성능이 더 뛰어나다. 특히 교란 방법의 경우 알려진 최소최대위험의 하한과 최대위험이 동일하며, 따라서 최소최대추정량이 된다.

참고로 Wasserman과 Zhou(2010)가 제안한 교란 방법을 간단히 기술하면 다음과 같다. 위치모수가 $\mu=0$이고 산포모수가 $\sigma=c_0m/n\alpha$인 이중지수분포에서 임의표본 $V=(V_1,...,V_{m_n})$을 추출하고 아래와 같이 계수를 조정한 밀도함수 $\tilde{f}_n$에서 차등정보보호 자료 $Z=(Z_1,...,Z_k)$를 임의로 추출한다.

$$\tilde{f}_n(x)=1+\sum_{j=0}^{m_n}(\hat{\beta}_{n,j}+V_j)\psi_j(x).$$

2. 국소차등정보보호

국소차등정보보호(local differential privacy)는 차등정보보호보다 좀 더 강력한 의미의 정보보호를 요구하는 방법이다. 이 절의 서두에서 설명한 바와 같이 수학적 정의를 통해 차등정보보호와의 차이를 정확하게 파악할 수도 있겠지만, 국소차등정보보호의 경우 자료의 수집 단계에서 직접 정보보호 처리가 진행된다는 점을 상기하기 바란다. 다음 [그림 5-10]은 이러한 차이를 나타낸다.



출처: https://www.accessnow.org/understanding-differential-privacy-matters-digital-rights/

[그림 5-10] 국소차등정보보호와 차등정보보호의 비교

국소차등정보보호는 구글(Erlingsson et al., 2014; Kairouz et al., 2014; Bittau et al., 2017), 애플(Apple Team, 2017), 마이크로소프트(Ding et al., 2017) 등의 대형 기업에서 현재 운영하고 있거나 활발히 연구 중이다. 본 보고서는 Duchi et al. (2018)에 소개된 내용을 중심으로 국소차등정보보호를 정의하고, 국소차등정보보호 자료를 사용하여 분포모수를 추정하는 과정을 살펴본다.

가. 국소차등정보보호의 정의

국소차등정보보호를 정의하기 전에 Warner(1965)에서 제안한 비식별 데이터 구성방법을 간단한 예를 들어 소개한다. Warner(1965)의 방법은 국소차등정보보호의 기본 개념을 충실하게 반영하고 있으며, 또한 그 개념을 이해하는 과정에서 반드시 필요한 방법이다.

[예제 3] 무단횡단을 하는 확률을 추정하기 위하여 자료를 수집하고자 한다. 자료를 수집하는 과정은 다음과 같이 진행된다.

1) 다음과 같이 두 가지 질문을 준비한다.
   A) 무단횡단을 한 적이 있다.
   B) 무단횡단을 한 적이 없다.

2) 자료를 제공하는 사람은 주사위를 던져 1, 2, 3, 4가 나오면 A 질문에 대한 답을 하고 5, 6이 나오면 B 질문에 대한 답을 한다. 자료를 수집하는 사람은 어떤 질문에 답을 했는지 알 수 없고, 질문에 대한 답 "예"와 "아니오"만 수집한다. ■

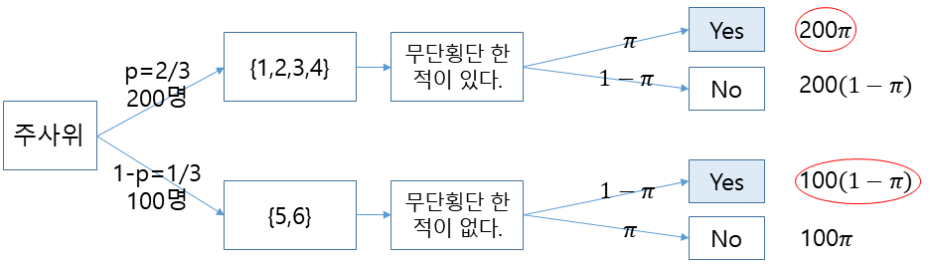
위 [예제 3]에서 실제 무단횡단을 한 사람의 비율이 π이고 A 질문에 답할 확률이 p라고 하자. 만약 n명의 사람을 조사하여 m명이 "예"라고 대답했다면 다음과 같이 π를 추정할 수 있다.

$$\hat{\pi} = \frac{p-1}{2p-1} + \frac{m}{(2p-1)n}.$$

위 추정량은 질문의 답이 "예"일 확률이 $\pi p + (1-\pi)(1-p)$이고 베르누이 분포를 따른다는 사실을 사용하여 최대가능도 추정법(maximum likelihood estimation)을 적용하여 얻은 것이다. 위 추정량의 기댓값을 계산하면 다음과 같이 $\hat{\pi}$이 π의 불편추정량(unbiased estimator)임을 확인할 수 있다.

$$E(\hat{\pi}) = \frac{1}{2p-1} + [p-1+\pi p+(1-\pi)(1-p)] = \pi.$$

다음 [그림 5-11]은 이러한 과정을 정리해서 보여준다.



[그림 5-11] Warner(1965) 방법을 따른 비율 추정 예제

위 [예제 3]은 질문에 답을 하는 사람이 주어진 질문에 정직하게 답한다면 자료를 조사하는 사람에게 자신의 무단횡단 경험을 공개하지 않아도 조사에서 원하는 확률을 추정할 수 있다는 사실을 보여준다. 이처럼 자료 제공자가 자료 수집가에게 정확한 정보를 제공하는 것을 거부할 필요가 있을 때 국소차등정보보호의 개념이 필요하다.

1) 국소차등정보보호를 위한 비식별 데이터의 구성

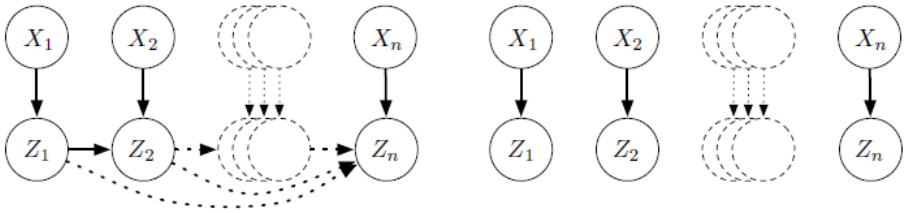
원자료 $X=(X_1,\ldots,X_n)$로부터 비식별 데이터 $Z=(Z_1,\ldots,Z_n)$를 구성하는 과정은 크게 두 가지 경우로 나눌 수 있다. [그림 5-12]는 이 두 경우를 비교하여 보여준다. 첫 번째 자료 생성 방법은 다음과 같이 순차적인 방식으로 이루어지는 것으로, 이를 순차적 교호 자료 생성 방법(sequentially interactive mechanism)이라고 한다.

$$\{X_i, Z_1, \ldots, Z_{i-1}\} \rightarrow Z_i \text{ 그리고 } Z_i \perp X_j | \{X_i, Z_{1:i-1}\},\ j \neq i.$$

위 식을 좀 더 자세히 설명하면 다음과 같다. 먼저 X_1으로부터 Z_1을 생성하고 그 다음 X_1과 Z_1으로부터 Z_2를 생성한다. 이와 같은 방식을 순차적으로 적용하여 X_n과 $Z_1,\ldots,Z_{n-1}$으로부터 Z_n를 생성한다. 단, 비식별 데이터 생성 과정은 반드시 $Z_i \perp X_j | \{X_i, Z_{1:i-1}\}$, $j \neq i$를 만족해야 한다. 두 번째 자료 생성 방법은 다음과 같이 X_i만 사용하여 Z_i를 생성하는 것이며 이를 순차적 비교호 자료 생성 방법(sequentially non-interactive mechanism)이라고 한다.

$$X_i \rightarrow Z_i \text{ 그리고 } Z_i \perp X_j | X_i,\ j \neq i.$$

다음 [그림 5-12]는 위 두 가지 자료생성 방법을 비교하여 보여준다.



출처 Duchi et al.(2018)

[그림 5-12] 국소차등정보보호 자료 생성 : 순차적 교호 방법 vs. 순차적 비교호 방법

2) 국소차등정보보호 자료와 방법의 정의

위에서 소개한 차등정보보호를 만족하는 자료 생성 과정은 Duchi et al.(2018)에서 기술된 내용이다. 자료 생성 과정에서 필요한 각각의 조건부 밀도함수를 모두 사용하면 차등정보보호 자료 Z의 조건부 밀도함수 $f_{Z|X}$를 정확하게 계산할 수 있다. 이러한 사실을 바탕으로 Duchi et al.(2018)은 다음과 같이 α-국소차등정보보호를 정의하였다.

[정의 2] Z_i를 원자료 X_i와 비식별 데이터 $Z_1,...,Z_{i-1}$를 사용하여 순차적 교호 방법으로 생성한 자료라 하고 Z_i의 조건부 밀도함수를 $f_{Z_i|X_1,Z_1,...,Z_{i-1}}$이라고 하자. 만약 $f_{Z_i|X_1,Z_1,...,Z_{i-1}}$가 적당한 $\alpha>0$에 대하여 아래의 부등식을 만족하면 $f_{Z_i|X_1,Z_1,...,Z_{i-1}}$을 α-국소차등정보보호 방법(α-differential privacy mechanism)이라고 하고 Z_i를 X_i에 대한 α-국소차등정보보호 자료(α-differential privacy view)라고 한다.

$$sup_{u_i \neq v_i} sup_{z_i} \frac{f_{Z_i|X_i,Z_1,...,Z_{i-1}}(z_i|u_i,z_1,...,z_{i-1})}{f_{Z_i|X_i,Z_1,...,Z_{i-1}}(z_i|v_i,z_1,...,z_{i-1})} \leq exp(\alpha). \quad \blacksquare$$

만약 비식별 데이터의 구성 과정이 순차적 비교호 방법인 경우에는 다음과 같이 좀 더 간단하게 α-국소차등정보보호를 정의할 수 있다.

$$sup_{u_i \neq v_i} sup_{z_i} \frac{f_{Z_i|X_i}(z_i|u_i)}{f_{Z_i|X_i}(z_i|v_i)} \leq exp(\alpha).$$

차등정보보호는 비식별 데이터 전체에 대한 조건부 결합분포를 고려하지만 [정의 2]에서 기술한 국소차등정보보호의 정의는 개별 비식별 데이터에 대한 조건부 결합분포를 고려하고 있다. 또한 차등정보보호가 비식별 데이터의 수에 특별한 제한을 두지 않는 반면 국소차등정보보호의 경우 원자료의 개수와 동일한 개수의 비식별 데이터만 구성할 수 있다는 차이점이 있다. 특히 순차적 비교호 생성 방법을 적용한 경우 α-국소차등정보보호 자료는 $n\alpha$-차등정보보호 자료가 된다는 사실을 쉽게 확인할 수 있다.

나. 국소차등정보보호를 활용한 분포모수의 추정

α-국소차등정보보호 방법을 적용하여 분포모수를 추정하는 방법을 소개하고 각각의 방법에서 어떠한 방식으로 α-국소차등정보보호 자료를 생성하는지에 대하여 간단하게 살펴본다.

[정리 8] $W=(W_1,...,W_n)$가 위치모수가 $\mu=0$이고 산포모수가 $2T$인 이중지수분포에서 추출한 임의표본이라고 하자. 원자료 $X=(X_1,...,X_n)$가 $E(X_i)\in[0,1]$, $E(|X_i|^k) \leq 1$, $i\in\{1,...,n\}$을 만족하면 다음과 같이 W로 교란하여 얻은 비식별 데이터 $Z=(Z_1,...,Z_n)$은 X의 α-국소차등정보보호 자료이다.

$$Z_i=[X_i]_T+W_i,\ i\in\{1,...,n\},\ T=(n\alpha^2)^{1/2k},\ k>1.\ \blacksquare$$

위 [정리 8]은 Duchi et al.(2018)에서 소개되었으며 원자료 X의 크기를 T로 제한하고 이중지수분포에서 추출한 오차를 직접 더하면 α-국소차등정보보호 자료를 생성할 수 있음을 의미한다. Duchi et al.(2018)은 이러한 국소차등정보보호 자료를 활용하여 원자료 X의 평균을 다음과 같이 추정하고 있다.

$$\hat{\theta}=\frac{1}{n}\sum_{i=1}^{n} Z_i\ .$$

Duchi et al.(2018)은 적당한 정칙조건 하에서, 국소차등정보보호 자료를 사용하여 위치모수를 추정하는 경우 α-국소차등정보보호 최소최대위험이 $(n\alpha^2)^{-(k-1)/k}$임을 증명하였다. 따라서 α가 0에 가까워질수록, 즉 정보보호가 강력해질수록 최소최대위험은 증가하며 반대의 경우 감소하게 된다. 정보보호를 강화할수록 국소차등정보보호 자료의 정보 손실이 많아지므로 위 결과는 직관적인 예측과 일치한다. 하지만 $k\to\infty$이고 $\alpha\in[0,1]$인 경우, 원자료를 사용하는 방법의 최소최대위험이 $n^{-min\{1,2(k-1)/k\}}=n^{-1}$이고 국소차등정보보호 자료를 사용하는 방법의 최소최대위험이 $(n\alpha^2)^{-1}$이다. 따라서 국소차등정보보호 자료를 사용하는 방법에서 α^2만큼의 손실을 갖는다는 것을 확인할 수 있다.

위 추정량 $\hat{\theta}$를 [예제 2]에서 제안한 추정량 $\tilde{\theta}$와 비교하면 차등정보보호와 국소차등정보보호의 차이를 확인 할 수 있다.

$$\tilde{\theta}= Z/n=(\sum_{i=1}^{n}X_i+e)/n.$$

$\hat{\theta}$의 경우 원자료 X의 값 대신 국소차등정보보호 자료의 값을 사용하여 평균을 추정하고 있지만 $\tilde{\theta}$의 경우 원자료 X의 합을 정확하게 알아야 추정량을 계산할 수 있다.

다음 정리는 Duchi et al.(2018)이 증명하였으며 원자료의 중앙값(median)을 α-국소차등정보보호 자료를 사용하여 추정하는 방법을 소개하고 있다.

[정리 9] $W=(W_1,...,W_n)$가 $W_i\in\{-1,+1\}, P(W_i=1)=e^{\alpha}/(e^{\alpha}+1), i\in\{1,...,n\}$을 만족하는 임의표본이라고 하자. 원자료 $X=(X_1,...,X_n)$가 $med(X_i)\in[-1,1]$, $E(|X_i|)<\infty$, $i\in\{1,...,n\}$이면 다음과 같은 비식별 데이터 $Z=(Z_1,...,Z_n)$는 X의 α-국소차등정보보호 자료이다.

$$Z_i=\frac{e^{\alpha}+1}{e^{\alpha}-1}W_i\ sign(\theta_i-X_i).$$

단, $\theta_{i+1}=[\theta_i-\eta_i Z_i]_r\ \ i\in\{1,...,n\},\ \eta_i>0.$ ■

원자료 X의 중앙값 $\theta = med(X_i)$를 추정할 때 X의 순서통계량을 사용하는 것이 가장 간단한 방법이지만 결정이론(decision theory)에 근거하여 경사하강 알고리즘(gradient descent algorithm)을 사용할 수도 있다. 위 [정리 9]에서 생성한 자료 Z_i는 이러한 알고리즘에 사용되는 경사(gradient)를 확률적 오차를 허용하도록 개선한 확률적 경사(stochastic gradient)이며, α-국소차등정보보호 자료가 되도록 Warner(1965)에서 제안한 재현자료 구성방법을 추가로 적용한 것이다. Duchi et al.(2018)은 이러한 재현자료를 사용하여 X의 중앙값을 다음과 같이 추정하고 있다.

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\theta_{i.}.$$

선형회귀 모형의 기울기를 추정하는 경우에도 [정리 9]와 비슷한 방법을 적용할 수 있으며 일반화 선형모형(generalized linear model)으로 쉽게 확장할 수 있다.

3. 차등정보보호의 방법의 평가

통계학에서 추정량의 성능을 평가하는 방법 중 가장 기본적인 방법은 일치성(consistency)을 확인하는 방법이다. 또한 추정량이 일치성을 가지는 경우 모수와 가까워지는 수렴 속도를 계산하여 추정량의 성능을 좀 더 구체화 할 수 있다. 또 다른 평가 방법은 추정량의 최소최대위험(mini-max risk)을 계산하고 이것이 알려진 하한 혹은 최적의 하한과 일치하는지를 파악하는 것이다. 여기에서는 α-차등정보보호와 α-국소차등정보보호 방법에 대하여 일치성과 최소최대위험을 정의하고, 지금까지 살펴본 다양한 방법에 대하여 수렴 여부와 수렴 속도 등을 소개한다. 본 내용을 기술하기 전에 다음과 같이 몇 가지 기호를 정의한다.

- χ : 원자료의 표본공간(sample space)
- $\boldsymbol{P}$: χ 위에서 정의된 분포 P의 집합
- Θ : 모수공간(parameter space)
- θ : $\boldsymbol{P} \rightarrow \Theta$: P에서 정의된 모수(paramter)
- ρ : Θ 위에서 정의된 거리(metric)
- $\Phi : R_+ \rightarrow R_+$: $\Phi(0) = 0$이고 증가하는 함수
- χ^n : χ의 n차 곱공간(n-fold product space)
- $\hat{\theta}_n$: $\chi^n \rightarrow \Theta$: 추정량(estimator)

가. 차등정보보호 방법의 평가

차등정보보호 방법을 평가하려면 일치성과 최소최대위험을 차등정보보호의 개념에 맞도록 수정해야 한다. 수정된 평가 측도를 소개하기 전에 먼저 일치성과 최소최대위험을 계산하는 절차에 대하여 간단히 알아보고 손실함수, 위험함수, 최대위험, 최소최대위험 등의 측도들을 정의한다.

1) 위험함수와 최대최소위험

추정량 $\hat{\theta}_n = \hat{\theta}(X_1, \ldots X_n)$에 대하여 아래와 같이 정의한 함수 L을 손실함수(loss function)이라고 한다.

$$L(\hat{\theta}_n, \theta) = (\phi \circ \rho)(\hat{\theta}_n, \theta).$$

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

$$R(\hat{\theta}_n, \theta) = E_P(L(\hat{\theta}_n, \theta)).$$

손실함수가 현재 수집된 자료를 사용하여 추정량을 만들었을 때 발생하는 손실의 크기를 나타낸다면 위험함수는 수집 가능한 모든 자료를 사용하여 추정량을 만들었을 때 발생하는 손실의 크기를 나타낸다. 즉 위험함수는 분포에서 발생 가능한 손실의 기댓값을 의미한다.

추정량 $\hat{\theta}_n$이 고정된 경우, 위험함수를 분포집합 $\boldsymbol{P}$에서 모수공간 Θ를 거쳐 양의 실수 집합으로 가는 하나의 함수로 이해할 수 있으며, 다음과 같이 최대위험 R_{max}를 정의할 수 있다.

$$R_{max}(\hat{\theta}_n, \boldsymbol{P}) = sup_{P \in \boldsymbol{P}} E_P(L(\hat{\theta}_n, \theta)).$$

만약 최대위험이 0으로 수렴하면 추정량 $\hat{\theta}_n$이 일치성을 가진다고 말하며 고정된 모수에 대한 추정량의 평가 측도로 사용된다. 최대위험은 고정된 하나의 추정량에 대한 위험을 의미하므로 어떠한 추정량이 최적인가를 평가하기 위하여 주어진 최대위험을 추정량 $\hat{\theta}_n$의 함수로 간주하여 다음과 같이 최소최대위험(minimax risk) M을 정의한다.

$$M(\boldsymbol{P}) = inf_{\hat{\theta}_n} R(\hat{\theta}_n, \boldsymbol{P}).$$

최소최대위험은 가능한 모든 추정량에 대하여 발생할 수 있는 최대위험을 의미한다. 따라서 주어진 추정량 $\hat{\theta}_n$이 최소최대위험보다 더 큰 최대위험을 갖는다면 다른 추정량을 구하여 최대위험을 줄일 수 있다는 것을 의미하며 최소최대위험과 같은 최대위험을 갖는다면 이를 최소최대추정량(minimax estimator)이라고 한다.

최대최소위험은 어떠한 분포집합을 고려하는가에 따라 결정되며 거의 대부분의 문제에서 잘 알려져 있다. 따라서 보통의 경우 주어진 추정량 $\hat{\theta}_n$의 최대위험을 계산하여 최대최소위험과 비교를 하는 과정을 통해 추정량을 평가한다. 만약 최소최대위험이 알려지지 않은 경우라면 다음과 같은 절차를 진행한다.

① 모수공간 Θ의 2-패킹 구분기호 집합(2-packing index) V를 구성한다.

$$\rho(\theta_\nu, \theta_{\nu'}) \geq 2\delta,\ \nu \neq \nu' \in V.$$

② 유한 집합 V에 대하여 $\{P_\nu : \nu \in V\} \subset \boldsymbol{P}$이면

$$M(\boldsymbol{P}) \geq \sum_{\nu \in V} M(P_\nu).$$

③ $\Psi = \{\psi : \chi^n \rightarrow V\}$가 모든 검정함수의 집합이고 U가 V 위에서 균등분포이면

$$M(\boldsymbol{P}) \geq \phi(\delta)\, inf_{\psi \in \Psi} P(\hat{\psi} \neq V).$$

④ $|V| = 2$이면 르캄의 부등식(Le cam's ineqaulity)을 적용한다.

$$\phi(\delta)\, inf_{\psi \in \Psi} P(\hat{\psi} \neq V) \geq \left\{1 - \| M_\nu^n - M_{\nu'}^n \|_{TV}\right\} / |V|.$$

⑤ $|V| \geq 3$이면 파노의 부등식(Fano's ineqaulity)을 적용한다.

$$\phi(\delta)\, inf_{\psi \in \Psi} P(\hat{\psi} \neq V) \geq 1 - \{I(X;V) + \log 2\} / \log |V|.$$

손실함수와 위험함수의 개념은 추정량을 구현하는 과정에도 사용된다. 손실함수를 현재 수집된 자료 각각에 적용하여 평균을 취하면 일종의 평균손실을 얻을 수 있고 대수의 법칙(Law of large numbers)에 근거하여 위험함수를 대체하여 사용할 수 있다. 이러한 위험함수를 경험위험함수(empirical risk function) $\hat{R}_n$이라고 한다.

$$\hat{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} L(\hat{\theta}(X_i), \theta).$$

경험위험함수를 모수 θ의 함수로 간주하고 아래와 같이 이를 최소로 하는 θ의 값 $\hat{\theta}_n$을 하나의 추정량으로 사용할 수 있으며 이를 경험위험을 최소화 추정량(empirical risk minimizer)이라고 한다.

$$\hat{\theta}_n = argmin_\theta \hat{R}_n(\theta).$$

지금까지 논의한 내용을 바탕으로 추정량의 평가 과정을 간단하게 요약하면 다음과 같다.

① 최소최대위험의 하한 L_n을 계산한다.

$$L_n \leq M(\boldsymbol{P}).$$

② 주어진 추정량 $\hat{\theta}_n$의 위험함수의 상한 U_n을 계산한다.

$$R(\hat{\theta}_{n,}\theta) \leq U_n.$$

③ L_n과 U_n의 크기를 비교한다.

마지막으로 임의표본 $X = (X_1, \ldots, X_n)$에 대하여 $Var(X_1) = \sigma^2$이 알려져 있는 경우 표본평균 $\overline{X}$를 평가하는 과정을 간략히 소개한다. 손실함수를 $L(\hat{\mu}, \mu) = (\hat{\mu} - \mu)^2$로 두면 위험함수는 $R(\hat{\mu}, \mu) = E_\mu(\hat{\mu} - \mu)^2$이고, 표본평균 $\hat{\mu}_n = \overline{X}$가 다음과 같이 경험위험 최소화 추정량이 된다는 것을 쉽게 확인할 수 있다.

$$\overline{X} = argmin_\mu \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

표본평균의 최대위험은 $E(\overline{X}-\mu)^2 = Var(X_1)/n = \sigma^2/n$ 이고, 이 값은 모수 μ에 의존하지 않으며 알려진 최소최대위험과 동일하다. 따라서 표본평균은 최소최대위험 추정량이다.

2) 차등정보보호 방법과 최대최소위험

위에서 기술한 것처럼 차등정보보호 방법을 평가하는 과정은 일치성과 최소최대위험을 차등정보보호의 개념에 맞도록 수정해야 한다. 마지막으로 Wasserman과 Zhou(2010)과 Duchi et al.(2018) 등이 기술한 차등정보보호 방법에 대한 최대위험과 최소최대위험의 정의를 간단히 소개하고 이 절을 마친다.

[정의 3] 원자료 $X=(X_1,...,X_n)$가 주어져 있을 때 비식별 데이터 $Z=(Z_1,...,Z_k)$의 조건부 분포를 Q_α라고 하자. Z가 X에 대한 α-(국소)차등정보보호 자료일 때 추정량 $\hat{\theta}_{n,k}=\hat{\theta}_{n,k}(Z|X)$에 대하여 다음과 같이 정의된 위험함수를 α-(국소)차등정보보호 위험함수라고 말한다.

$$R(\hat{\theta}_{n,k}, Q_\alpha, P) = E_{P,Q_\alpha} L(\hat{\theta}_{n,k}, \theta). \quad \blacksquare$$

[정의 4] 원자료 $X=(X_1,...,X_n)$가 주어져 있을 때 비식별 데이터 $Z=(Z_1,...,Z_k)$의 조건부 분포를 Q_α라고 하자. Z가 X에 대한 α-(국소)차등정보보호 자료일 때 추정량 $\hat{\theta}_{n,k}=\hat{\theta}_{n,k}(Z|X)$에 대하여 다음과 같이 정의된 위험함수를 (국소)차등정보보호 최소최대위험이라고 말한다.

$$M(\boldsymbol{P}, \boldsymbol{Q}_\alpha) = inf_{Q_\alpha \in \boldsymbol{Q}_\alpha} inf_{\hat{\theta}_{n,k}} sup_{P \in \boldsymbol{P}} R(\hat{\theta}_{n,k}, Q_\alpha, P). \quad \blacksquare$$

제4절 활용 사례

이 절에서는 차등정보보호의 기준을 만족하도록 자료[15]를 생성한 사례들을 소개한다. 이를 통해 바람직한 실무 적용을 위한 차등정보보호 연구 방향을 모색하고자 한다. 참고로 용어를 정리하면 다음과 같다. 전통적인 매스킹 기법을 적용한 자료는 국내에서 비식별 데이터(masked data)라고 불리고 있다. 3절에서는 차등정보보호의 기준을 만족하도록 처리된 자료에 대해 차등정보보호 자료(differential privacy view)라고 표현했다. 이 절의 마지막 소절에서는 통계적 비밀보호 분야에서 다루는 재현자료(synthetic data)를 다룬다. 이들은 모두 각 방법에 따라 개별정보 노출이 제한되도록 변형한 자료를 일컫는다. 영어로 쓰인 문헌에서는 이들의 구분이 명확하나, 국내에는 용어에 대한 합의가 아직 명확하지 않은 실정이다. 이에 본 보고서에서는 다음 <표 5-3>과 같이 용어를 사용하고 있으며, 보고서에서 비식별 데이터라는 용어는 모두 광의의 의미(sanitized data)임을 밝혀둔다.

〈표 5-3〉 용어 구분

비식별 데이터(sanitized data)		
비식별 데이터 (masked data)	차등정보보호 자료 (differential privacy view)	재현자료 (synthetic data)

1. 직접적인 잡음 첨가에 의한 차등정보보호 자료 생성

차등정보보호를 적용하려할 때 가장 자연스러운 시도는 라플라스 잡음을 추가해서 차등정보보호의 기준을 만족하도록 자료를 변형하는 것이다. 차등정보보호와 관련된 이론들은 어렵지만, 실제 사용할 때는 단순히 어떤 적절한 메커니즘(machine, algorithm)을 통해 생성시킨 잡음을 더하는 것만으로도 차등정보보호 자료를 얻을 수 있다. 따라서 차등정보보호 자료를 생성하는 어떤 매커니즘을 잘 만들면, 차등정보보호 자료를 만드는 것이 실무적으로는 간단한 작업이 될 수도 있다. 차등정보보호의 정의를 따라 가장 먼저 생각할 수 있는 메커니즘은 라플라스 머신이다. 라플라스 분포[16]를 따르는 잡음을 원자료에 더해서 차등정보보호 자료를 만드는 사례들은 여럿 존재한다. 이 소절에서는 두 개의 석사논문 사례를 살펴보도록 한다.

15) 차등정보보호일 경우에는 요청받은 쿼리에 대한 분석 결과물이며, 국소차등정보보호의 경우에는 개별 수준의 자료가 된다.

16) 물론 라플라스 분포의 모수는 ϵ 및 쿼리의 민감도에 따라 결정된다.

먼저, 잘 알려진 Adult dataset[17]이라는 자료에서 인종 및 교육기간 변수에 차등정보보호를 적용한 사례(Seleshi와 Asseffa, 2017)를 소개한다. 약 3만 명으로 구성된 이 데이터베이스를 이용해 인종별로 교육받은 기간에 대한 빈도표를 만들 수 있다. 라플라스 잡음을 추가할 때, 다양한 ϵ값에 대해서 빈도표를 만들면 <표 5-4>와 같은 결과를 얻는다. ϵ값이 작을수록 (혹은 원자료에서 발생한 작은 차이에 대한 결과의 변화를 허용하지 않을수록) 원래 빈도값과 많이 다른 (혹은 자료 유용성이 떨어지는) 자료가 생성되는 것을 알 수 있다.

〈표 5-4〉 빈도표에 대한 차등정보보호 자료 생성 사례

원래 빈도값	$\epsilon = 2$	$\epsilon = 1$	$\epsilon = 0.5$	$\epsilon = 0.1$	$\epsilon = 0.01$	$\epsilon = 0.001$
38	38.143	37.949	38.004	35.391	35.391	27.812
129	1288.999	128.988	128.904	129.887	129.887	114.259
267	266.913	267.005	266.991	264.811	264.811	376.422
515	515.018	515.110	515.433	514.276	521.427	508.410
381	380.949	381.026	381.234	380.454	367.728	318.299
708	707.936	708.002	707.957	708.280	710.750	633.458
927	927.001	927.043	926.672	927.737	912.219	905.903
308	308.059	308.102	308.206	307.251	310.304	229.256
7362	7361.985	7362.021	7361.968	7365.090	7363.949	7361.250
4952	4951.943	4951.641	4951.469	4951.782	4986.021	4766.830
874	874.015	874.008	873.742	874.203	889.217	999.507
680	680.025	679.971	679.928	678.678	680.438	668.460
2667	2666.956	2666.996	2666.966	2667.153	2675.991	2745.337
666	666.002	666.140	666.026	665.518	672.842	643.887
132	132.036	132.031	131.979	131.312	141.364	93.348
93	92.968	93.167	93.118	92.744	89.435	114.495

이번에는 9개 범주의 직업군에 대해서 주당 노동시간 평균값을 나타내보자. 그러면 다음 <표 5-5>의 결과를 얻을 수 있으며, 앞의 빈도표의 경우와 동일한 경향을 볼 수 있다.

17) http://archive.ics.uci.edu/ml/datasets/Adult

〈표 5-5〉 평균에 대한 차등정보보호 자료 생성 사례

원자료	$\epsilon = 2$	$\epsilon = 1$	$\epsilon = 0.5$	$\epsilon = 0.1$	$\epsilon = 0.01$	$\epsilon = 0.001$
39.032	39.062	39.069	38.880	39.458	35.839	131.613
41.379	41.290	41.126	41.270	42.600	25.580	293.379
40.267	40.325	40.354	40.335	41.172	30.571	60.371
40.983	41.125	40.803	41.516	37.189	55.854	-121.157
48.818	48.783	48.975	48.997	45.875	51.524	-124.318
44.422	44.453	44.178	43.374	43.694	63.597	-41.978
31.919	32.051	31.890	31.857	32.276	85.032	362.218
28.429	28.445	28.361	30.040	26.414	6.069	104.168
32.714	32.847	32.713	32.852	29.471	47.370	119.513

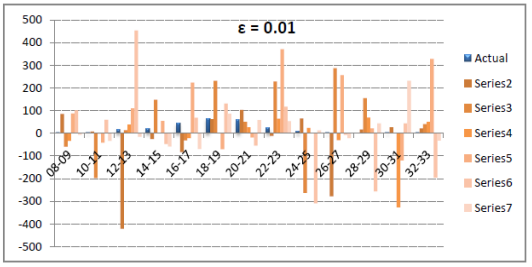
빈도수와 평균값에 대해 랜덤하게 생성한 잡음을 첨가해서 차등정보보호의 기준을 만족시킨 위의 실험들은 각각 1회 이루어졌다. 실험할 때마다 생성된 차등정보보호 자료의 값이 다를 수 있겠지만, ϵ값이 동일하면 생성된 자료의 분포는 유사하다고 볼 수 있다.

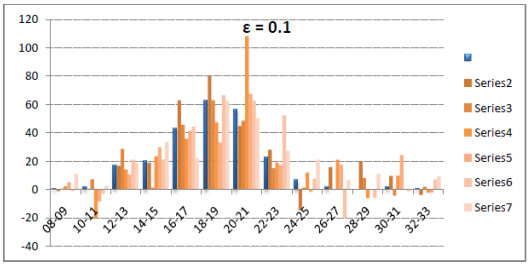
차등정보보호에서 ϵ값이 작으면 강한 정보보호를, 크면 약한 정보보호를 의미한다. 자연스럽게 ϵ값이 작으면 첨가하는 잡음의 크기가 클 것이고, 많은 정보손실을 가지는 자료를 제공하게 된다. <표 5-4> 및 <표 5-5>에서도 이런 경향을 확인할 수 있다. 위의 예들에서는 ϵ값이 0.01인 수준까지는 정보손실의 변화가 크지 않지만, 0.001이 되면서 정보손실이 급격하게 커지는 것을 볼 수 있다.

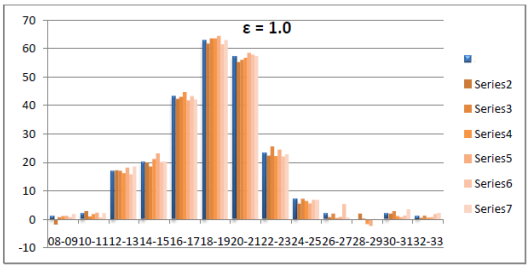
다음으로 보건의료 분야의 Woolson dataset이라는 자료에 차등정보보호를 적용하는 사례(Lee, 2008)를 살펴보자. 이 데이터에서 안압 범위별 환자수를 빈도표로 나타내면 <표 5-6>과 같다. 몇몇 ϵ값에 대해서 <표 5-6>의 빈도값들에 대한 차등정보보호 자료를 생성해서 <표 5-6>의 오른쪽에 그림으로 나타냈다. 이번에는 차등정보보호 자료를 여러 개 생성했다. ϵ값이 클 때는 참값과 가까운 값들이 반복적으로 생성되는 것을 확인할 수 있다. 즉 ϵ값이 상대적으로 크면($\epsilon = 1$), 반복적으로 차등정보보호 자료를 생성해도 원래와 유사한 자료가 안정적으로 만들어지는 것을 알 수 있다. 반면 ϵ값이 작을 때는 ($\epsilon = 0.01$) 원래의 값과 상당히 다른 값들이 생성되는 것도 확인할 수 있다.

〈표 5-6〉 빈도표에 대한 반복적인 차등정보보호 자료 생성 사례

IOP	# of patients
08-09	1
10-11	2
12-13	17
14-15	20
16-17	43
18-19	63
20-21	57
22-23	23
24-25	7
26-27	2
28-29	0
30-31	2
32-33	1







만약 이러한 사례들과 같이 원자료에 라플라스 잡음을 첨가하는 방식으로 차등정보보호를 적용한다면 우리는 매스킹 기법을 사용할 때와 동일한 문제에 봉착하게 된다. 그것은 자료를 많이 변형하면 정보손실이 커지고 노출위험이 낮아지므로, 적절한 수준에서 타당한 자료를 선택해야 한다는 것이다. 따라서 적절한 ϵ값을 선택하는 것이 주요 이슈가 된다.

그렇다면 어떤 ϵ값을 사용하는 것이 타당할까? 정보손실에 지나치게 민감하지 않은 경우라면 첫 번째 사례에서는 아마도 0.01의 값을 선택하는 것이 타당해 보인다(<표 5-4>, <표 5-5>). ϵ값이 2여도 충분한 노출제어가 된다고 생각되는 경우라면, ϵ값이 2일 때 생성한 잡음을 첨가하여 정보손실이 더 적은 자료를 제공할 수 있다. 한편 데이터의 크기가 훨씬 작은 두 번째 사례에서 ϵ값이 1인 경우는 원래 자료와 너무 차이가 없어서 노출제어가 되지 않는다고 판단할 수 있다(<표 5-6>). 또 0.01인 경우는 원자료와 너무

달라져서 차등정보보호 자료를 쓸 수 없다고 생각할 수 있다. 결론적으로 0.1을 선택해서 차등정보보호 자료를 제공할 수 있다. 물론 ϵ값이 0.1이고 랜덤하게 생성된 잡음을 더할 때도, 우연히 정보손실이 너무 큰 경우가 생길 수도 있다.

이렇듯 적절한 ϵ값을 결정하는 것은 또다시 정보손실 파악을 위한 여러 측면의 자료 분석을 요구한다. 그렇다고 그 타당성이 명확하게 결정된다고 말하기도 어렵다. ϵ의 의미와 값 결정에 대한 문제는 차등정보보호 연구에 있어 중요한 이슈이며, 그 연구 방향에 대한 논의는 5절 결론에서 다루기로 한다.

이 소절에서 내리는 결론은 두 가지이다. 첫째는 라플라스 잡음을 자료에 직접 더하는 단순한 방법으로는 매스킹과 유사한 의사결정의 어려움에 빠진다는 것이다. 이를 해결하기 위해서는 ϵ을 얼마로 하는 것이 적절한가를 결정하는 문제를 다루어야 한다. 두 번째는 ϵ이 주어졌다고 하더라도 정보손실이 너무 커서 쓸모없는 자료가 만들어질 수 있다는 것이다. 첫째 결론은 "formal privacy" 관점에서 장기적인 연구가 필요하므로 5절에서 논의하면서 후속 연구로 남긴다. 이 절의 남은 부분에서는 두 번째 내용을 논의하도록 한다. 이제 주어진 ϵ에서 정보손실을 최대한 줄여서 차등정보보호 자료를 만들기 위한 노력들, 즉 정보손실을 최소화하는 알고리즘 연구 사례들을 계속해서 소개하겠다.

2. 통계적 모형을 이용한 차등정보보호 자료 생성

이 소절에서는 Jiang 외(2017)의 사례를 소개한다. Jiang 외(2017)는 2절에서 언급되기도 했던 뉴욕 택시 운행 자료[18]로 차등정보보호를 적용하는 방안을 논의했다. 주어진 ϵ에 대해서 단순히 잡음을 추가하는 방법을 사용하는 것보다 정보손실을 줄이기 위한 주요 아이디어로, 자료에 통계적인 모형을 적용하고 특정 단계에서 라플라스 잡음을 추가하는 방법을 제안했다.

참고로 뉴욕 택시 데이터를 제공할 때 단순한 잡음 추가 방식으로 차등정보보호의 기준을 만족시키는 문제에 관해서는 관련 블로그[19] 및 박민정(2016)에 자세히 설명되어 있다. 이 데이터에 차등정보보호를 단순하게 적용할 때 주요 아이디어는 지역을 그리드로 나누고, 발생 빈도에 관한 확률을 색의 진하기를 달리해서 각 셀들에 표현한다는 것이었다. 앞 소절에서 설명한 것처럼 ϵ값에 따라서 제공되는 분석 결과의 정확도가 달라진다.

---

18) NewYork City Taxi Trip Data (2010-2013). Available from: https://databank.illinois.edu/datasets/IDB-9610843.
19) Neustar (2014), Company Blog(research.neustar.biz/author/atockar)

[그림 5-13]은 기존의 잡음첨가 방법으로 차등정보보호를 만족하여 자료를 제공할 때 그 결과를 보여준다. 제공하는 자료는 지역별 택시 승하차 통계이며, 외부인이 알아내려는 민감정보는 어느 개인의 이동위치이다. 차등정보보호를 적용하지 않는 기존 자료제공 방식에서는, 뉴욕 택시 데이터를 분석하는 외부인은 데이터에 포함된 어느 개인이 이동한 지역을 특정할 수 있다. [그림 5-13]에서 파란 선으로 표시된 구역은 외부인 노출시키는데 성공한 어떤 개인의 이동 정보를 나타낸다. 한편 다양한 농도의 초록색 표시는 차등정보보호를 적용한 자료제공 결과를 나타낸다. 이 경우 외부인은 노출시키려는 개인의 이동에 대한 정보를 확률로 얻을 수 있으므로 어느 개인이 이동한 지역을 특정하기 어렵다.



출처: Neustar, 2014, Company Blog

[그림 5-13] 차등정보보호를 만족하는 자료제공 사례

이제 Jiang 외(2017)의 내용을 자세히 설명하자. 이 논문은 뉴욕 택시 데이터 중 택시 승차 기록 항목만을 분석의 대상으로 삼았다. 따라서 각 장소별로 택시의 승차 빈도가 차등정보보호의 기준을 만족시켜 보호해야 할 자료가 된다. 통계적 모형을 적용할 균일한[20] 데이터 세트를 만들기 위해 8월, 화~목요일, 오후 2시~4시 30분까지, 5분 간격으로 택시의 승차 빈도를 얻는다. 가장 중요한 아이디어는 5분 간격으로 빈도를 측정할 때,

20) 여기서 통계적 모형을 적용할 때는 하나의 모수로 설명되는 분포에서 자료가 생성되었다고 가정한다. 따라서 자료가 최대한 균일할(homogeneous) 필요가 있다.

이 값이 대부분 0 또는 1이 되도록 지역을 잘게 쪼개는 것이다. 즉, 기존의 단순 잡음 추가 방식의 활용 사례에서 제안된 넓은 그리드를 이런 조건에 맞게 쪼개는 것이 이의 주요 제안 중 하나이다. 또한 1이 최대한 많이 생기도록 셀의 크기를 정해야 한다. 논문에서 제안된 크기는 $5m \times 5m$이며, 이를 최소셀이라 한다. [그림 5-13]에서 네모로 표현된 각 구역을 이러한 최소셀들로 다시 나눈다. 이제 각각의 최소셀에서 택시 승차가 있는지 없는지를 총 T회 관찰하는 문제를 다루면 되고, 8월에는 각 셀별로 $T = 660$개의 관측값을 얻는다. 각 ij번째 셀에서 관찰하는 총 T회의 사건 중, 0이 일어나는 횟수와 1이 일어나는 횟수를 $m_{ij}(0)$ 및 $m_{ij}(1)$로 표현하면 $m_{ij}(0) + m_{ij}(1) = T$가 된다.

다음으로 최소셀에 베르누이 분포를 가정한다. 각 ij번째 셀에서 택시 승차 사건이 발생할 확률을 p_{ij}라고 표현하면, 데이터로부터 이 확률의 추정값은 $\tilde{p}_{ij} = m_{ij}(1)/T$가 된다. 이제 차등정보보호의 기준을 만족시키기 위해, 이 확률을 아래와 같이 라플라스 잡음 e_{ij}을 추가한 빈도를 이용해 계산한다.

$$\tilde{p}_{ij} = \frac{m_{ij}(1) + e_{ij}}{T} .$$

이제 최소셀에서 택시 승차가 일어나는 확률은 $\tilde{p}_{ij}$를 이용하고, 빈도는 $\tilde{p}_{ij}$를 모수로 하는 베르누이 분포에서 자료를 생성한 후 이를 이용해서 결정한다. 이 때 빌딩 안과 같이 택시 승차가 절대 일어나지 않는 셀은 강제로 0의 값을 할당한다.

이제 이용자가 사용하려고 요구하는 더욱 큰 단위, 즉 $50m \times 50m$, $100m \times 100m$, $200m \times 200m$ 단위의 셀들에서 정보를 제공하는 문제를 생각하자. [그림 5-13]에서 네모로 표현된 구역들이 이러한 큰 단위 셀들이다. Jiang 외(2017)에서는 세 가지 방안을 제시하고 비교한다. 방법 ①은 지금까지 설명한 방식으로 얻은 최소셀마다 생성된 빈도를 모두 더해서 큰 셀의 빈도를 얻는 것이다. 이는 다음의 세 단계로 정리할 수 있다. ㉮ 1이나 0만이 관측되는 최소셀을 T회 반복해서 관측하고, 1이 나온 횟수에 라플라스 잡음을 더한다. ㉯ 잡음이 더해진 1이 나온 횟수를 이용해 최소셀에 대한 확률 추정값을 얻고, 추정된 확률의 베르누이 분포에서 생성시킨 빈도값을 최소셀의 빈도값으로 삼는다. 이 빈도값이 최소셀에 대한 차등정보보호 자료가 된다. ㉰ 큰 셀의 빈도값은 최소셀의 차등정보보호 자료들을 단순 집계해서 얻는다.

방법 ②는 다음과 같다. 만약 p_{ij}가 모든 최소셀들에서 동일하다면 큰 셀들에 관한 정보는 이항분포를 이용해서 간단히 제공할 수 있다. 그러나 최소셀들마다 택시 승차 확률이 모두 다르므로 포아송-이항 분포를 이용한다. 포아송-이항 분포의 누적분포함수(cumulative distribution function)는 포아송 근사나 정규 근사를 통해 얻을 수도 있지만

최근 연구결과에 의해 수식(closed form)으로 얻을 수도 있다. 어느 큰 셀에 최소셀이 총 M개 존재해서 큰 셀이 0에서 M사이의 값을 가진다면 포아송-이항분포의 누적분포함수 수식은 다음과 같다.

$$\Pr(Y \le y) = \frac{1}{M+1} \sum_{j=1}^{M} \frac{[1 - exp(-iwj(y+1))]\tilde{\gamma}_j}{1 - exp(-iwj)} \quad .$$

단, $\tilde{\gamma}_j = \prod_{l=1}^{M} [1 - \tilde{p}_l + \tilde{p}_l \times exp(iwj)]$, $w = \frac{2\pi}{M+1}$ 및 $j = 1, ..., M$이다.

이 누적분포함수를 직접 계산하기는 어려워 최근 제안된 푸리에 변환(Fast Fourier Transform)을 이용하는 방법으로 얻는다. 즉, 최소셀에서 차등정보보호의 기준을 만족하도록 얻어진 교란된 확률 $\tilde{p}_l$을[21] 이용해, 큰 셀의 빈도를 나타내는 변수 Y에 대해 교란된 누적확률분포를 얻고, 그 분포에서 빈도값을 생성할 수 있다. 이러한 과정을 거쳐 얻은 빈도값들도 3절의 [정리 2]를 따라 여전히 차등정보보호가 만족된다.

마지막으로 최소셀과 같은 방식으로 큰 셀에서도 확률과 빈도를 얻을 수 있다. 이를 방법 ③이라 하자. 어느 큰 셀에서 5분마다 관측할 때 얻는 빈도가 0부터 b까지 있다고 하자. 또 총 T회 관측할 때 k번 사건이 일어나는 횟수를 $m(k)$라고 하자. 그러면 최소셀의 경우와 마찬가지로 $m(0) + m(1) + \cdots + m(b) = T$를 만족한다. 이제 $m(k)$에 라플라스 잡음을 더해서 $\widetilde{m}(k)$을 얻으면, 다음과 같이 큰 셀에서의 확률이 계산된다.

$$\tilde{p}(k) = \frac{\widetilde{m}(k)}{\sum_{k=0}^{b} \widetilde{m}(k)} \quad .$$

이는 큰 셀에서 관측된 빈도값들에 잡음을 추가하고 나서, 이를 이용해 사건 발생 확률 추정값을 얻는 것이다. 따라서 기존의 단순 적용 방식과 유사하다. 다만 Jiang 외(2017)에서는 이 확률추정값을 이용해 큰 셀에 해당하는 빈도를 다시 생성할 것을 제안하고 있다.[22]

차등정보보호가 적용된 큰 셀의 빈도를 얻는 세 방법을 비교하기 위해, Jiang 외(2017)은 9월의 동일한 날짜 및 시간대의 자료를 사용했다. 9월 자료에서는 $T = 630$이다.

---

21) 앞에서 설명한 $\tilde{p}_{ij}$을 의미한다.

22) Jiang 외(2017)에서 생성 방법을 명확히 제시하지는 않았다. 아마도 교란된 확률을 가지는 다항(multinomial) 분포를 이용해 빈도를 생성할 것으로 생각된다.

큰 셀에 대해 8월의 자료를 이용해 차등정보보호 자료를 100번 생성해 순서대로 세로축에 정렬하고, 9월 자료에서 100개의 분위값을 추출해 가로축에 나타낸다. 자료 생성 방법이 좋을수록, 즉 차등정보보호를 만족시키는 알고리즘이 좋을수록 100개의 값들이 서로 일치해서 $y = x$에 나타날 것이라고 간주한다. 일치하지 않아 $y = x$에서 벗어나는 정도로 정보손실을 정의한다. 논문은 이용자가 분석을 요청한 큰 셀들에 대해, 차등정보보호를 단순히 적용하는 방법 ③에 비해서 통계적 모형을 활용해 자료를 생성하는 방법 ① 및 ②가 좀 더 정보손실이 적은 경향을 나타냄을 보이고 있다.

결론적으로 자료에 직접 잡음을 더하기만 해서 제공하기보다는, 적절한 통계적 모형을 이용해 자료를 여러 단계로 설명하고, 특정 단계에서 잡음을 추가해서 차등정보보호를 만족시키는 것이 최종 결과에서 정보손실이 작아짐을 확인할 수 있다. 특히 빈도표의 경우, 잡음을 바로 더하는 경우 정보손실이 너무 커서 자료가 쓸모없어지는 경향이 있다는 것이 잘 알려져 있다(Duchi 외, 2018). 실제 자료제공에서는 정보손실이 중요한 문제이므로 차등정보보호를 만족시키는 알고리즘을 정교하게 만들 필요가 있다.

참고로 라플라스, 이중지수 분포 등을 이용해 잡음을 추가하는 R패키지로 “diffpriv” 등이 있으니 차등정보보호 자료를 직접 만들어 보는데 활용할 수 있다.

3. 차등정보보호를 만족하는 재현자료 생성

재현자료(synthetic data)는 원자료의 분포를 따르지만 원자료는 포함하지 않는 것을 목표로 시작되었다. 재현자료를 생성할 때는 주로 베이지안 다중대체 기법을 이용한다. 목적대로 된다면 노출위험이 없는[23] 하나의 재현자료로 다양한 분석을 수행할 수 있으므로, 정보손실을 줄이는 재현자료 생성 모형에 관한 연구에만 집중하면 된다. 실제로 미국 센서스국에서는 원자료는 전혀 제공하지 않고 재현자료만을 제공하는 사례도 있다.[24] 그러나 정말로 재현자료에 노출위험이 없다고 할 수 있는지에 대하여 엄밀한 논의가 이루어진 적은 거의 없었다고 볼 수 있다. 재현자료로 얻는 추정량의 분산의 정확한 추정 방법 등, 재현자료의 유용성에 대한 논의가 활발했을 뿐이다. 그러므로 차등정보보호가 제안되면서 연구자들이 재현자료에 대한 노출위험을 엄밀하게 설명하고자 하는 것은 자연스러워 보인다. 그러한 맥락에서 재현자료가 차등정보보호를 만족해서 안전하다고 할 수 있는지에 대해 의문을 제기하고 답한 논문이 있어 여기에 소개한다.

---

23) 완전 재현자료에는 개념상 원자료가 없으므로 노출위험이 0이라고 쉽게 단정할 수도 있지만, 재현자료의 노출위험 측정은 아직 좀 더 연구되어야할 주제이다(Ruiz 외, 2018).

24) 연구자의 요청이 있는 경우, 연구자의 연구 방법을 따라 재현자료를 분석한 결과와 원자료를 분석한 결과의 차이, 예를 들면 추정한 통계값의 차이를 알려 주기도 한다.

Machanavajjhala 외(2008)는 재현자료를 생성하는 기법을 차등정보보호 자료를 만들어내는 여러 알고리즘 중의 하나로 취급한다. 즉, 재현자료 생성 기법을 마치 차등정보보호 자료를 생성하는 라플라스 메커니즘과 같이 생각하고, 재현자료 생성 알고리즘이 정말 차등정보보호를 만족하는지 질문했다. 논문에 따르면 베이지안 방식으로 사후예측분포에서 생성된 재현자료는 ϵ-차등정보보호 및 (ϵ,δ)-차등정보보호 모두를 만족시키지 않는다. 논문은 두 차등정보보호를 만족시키기 위해서는 재현자료의 정보손실이 커질 수밖에 없음을 논하고, (ϵ,δ)-확률적 차등정보보호(probabilistic differential privacy)라는 새로운 정의를 제안했다. 나아가 제안한 (ϵ,δ)-확률적 차등정보보호를 만족시키는 재현자료 생성 알고리즘을 제시하고, 정보손실을 추가로 더 낮출 수 있는 아이디어들을 제안했다.

논문의 내용 일부를 자세히 살펴보자. 논문은 미국 OnTheMap[25]에서 제공하는 LEHD (Longitudinal Employer-Houshlod Dynamics Program) 자료를 이용했다. 이 데이터는 센서스구(census block)라는 작은 지역 단위마다 근로자들의 출퇴근 정보로 구성되어 있다. 즉, 각 개인별로 거주하는 센서스구(origin)와 출근하는 센서스구(destination)가 나열된 마이크로데이터가 있고, OnTheMap에서는 이 데이터를 가공해서 각 센서스구마다 출퇴근에 관한 통계를 제공한다. 미국 전역에 분석 대상이 되는 센서스구는 총 $B=8,206,261$개이다. 이때 차등정보보호를 만족시켜 노출을 제어해야 하는 자료는 각 센서스구마다 작성할 수 있는 거주지에 관한 빈도표가 된다. 다음 <표 5-7>은 OnTheMap이 사용하는 빈도표를 보여준다.

〈표 5-7〉 전국 거주지별 목적지 d로 출퇴근하는 인구수

	목적지 d				
거주지	1	2	…	K	합
빈도	n_1	n_2	…	n_K	$n(d)$

전국의 센서스구마다 목적지 d로 출퇴근하는 사람들의 자료를 빈도표로 만들면 대부분의 빈도값들은 0일 것이다. 예를 들어 목적지 센서스구가 서부의 샌디애고라면, 출퇴근자의 거주지 센서스구가 동부의 뉴욕이기는 어렵기 때문이다. 따라서 어느 목적지 센서스구 d에 대해서 총 k개(k≤K)의 고려 대상이 되는 거주지들이 존재하고 이 빈도표를 OnTheMap에서 사용한다고 생각하자. 다음 <표 5-8>은 논문이 다루는 빈도표를 보여주며, 여기서 o_i는 i번째 거주지 센서스구에 해당한다.

25) 이용자가 선택한 통계를 지도 위에 제공하는 미국 센서스국의 홈페이지

〈표 5-8〉 목적지 d에 대해, 거주지별 출퇴근하는 인구수

센서스구	o_1	o_2	$\cdots$	o_k	d
빈도값	n_1	n_2	$\cdots$	n_k	n

한편 OnTheMap에서는 개인정보 노출제한을 위해 이미 재현자료를 제공하고 있다. 논문은 OnTheMap이 제공하는 재현자료를 원자료라고 간주하고, 공식자료인 재현자료에서 새로운 재현자료를 다시 만들고 비교했다. 이를 통해 새로운 재현자료가 차등정보보호를 만족하는지 연구했다. 본 보고서에서도 공식자료인 재현자료를 원자료라고 말하도록 하겠다. 이제 원자료에서 재현자료를 생성하는 과정을 살펴보자. 데이터 $\vec{n}=(n_1,n_2,\ldots,n_k)$에 대한 재현자료를 $\vec{m}=(m_1,m_2,\ldots,m_k)$이라 하자. 우선, <표 5-7>과 같은 빈도표는 다항분포를 따른다고 가정한다. 다항분포의 정의는 다음과 같다.

[정의 5] $p_1+\cdots+p_k=1$을 만족하며 음이 아닌 값으로 구성된 확률 벡터 $\vec{p}=(p_1,p_2,\ldots,p_k)$에 대해, 크기 m인 벡터가 다항분포 $M(\vec{p},m)$을 따른다는 것은 다음과 같은 확률밀도함수를 가짐을 의미한다.

$$P(m_1,\ldots,m_k)=\frac{m!}{\prod_{i=1}^{k}m_i!}\prod_{i=1}^{k}p_i^{m_i}. \blacksquare$$

이제 재현자료를 생성하기 위해 사전분포로 디리슈렛(Dirichlet) 분포를 사용한다. 디리슈렛 분포의 정의는 다음과 같다.

[정의 6] $|\alpha|=\alpha_1+\cdots+\alpha_k$를 만족하며 양수로 구성된 벡터 $\vec{\alpha}=(\alpha_1,\alpha_2,\ldots,\alpha_k)$에 대해, $p_1+\cdots+p_k=1$을 만족하며 음이 아닌 값으로 구성된 벡터 $\vec{p}=(p_1,p_2,\ldots,p_k)$가 디리슈렛 분포 $D(\vec{\alpha})$를 따른다는 것은 다음과 같은 확률밀도함수를 가짐을 의미한다.

$$P(p_1,\ldots,p_k|\vec{\alpha})=\frac{\Gamma(|\alpha|)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}p_i^{\alpha_i-1}.$$

이때 벡터 $\vec{\alpha}=(\alpha_1,\alpha_2,\ldots,\alpha_k)$는 사전 샘플(prior sample)을, $|\alpha|$는 사전 샘플의 크기를 나타내며, 벡터 $(\alpha_1/|\alpha|,\alpha_2/|\alpha|,\ldots,\alpha_k/|\alpha|)$는 사전분포의 모양(shape)이 된다. ■

데이터의 정보 없이 사전분포만을 고려할 경우, 재현자료 생성과정은 다음과 같은 두 단계로 정리할 수 있다.

① $D(\vec{\alpha})$에서 벡터 $\vec{p}=(p_1,p_2,...,p_k)$를 생성한다.
② $M(\vec{p},m)$에서 크기 n인 벡터를 생성한다.

이제 관측한 데이터의 정보를 반영해서 사후분포를 얻어 재현자료를 생성하자. 먼저 잘 알려진 다음 정리를 생각한다.

[정리 10] 자료 벡터 $\vec{n}=(n_1,n_2,...,n_k)$가 다항분포 $M(\vec{p},m)$을 따르고 디리슈렛 분포 $D(\vec{\alpha})$를 사전분포로 가진다면, 사후분포 $P(\vec{p}|\vec{\alpha})$는 $D(\vec{n}+\vec{\alpha})$가 된다. ■

[정리 10]을 따라 재현자료 생성과정은 다음과 같이 간단히 정리된다. 관측한 데이터를 $\vec{n}=(n_1,n_2,...,n_k)$이라고 할 때,

① 적절한 사전 샘플 $\vec{\alpha}=(\alpha_1,\alpha_2,...,\alpha_k)$를 선택한다.
② 재현자료의 크기 m을 선택한다.
③ $D(\vec{n}+\vec{\alpha})$를 사후분포로 사용해서 모수 $\vec{p}$를 생성하고, 이 모수를 가지는 다항분포에서 크기 m인 자료를 생성한다.

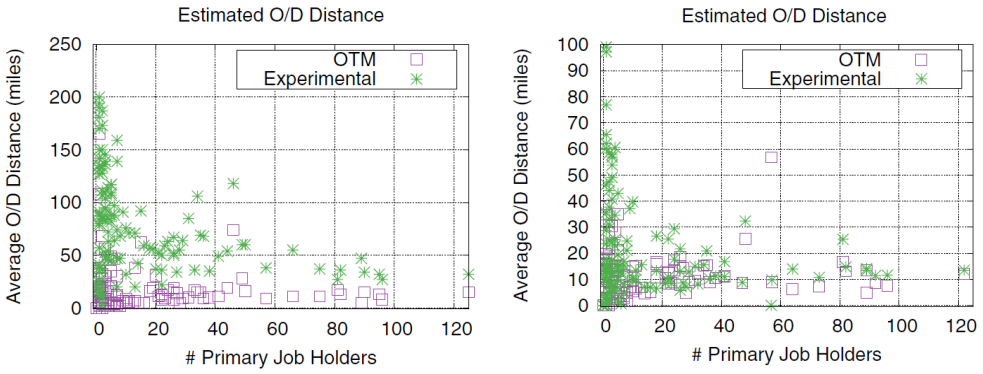
Machanavajjhala 외(2008)는 이렇게 생성한 재현자료 $\vec{m}=(m_1,m_2,...,m_k)$가 기존의 차등정보보호를 만족하지 못한다는 것을 증명하고, 새로운 확률적 차등정보보호, (ϵ,δ)-확률적 차등정보보호를 제안했다. 그리고 (ϵ,δ)-확률적 차등정보보호를 만족시키기 위해, 위의 재현자료 생성 알고리즘에 수용/거절(accept/reject) 단계를 추가하여 (ϵ,δ)-확률적 차등정보보호를 만족시키지 못하는 자료의 생성을 제한했다. 즉, 논문이 제안하는 통계량들을 사용해서 수용/거절 단계를 추가한 재현자료 생성 알고리즘[26]을 사용하면, (ϵ,δ)-확률적 차등정보보호를 만족하는 재현자료가 생성된다.

한편 정보손실을 축소하기 위해 Machanavajjhala 외(2008)는 CTPP[27] 자료를 이용해 센서스구를 적절히 묶어주는(shrinking the domain) 방안도 제시했다. 관련된 이론적인 내용을 더 구체적으로 설명하는 것은 활용사례를 주제로 하는 본 절에서 다루기는 적합하지 않으므로 생략하도록 한다. 다만, 이 논문의 내용처럼 확률적 차등정보보호를 만족

26) 확률적 차등정보보호를 만족시키는 알고리즘의 내용 및 관련 이론이 너무 복잡하여, 차등정보보호 활용을 이해하기 위한 4절의 흐름에 어울리지 않아 생략한다.
27) Census Transportation Planning Package

시키는 재현자료를 생성하는 알고리즘을 제안하고, 추가로 정보손실을 감소시키는 일련의 아이디어들을 사용해서 알고리즘을 개선하는 연구가 향후 자료제공과 관련해서 나아가야할 방향이라고 생각된다.

다음 [그림 5-14]는 LEHD자료를 이용해 OnTheMap이 제공하는 센서스구별 평균 통근 거리 일부를 보여준다. 가로축은 센서스구의 15개 주요 직군의 평균 종사자수(averaging 15 primary job holders)를 나타내고, 세로축은 평균 통근거리를 표시하고 있다. 총 120개 목적지 센서스구에 대한 결과를 나타냈다. 네모는 원자료 분석 결과, 별모양은 재현자료 분석결과를 나타낸다. 왼쪽의 그림은 (ϵ,δ)-확률적 차등정보보호를 만족시키는 재현자료에 대한 결과이다. 원래 통근거리에 비해 과대추정 경향을 보임을 알 수 있다. 오른쪽의 그림은 정보손실을 줄이기 위해 센서스구를 재편성하는 단계를 추가해서 생성한 재현자료이다. 왼쪽 그림에 비해 정보손실이 적은 것을, 즉 자료의 정확도가 더 높은 것을 확인할 수 있다.



[그림 5-14] 평균 출퇴근 거리에 대한 원자료(네모)와 차등정보보호 자료(별)

이제 OnTheMap에서 자료를 제공하는 방식을 정리하면 다음과 같다.

① (ϵ,δ)-확률적 차등정보보호를 만족하는 전년도 자료를 바탕으로 사전분포를 형성한다. 이는 전년도 자료를 $\vec{\alpha}=(\alpha_1,\alpha_2,...,\alpha_k)$로 사용한다는 것을 의미한다.

② 위의 전년도 자료와 차등정보보호를 만족하지 않는 올해 자료를 사후분포를 형성하는데 사용한다. 이는 올해 자료 $\vec{n}=(n_1,n_2,...,n_k)$와 ①의 사전분포로부터 사후분포 $D(\vec{n}+\vec{\alpha})$를 얻는다는 것을 의미한다.

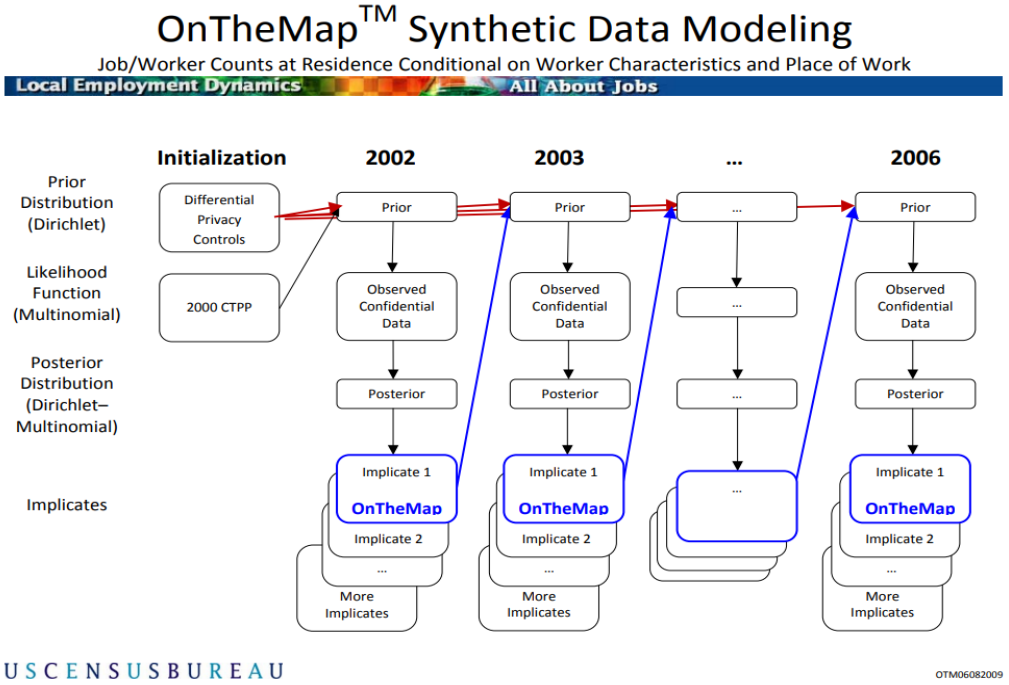
③ 앞에서 설명한 재현자료 생성과정을 따라 재현자료를 얻는다.

참고로 4.1절의 내용과 같은 방식을 사용한다면 올해 자료 $\vec{n}=(n_1,n_2,...,n_k)$에 대해 단순히 라프라스 잡음을 첨가한 빈도값들을 사용하게 될 것이다. 이는 차등정보보호를 만족할지는 모르지만, 자료의 유용성이 너무나 많이 떨어지게 된다.

OnTheMap의 방식은 새로 제안한 확률적 차등정보보호를 만족시킬 뿐만 아니라 정보손실을 줄이고 자료의 유용성을 개선하는 알고리즘이다. 또한 전년도 자료는 (ϵ,δ)-확률적 차등정보보호로 보호하고, 올해 자료는 추정된 사후분포에서 재현자료를 생성해서 보호한다. 이 자료는 완전재현자료(fully synthetic data)가 되므로 차등정보보호를 논할 여지가 없다. 그래서 매뉴얼에는 다음과 같은 문구가 포함되어 있다.

> "there are no actual worker's data involved in the tabulation of statistical results by home and work location pairs for OnTheMap"
> "No actual worker data are released in the synthetic data modeling approach. Therefore, re-identification, if it occurs, applies only to a synthetic worker, not an actual worker. A synthetic worker has no actual personal information to reveal or to re-identify."

이상의 자료 생성 방식은 OnTheMap의 홈페이지에 [그림 5-15]와 같이 정리되어 있다.



[그림 5-15] OnTheMap의 통계적 비밀보호 모형

제5절 맺음말

안전하고 유용하며 편리한 데이터 활용을 위한 통계적 비밀보호 기술을 확보하는 것은 현 시점에서 절실한 작업이다. 전통적인 매스킹 기법의 한계를 벗어나기 위해 재현자료(synthetic data)와 차등정보보호(differential privacy)가 제시되었지만, 이들을 바르게 이해하고 적절히 활용하는 데는 상당한 전문 지식이 필요하다. 게다가 두 접근 방식은 더 연구되고 발전되어야 할 영역이 많다. 또한 비밀보호 이론과 알고리즘이 전문적으로 발전하는 동시에, 비밀보호 확보를 위한 구체적인 기술들은 실무적으로 쉽게 사용 가능한 형태로 보급되어야 한다. 현재 국내 비밀보호 분야 연구자 부족 현상까지 고려할 때, 새로운 기술의 발전을 촉진하기 위한 노력이 필요하다. 본 연구는 국내에 데이터 활용을 위한 통계적 비밀보호 방법의 발전을 촉진하기 위해 기획되었다. 차등정보보호에 대해서 개념, 이론, 활용 사례를 소개한 본 연구가 더 많은 연구자들로 차등정보보호 연구를 시작할 수 있게 하는 토대가 되기를 기대한다.

본 보고서의 서론에서는 데이터를 통계적으로 비밀보호(keep confidentiality) 처리하기 위한 각 방법들을 위험-유용성 관점에서 비교했다. 위험-유용성 관점에서 노출위험과 정보손실이 최소화되도록 처리된 데이터(sanitized data)가 바람직하게 비밀보호 처리된 데이터라고 할 수 있다. 그러나 기존의 매스킹 기법만으로는 상충관계에 놓인 노출위험과 정보손실을 동시에 축소시키는 것은 거의 불가능했다. 패러다임을 바꾸어 매스킹 기법이 봉착했던 의사결정의 문제를 완화시킴으로써, 재현자료와 차등정보보호 관련 연구는 활발하게 진행되고 있다. 재현자료는 공표 자료에 원자료가 없는 것을 목표하며 시작되었고, 차등정보보호는 정보보호의 개념과 정의를 새롭게 제안하여 패러다임의 변화를 일으켰기 때문이다.

차등정보보호가 패러다임을 바꾸었다는 것은 노출위험을 측정하는 시점의 변화에서도 알 수 있다. 기존 노출위험 측도들이 비밀보호 기법을 적용해 이미 변형된 데이터의 노출위험을 측정하는데 집중했다면, 새로운 노출위험 정의인 차등정보보호는 차등정보보호를 만족하도록 데이터를 생성할 것을 요구한다. 즉 아예 노출위험을 일정 수준 이하로 제어해서 생성한 자료이므로, 정보손실을 더 줄여나갈 알고리즘을 개발하는 문제로 연구 역량이 집중될 수 있다.

차등정보보호와 관련하여 정보손실 측도 역시 과거의 패러다임에서 벗어나 통계적 추정량의 효율성을 논의하는 것으로 방향이 전환되었다.[28] 즉, 정보손실을 측정하기 위해, 원자료로 얻은 추정량과 차등정보보호 자료로 얻은 추정량의 효율성을 최소최대 위험(minimax risk)[29]을 이용해 측정하는 것이다. 이와 같은 정보손실 측도를 통해 도달

28) 차등정보보호는 쿼리(통계량)마다 잡음을 생성하는 머신(알고리즘)이 달라지는 것이므로 통계 추정량의 효율성을 통해 정보손실을 측정하는 것이 타당한 측면이 있다. 단순히 자료 값의 변화를 정보손실 기준으로 삼는 기존의 접근방식과 다르다.

가능한(optimal) 최소의 정보손실, 즉 최소최대위험을 구하고, 새로운 알고리즘의 효율성은 최적의(optimal) 최소최대위험을 달성하는가를 통해 평가할 수 있다. 나아가 이러한 수학적 엄밀성을 이용해 새로운 알고리즘을 연구해 볼 수도 있다.

본 연구를 통해 기대하는 향후 연구 방향은 다음과 같다. 첫째, 차등정보보호의 노출위험 기준을 나타내는 ϵ(또는 α)의 실제적인 의미에 관한 연구가 이루어지기를 바란다. 그런 의미에서 3절의 [정리 1]은 대단히 중요하다. $\alpha = 0$이라는 것은 검정 오차(testing error)가 0.5라는 것을 의미한다. 이는 귀무가설을 기각하지도 대립가설을 채택하지도 못하므로, 노출위험이 없다고 해석할 수 있다. 의미 있는 노출위험 측도로써 차등정보보호가 더 잘 이해될 수 있도록 통계학 분야에서 지속적인 뒷받침이 있기를 기대한다.

둘째, 차등정보보호를 만족시키면서 정보손실을 축소시키는 새로운 알고리즘들이 연구되기를 바란다. 다양한 통계량(쿼리)들에 대해서 효율적인 알고리즘이 제안되어서 자료제공에 쓰이기를 기대한다. 이를 위해 이론적으로 어렵더라도 최소최대위험 등에 관한 심도 깊은 연구가 필요할 것이다. 이상적으로는 자료 제공 기관이 손쉽게 사용할 수 있도록 간단한 함수나 프로그램의 형태로도 연구 결과가 제공된다면 무척 바람직할 것이다.

마지막으로 국내에서 국소차등정보보호 관련 이론과 실무 적용 연구가 더욱 발전하기를 기대한다. 이미 구글에서는 “RAPPOR”라는 알고리즘을 사용하고 있고, 애플도 국소차등정보보호를 적용해 자료를 수집하는 것으로 알려져 있다. 국가통계기관도 자료를 제공하는 방법에 대해 고민을 할 때, 국소차등정보보호를 활용할 수 있기를 바란다.

향후 발전 방향뿐만 아니라 지금 당장 차등정보보호를 통해 바뀌어 나가야 할 부분도 있다. 예를 들면 차등정보보호는 통계 자체가 아니라 통계를 제공하는 프로세스가 안전할 것을 요구한다. 당장 연구자들이 데이터센터를 이용하고 분석결과를 반출할 때부터, 안전성에 대해, 개인정보보호에 대해, 차등정보보호에 입각한 검토를 시작할 필요가 있다. 이를 위해서는 기술적인 부분뿐만 아니라, 법이나 철학에 대한 고민도 필요할 것이다. 행정적으로 데이터 표준화 등에 대한 노력도 중요하다.

2018년 5월에 유럽에서는 일반개인정보보호법(General Data Protection Regulation)이 시행되기 시작했다. 비밀보호 방법론의 입장에서 바라볼 때, GDPR의 내용 중에서 인상적인 요구는 25번의 “Data protection by design and by default”라고 하겠다. 이는 자료를 제공하는 플랫폼을 구축할 때, 개인정보보호를 위한 기술적인 조치가 플랫폼에 내재되어 있기를 요구하기 때문이다. GDPR을 준수하기 위해서는 통계적 비밀보호 방법론의 발전이 필수적임을 직감할 수 있다. 데이터의 활용과 개인정보보호를 동시에 성취하기 위해서, 또 법을 준수하기 위해서도 향후 차등정보보호 관련 연구가 발전해가기를 기대한다.

---

29) 노출위험과 동일하게 위험(risk)이라는 용어를 사용하지만 최소최대위험은 정보손실을 측정하는데 사용되는 것이므로 주의할 필요가 있다.

참고문헌

박민정(2014), “가계금융·복지조사 마이크로데이터 매스킹 방안 평가”, 통계개발원.
박민정(2015), 『마이크로데이터 매스킹 기법 실무 활용 안내서』, 통계개발원.
박민정(2016), “통계적 노출제어 최신동향 검토 - 차등정보보호를 중심으로-”, 통계개발원.
박민정, 김정연(2017), “재현자료 작성 방법론 검토”, 통계개발원.
박민정, 김정연(2019), “다중대체와 재현자료”, 응용통계연구, 발간예정.
박민정, 김항준(2016), “마이크로데이터 공표를 위한 통계적 노출제어 방법론 고찰”, 응용통계연구, 39(6), 1041-1059.
천계영(2015), 좋아하면 울리는, 다음 웹툰. http://webtoon.daum.net/webtoon/view/joalarm.

Abowd, J. M. and Schmutte, I. M.(2017), Revisiting the economics of privacy: Population statistics and confidentiality protection as public goods, CES, 17-37.
Agrawal, R. and Srikant, R.(2000), Privacy-preserving data mining, 29(2), 439-450, ACM.
Bertino E., Byun JW., Li N.(2005), Privacy-Preserving Database Systems, In: Aldini A., Gorrieri R., Martinelli F. (eds) Foundations of Security Analysis and Design III. FOSAD 2005, FOSAD 2004. Lecture Notes in Computer Science, vol. 3655, Springer, Berlin, Heidelberg.
Bittau, A., Erlingsson, U., Maniatis, P., Mironov, I., Raghunathan, A., Lie, D., Rudominer, M., Kode, U., Tinners, J. and Seefeld, B.(2017), Prochlo: Strong privacy for analytics in the crowd, In Proceedings of the 26th Symposium on Operating Systems Principles, 441-459, ACM.
Dalenius, T.(1977), Towards a methodology for statistical disclosure control, Statistik Tidskrift, 15, 429-444.
Ding, B., Kulkarni, J., & Yekhanin, S.(2017), Collecting telemetry data privately, In Advances in Neural Information Processing Systems, 3571-3580.
Duchi, J. C. Jordan, M. I. and Wainwright, M. J.(2013), Local privacy and statistical minimax rates. In Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium, 429-438. IEEE.
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.
Duncan, G., T. Elliot. M., and Slazar-Gonzalez, J.(2011), Statistical confidentiality principles and practice statistics for social and behavioral sciences, Springer, New York, NY, USA.
Drechsler, J.(2018), Some Clarifications Regarding Fully Synthetic Data, In: Domingo-Ferrer, J., Montes, F. (eds) Privacy in Statistical Databases 2018, Lecture Notes in Computer Science, vol 11126, 109-121, Springer, Cham.
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.(2008), Differential privacy: A survey of results, In International Conference on Theory and Applications of Models of Computation, 1-19, Springer, Berlin, Heidelberg.

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.

Erlingsson, Ú., Pihur, V., & Korolova, A.(2014, November), Rappor: Randomized aggregatable privacy-preserving ordinal response, In Proceedings of the 2014 ACM SIGSAC conference on computer and communications security, 1054-1067, ACM.

Franconi, L. and Polettini, S.(2004), Individual Risk Estimation in μ-Argus: A Review, In Domingo-Ferrer(ed.), Privacy in Statistical Databases, Lecture Notes in Computer Science, vol. 262–272. Springer-Verlag.

Jiang, J., Xi, B. and Kantarcioglu, M.(2017), Spatial counts under differential privacy mechanism on changing spatial scales, Computers & Security.

Kairouz, P., Oh, S., & Viswanath, P.(2014), Extremal mechanisms for local differential privacy, In Advances in neural information processing systems, 2879-2887.

Kenthapadi, K., Mishra, N. and Nissim, K. Denials.(2013), Leak Information: Simulatable Auditing, Journal of Computer and System Sciences, 79, 1322-1340.

Kifer, D., and Machanavajjhala, A.(2011), No free lunch in data privacy, In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, 193-204, ACM.

Differential Privacy Team, Apple(2017), Learning with Privacy at Scale, Apple Machine Learning Journal.

Lee, D. G.-Y.(2008), Protecting patient data confidentiality using differential privacy, Master's thesis, Oregon Health & Science University.

Lee, Y.(2013), Review on statistical methods for protecting privacy and measuring risk of disclosure when releasing information for public use, Journal of the Korean Data and Information Science Society, 24, 1029-1041.

Li, N., Li, T. and Venkatasubramanian, S.(2007), t-closeness: Privacy beyond k-anonymity and l-diversity, In ICDE 2007, IEEE 23rd International Conference on Data Engineering, 106-115.

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M.(2006, April), l-Diversity: Privacy Beyond Kappa-Anonymity, in Proceedings of the 22nd International Conference on Data Engineering (ICDE '06), IEEE.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L.(2008), Privacy: Theory Meets Practice on the Map, in Proceedings of the 24th International Conference on Data Engineering, Istanbul, Turkey, pp. 277－286.

McSherry, F., & Talwar, K.(2007, October), Mechanism design via differential privacy, In Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on (pp. 94-103). IEEE.

Narayanan, A., Huey, J. and Felten. E. W.(2015), A Precautionary Approach to Big Data Privacy, Working Paper.

Neustar.(2014), Company Blog.

https://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/ .

Nissim, K., Steinke, T., Wood, A., Altman, M., Bembenek, A., Bun, M., and Vadhan, S.(2017), Differential Privacy: A Primer for a Non-technical Audience, Working Group Privacy Tools Sharing Res. Data, Harvard Univ., Boston, MA, USA, Tech. Rep. TR-2017-03.

Ruiz N., Muralidhar, K, and Domingo-Ferrer, J.(2018), On the privacy guarantees of synthetic data: A reassessment from the maximum-knowledge attacker perspective, In: Domingo-Ferrer, J., Montes, F. (eds) Privacy in Statistical Databases 2018, Lecture Notes in Computer Science, vol. 11126, pp. 59~74. Springer, Cham.

Seleshi, B. and Asseffa, S. A Case study on differential privacy, Master's thesis, Umea University.

Sweeney, L.(1997), Guaranteeing anonymity when sharing medical data, the Datafly system. Proceedings, Journal of the American Medical Informatics Association. (AMIA). Washington, DC: Hanley and Belfus, Inc.

Sweeney, L.(2002), k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, 10, 557-570.

Vilhuber, L. and Schmutte, I. M.(2017), Proceedings from the 2016 Cornell-Census-NSF-Sloan workshop on practical privacy.

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

Warner, S. L. (1965). Randomised response: a survey technique for eliminating evasive answer bias, Journal of the American Statistical Association, 60(309), 63–69.

Zhu, T., Li, G., Zhou, W. and Yu, P.(2017), Differential Privacy and Applications, Springer.

제6장

차등정보보호 적용 실험 연구

박민정 · 권성훈 · 정재호

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

✍ 이 연구는 2019년 하반기 통계개발원 연구보고서에 수록된 자료입니다.

제1절 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 정책 수행에 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 개인정보 보호이다. 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 6-1>은 센서스 자료제공에서 이러한 딜레마를 해결하기 위한 미국의 정책적 노력을 보여주고 있다(Ruggles et al., 2018). 시대 상황에 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 6-1〉 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 전략이 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 면제국(GDPR 적정성 결정 획득)으로 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.

제2절 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 ϵ 으로 표현한다. 반면 차등정보보호를 연구한 통계학 분야의 주요 문헌에서는 α 를 사용한다. 아마도 통계학에서 주로 잡음을 표현할 때 ϵ 을 사용하고, 유의수준을 표현할 때는 α 를 사용하기 때문일 것으로 추측해 본다. 차등정보보호 모수 역시 통계학적으로 다시 정의되는 노출위험 수준을 표현하는 것이므로(Wasserman과 Zhou, 2010; Duchi et al., 2018), ϵ 이 아닌 α 를 사용한다고 생각하는 것이 차등정보보호를 이해하는데 도움이 될지도 모르겠다. 따라서 차등정보보호를 논할 때 ϵ 을 사용하는 것이 독자들에게 친숙하겠지만, 본 연구는 통계학 기반에서 제시된 이론적 설명을 따르므로 ϵ 을 대신하여 α 를 사용하겠다.

1. 노출위험과 정보손실

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

가. 매스킹 처리에서 노출위험-정보손실 관계

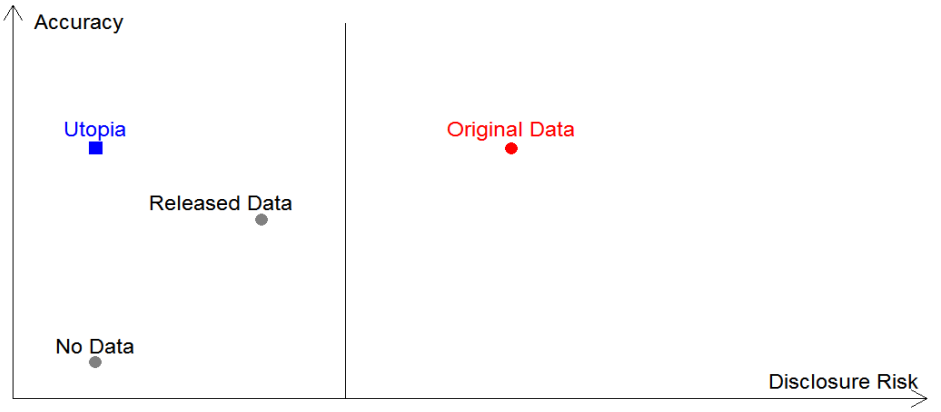
최근까지 자료를 제공할 때 주로 매스킹(masking) 기법을 이용해 개별 정보의 노출위험을 제어해 왔다. 매스킹 기법은 세부 종류가 다양하여 하나의 자료에 대해서도 수십 개 이상의 매스킹 처리 결과를 생성할 수 있다. 따라서 원래 자료를 대신할 수 있는 다양한 방안의 후보군 중에서 선택을 위한 적절한 판단 기준이 필요하다. 이러한 판단의 기준으로 사용하는 것이 노출위험과 정보손실 측도이다. 노출위험과 정보손실(자료 유용성) 사이의 관계를 종합적으로 판단하기 위해 각 측도의 계산 결과를 평면에 표현하는 것을 위험-유용성 지도(risk-utility map)라고 부른다. [그림 6-1]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.



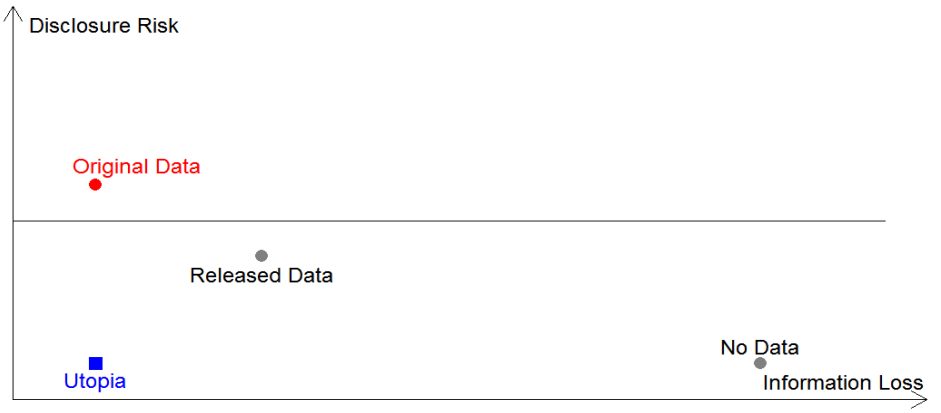
[그림 6-1] 위험-유용성 지도(노출위험-자료유용성)

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 6-2] 및 [그림 6-3]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 6-2] 위험-유용성 지도(정확성-노출위험)



[그림 6-3] 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 노출

2) 응답자(레코드)의 특성을 나타내는 변수

위험 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.

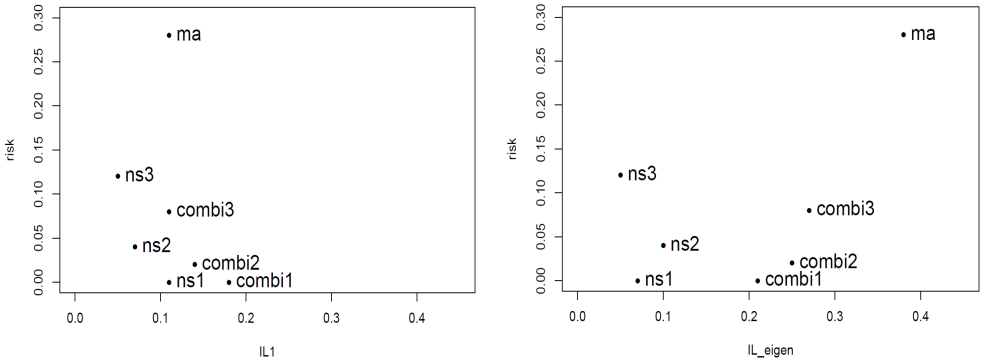
실무적으로 키변수와 (연속형인 경우가 많은) 민감변수를 모두 고려해서 노출위험을 계산하기 위해서는, 단순히 매스킹 전후의 거리를 고려하는 경우가 많다. 예를 들면, 먼저 원자료에서 레코드마다 순위에 근거해 어떤 구간을 만들고, 매스킹 처리 전후 각 레코드의 값이 그 구간 내에 있는지 없는지를 살펴본다. 다음으로 전체 레코드 중 각 구간 내에 원자료가 있는 비율을 이용해 파일 단위의 노출위험을 계산할 수 있다(Templ과 Meindl, 2008).

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 정보손실이 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 6-4]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 6-4]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 6-4] 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.

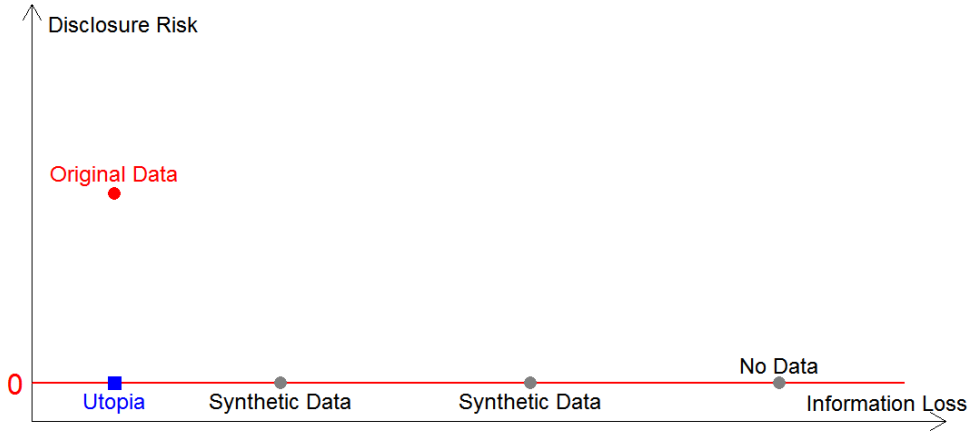
3) 베이지안 관점의 다중대체 방식으로 생성한 재현자료를 베이지안 재현자료라 부르도록 한다. 다음 소절의 <표 6-2> 용어 정리에서 관련 논의가 제시된다.

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 6-5]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 6-5]의 원점에서 먼 재현자료를 생성하게 된다.

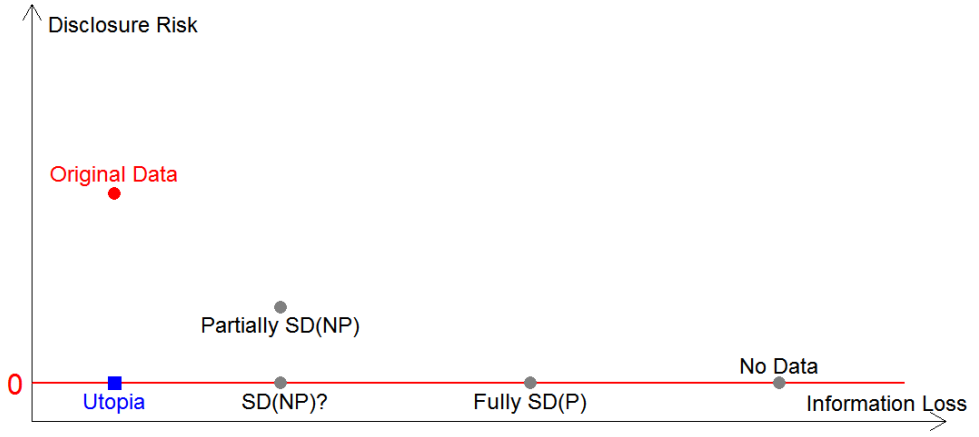
이를 개선하여 자료 유용성을 높이기 위해서는 붓스트랩과 같은 비모수적인 방법을 사용하여 완전 재현자료를 생성한다. 그러면 실제자료를 직접 사용하게 되므로, 실제자료가 재현자료에 나타날 수 있다. 따라서 완전 재현자료가 제안된 원래 의도와 다르게 노출위험이 없이 안전하다고 말하기 어렵다. 또한 충분한 자료 유용성을 확보하는 것도 쉽지 않으므로 실제자료에 완전 재현자료 생성 방식을 적용한 사례를 찾기 어렵다.



[그림 6-5] 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 6-6]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.

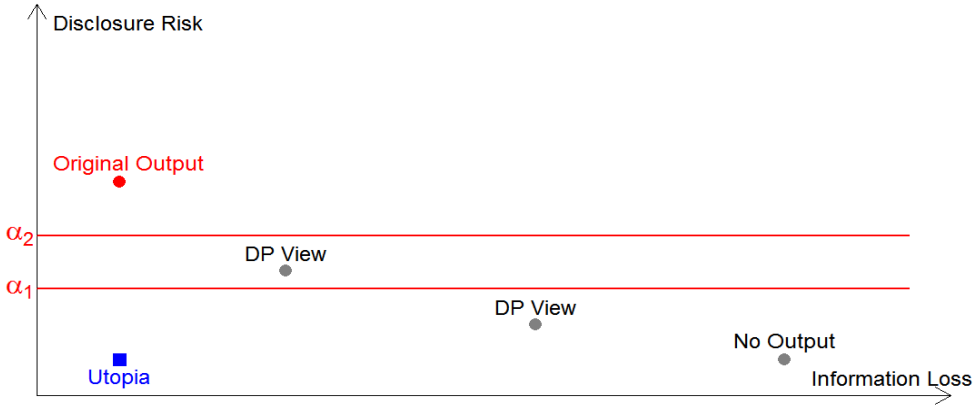


주: Fully SD(P)는 모수적으로 생성한 완전 재현자료를, Partially SD(NP)는 비모수적으로 생성한 부분 재현자료를 나타낸다.

[그림 6-6] 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 6-7] 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 이용할 때처럼 노출위험과 정보손실 사이의 상충관계가 여전하나, 차등정보보호라는 노출위험

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.

정의를 사용하므로 노출위험 측도에 수학적 엄밀함과 일관성이 있다는 장점이 있다. 차등정보보호를 만족시키도록 생성한 자료의 위험-유용성은 [그림 6-7]과 같이 표현된다. 한편, 차등정보보호를 만족시키도록 변조된 자료를 차등정보보호 자료(differentially private view, DP view)라고 부른다(Duchi et al., 2018).

지금까지 노출위험과 정보손실 측도를 살펴보고, 매스킹 기법 활용, 베이지안 재현 자료 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 6-8]은 통계학 연구자와 공학 분야 연구자가 가지는 통계학에 대한 관점의 차이를 보여준다.

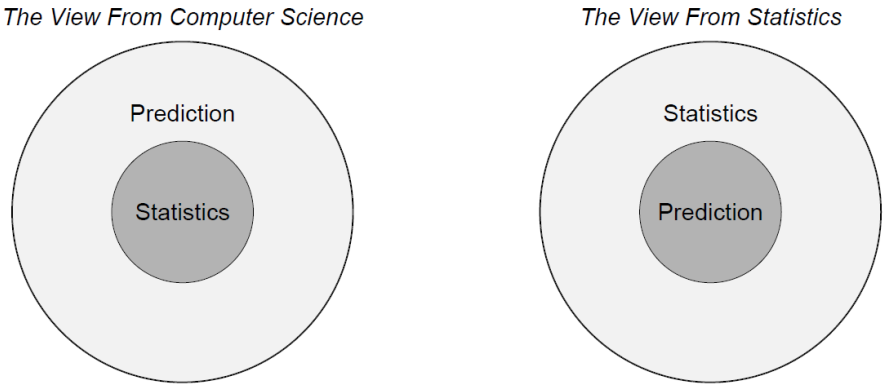


Figure 1: Statisticians view prediction (classification, regression, etc.) as a part of statistics. Some in CS see statistics as a subset of prediction.

[그림 6-8] 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 개인정보 침해가 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 6-2〉 용어 정리

데이터 정보보호 방식	기존 국내(국외) 문헌	신규 제안
매스킹 기법 적용	비식별 데이터 (masked data)	비식별 데이터 (masked data)
다중대체 기법으로 재현자료 생성	재현자료 (synthetic data)	베이지안 재현자료 (Bayesian synthetic data)
차등정보보호 적용 - 교란 방식	비식별 데이터 (sanitized database) (synthetic data)	차등정보보호 자료 (DP view)
차등정보보호 적용 - 재현 방식	차등정보보호 자료 (DP view)	차등정보보호 재현자료 (synthetic data)

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,

통계학적 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.

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

[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_ϵ에서 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. ■

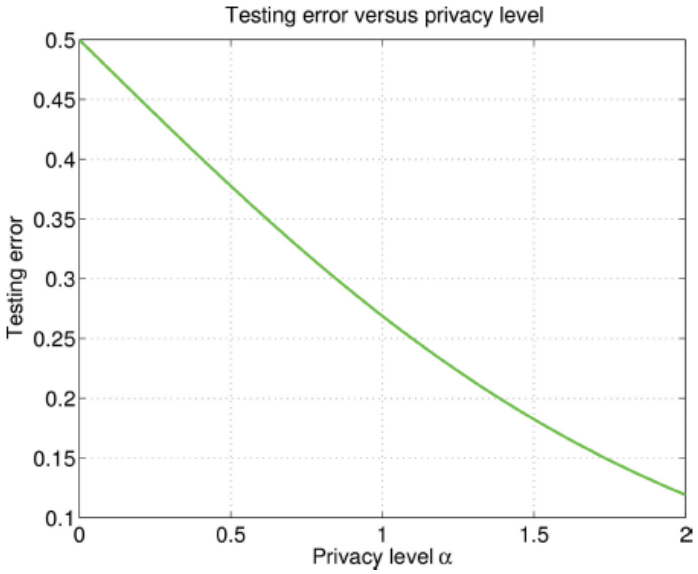
위 정리는 주어진 재현자료 Z를 사용하여 원자료에서 특정 레코드 X_i의 값을 식별(identification)할 수 있는가를 설명하고 있다. [정리 2]의 결과에 따르면 $\alpha \to 0$인 경우 $\gamma exp(\alpha) \to \gamma$가 되므로 검정법의 검정력이 유의수준과 거의 같아진다. 유의수준 γ는 0.1 이하인 경우가 대부분이므로, 결국 강한 차등정보보호($\alpha \to 0$)가 적용된 재현자료를 사용하면 특정 레코드의 값에 관하여 어떤 판단을 내리기가 어려워진다는 것을 알 수 있다.

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 6-9]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 6-9] 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)
$\theta : \boldsymbol{P} \rightarrow \Theta$	P에서 정의된 모수(parameter)
ρ	Θ 위에서 정의된 거리(metric)
$\hat{\theta}_n : \chi^n \rightarrow \Theta$	추정량(estimator)
$\hat{\theta}_n : \aleph^k \rightarrow \Theta$	재현자료 추정량(estimator)
$\Phi : R_+ \rightarrow R_+$	$\Phi(0) = 0$ 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 θ 와 $\hat{\theta}_k$ 사이의 차이(difference)라고 생각해도 좋다. $\hat{\theta}_k$ 를 평가하는 경우 손실함수를 측도로 직접 사용하지는 않으며, 다음과 같이 위험함수(risk function) R 을 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 P 와 Q 에서 발생하는 손실의 기댓값을 의미한다. 만약 위험함수가 0 으로 수렴하면 추정량 $\hat{\theta}_k$ 가 모수 θ 에 대하여 일치성(consistency)을 가진다고 말하며, 이를 고정된 분포 P 와 Q 에서 모수 θ 에 대한 추정량의 평가 측도로 사용할 수 있다.

추정량 $\hat{\theta}_k$ 와 조건부분포 Q 가 고정된 경우, 위험함수를 분포집합 $\boldsymbol{P}$ 에서 모수공간 Θ 를 거쳐 양의 실수 집합으로 가는 하나의 함수로 이해할 수 있으며, 다음과 같이 최대위험 R 을 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 Q 의 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 최소최대 위험이 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 6-10]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 6-10] 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

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.

기법을 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 $f_e(x) \propto exp(-|x-\mu|/\sigma)$[10]이고 X와 독립이라고 하자. X와 e을 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.

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

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

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 $X = (X_1, \ldots, X_n)$이 $X_i \in [0, 1]$, $i \in \{1, \ldots, n\}$이라고 하자. 임의의 k에 대하여 아래와 같은 밀도함수에서 추출한 임의표본 $Z = (Z_1, \ldots, Z_k)$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.

제3절 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

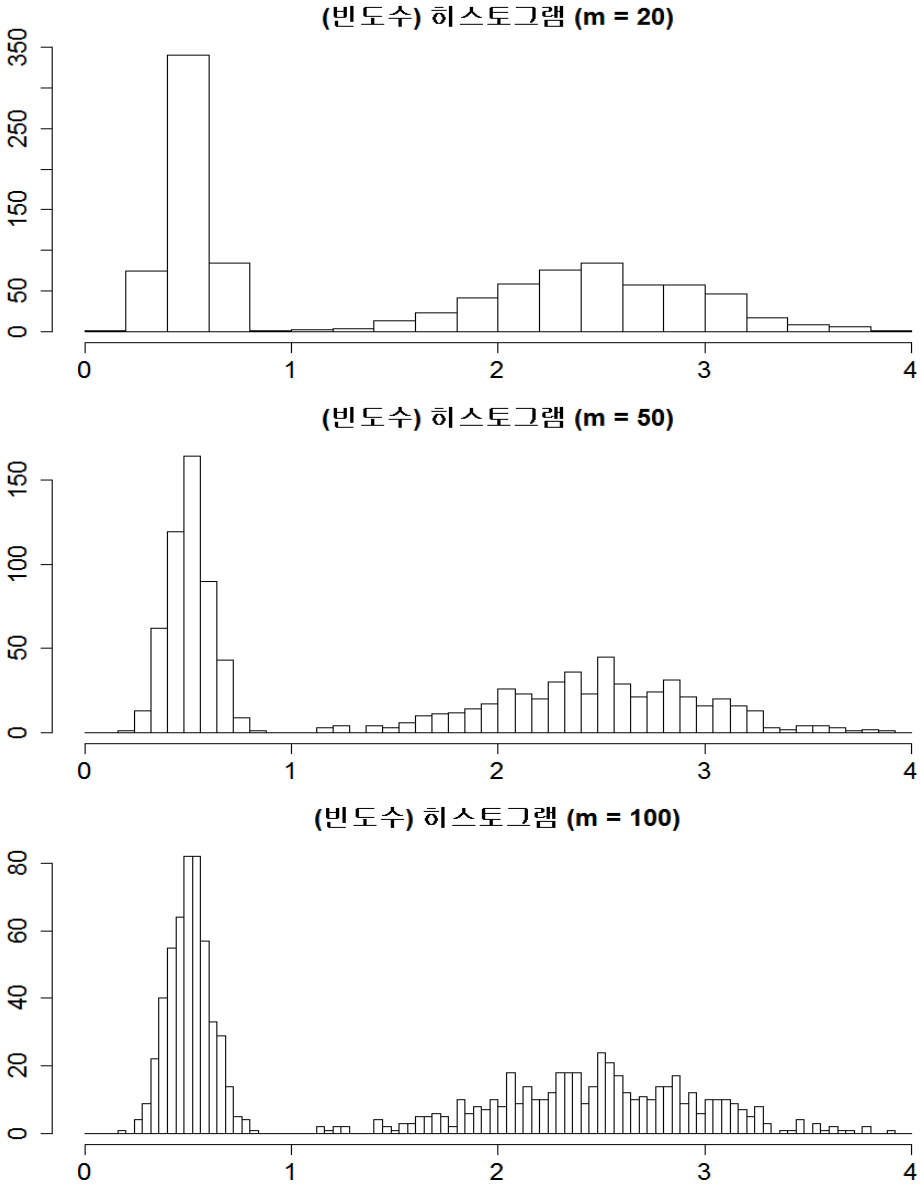
(요약) 통계학에서는 보통 동일한 크기의 막대를 사용하여 히스토그램을 작성한다. 적당한 조건을 만족시키는 히스토그램 추정량은 최소최대 추정량이며, 이는 작성된 히스토그램이 효율적이라는 것을 의미한다. 이제 차등정보보호를 만족하면서 동일한 효율성을 가지는 히스토그램을 생성하는 알고리즘을 찾는 것이 주요 논의 사항이 된다.

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.

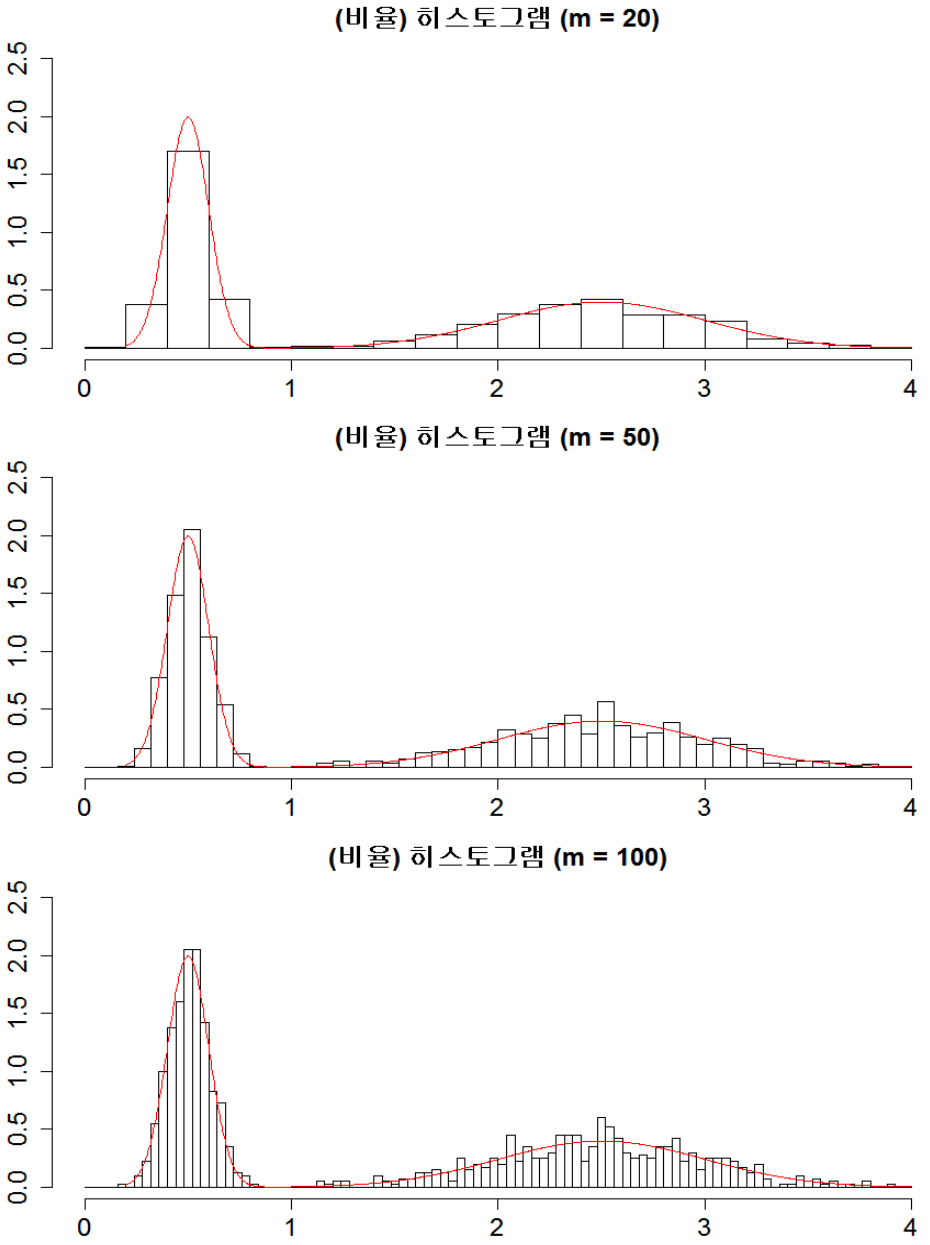
가. 히스토그램의 특징

히스토그램의 특징을 설명하기 위해 두 정규분포 $N(0.5, 0.1^2)$ 과 $N(2.5, 0.5^2)$ 의 혼합분포(normal mixture)를 고려하였다. [그림 6-11]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 6-11] 빈도수를 이용해 작성한 히스토그램: 막대 개수에 따른 변화



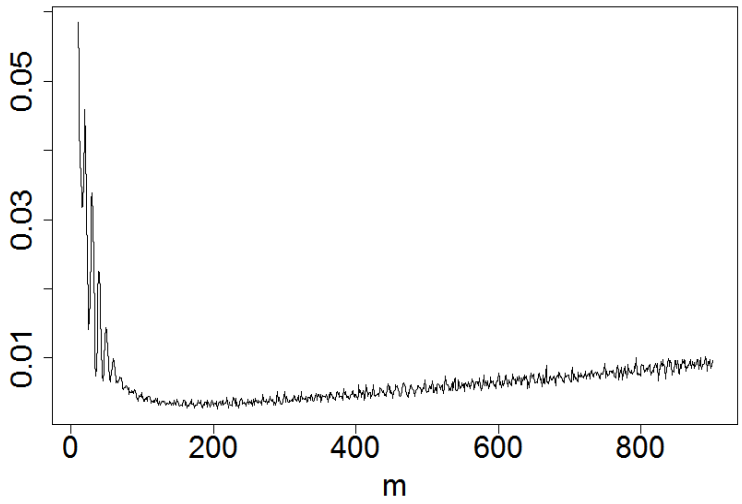
[그림 6-12] 비율을 이용해 작성한 히스토그램: 막대 개수에 따른 변화

빈도수 대신에 비율을 기준으로 히스토그램을 작성하면, [그림 6-12]와 같다. 곡선은 이 혼합분포의 밀도함수를 나타낸다. [그림 6-12]는 혼합분포의 밀도함수와 비교하기

위하여 [그림 6-11]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 6-13]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 6-13] 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 6-3>은 원자료를 사용한 히스토그램의 작성 절차이며, 본 연구에서는 이 히스토그램을 "기본 히스토그램"이라고 한다.

〈표 6-3〉 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.

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

표본의 수 n이 무한히 증가하면, 밀도함수 공간에 정의된 적당한 거리(distance)에

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.

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. ■

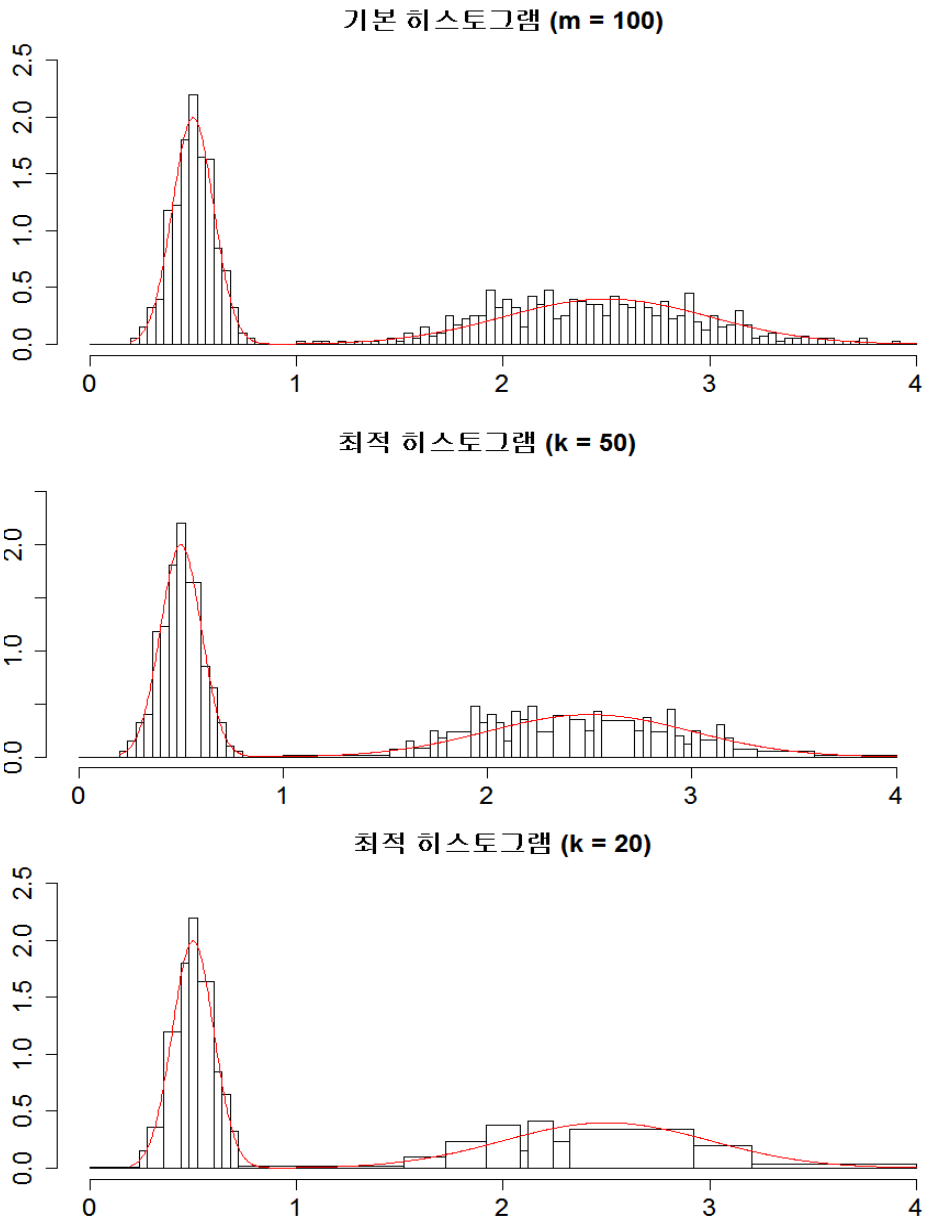
즉, L_2-거리 함수를 사용하여 최소최대 위험을 정의하는 경우 기본 히스토그램은 원자료 밀도함수의 최소최대 추정량이 된다. 따라서 기본 히스토그램은 α-차등정보보호 히스토그램을 평가하는 과정에서 하나의 비교 기준으로 사용 가능하다.

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 히스토그램의 막대를 병합하여 $k(\le m)$개의 막대를 가지는 새로운 히스토그램으로

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 1 만큼 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.

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 6-14] 기본 히스토그램 및 최적 구조 히스토그램

[그림 6-14]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 6-14]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. 이상 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 6-4〉 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 6-5〉 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 6-6〉 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 0 인 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 0 이 아닌 막대의 경우 교란된 히스토그램에서도 빈도가 0 이 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 n 인 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 6-7〉 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 6-8〉 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 6-9〉 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.

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

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

제6장

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 6-10〉 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 수준과

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

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

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.

제4절 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 : $d(f, g) = \int_R |f(x) - g(x)|\, dx$

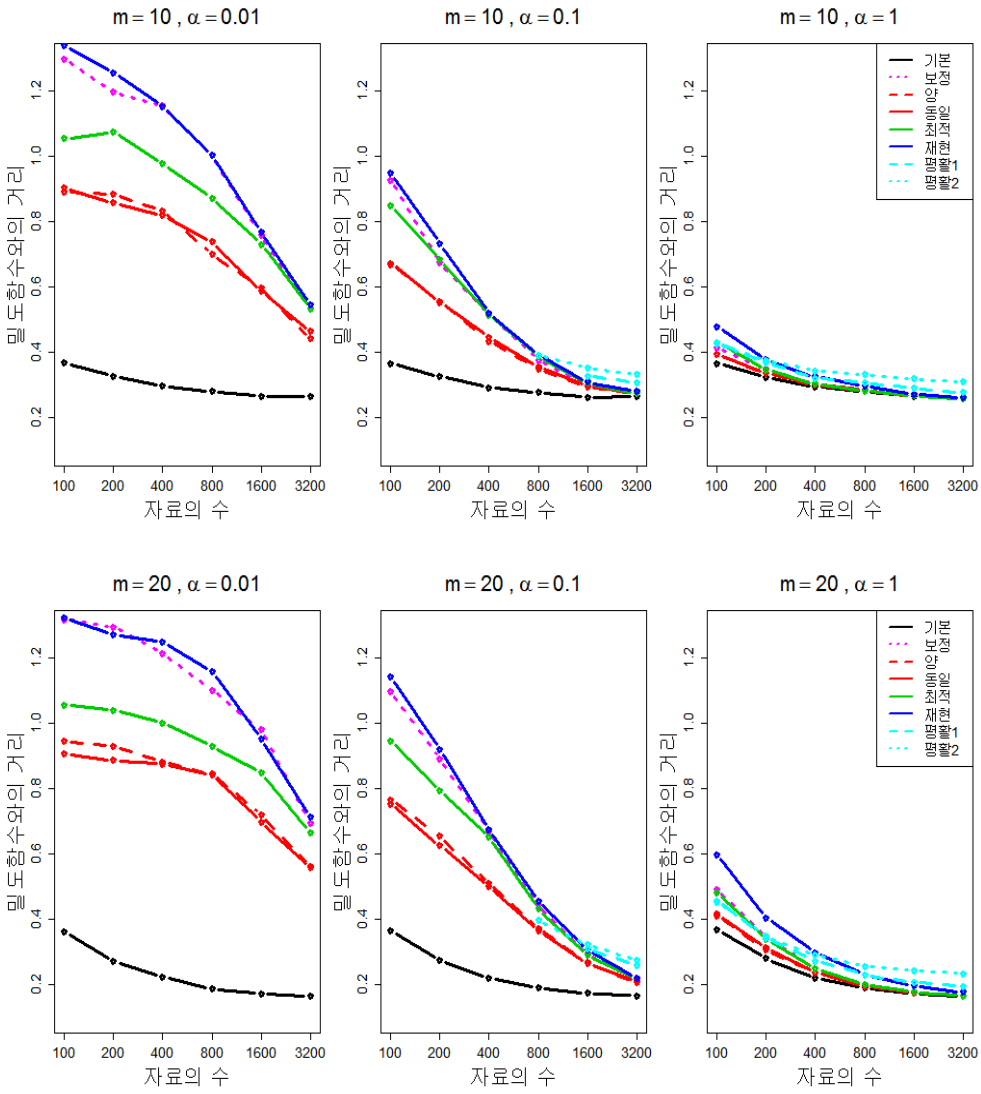
고정된 실험 조건에서 원자료 밀도함수로부터 n개의 임의표본을 생성하고, 이 임의표본을 사용하여 α-차등정보보호 히스토그램을 작성하고, 원자료 밀도함수와 기본 히스토그램에 대한 거리를 각각 계산하였다. 각각의 실험은 100회씩 반복하였으며 계산된 거리의 평균을 표와 그림으로 정리하였다. 교란 히스토그램은 밀도함수 추정량이 될 수 없는 경우가 발생하므로 실험 대상에서 제외하였다.

〈표 6-11〉 차등정보보호 히스토그램 생성 알고리즘

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

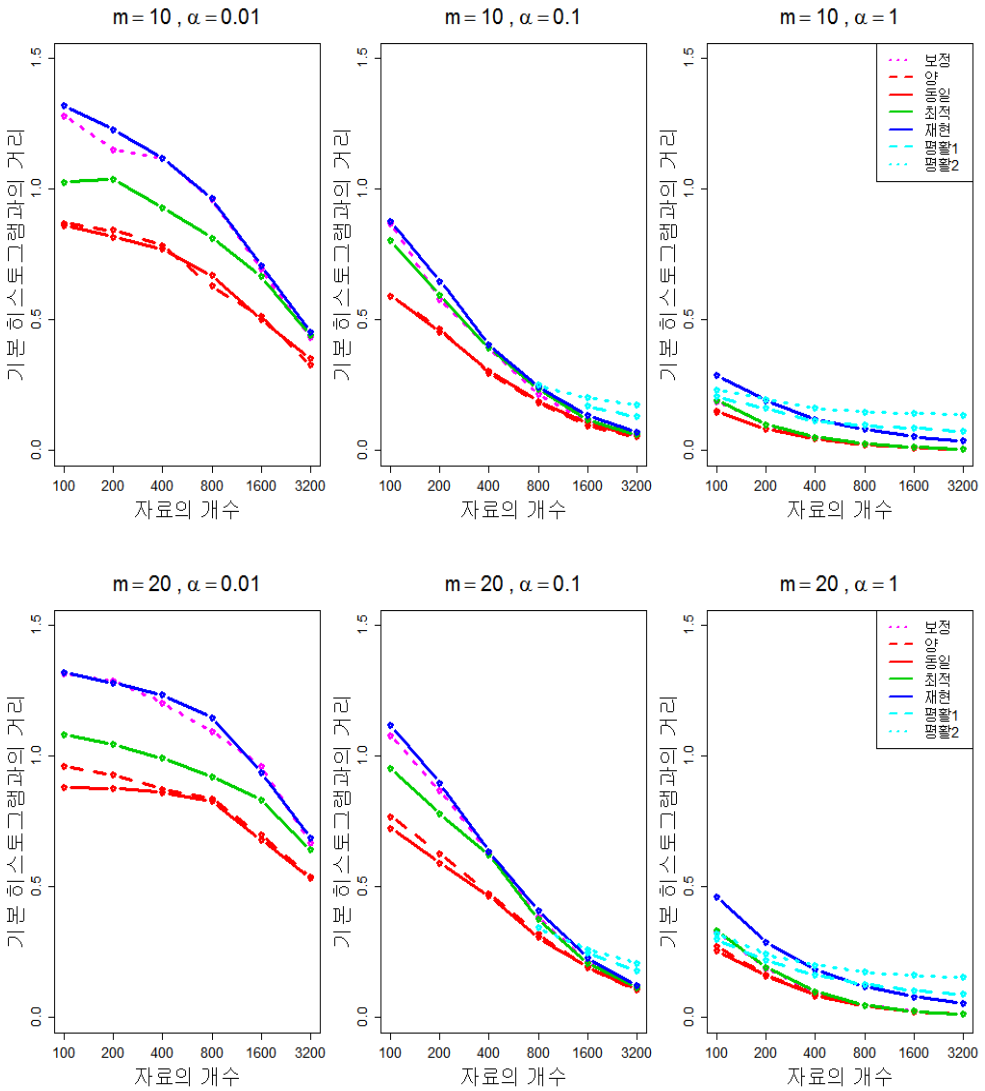
[그림 6-15]는 원자료 밀도함수와 차등정보보호 히스토그램 사이의 거리를 나타낸다. 정보보호 수준이 강할수록 (α 값이 작을수록) 정보손실이 크지만, 자료의 수가 많으면 정보손실이 작아지면서 각 방법별로 정보손실 차이가 줄어든다.

[그림 6-15]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 6-15] 원자료 밀도함수와 차등정보보호 히스토그램 사이의 거리

[그림 6-15]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 6-15]와 관련된 표는 <부록>의 <부표 6-1> 및 <부표 6-2>에 수록되어 있다.



[그림 6-16] 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 6-16]은 기본 히스토그램을 기준으로 차등정보보호 히스토그램이 가지는 정보손실을 계산한 결과이며 [그림 6-15]와 유사한 결과를 나타낸다. [그림 6-15]와 [그림 6-16]과 관련된 표는 <부록>의 <부표 6-3> 및 <부표 6-4>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.

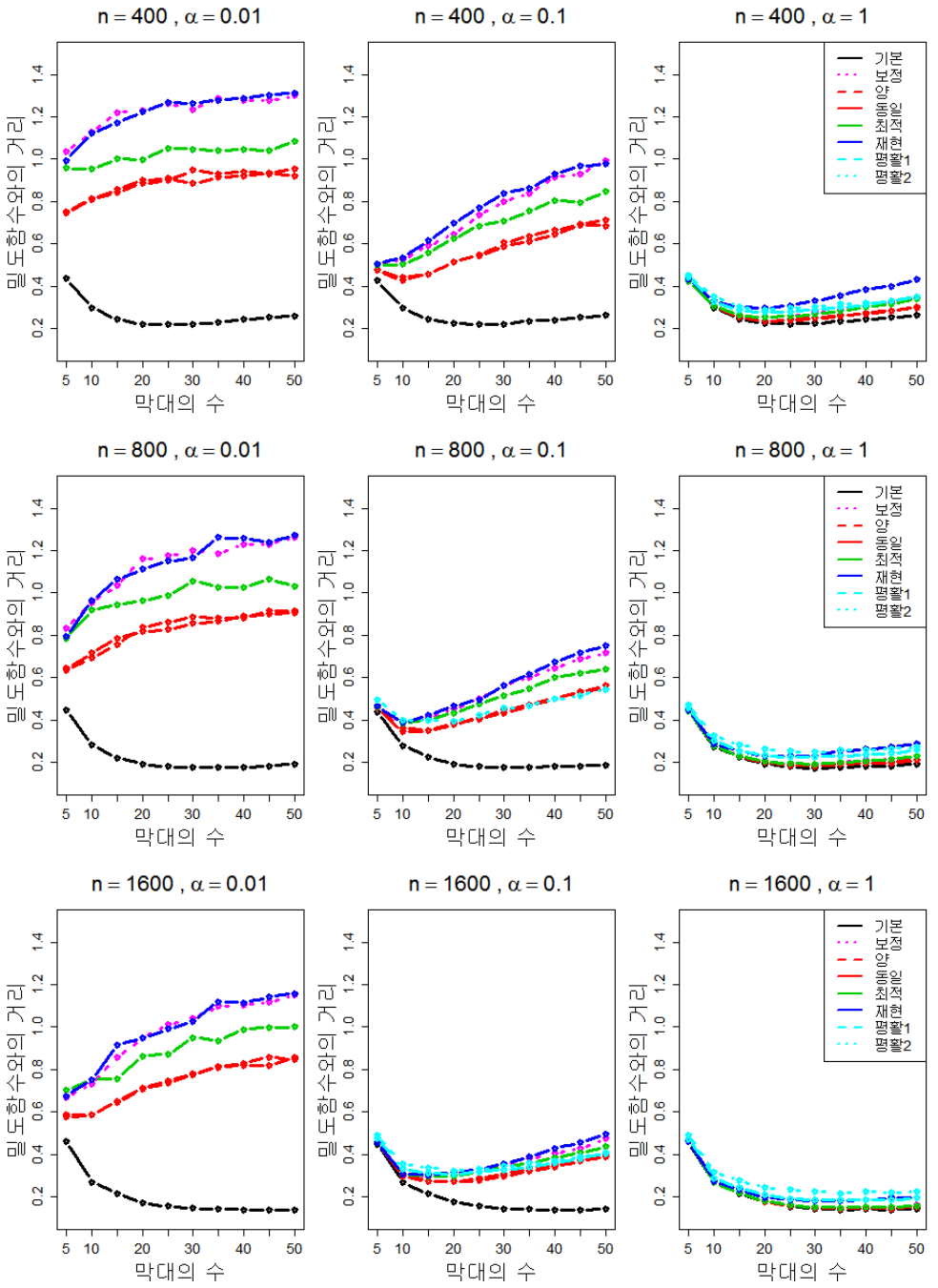
- 막대의 수 : $m = 5k,\ k \in \{1, 2, ..., 10\}$

[그림 6-17] 및 [그림 6-18]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.

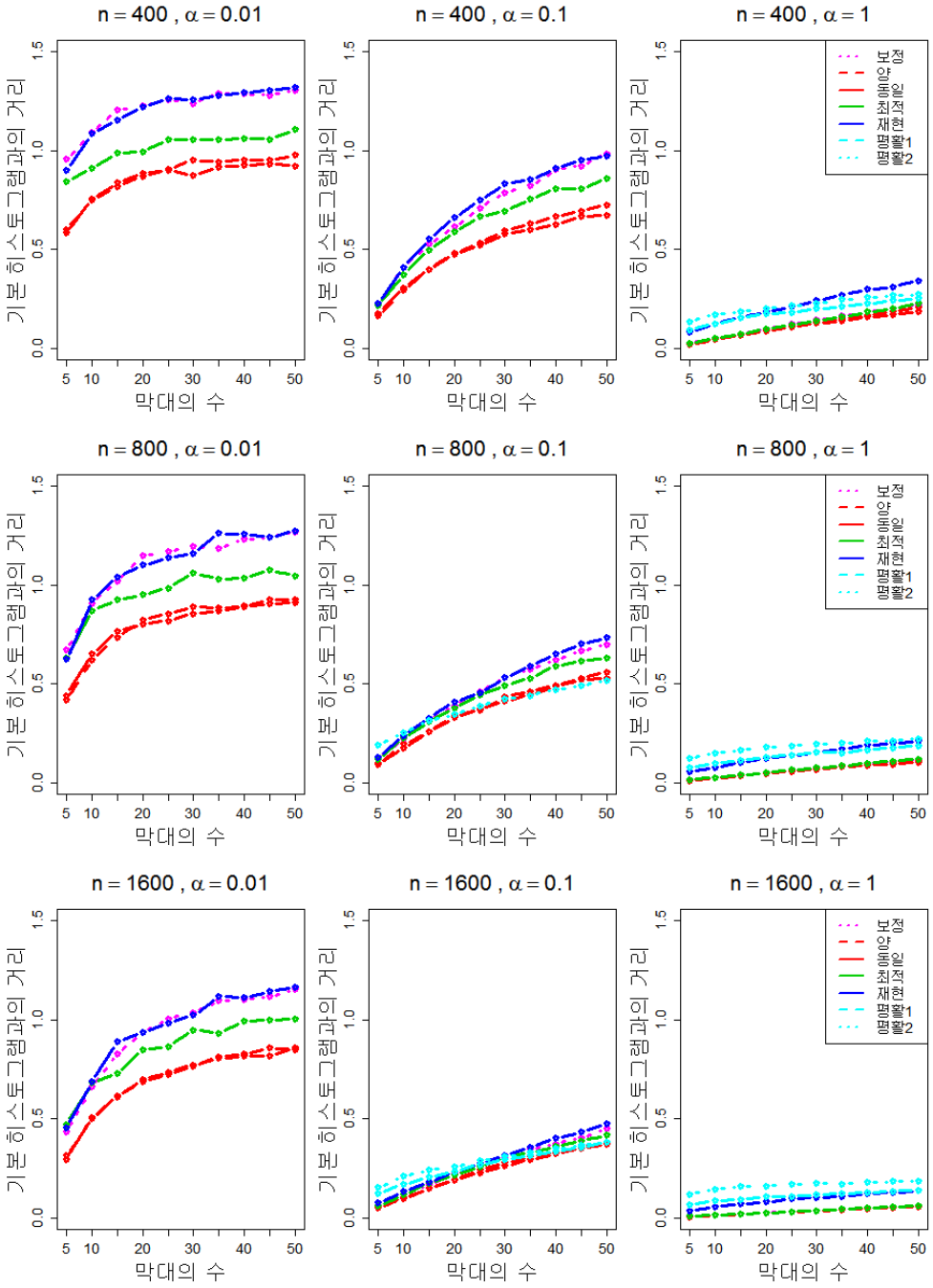
- $\alpha \leq 0.1$ 인 경우, n 이 작으면 막대의 개수가 많을수록 정보손실이 증가하는 경향을 보인다. 하지만 n 이 커지면 증가의 속도가 둔화되어 최적의 막대 개수가 발생하는 것을 확인할 수 있다.
- $\alpha \geq 1$ 인 경우, n 이 작으면 최적의 막대 개수가 발생하지만 n 이 커지면서 막대 개수가 증가함에 따라 정보손실이 계속 감소하는 것을 확인할 수 있다.

원자료의 히스토그램과 차등정보보호 히스토그램 사이의 정보손실은 [그림 6-18]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



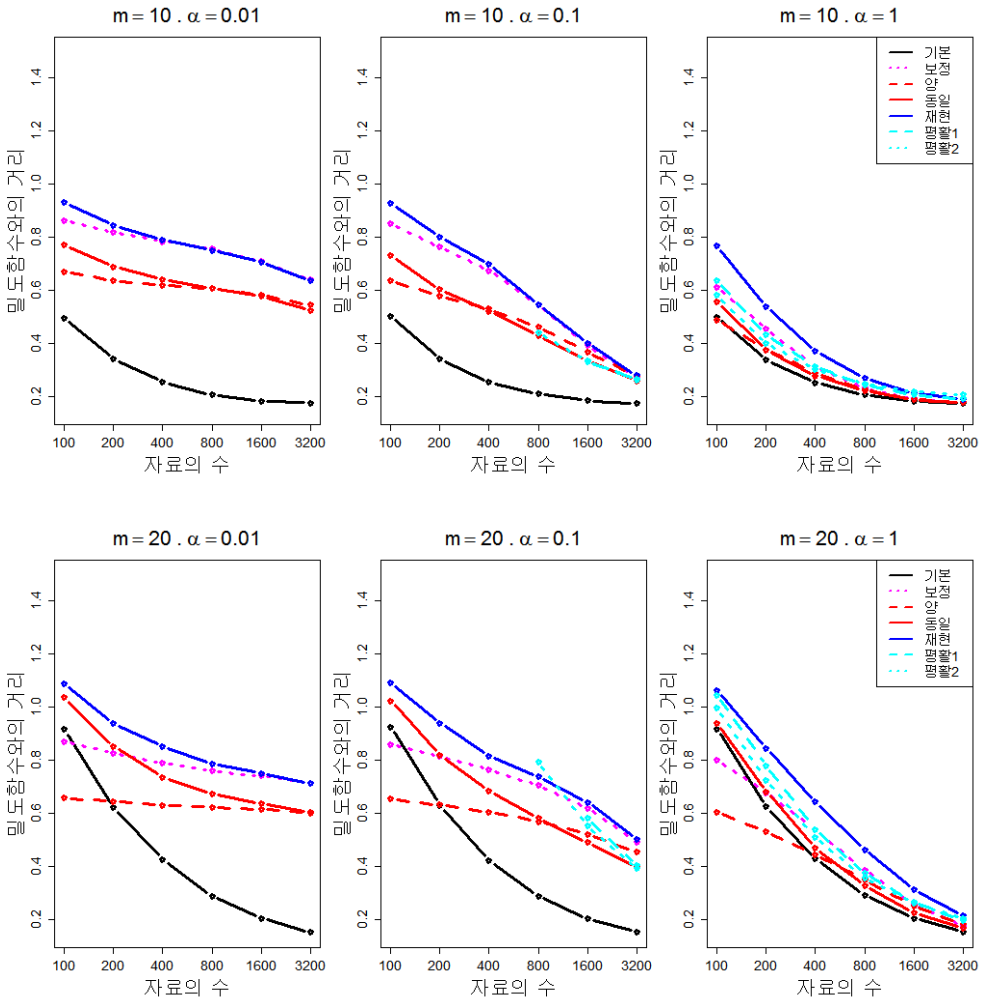
[그림 6-17] 막대 개수에 따른 원자료 밀도함수와 차등정보보호 히스토그램 사이의 거리



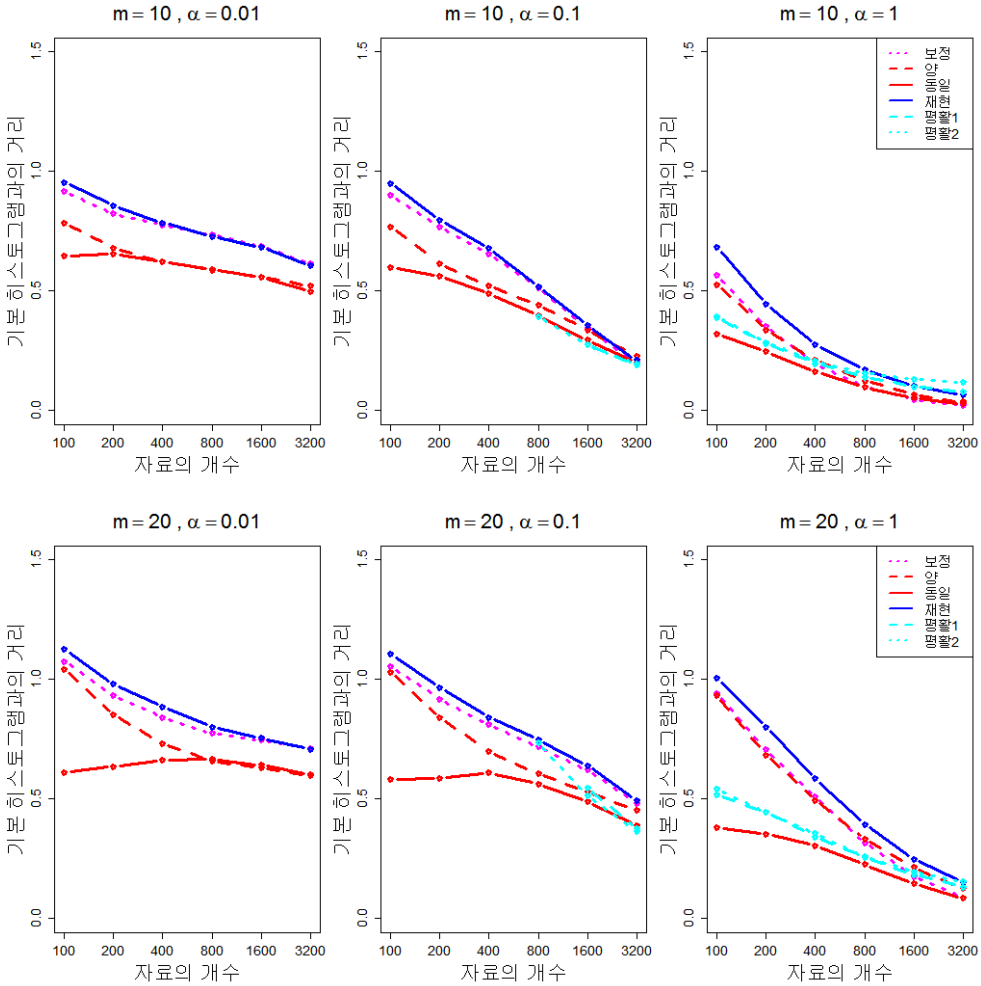
[그림 6-18] 막대 개수에 따른 기본(원자료) 히스토그램과 차등정보보호 히스토그램 사이의 거리

개수 m에 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 6-19] ~ [그림 6-22]에 도시하였다. 관련된 표는 <부록>에 수록한다. 실험의 결과는 $d=1$인 경우와 매우 유사하며 따라서 자세한 기술은 생략한다. 다만 재현 방식에서



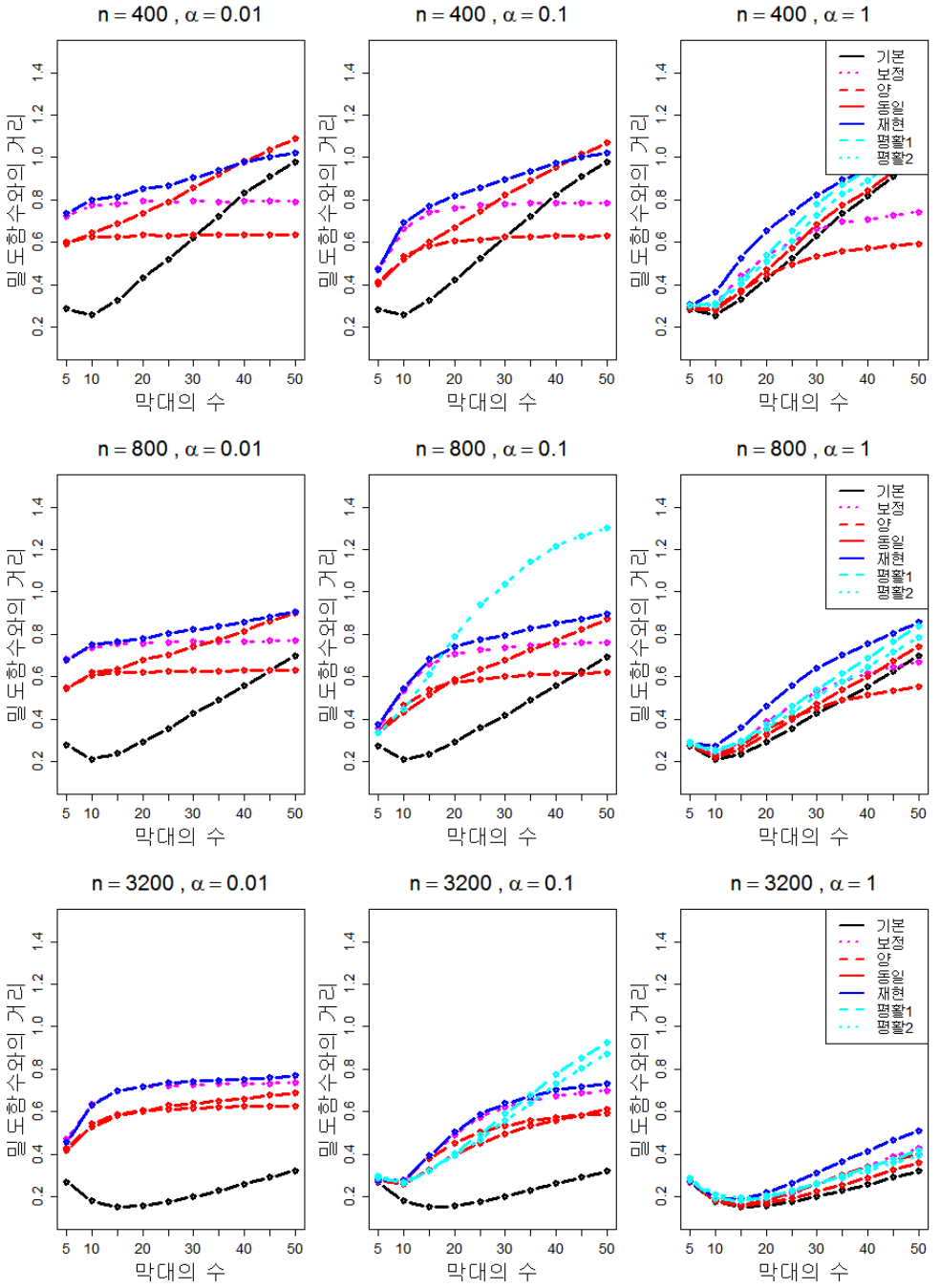
[그림 6-19] 원자료 밀도함수와 차등정보보호 히스토그램 사이의 거리($d=2$인 경우)



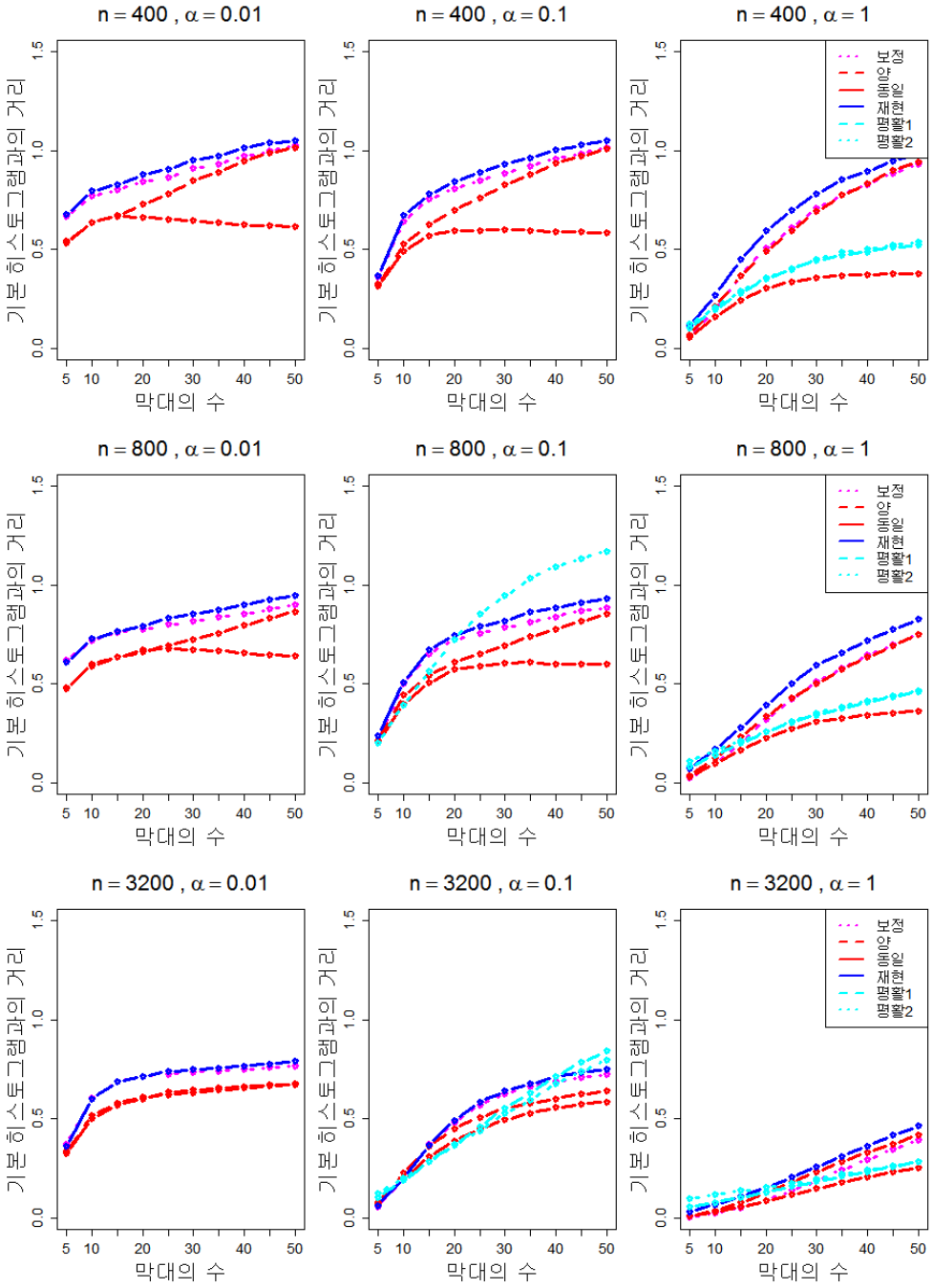
[그림 6-20] 기본(원자료) 히스토그램과 차등정보보호 히스토그램 사이의 거리($d = 2$인 경우)

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 6-21] 막대 개수에 따른 원자료 밀도함수와 차등정보보호 히스토그램 사이의 거리 ($d = 2$인 경우)



[그림 6-22] 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 6-12〉 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 6-23] ~ [그림 6-28] 및 <표 6-16>에서 정리하였다.

[그림 6-23]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 6-13>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 6-13〉 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 6-24]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 6-14>의 연령별 빈도표를 참고하면 전체 자료에 비하여 특히 30세 미만 가구주의 수가 상당히 적은 것을 확인할 수 있다.

〈표 6-14〉 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 6-24]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 6-14>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 6-15〉 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

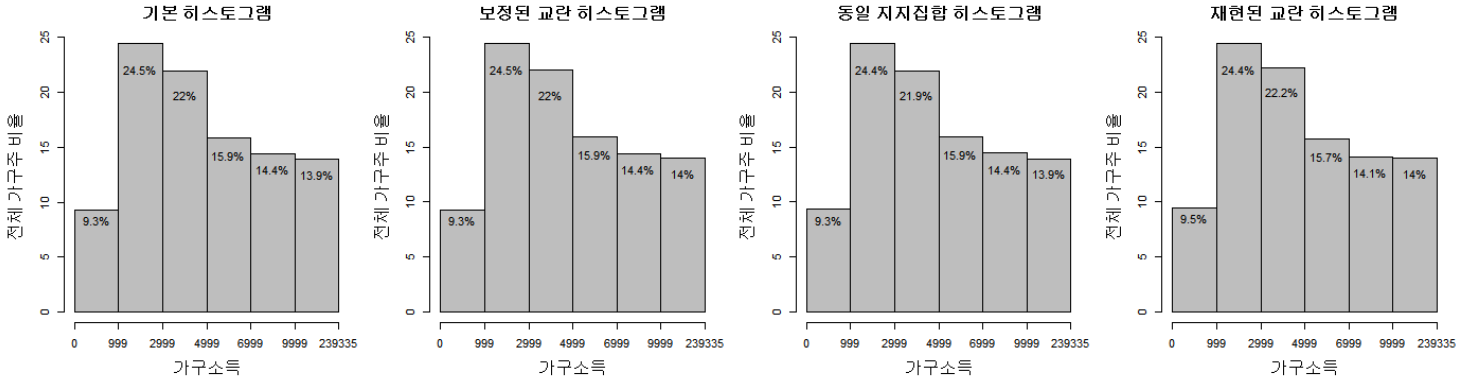
설명한 대로 α-차등정보보호 히스토그램에서 때때로 0에 아주 가까운 밀도값을 가지는 막대가 존재할 수 있다. 재현된 교란 히스토그램의 경우 이 구간에서 재현자료가 추출되지 않는 현상이 발생한다. 특히 10천만 이상 구간에 포함되는 가구소득의 최대값은 약 239천만이며 따라서 구간의 길이가 약 229천만이므로 밀도함수는 거의 0의 값을 가진다. 실험을 반복하면 [그림 6-24]에서 마지막 막대의 밀도값이 0에 가까운 경우가 더 많이 발생한다. 이러한 경우를 방지하려면 밀도값 대신 빈도값을 사용하여 임의 추출을 하는 것이 좀 더 현실적인 히스토그램을 제공할 것이라고 판단된다.

[그림 6-25] ~ [그림 6-28]의 다른 연령대에서는 [그림 6-24]와 달리 전체 연령의 히스토그램과 같이 기본 히스토그램과 유사한 차등정보보호 히스토그램이 생성되는 것을 확인할 수 있다. <표 6-14>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.

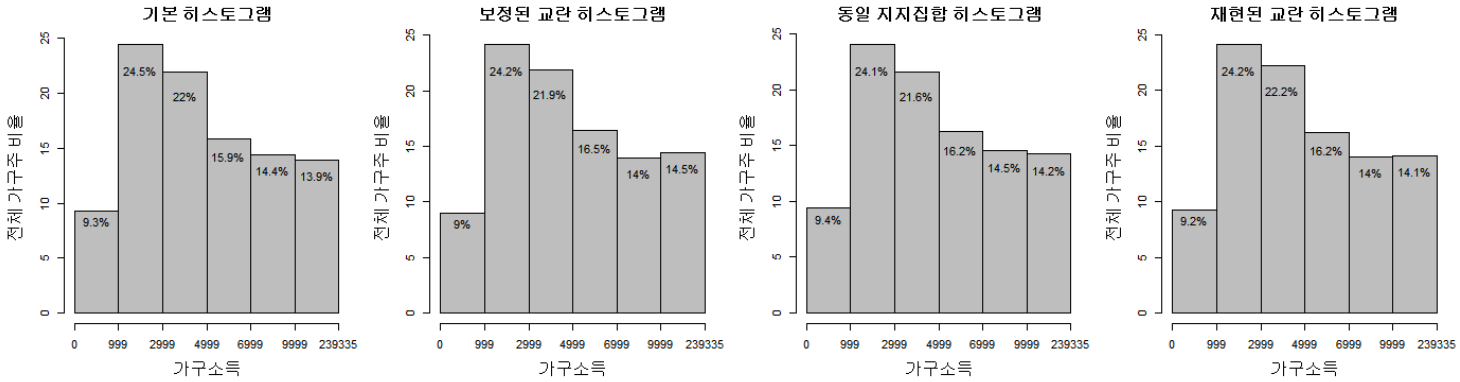
---

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

$\alpha = 1$ 인 경우

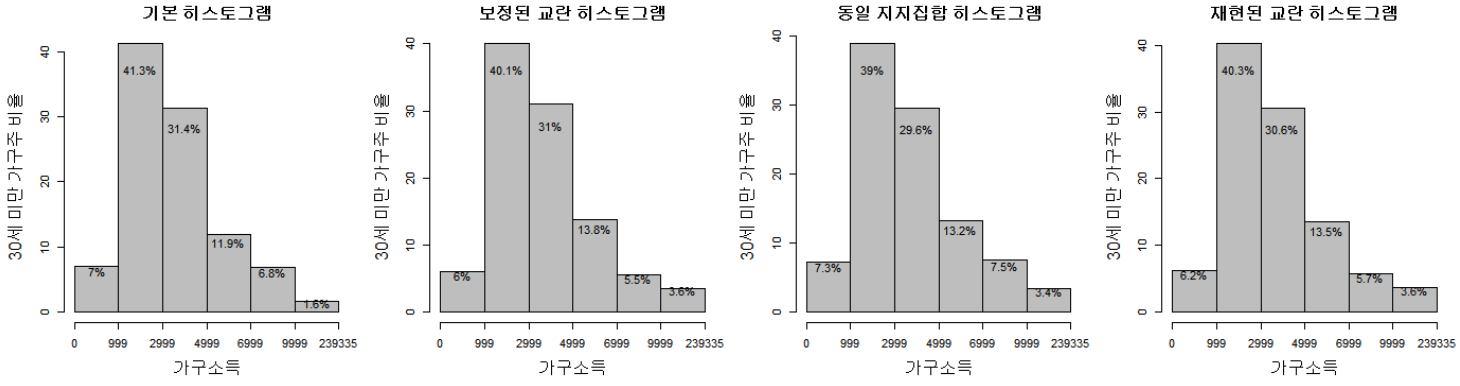


$\alpha = 0.1$ 인 경우

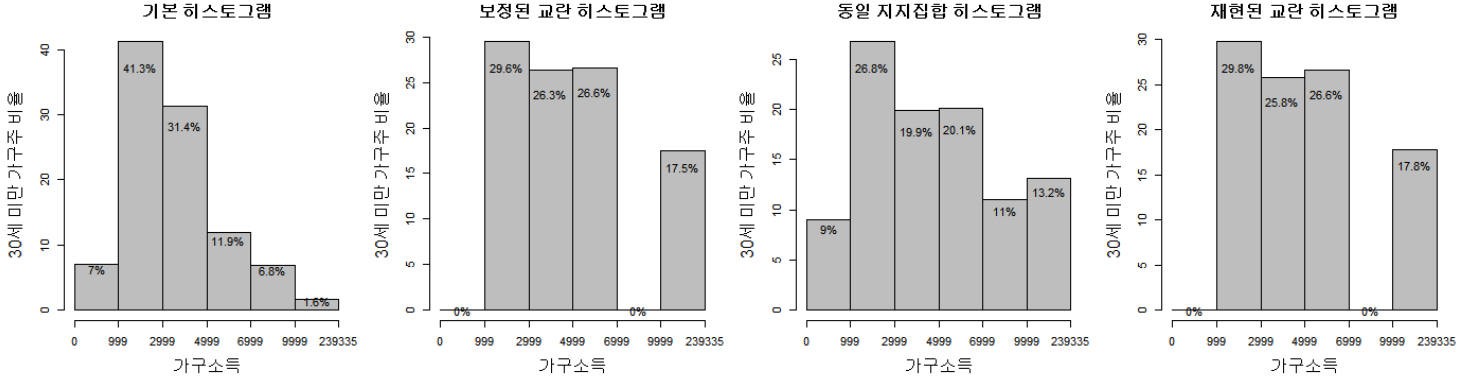


[그림 6-23] 가구소득 분포(전체 가구주)

$\alpha = 1$ 인 경우

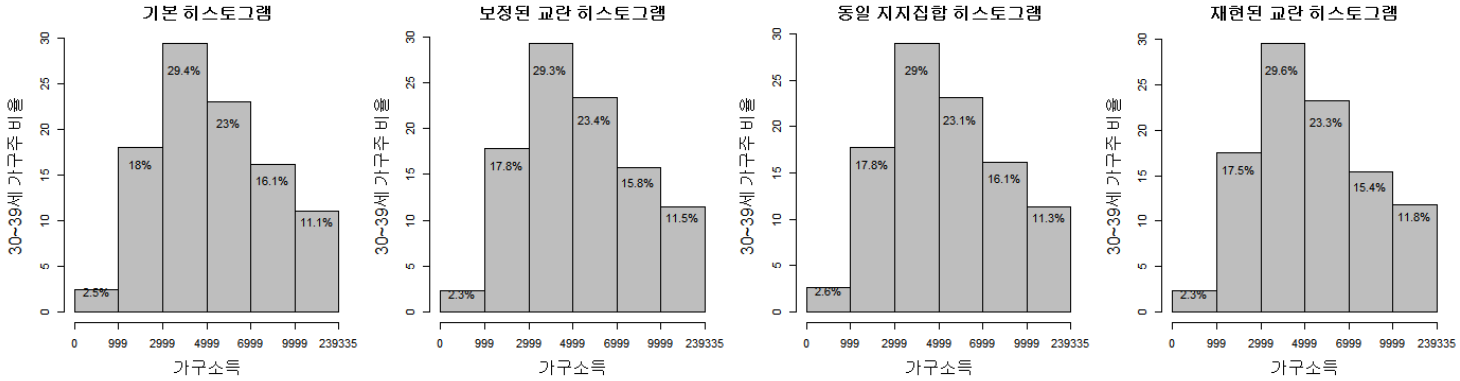


$\alpha = 0.1$ 인 경우

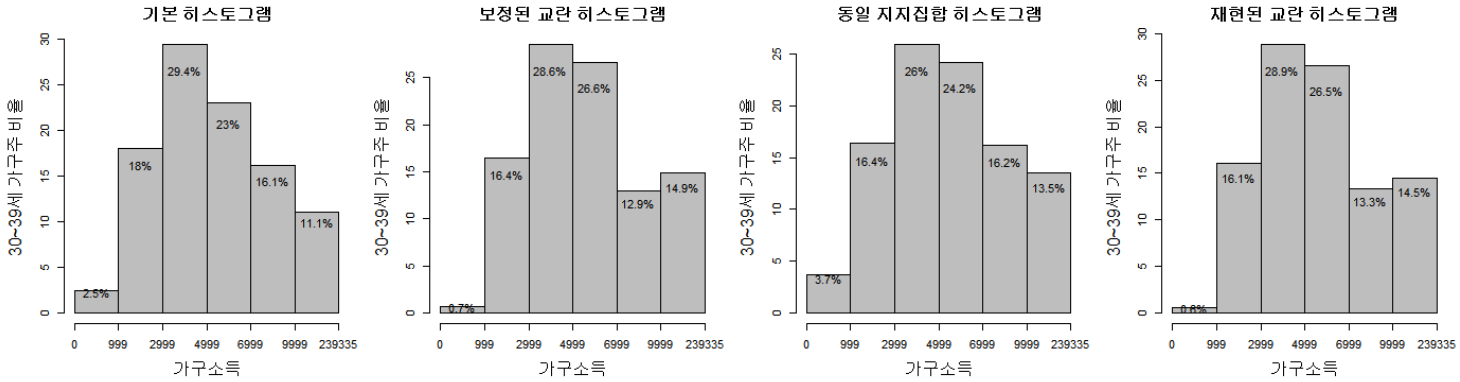


[그림 6-24] 가구소득 분포(30세 미만 가구주)

$\alpha = 1$ 인 경우

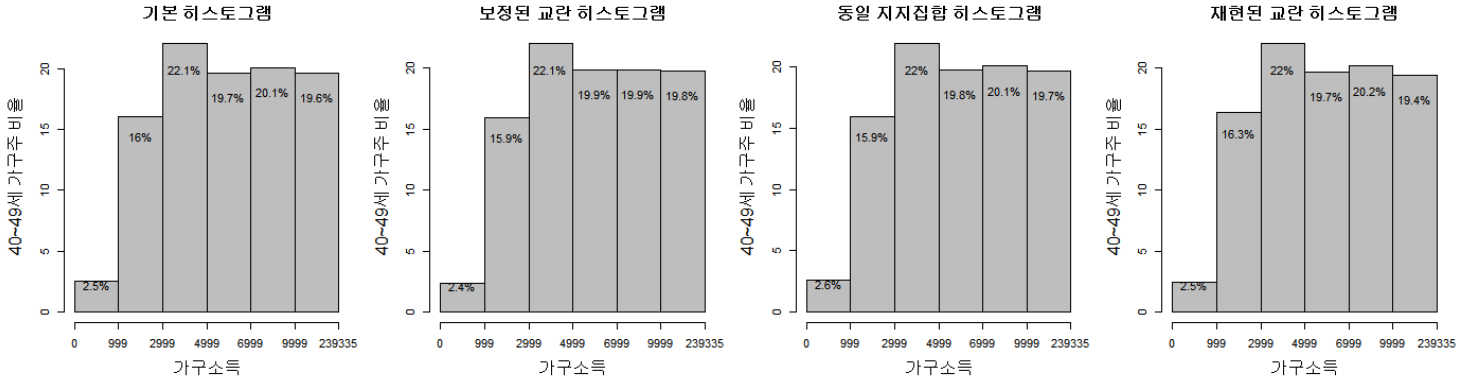


$\alpha = 0.1$ 인 경우

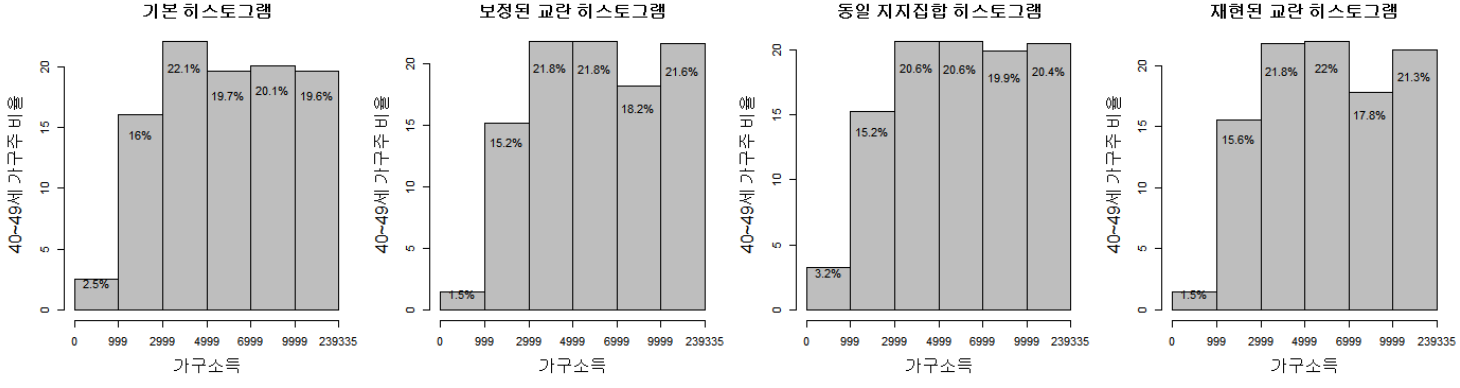


[그림 6-25] 가구소득 분포(30~39세 가구주)

$\alpha = 1$ 인 경우

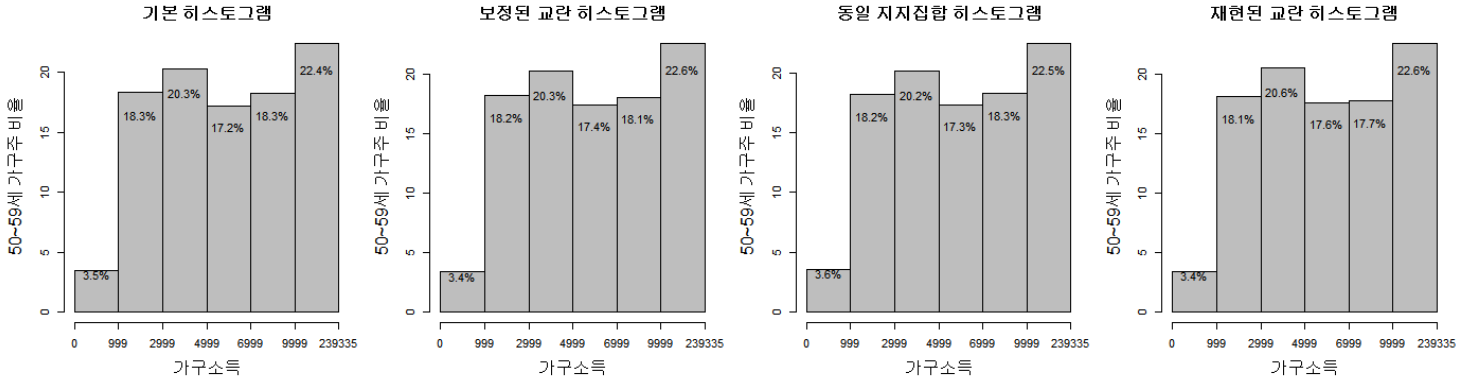


$\alpha = 0.1$ 인 경우

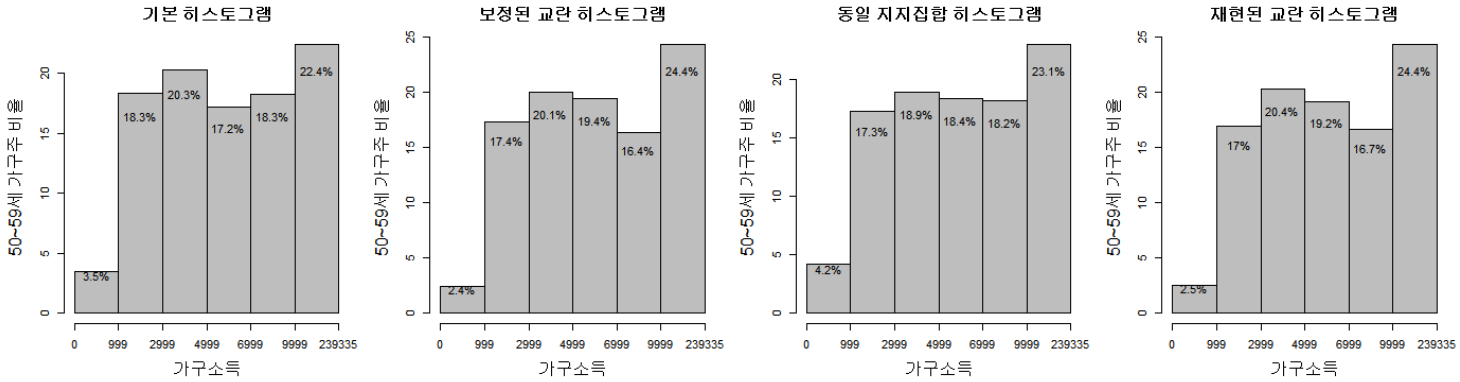


[그림 6-26] 가구소득 분포(40~49세 가구주)

$\alpha = 1$ 인 경우

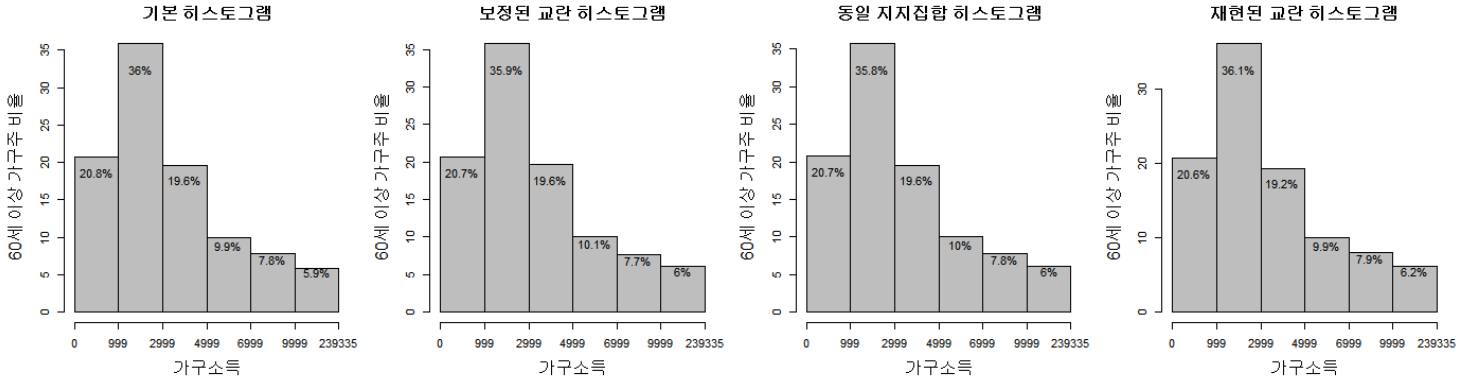


$\alpha = 0.1$ 인 경우

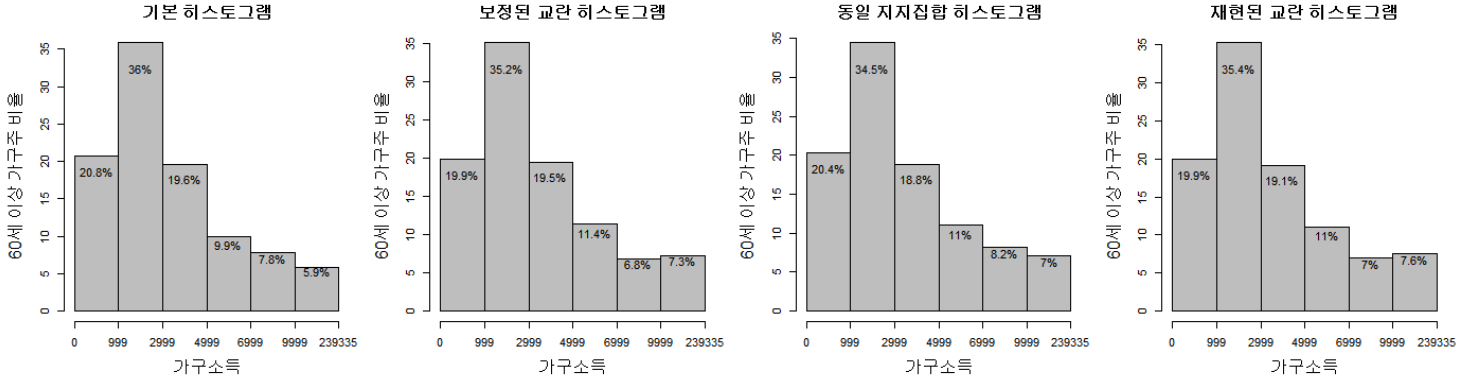


[그림 6-27] 가구소득 분포(50~59세 가구주)

$\alpha = 1$ 인 경우



$\alpha = 0.1$ 인 경우



[그림 6-28] 가구소득 분포(60세 이상 가구주)

〈표 6-16〉 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 6-16〉 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 6-17>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 6-17〉 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 6-18〉 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 6-19>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 6-20>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.

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

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

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

〈표 6-19〉 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	

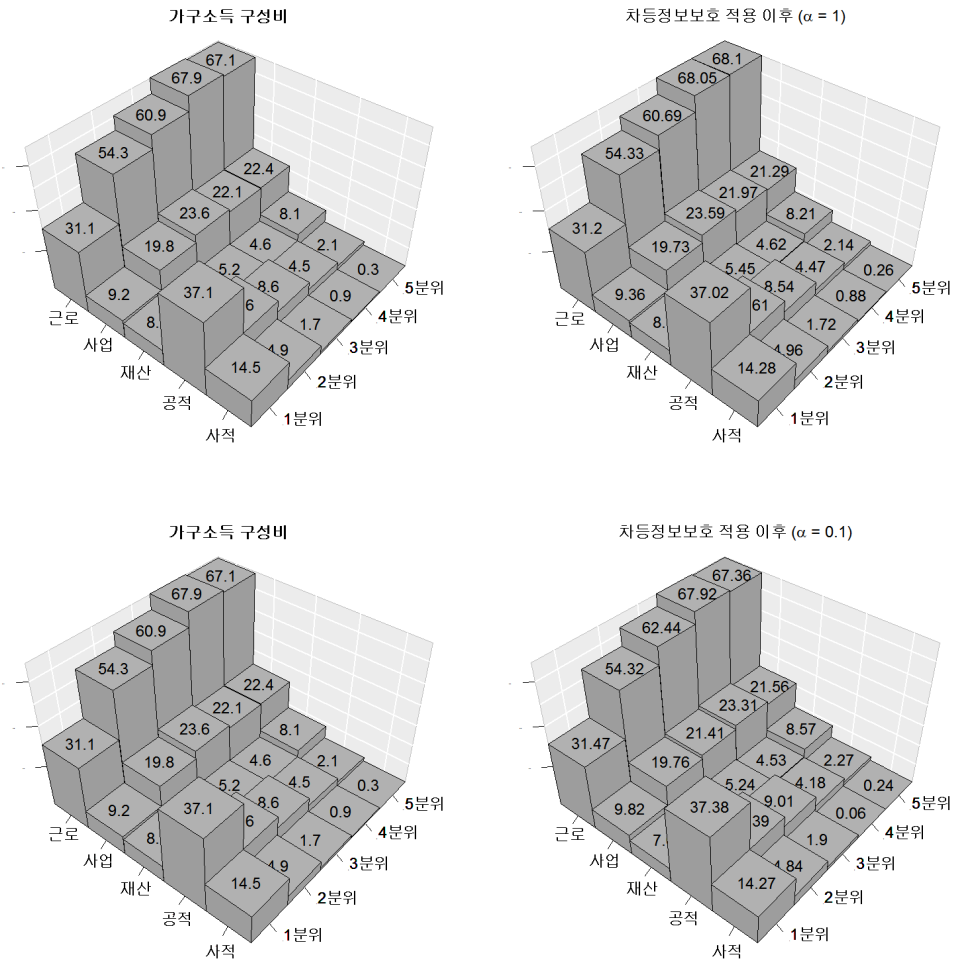
〈표 6-20〉 소득 5분위 및 소득원천별 가구소득 평균 : 정보손실(차이의 절대값) 비교

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 6-29]는 소득원천별 가구소득 구성비를 그림으로 나타내었다. 평균과 마찬가지로 소득 4분위의 사적이전소득 항목에서 정보손실이 큰 것을 확인할 수 있다. <표 6-21>은 소득원천별 가구소득을 구성비로 나타낸 보도자료에 대해 차등정보보호 적용 결과를 보여준다. <표 6-22>는 원래의 구성비과 차등정보보호 구성비의 차이를 나타내며 참고로 제시하였다. 정보보호를 위해 차등정보보호를 적용한 결과를 수용할 수 있을 것인가에 관하여는 별도의 논의가 필요하다.



[그림 6-29] 소득원천별 가구소득 구성비의 변화

〈표 6-21〉 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 6-22〉 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 6-23〉 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 6-25> 및 <표 6-26>에 정리하였다. <표 6-26>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 6-24>는 자산분위 및 소득분위의 빈도표이며, 차등정보보호 적용 결과를 이해하는데 도움을 주기 위하여 제시하였다.

〈표 6-24〉 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 6-25〉 소득분위별 순자산 분포 결과 비교($\alpha = 1$)

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	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

순자산 \ 소득	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

순자산 \ 소득	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 6-26〉 소득분위별 순자산 분포 결과 비교($\alpha = 0.1$)

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	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

소득 / 순자산	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

소득 / 순자산	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 6-27〉 소득분위별 순자산 분포 결과 : 정보손실(차이의 절대값) 비교

보정 ($\alpha = 1$)

순자산 \ 소득	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 ($\alpha = 0.1$)

순자산 \ 소득	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 ($\alpha = 1$)

순자산 \ 소득	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 ($\alpha = 0.1$)

순자산 \ 소득	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 ($\alpha = 1$)

순자산 \ 소득	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 ($\alpha = 0.1$)

순자산 \ 소득	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 6-28>과 같이 나타난다면, 재현자료는 <표 6-29>와 같이 나타난다. 가중치를 고려하지 않은 평균은 각 표의 맨 아랫줄에 수록하였다. 이 차등정보보호 재현자료는 <표 6-12>에서와 같이 6개의 막대를 가지는 히스토그램으로부터 생성되었다.

〈표 6-28〉 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 6-29〉 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 6-30> 및 <표 6-31>과 같이 나타난다. 각 변수별로 차등정보보호를 적용하여 생성한 재현자료에서 상관관계가 거의 사라지는 것을 알 수 있다. 처음부터 상관관계를 고려하지 않았으므로 이는 당연한 결과이다.

〈표 6-30〉 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 6-31〉 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 6-32>의 결과를 보면 정보손실이 크다는 것을 알 수 있다. 상관계수에 관하여는 <표 6-33>에서 재현자료의 상관계수가 자료를 생성할 때마다 달라지는 것을 보여준다.

〈표 6-32〉 가구소득 세부항목 평균

	근로	재산
원자료	3288.2	407.8

	근로	재산
재현자료	83154.4	48129.1

〈표 6-33〉 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.

제5절 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. 자료분석 목적이 명확한 경우 요청받은 쿼리에 적합한 차등정보보호 재현자료를 제공할 수 있다. 단계1은 탐색용 자료를 제공하는 절차에 해당된다.

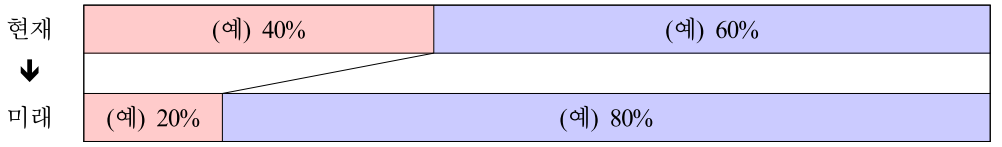
단계2. 분석계획 수립을 마치고 본격적인 분석을 요구하는 이용자에게 선택한 변수 범위에 대하여 베이지안 재현자료를 생성하여 원격 분석용 등으로 제공한다.

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 6-34>와 같이 나누어 생각할 수 있다. 효율적 자료제공을 위하여 통계적 노출제어만으로 데이터를 공유할 수 있는 플랫폼이 많아지는 것이 바람직하다고 판단된다.

〈표 6-34〉 데이터 정보보호 방안 분류

<table>
<tr><td>이용자</td><td colspan="2">심층 이용자</td><td colspan="4">불특정 다수</td></tr>
<tr><td>종류</td><td colspan="2">물리적 노출제어</td><td colspan="4">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="3">macrodata
(analysis output)</td></tr>
<tr><td>인가파일</td><td>가공 통계표
원격 접속(RAS)
데이터센터(RDC)</td><td colspan="2">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(~'15년) → 통계적 노출제어(~'17년) → 통계적 비밀보호(~'19년) → 데이터 정보보호('20년~)
[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 위해서는 중장기 계획을 세워 각 주제별로 세부 과제를 정교하게 선정하여 수행할 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.

<부 록>

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.

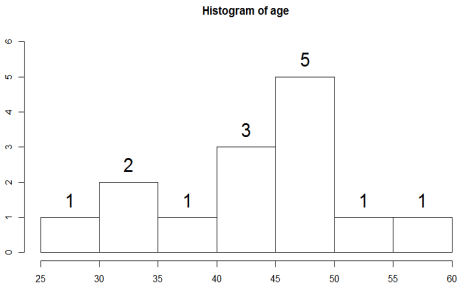
$$SSE(H_k, C_m) = \sum_{j=1}^{k} \sum_{i=l_j}^{i=r_j} (h_j - c_i)^2.$$

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.

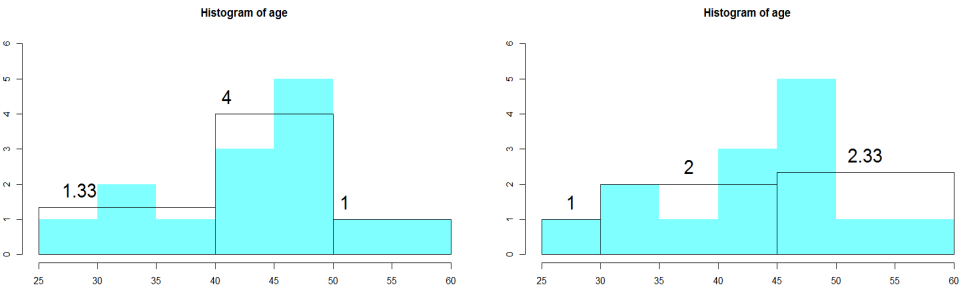
$$T(i, k) = min_{k-1 \leq j \leq i-1} \{T(j, k-1) + SSE(C_m, j+1, i)\}.$$

단, $SSE(C_m, l, r) = \sum_{i=l}^{r}(c_i - \overline{c}(l,r))^2$ 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 음영으로 표현되어 있다.



일반적으로 k값이 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($d=1$, 일차원 히스토그램인 경우)

〈부표 6-1〉 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 6-2〉 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 6-3〉 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 6-4〉 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($d = 2$, 이차원 히스토그램인 경우)

〈부표 6-5〉 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 6-6〉 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 6-7〉 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 6-8〉 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

- 김경미(통계청 사무관)
- 박민정(통계청 사무관)
- 정재호(통계청 주무관)
- 정동욱(통계청 과장)
- 신우람(통계청 사무관)
- 김정연(통계청 주무관)
- 박천영(㈜네이버 수석연구원)
- 서영오(서울대학교 석사과정)
- 이용희(서울시립대학교 교수)
- 권성훈(건국대학교 교수)

개인정보보호와 데이터 활용을 위한 통계적 방법론(Ⅱ)
– 데이터 정보보호를 위한 최신 기법의 활용 –

■발행연월일	2020년 4월 21일 인쇄
	2020년 4월 22일 발행
■발 행 인	통계개발원장 전 영 일
■발 행 처	통계청 통계개발원
	35220 대전광역시 서구 한밭대로 713
	(월평동 282-1) 통계센터 6~8F
	☎ 대표 (042) 366-7100 Fax (042) 366-7123
■인 쇄 처	강문인쇄사 (042) 226-4722~3