

2020년 연구보고서

통계데이터센터 반출 결과물의 통계적 노출제어 가이드라인(안)

2021. 4.



<http://kostat.go.kr/sri>



ISSN 2288-1166(Print)
ISSN 2733-4120(Online)



통계청
통계개발원

통계데이터센터 반출 결과물의 통계적 노출제어 가이드라인(안)

박민정 · 박영옥 · 황영자



Statistics Korea
Statistics Research
Institute

발간사

통계개발원(Statistics Research Institute, SRI)은 2006년에 개원한 이래 국내에 유일한 「국가통계 전문연구기관」으로서 국가통계·조사 방법론의 발전과 데이터기반 정책 연구·개발 분야에서 선도적인 역할을 하고 있습니다. 지난 한 해 동안 SRI는 “광출어혁(光出於革)”의 정신, 즉 혁신에(革) 기반한 실용적인 「팀연구」를 통해서 국가통계·데이터과학의 미래를 밝히고자(光) 노력하였습니다.

SRI의 「2020년 연구보고서」는 이러한 혁신연구의 결과물입니다. 특히 코로나19의 경제·사회 영향을 측정하는 지표 설계 연구를 비롯해서 「데이터기반 인구·사회·경제정책」을 뒷받침하는 심도 깊은 연구가 진행되었습니다. 예를 들어서 저혼인·저출산·초고령 사회에 대응하기 위한 데이터 심층 분석, 균형발전 정책수립을 지원하는 지역사회지표 체계구축, 그리고 SDG를 활용한 남북한 통합통계 방안 등입니다. 또한 2020년 2월에 출범한 「SDG 데이터연구센터」는 지속가능발전목표(SDG)의 이행 현황 점검과 SDG 대국민 데이터 서비스를 준비하는 실용적인 연구를 진행하였습니다.

국가통계·데이터과학·조사방법론 분야에서는 「데이터 정보보호센터」를 중심으로 개인정보보호 방법의 통계지리정보서비스(SGIS) 적용 등 실용적인 연구가 진행되었습니다. 또한 「조사표연구센터」를 활용하여 빅데이터 활용 홈페이지 개선을 위한 사용성 평가를 비롯해서 「조사표 인지실험」을 적용한 국가통계의 품질개선이 이루어졌습니다. 특히 지난해에는 데이터과학(Data Science)을 국가통계에 활용하고자하는 기초 연구를 추진함으로써 「데이터 혁신 방법론」의 새로운 전기를 준비하였습니다.

금년에 개원 15주년을 맞는 SRI는 본 연구보고서가 증거기반정책 입안자의 데이터 활용에 실제적인 도움이 될 뿐만 아니라 국가통계 생산자의 혁신적인 조사방법론 개발에 유용한 자료로 활용될 수 있기를 기대합니다. SRI가 「국가통계 싱크탱크」로서 국내외적으로 선도적인 역할을 할 수 있도록 독자 여러분의 지속적인 격려를 부탁드립니다. 국가통계의 도약과 혁신적 개발을 위하여 제언이 있다면 언제든지 말씀하여 주십시오. 겸허히 귀 기울이겠습니다.

이번 「2020년 연구보고서」를 위하여 실용적이고 품질 높은 연구 결과를 도출하기 위해 광출어혁(光出於革)의 정신으로 최선을 다한 연구진과 대내외적으로 협력·공동 연구에 참여한 민·관의 연구자들에게 따스한 감사를 전합니다. 끝으로 본 연구에서 제시된 내용 및 결과는 저자 개인의 의견이며, 통계청 또는 통계개발원의 공식적인 견해가 아님을 밝혀둡니다.

2021년 4월

통계청 통계개발원장



전 영 일

목 차

제1장 서론	1
1. 추진배경	5
제2장 통계데이터센터 SDC 기본 원칙	8
1. 표준화	8
2. 노출위험	8
3. 정보손실	9
4. 체계적인 운영·관리	10
5. 반출유형별 반출 원칙	10
제3장 관리자 가이드라인	14
1. 노출위험 및 정보손실 이해	14
2. 자료이용 신규신청 접수	14
3. 자료이용 수정신청 접수	15
4. 분석결과 반출신청 접수	15
제4장 이용자 가이드라인	17
1. 성실한 자료 이용 신청서 작성	17
2. 반출유형별 노출제어 처리	19
3. 추론에 의한 노출 가능성 검토	20
4. 최종 성과물 입증(개별정보 비식별 입증)	20
5. 이용 요청 이외 목적 사용 금지	20
제5장 향후 발전방향	21
1. 최신 기법 활용 기대 효과	21
2. 중장기 연구 방향	21
3. 가이드라인 업데이트 지속	22
참고문헌	23
부 록	24

요 약

통계기관은 불특정 다수에 완전히 공개하기에는 민감한 통계자료를 데이터센터를 통하여 제공한다. 전수 자료와 같이 연구 이외의 목적으로 악의적으로 사용될 위험이 있는 자료가 그 대상이다. 데이터센터는 학술적, 정책적으로 분명한 목적을 가지는 자료 분석 활동을 위하여 자료를 제공하며, 이용자는 데이터센터를 방문하여 원자료를 분석할 수 있다. 분석 결과물은 위원회의 심의를 거쳐 이용자에게 제공된다. 다만 대용량 자료를 분석한 결과물 역시 대용량인 경우가 많으며, 분석 결과물을 통한 개인정보 침해 등의 위험을 점검하는 것은 쉬운 일이 아니다.

본 보고서는 데이터센터 반출물에 대한 노출제어 가이드라인을 작성하기 위한 첫걸음을 담고 있다. 국내외 사례를 검토하고, 각종 문헌을 참고하여 가이드라인 초안을 제시하고 있다. 먼저 반출물 평가를 위한 노출제어 기본 원칙을 정리하고, 데이터센터 운영에 있어 실무적인 측면을 반영하여 관리자 및 이용자 각각을 위한 지침을 정리하였다. 특히 반복적으로 사용할 자료이용 신청 서식을 제안하였고, 관리를 위하여 필요한 이론적 내용을 간략히 소개하였다. 본 초안을 바탕으로 데이터센터 운영 결과를 반영하여 가이드라인을 지속적으로 업데이트할 필요가 있으며, 이는 효율적이고 안전한 데이터센터 운영에 필수적인 것으로 판단된다.

주요용어 : 통계적 노출제어, 매크로데이터, 데이터센터, 가이드라인

제 1 장

서 문

- 통계청의 주요 역할은 우리 사회를 설명하기에 적절한 통계를 생산하고 제공하는 것이다. 통계청이 만든 통계는 공공재로서 널리 활용되어 왔으며, 주요 이용 사례로는 정책 결정 근거 자료, 학술 논문용 데이터, 언론 보도용 각종 수치 등이 있다.
 - 이를 위해 통계청은 분야별, 항목별로 많은 통계표들을 만들어 KOSIS 등을 통해 배포해 왔다.
 - 또한 직접 분석을 원하는 이용자들을 위해서는 MDIS 등의 시스템을 구축하여 통계표를 작성하는데 쓰인 원천 자료 마이크로데이터를 제공해 왔다.
- 한편, 요즘 언론에서 이슈로 매우 빈번하게 다뤄지는 단어는 ‘빅데이터’와 ‘4차 산업혁명’이다. 이러한 키워드들을 관통하는 핵심은 여기저기서 생겨나는 대용량 자료를 어떻게 얻고, 관리하고, 분석하여 가치를 창출할 것인가이다.
 - 이러한 대용량 자료의 유형은 비정형과 정형으로 나눌 수 있다. 통계청의 입장에서 생각해 보면, 조사를 통해서 생산하던 통계를 더 빠르고 정확하게 생산하고, 나아가 기존에 없던 새로운 통계를 생산하기 위해, 비정형 대용량 자료를 활용하는 방안을 찾아야 한다는 도전을 받게 된다.
 - 이미 통계청은 ‘온라인 물가지수’ 등을 만드는데 비정형 대용량 자료를 사용하기 시작했다. 물론 비정형 대용량 자료를 활용해 새로운 가치를 창출하는 작업은 민간에서도 매우 활발하게 이루어지고 있다.
- 이처럼 모으고, 정리하고, 분석하기 까다로운 비정형 대용량 자료를 활용하기 위해 많은 노력을 들이는 시대에, 수집과 관리가 상대적으로 용이한 정형 대용량 자료를 활용하려는 시도는 매우 자연스럽다.
 - 통계청과 밀접한 관련이 있는 정형 대용량 자료로는 센서스와 행정자료가 있다. 우리 사회 다양한 분야에서 가치 창출을 뒷받침하는 공공재로서 센서스와 행정

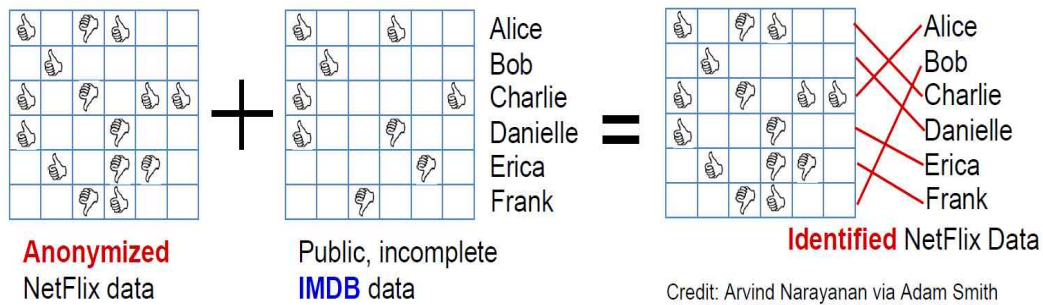
자료가 널리 활용될 수 있다면, 통계청은 본연의 역할을 더욱 충실히 수행하는 것이 된다.

- 또한 이러한 정형 대용량 자료 이용을 요구하는 수요는 통계청 내·외부에서 계속해서 증가하고 있다.

□ 하지만 여기에 넘어서기 무척 어려운 걸림돌이 존재한다. 그것은 바로 ‘개인정보 보호’이다. 얼핏 생각하면 가치 창출을 위해 대용량 자료를 사용하는데 개인정보 보호가 왜 문제가 되는지 이해되지 않을 수 있다. 너무나 큰 자료에서 개인이 드러날 수 있을까? 또 이름, 주소 등의 개인정보를 삭제하고 제공하는 자료에서 개인이 드러난다는 것은 무슨 의미일까?

- 이해를 돕기 위해 유명한 넷플릭스(Netflix)사의 사례를 간단히 살펴보자.
- 과학적인 조사표 설계의 발전, 데이터 수집 및 분석기술의 고도화, 인터넷 및 웹 기반의 조사환경 확산 등 긍정적인 조사환경 변화가 지속되고 있다.
- 넷플릭스는 미국의 유명한 동영상 스트리밍 서비스 회사이다. 세계 최대의 업체로 이용자가 원하는 영화를 추천하는 기능을 주요 강점으로 홍보하고 있다. 넷플릭스는 자사의 영화 추천 알고리즘을 홍보하면서 동시에 추천 알고리즘을 개선하기 위해 경연대회를 개최했다. 이를 위해 개인 식별정보를 모두 삭제한 회원별 영화 평가 자료를 만들었는데, 이 자료는 넷플릭스 이용자 약 50만 명의 6년간(1999년 12월 ~ 2005년 12월) 영화 스트리밍 신청 이력으로 약 1억 건이며, 비식별화 변수, 영화평가 내용, 평가일시 등이 포함되어 있다. 대회 참가자에게 이 자료를 제공하여 영화를 추천하는 새로운 알고리즘을 만드는 첫 번째 대회가 열렸다.
- 그러나 텍사스 대학의 한 연구진은 넷플릭스의 비식별화 처리된 자료를 다른 자료와 결합하여 일부 회원의 민감한 정보가 노출될 수 있음을 보였다. 결합에 사용된 자료는 또 다른 영화평가 제공사인 IMDB(Internet Movie Database)의 아이디와 영화평가점수로 구성된 공개된 회원자료이다. 특히 문제가 된 부분은 일부 개인의 성인물 등에 대한 영화 취향이 노출될 수 있다는 것이었으며, 결국 미국 연방거래위원회(FTC)가 프라이버시 침해 문제를 지적해 2회 경연은 중지되었다. 다음 <그림 1-1>은 이러한 내용을 요약하여 보여주고 있다.
- 이 사례는 비식별화 처리하여 제공한 대용량 자료에서도 개인의 민감한 정보가 드러날 수 있다는 경각심을 일으키며 널리 알려졌다. 참고로, 이러한 사례는 가치창출을 예상하고 널리 활용되도록 하는 목적을 가지고 제공하는 자료에서 일

어날 수 있는 사고이므로, 철저한 보안 속에 지켜지던 개인정보 파일이 유출되는 것과는 다른 영역에 속하는 문제이다.



출처: A Tutorial on Differential Privacy, G. Rothblum

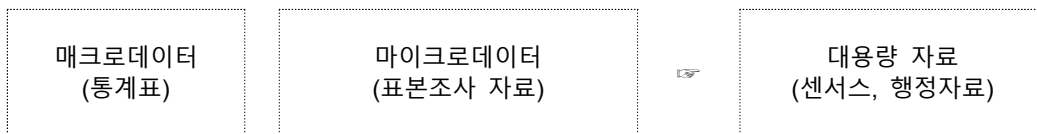
<그림 1-1> 넷플릭스 경연대회의 프라이버시 노출 사례

- 통계청은 정형 대용량 자료 역시 널리 활용될 수 있도록 제공 자료의 범위를 확대해 나가야 하며, 동시에 개별정보가 노출되지 않도록 조치해야 한다.
 - 자료 유용성은 높게 유지하면서 개별정보 노출은 제한하는 방법을 찾는 것은 과학적으로도 쉽게 풀리지 않는 문제로 이를 다루는 통계학 분야를 통계적 노출제어(Statistical Disclosure Control)라 부른다.
 - 지금까지 각종 통계표들과 마이크로데이터를 제공할 때 통계기관들이 사용해온 다양한 매스킹 기법들, 최근 미국을 중심으로 사용되기 시작한 재현자료 활용, 차등정보보호 적용 방법 등이 SDC 기법에 속한다. 즉, SDC 기법은 마이크로데이터가 통계기관을 벗어나기 전에 자료 자체에 시행되는 비밀보호 처리 방법을 말한다.
- SDC 기법이 아무리 발전한다고 해도, 원자료를 직접 분석해야 하는 수요는 늘 존재하기 마련이다.
 - 이러한 심층 분석 수요를 충족하기 위해 각국의 통계기관들은 데이터센터(Resrach Data Center)를 운영하고 있다. RDC를 방문한 이용자들은 철저한 보안 지침을 준수하며 원자료를 직접 분석할 수 있다. 분석결과는 RDC의 승인 절차를 거쳐 반출이 허용된다.
 - 이처럼 자료는 그대로 두고, 이용자를 규제하여 비밀보호를 달성하는 전략을 물리적 노출제어라 부르도록 하겠다.

- 이용자가 RDC를 방문하여 자료를 이용하면 충실한 자료제공을 할 수 있지만, 여기에는 경제적·행정적 비용 발생의 문제가 따른다. 이용자는 센터를 방문하기 위해 이동 시간과 비용을 지불해야 하고, 자신의 업무 환경에서 작업하지 못하는 불편함을 감수해야 한다. 통계기관은 결과물을 일일이 검열해야 하므로 상당한 업무 부담을 가지게 된다. 특히, 민감한 정보를 포함한 결과물들이 많거나 연간 수천 명의 이용자가 방문한다면 RDC 운영에만 상당한 인력이 필요하여 적지 않은 행정적·경제적 비용을 부담해야 된다.
 - 때문에 각국의 통계기관들은 물리적 노출제어를 통한 자료제공 비중을 줄이고, 통계적 노출제어를 통한 자료제공 비중을 높이기 위한 노력을 하고 있다.
 - 다만 물리적 노출제어를 통한 자료제공이 없어질 수는 없다는 점에서, 효율적인 RDC 운영 방안을 수립하는 것은 반드시 필요한 일이다.
- RDC 운영을 논의하는데 있어 RDC를 통해 분석결과를 반출할 때도 SDC 처리가 필요하다는 점은 주지할 필요가 있다. 분석결과에서도 개별정보 노출이 일어날 수 있기 때문이다.
 - 더구나 대용량 자료를 분석한 결과가 역시 대용량인 경우, 비밀보호 처리는 생각보다 복잡하여 만족스러운 솔루션을 마련하는 것은 여전히 연구 진행 중인 상태이다.
- 본 가이드라인은 RDC를 통해 정형 대용량 자료를 제공할 때 반출물에 대한 비밀 보호 처리 문제를 다루고 있다. 통계청이 KOSIS를 통해 통계표를 제공할 때와 달리, RDC의 분석결과는 다양성을 예측할 수 없으므로 기존의 매스킹 기법들을 기계적으로 적용하는 것으로는 비밀보호 문제를 해결하기 어렵다.
 - 반면, 안전성에 크게 문제가 없는 경우도 많으나, 통계 전공자가 아니어도 언제 안전한 것인지 쉽게 알 수 있는 가이드라인 또한 효율적인 센터 운영을 위해 필요하다.
 - 본 과제에서는 해외 사례 분석과 관계자 회의 등을 통해 분석결과의 종류를 나누어, 유형별로 어떻게 SDC 처리를 할 수 있는지 정리하였다.
 - 나아가 가장 문제가 되는 고차원 빈도표를 제공해야 하는 경우에 사용할 수 있는 SDC 프로그램을 작성하였다. 우선 하반기에 통계데이터센터에서 본 가이드라인을 베타 버전으로 운영하여 가이드라인의 효율성을 점검하고, 향후 지속적으로 업데이트해 나갈 계획이다.

1. 추진배경

1.1 자료제공 범위 확대



□ 과거 통계청은 직접 작성한 통계표나 표본조사 마이크로데이터를 제공하였으나, 최근 **대용량 전수자료(센서스 및 행정자료) 이용 요청이 증가**하여 자료제공 범위를 확대할 필요

- 표본에 비하여 전수자료는 개별정보 노출위험이 상대적으로 높음
 - 일반적인 표본조사 마이크로데이터는 표본추출 과정에서 매스킹 효과가 발생하여 노출위험이 감소하나, 전수자료는 그러한 보호 효과가 없음
- 대용량 자료제공으로 발생할 수 있는 개별정보 노출^[부록1]을 차단할 필요
 - 넷플릭스 사례^[1] 등, 가치 창출을 예상하고 자료를 제공할 때 뜻밖의 프라이버시 노출이 가능하므로 안전한 자료제공을 위한 적절한 대처 필요

미국 넷플릭스(Netflix)사의 사례

- ▶ 자사의 영화 추천 알고리즘 홍보와 정확성 개선을 위한 경연 대회 개최
- ▶ 식별정보 삭제 후 회원별 영화 평가 자료를 공개
 - : 대회 참가자에게 넷플릭스 이용자 약 50만 명의 1999년 12월~2005년 12월, 약 1억 건 영화신청이력을 제공 (비식별화 변수, 영화평가내용, 평가일시 포함)
- ▶ 일부 회원의 민감한 정보가 노출됨
 - : 텍사스 대학의 연구진은 넷플릭스의 비식별화 처리된 자료를 다른 영화평가 제공사인 IMDB(Internet Movie Database)의 회원 자료(아이디와 영화평가점수)와 결합하여 일부 회원의 민감한 정보가 노출될 수 있음을 보임
- ▶ 미국연방거래위원회(FTC)가 프라이버시 침해 문제를 지적해 2회 경연 중지됨

1.2 자료제공 방식의 변화

- 이용자의 마이크로데이터 사용 요청 시, 개별정보 보호를 위해 지금까지는 주로 **마스킹 처리**로 약간 변형된 자료를 **파일 형식으로 제공**
- 전수자료는 개인정보보호 및 대용량 문제로 파일 형태 제공은 어려우며, 요청받은 쿼리에 따른 **분석결과(Output)**를 제공하는 시스템 방식이 필요
 - 통계데이터센터 이용자의 자료 분석도 예측할 수 없는 쿼리에 포함됨
 - 다양한 쿼리에 대해 분석결과를 얻고, 다량의 분석결과에서 개별정보를 파악하는 것이 가능 (통계기관이 직접 작성한 통계표 제공과 다른 환경)
 - 기존의 매크로데이터를 위한 마스킹 기법의 단순 적용은 부적절하며, 장기적으로 시스템 방식에 대한 범용적인 SDC 방법론 수립 필요

1.3 심층 분석 지원 필요

구분	물리적 노출제어	통계적 노출제어(SDC)
이용자	심층 이용자	불특정 다수
배포형식	▶ 인가파일¹⁾ ▶ 데이터센터(RDC)	▶ 가공 통계표 ▶ 원격접속(RAS)
		▶ 공공이용파일

- 현재 대다수 국가에서는 공공이용을 위해 마스킹 처리로 노출위험을 낮춘 파일을 제공하나, 마스킹 처리는 정보손실을 발생시킴
- **노출위험이 높으나 정보손실이 적은 심층 분석용 자료를 제공**하기 위해서는 이용자를 규제하는 전략, 물리적 노출제어^[부록2] 시행 필요
 - 이용자의 경제적 비용(센터방문)과 제공자의 행정적 비용(결과검열)이 발생하나^[2], 가장 안전하게 정보손실이 적은 자료를 제공할 수 있음

1) 인가파일(Licensed File)은 자료 이용자 소속기관과 자료 제공기관 사이의 법적 계약을 통해 자료 제공기관 외부로 반출을 허용하는 자료를 말함.

1.4 가이드라인 부재

- 통계적 노출제어는 해외에서도 연구 인력이 적으며, 노출위험과 정보손실을 동시에 낮추는 **최적의 방안을 찾는 연구가 여러 관점에서 진행 중임**
 - 현재 대부분 통계기관의 가이드라인은 사안별로 데이터센터 관리자의 판단을 따르도록 작성됨
 - 미국 보건통계본부(NCHS) RDC의 경우, 담당자 가이드라인, 이용자 가이드라인 및 통계 프로그램 사용 규칙을 운영 중이나, 빈도표 등의 기술통계는 미반출을 권고할 뿐 명확한 제공 기준은 없음^[3]
 - * 센서스 빈도표처럼 하나의 마이크로데이터에서 많은 빈도표를 작성하는 경우 통계적 노출제어 방안은 여전히 만족스러운 해법을 찾는 연구 진행 단계^{[4][5]}
 - 데이터센터는 다양한 경우에 대해 ‘현재’ 활용할 SDC 방안이 필요
- 국내에서 2016년 「개인정보 비식별화 조치 가이드라인」이 발표되었으며, k -의 명성 위주의 노출위험 기준과 몇몇 매스킹 방안을 제시^[1]
 - 수용하기 어려운 수준의 정보손실이 발생할 수 있어 추가 연구 필요
 - 통계데이터센터의 분석결과 반출을 위한 맞춤형 가이드라인 필요
 - * 마이크로데이터의 통계적 노출제어는 본 가이드라인의 대상이 아님

제 2 장

통계데이터센터 SDC 기본 원칙

1. 표준화

- ☐ 여러 자료에서 사용되는 범주형 변수는 **가급적 동일한 범주**를 사용^[4]
 - 예. 연령 변수는 0세부터 5세 단위로, 85세 이상은 한 범주로 처리 등
- ☐ 연속형 변수의 경우, 반출물의 최소 단위를 통일시키는 것이 바람직
 - 예. A자료 총계표에서 매출액 총액은 가급적 백만 원 단위로 반출 등
- ☐ 빈도표/총계표에서 셀 값이 0인 경우 안전하다고 판단하여 제공

2. 노출위험

- ☐ 일반적인 **통계적 노출**은 다음 세 종류로 구분

식별	제공된 자료에서 모집단의 특정 개체가 파악되는 현상
속성 노출	특정 개체의 민감한 속성이 드러나는 현상
추론에 의한 노출	특정 개체의 정보가 상당히 정확하게 추정되는 현상

- ☐ 빈도표에 대한 **통계적 노출**은 다음 세 종류로 구분^[4]

개별 속성 노출	빈도표의 어떤 부분 합(margin)이 1인 경우
그룹 속성 노출	빈도표에서 열이나 행이 대부분 0으로 구성된 경우
차이에 의한 노출	두 빈도표의 차이로 새 빈도표가 작성되는 경우

- **노출위험** 측정 기준으로 유일성 기반 측도 중 k -**익명성**을 활용^[부록3]
 - k -익명성이 확보되면 식별 및 속성 노출 위험이 제어된 것으로 판단
 - 향후 SDC 방법론 발전에 따라 기준 추가·변경 가능^[6]
- 이용자가 직접 개발한 고유한 모형을 사용하는 경우 등, 관리자가 모든 분석 기법을 숙지하는 것은 불가능하므로, 추론에 의한 노출 발생 가능성은 자료 이용자가 검토하는 것을 원칙으로 함

k -익명성

- ▶ 동일한 키변수 조합을 가지는 개체가 k 개 이상이면, k -익명성을 획득했다고 함

예. 자료에서 키변수가 성별(남, 여) 및 교육수준(고교 이하, 대학, 대학원)인 경우에, (남, 고교 이하), (남 대학), (남, 대학원), (여, 고교 이하), (여, 대학), (여 대학원) 이라는 총 6가지 성별·교육수준별 키변수 조합이 존재. 이 자료에서 각 조합마다 최소 3개의 개체가 있으면 이 자료는 익명성 3을 획득했다고 말함

3. 정보손실

- 정보손실은 파일 단위로 측정하여 제공^[부록4]하는 것을 원칙으로 함
- 노출위험과 정보손실은 상충(trade off) 관계를 가져 동시에 축소시키기 어려우므로, SDC 방법론이 정립되기 전까지는 k -익명성 위주로 관리
 - 원칙적으로 노출위험은 k -익명성 확보를, 정보손실은 최소화를 추구

4. 체계적인 운영 · 관리

☐ 반출목적에 따라 행정자료 DB^[부록5] 이용과 반출유형을 기록 · 관리

- 반출한 다양한 분석결과를 이용자가 비교 · 정리하여 뜻밖의 개별정보 노출이 발생하는 경우, 대책 마련을 위한 기본적인 조치
- [부록6]의 동일한 서식을 사용하여 일관성 있는 관리 추구

☐ 분석결과를 반출하는 목적은 다음과 같은 세 유형으로 구분

- 표본설계용이 아닌 경우, 최종 성과물만을 반출하는 것이 바람직

반출목적		
① 일반이용	② 표본설계	③ 기타
▶ 반출물을 활용해 보고서, 논문 등을 작성하는 경우 ▶ 반출물을 신규통계 공표에 사용하는 경우 등	▶ 반출물을 표본설계에 이용 하는 경우 ▶ 모집단 보완 및 구축 등 ▶ 반출물 직접 공표 금지	▶ 앞의 유형에 속하지 않는 경우 ▶ 심의위원회가 별도 관리

☐ 분석결과 반출물은 다음과 같은 다섯 종류의 형태로 구분

반출유형				
① 빈도표	② 총계표	③ 기술통계	④ 모형결과	⑤ 기타

5. 반출유형별 반출 원칙

☐ 반출목적이 ①일반이용일 때, 기본적인 유형별 반출 허용 기준

빈도표	각 셀 값이 k -익명성을 확보하면 허용
총계표	셀 값은 반올림 처리하여 허용
기술통계	▶ 속성 노출을 일으킬 수 있는 분석결과는 반출 불허 예. 최대값, 최소값, 상위 1%-분위수 등 ▶ 예외적인 반출 요청은 심의위원회에서 검토 ▶ 기타 기술통계 반출은 원칙적으로 제한 없음
모형결과	▶ 원칙적으로 제한 없으나, 직접 추론이 가능한 결과물 반출은 불허 예. 잔차 등의 마이크로데이터, 산점도, 잎-줄기 그림 등 ▶ 표본추출 및 잡음추가 처리하면, 관리자 승인 후 반출
기타	위 범주에 포함되지 않는 반출유형은 심의위원회에서 검토

□ 반출목적이 ②표본설계의 경우, 기본적인 유형별 반출 허용 기준

○ 모두 허용하나, 사용 후 파기 계약서 제출 및 파기 확인 절차 준수

□ 통계데이터센터에서 사용하는 k -익명성 기준은 다음과 같음

사업체 · 기업체 관련	인구 · 가구 관련
3	5

* 통계법 시행령 제47조 3항에 의거 사업체의 상호, 업종, 주소 및 전화번호 제공이 허용되므로, 일반적으로 최소한의 익명성 기준으로 통용되고 있는 3을 사용

* 인구 가구는 KOSIS 등에서 이미 5를 기준으로 사용 중이므로 5를 사용

k -익명성이 노출위험 측도일 때 빈도표 셀 값의 안전성 판단^[4]

- ▶ (셀 값이 0인 경우) 안전하다고 판단
- ▶ (셀 값이 0보다 크고 k 보다 작은 경우) 작은 셀(small cells)이라 부르며, 익명성을 충분히 확보하지 못해 노출위험이 있다고 판단
- ▶ (셀 값이 k 보다 크거나 같은 경우) 해당 셀에 속하는 개체는 k 이상의 익명성을 확보하므로 안전하다고 판단

빈도표에서 작은 셀이 k -의명성을 확보하는 방법^[4]

- ▶ (감추기 처리) 작은 셀을 NA 표기하여 미공표하면 해당 셀의 참값이 1부터 $(k-1)$ 중에서 어느 값인지 모르므로 $(k-1)$ -의명성 확보 (2차 감추기 추가 필요)
- ▶ (라운딩 처리) 작은 셀을 0 혹은 k 로 라운딩하여 k -의명성 확보
- ▶ (셀 병합) 셀 값이 k 이상이 되도록 주변 셀을 병합하여 k -의명성 확보

k 기준 라운딩과 k -의명성 확보

- ▶ (0으로 라운딩) 해당 셀의 참값이 0부터 $(k-1)$ 중에서 어느 값인지 모름
- ▶ (k 로 라운딩) 해당 셀의 참값이 1부터 k 중에서 어느 값인지 모름

□ 빈도표와 총계표를 함께 제공하는 경우

- 빈도표에서 k -의명성 확보 여부를 점검하고, 총계표는 반올림 처리

□ 총계표 작성 시 주의 사항

- 행이나 열에 사용한 **설명변수**와 셀의 값을 계산할 때 사용한 **산출변수가 연관성이 큰 경우**, 셀마다 k 개 이상의 개체를 포함하도록 셀 병합
- 예. 매출액 규모별 매출 총계표의 경우, 셀마다 k 이상의 빈도수를 확보

k -익명성 확보 vs. k 이상의 빈도수 확보

빈도표 참값	k -익명성 확보 : 감추기 처리								
<table><tr><td>셀 A</td><td>셀 B</td></tr><tr><td>2</td><td>1</td></tr></table>	셀 A	셀 B	2	1	<table><tr><td>셀 A</td><td>셀 B</td></tr><tr><td>NA</td><td>NA</td></tr></table>	셀 A	셀 B	NA	NA
셀 A	셀 B								
2	1								
셀 A	셀 B								
NA	NA								
▶ $k=3$ 기준 ☞ 값이 1 또는 2인 경우 작은 셀	▶ 3 미만 감추기 적용 ▶ 참값은 1, 2 중 하나 ☞ 셀은 익명성 2를 확보								
k -익명성 확보 : 라운딩 처리	k 이상의 빈도수 확보 : 셀 병합								
<table><tr><td>셀 A</td><td>셀 B</td></tr><tr><td>0</td><td>3</td></tr></table>	셀 A	셀 B	0	3	<table><tr><td>셀 A</td><td>셀 B</td></tr><tr><td colspan="2">3</td></tr></table>	셀 A	셀 B	3	
셀 A	셀 B								
0	3								
셀 A	셀 B								
3									
▶ 랜덤 라운딩 적용 (1 or 2 → 0 or 3)	▶ 인근 두 셀을 병합								
▶ 셀 A 참값은 0, 1, 2 중 하나 ▶ 셀 B 참값은 1, 2, 3 중 하나 ☞ 셀은 익명성 3을 확보	▶ (A, B) 참값은 (1,2) 또는 (2,1) ☞ 셀 값은 익명성 2를 확보 ☞ 셀에 속한 개체는 익명성 3을 확보								

제 3 장

관리자 가이드라인

1. 노출위험 및 정보손실 이해

- ☐ 관리자는 [부록1]의 사례 및 [부록3]의 유일성과 노출위험을 이해할 것
 - 분석 대상이 소지역인 경우, 식별 및 속성 노출 가능성에 특히 유의
- ☐ 관리자는 [부록4]를 통해 다양한 측면의 정보손실 발생을 이해할 것

2. 자료이용 신규신청 접수[부록6]

- ☐ (연번 부여) 반출목적에 따라 일반, 설계, 기타로 구분하여 연번 부여
 - 예. 일반201706_1, 설계201705_4, 기타201703_1 등
- ☐ (반출목적의 세부내용 검토) 이용자가 공공이익실현에 부합하는 프로젝트 진행을 위해 자료이용을 요청하였는지 검토
- ☐ (이용자료와 분석결과 구조 비교) 이용을 요청한 자료 및 변수와 도출하고자 하는 분석결과 구조를 확인하여 분석 계획의 구체성 검토
 - 이용자는 행정자료 레이아웃과 다른 변수 이름을 사용할 수 있음
- ☐ (반출유형별 세부사항) 대략적인 예상 내용 혹은 공란 접수 가능
- ☐ 신규 접수 내용을 반출 대장에 기록

3. 자료이용 수정신청 접수

- ☐ 수정사항이 최초 신청의 범위를 크게 벗어나지 않는지 검토
 - 최초 신청과 일관성이 떨어지는 경우, 이용자에게 재신청 권고
- ☐ 수정 접수 내용을 반출 대장에 기록

4. 분석결과 반출신청 접수

- ☐ 반출신청 내용이 최초 신청 내용의 범위를 벗어나지 않는지 확인
- ☐ (빈도표) 값이 k 보다 작은 셀이 없음과 각 셀의 k -익명성 확보를 확인
 - 필요한 경우, 그룹 속성 노출이 일어날 수 있는지 점검 [부록3]
- ☐ (총계표) 반올림 처리 등으로 셀의 정확한 값이 반출되지 않음을 확인
 - 필요한 경우, 총계표에 사용한 설명변수들로 작성된 빈도표를 제출받아 k -익명성 확보를 확인
- ☐ (빈도표와 총계표 동시 반출) 빈도표와 총계표를 각각 점검
 - 빈도표는 k -익명성 확보 여부를 점검하고, 총계표는 반올림 처리 확인
 - 표의 행이나 열에 사용한 변수와 셀의 값을 계산할 때 사용한 변수가 연관성이 큰 경우, 인근 셀과 병합하여 빈도가 k 이상이 되도록 처리
- ☐ (기술통계) 특정 개체를 식별하는 경우가 있는지 점검
 - 분석 대상이 되는 부분 모집단의 크기가 충분히 큰 경우, 평균 등의 일반적인 기술통계는 개별정보 노출위험이 거의 없다고 판단^[3]
 - 반출목적이 ①일반이용의 경우, 개별정보 노출 가능성이 있는 최대값 등의 기술통계는 반출 불허

- 예외적으로 반출을 요청하는 경우, 익명성을 확보하거나 잡음을 추가하여 식별이 일어나기 어려움을 확인하고 반출 허용

☐ **(모형결과)** 특정 개체를 식별하는 경우가 있는지 점검

- 분석 대상이 되는 부분 모집단이 충분히 큰 경우, 도출된 모형으로 예측한 결과에는 일반적으로 개별정보 노출위험이 거의 없다고 판단^[3]
- 서식에서 모형을 통한 개별정보 노출에 대한 이용자 의견 기재 확인

☐ 반출신청 접수 내용을 반출 대장에 기록

제 4 장

이용자 가이드라인

1. 성실한 자료 이용 신청서 작성^[부록6]

- 이용자는 ①신규신청, ②분석 대상 추가 등의 수정신청, ③분석결과 반출신청 시 동일한 서식을 사용하여 자료 이용의 일관성을 유지

자료이용 신규신청	
제목	반출하는 결과물의 사용처가 뚜렷이 드러나도록 작성
신청자정보	자료 분석자 전원을 대상으로 각각 작성
(반출목적) 사용목적	(일반이용) 신청한 자료를 분석하여 얻은 결과를 반출하여 보고서, 논문, 신규통계 등의 형태로 공표하는 경우 (표본설계) 반출하는 자료를 전혀 공표하지 않으며, 표본설계, 모집단 관리 등을 위해 내부적으로 이용하는 경우
(반출목적) 세부내용	반출하는 자료를 이용하여 수행하는 프로젝트가 공공의 이익실현에 유익함을 충분히 설명
반출유형	예상하는 분석결과 형태를 모두 선택
이용자료	사용하려는 DB, 자료명, 변수명을 기입 변수명이 센터 제공 행정자료 레이아웃과 동일할 필요는 없음
분석결과 구조	(분석 모집단) 분석결과를 얻으려는 모집단을 특정하여 기입 (설명변수) 빈도표/총계표의 경우 행과 열을 구성하는 변수 (산출변수) 빈도표/총계표의 경우 셀 값을 계산하는 변수
반출유형별 세부사항	반출결과로 예상되는 내용을 기입하며, 정확하지 않아도 됨
자료이용 수정신청	
수정사항	수정이 발생한 항목을 기입
각 항목	이용 계획이 수정된 부분을 선별하여 기입

분석결과 반출신청	
제목	반출하는 분석 결과물의 사용처가 뚜렷이 드러나도록 변경 가능
수정사항	최초 신청 내용과 달라진 사항을 기입 (별도 첨부 가능)
신청자정보	실제 자료 분석자 전원을 대상으로 각각 작성
(반출목적) 세부내용	반출 결과물을 이용하여 수행하는 프로젝트의 내용을 설명하고, 반출 결과물을 이용함으로써 창출되는 공공이익 실현 의의를 설명
반출유형	반출하는 분석결과 형태를 모두 선택
이용자료	사용하려는 DB, 자료명, 변수명을 기입 (별도 첨부 가능) 반출 파일에 사용된 변수명을 행정자료 레이아웃을 이용해 설명
분석결과 구조	반출결과 파일별로 작성 (별도 첨부 가능) (분석 모집단) 분석결과가 도출된 모집단을 특정하여 기입 (설명변수) 빈도표/총계표의 경우 행과 열을 구성하는 변수 (산출변수) 빈도표/총계표의 경우 셀 값을 계산하는 변수
추론에 의한 노출 가능성	반출 결과물이 여타 통계적 방법에 의해 개별정보 노출을 일으키기 어렵다는 의견과 근거를 제시
최종 결과물 입증	반출 결과물이 프로젝트에 사용될 최종 산출물이며, 개별정보 식별 등 다른 분석에 사용될 중간 산출물이 아님을 성실히 설명
목적 이외 사용 금지	신청 목적 이외의 용도로 반출물을 사용하는 경우, 향후 데이터센터 이용 금지 등의 조치를 감수할 것을 서약
반출유형별 세부사항	
빈도/총계표 구조	(기변수) 설명변수로 사용한 변수들을 기입 다차원 빈도/총계표의 경우, 빈도 혹은 총계를 산출한 변수를 제외하고 모든 변수를 기입 (산출유형) 셀 값의 형식을 기록하며 노출제어 기법을 명시 예. 빈도수(셀 병합), 평균(반올림), 총합(반올림) 등
기술통계	반출을 요청하는 기술통계 종류를 기입 예. 평균, 상관계수 등
모형	반출을 요청하는 모형 관련 분석결과를 설명
기타	기타 반출을 요청하는 결과물의 특징을 설명

2. 반출유형별 노출제어 처리

☐ 본 가이드라인의 통계데이터센터 SDC 기본원칙을 참고

☐ (빈도표) k -익명성을 확보하도록 처리

- 기법은 국소적 감추기 (2차 감추기 포함), 라운딩 등을 사용 가능
- 대량의 다차원 빈도표 반출은 Bounded SCA 프로그램 사용 가능
- * 통계개발원 통계적 노출제어 연구 담당자와 상의

☐ (총계표) 반올림 처리 등으로 정확한 값이 노출되지 않도록 주의

- 매출액은 백만 원 단위 제공 등으로 처리하여 반출 신청을 권고
- * 변수별 라운딩 단위는 표준화 차원에서 향후 결정
- 설명변수와 산출변수가 연관성이 크면, 모든 셀의 값이 k 이상이 되도록 셀 병합 등의 방법으로 처리하여 반출 신청

* 예. 다음과 같은 분석결과에서는 각 셀에 속하는 한 개체의 참값이 노출됨

매출액 규모 (설명변수)	...	80-90	90-100	100 이상
매출총액 (산출변수)	...	84	95	120

* 위 분석결과는 아래와 같이 셀을 병합하여 반출 신청할 수 있음

매출액 규모 (설명변수)	...	80 이상
매출총액 (산출변수)	...	299

- 노출 위험도가 반출 가능한 수준으로 낮다고 판단하기 어렵거나 추가로 검토가 필요한 경우, 동일한 설명변수로 작성한 빈도표 제출을 요구할 수 있음

☐ (빈도표와 총계표) 일반적으로 빈도표와 총계표 각각의 기준을 따름

☐ (기술통계) 원칙적으로 제한 없으나, 개별정보에 대해 직접적인 추론이 가능한 결과물* 반출은 불허

* 특정 집단의 최대값, 최소값 등

- ☐ **(모형결과)** 원칙적으로 제한 없으나, 직접적인 추론이 가능한 결과물* 반출은 불허, 익명성이 확보된 분석 결과물** 반출은 허용

* 잔차 마이크로데이터, 전체 산점도, 앞-줄기 그림 등 반출 불허

** 구간폭이 k 의 2배 이상인 히스토그램, 샘플링 및 잡음 처리한 산점도 등 반출 허용

- ☐ 반출목적이 ②표본설계이면, SDC 처리 없이 반출 가능하며 별도 관리

3. 추론에 의한 노출 가능성 검토

- ☐ 통계데이터센터 이용자는 자신의 분석 결과물이 반출 시 여타 통계적 방법에 의해 개별정보 노출을 일으키기 어렵다는 의견과 근거를 제시

- 예. 반출을 요청한 스트림 파일(*.str 형식)은 반출유형 ⑤기타에 해당하며, SPSS Modeler로 작성된 분석 알고리즘임. 수치적 데이터가 포함되지 않고 오차가 존재하여 개별정보의 정확한 노출 가능성이 희박

4. 최종 성과물 입증 (개별정보 비식별 입증)

- ☐ 통계데이터센터 이용자는 반출을 요청하는 분석 결과물이 논문, 정책 수립 등에 있어 사용될 최종 결과물이며, 개별정보 식별 등을 위한 다른 분석에 사용될 수 있는 중간 산출물²⁾이 아님을 설명

- 예. 반출을 요청하는 빈도표는 통계 공표에 사용할 최종 결과물이며, 익명성을 확보하여 개별정보 노출에 사용할 수 있는 중간 산출물이 아님

5. 이용 요청 이외 목적 사용 금지

- ☐ 의도적으로 이용 요청 이외의 목적으로 반출물을 활용하는 것을 금지

2) 이용기간 중에 반출을 요청하는 결과물을 말하는 것이 아니며, 반출물의 성격을 말함

제 5 장

향후 발전방향

1. 최신 기법 활용 기대 효과

- ☐ 매스킹의 대안으로 제시되고 있는 재현자료 및 차등정보보호 연구
- ☐ 차등정보보호 활용 기대 효과
 - 차등정보보호를 만족시키는 잡음을 추가한 결과물을 제공하면 개별정보 노출위험과 이용자의 센터 방문 없이 통계 분석결과 제공 가능 (중장기)
 - * 다만 사전에 정의된 분석기법에 대해서만 차등정보보호 처리 가능
 - 이는 물리적 노출제어를 축소하고 통계적 노출제어를 늘리는 방안 모색이 필요함
- ☐ 재현자료를 활용하여 마이크로데이터 제공 시 기대 효과
 - 캐나다 통계청 사례와 같이, 이용자가 사전에 충분히 자료를 탐색할 수 있도록 재현자료를 지원하면, 이용자 편의성 및 만족도 증진 가능
 - 센터 내에서는 마이크로데이터가 이용자에게 노출되므로, 재현자료를 활용하여 이용자의 마이크로데이터 접촉을 차단하는 방안 모색

2. 중장기 연구 방향

- ☐ 경제적·행정적 비용이 발생하는 물리적 노출제어의 비중을 줄이는 대신 자료 제공 및 분석 플랫폼을 구축하는 흐름을 대비
 - 최신 통계적 노출제어 기법 연구 및 자료제공 지원 방안 연구

3. 가이드라인 업데이트 지속

- ☐ 행정자료 DB별로 기초적인 자료 분석 및 노출위험 검토 결과 추가
- ☐ 2017년 가이드라인(안)을 베타 버전으로 운영하고, 향후 업데이트 지속

참고문헌

- 관계부처 합동(2016). **개인정보 비식별화 조치 가이드라인 -비식별 조치 기준 및 지원·관리 체계 안내-**, 관계부처 합동.
- 김경미, 이의규, 정미옥(2007). 마이크로데이터 제공에 관한 해외사례연구, **2007년 연구보고서**, 통계개발원.
- 김규영, 권대홍, 신지은, 이석훈(2011). **통계적 비밀보호방법 입문**, 자유아카데미.
- 박민정 (2014). 가계금융·복지조사 마이크로데이터 매스킹 방안 평가, **2014년 연구보고서**, 통계개발원.
- 박민정 (2015). 마이크로데이터 매스킹 실무 활용 안내서, **2015년 연구보고서**, 통계개발원.
- 박민정 (2016). 통계적 노출제어 최신동향 검토 -차등정보보호를 중심으로-, **2016년 연구보고서**, 통계개발원.
- 박민정, 권순필, 심규호 (2013). 가계금융·복지조사 마이크로데이터 제공을 위한 매스킹 방안, **2013년 연구보고서**, 통계개발원..
- Bethlehem, J. G., Keller, W. J., & Pannekoek, J.(1990). Disclosure control of microdata, *Journal of the American Statistical Association*, **85**(409), 38-45.
- Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & de Wolf, P. P.(2012). *Statistical Disclosure Control*, John Wiley & Sons Ltd.
- Meindl, B., Templ, M., & Kowarik, A.(2013). Guidelines for the Anonymization of Microdata Using R-package sdcMicro.
- NCHS(2010). *NCHS RDC Disclosure Manual: Rules for Researchers*, NCHS.
- Reiter, J. P.(2004). New approaches to data dissemination: A glimpse into the future. *Chance*, 17:3 (Summer 2004), 12 – 16.
- Shlomo, N., Antal, L., & Elliot, M.(2015). Measuring disclosure risk and data utility for flexible table generators, *Journal of Official Statistics*, **31**, 2, pp. 305-324.

부록 1 개별정보노출 가상사례

1. 센서스 빈도표에서 개별정보노출 가상사례

- 전국의 평균 500명 규모의 집계구마다, 이용자는 「성별, 각 세 연령, 교육수준, 혼인상태」, 총 4개 변수에 대해 원하는 범주를 선택하여 빈도표를 요청하고, 관리자는 5 미만의 자료는 NA 처리하여 제공할 때, 다음과 같은 노출 발생이 가능
- 해당 집계구의 남자 박사 50세가 1명이며, 이혼상태임이 알려짐

조건 선택	제공 빈도수	사용자 유추
남 + 박사 + 50~65세 + 기혼, 이혼	6	
남 + 박사 + 50~65세 + 기혼	5	
남 + 박사 + 51~65세 + 기혼, 이혼	5	
남 + 박사 + 50세 + 이혼	NA	1명

분석범위	감추기 처리		
집계구 : A 성별 : 남자 교육수준 : 박사		50세	51~65세
	기혼	0	5
	이혼	NA	0
		NA	5
			50~65세
			5
			6
6개 셀 값 제공 NA 처리된 셀 값이 1이라고 계산됨			

- 즉, 하나의 마이크로데이터에서 무수히 많은 빈도표를 만들어 낼 수 있을 때, 5 미만 NA처리와 같은 단순 감추기 처리로는 노출제어 불가능

□ 반면, 노출 방지를 위해 2차 감추기*를 하면 이용자의 개별정보 유추가 불가능하나 너무 많은 정보손실 발생

* 제공된 표의 전체합에서 부분합(margin)을 빼면 NA처리된 셀의 값을 알 수 있으므로 (아래표 감추기 처리 참조), 충분히 안전한 셀들 중 하나를 더 감추는 것을 2차 감추기 (secondary/complementary suppression)라 함

분석범위	감추기 처리			2차 감추기 처리				
집계구 : A 성별 : 남자 교육수준 : 박사		50	51~65	50~65		50	51~65	50~65
	기혼	0	5	5	기혼	0	NA	NA
	이혼	NA	0	NA	이혼	NA	0	NA
		NA	5	6		NA	NA	6
	6개 셀 값 제공 NA 처리된 셀 값이 계산됨			3개 셀 값 제공 NA 처리된 셀 값을 알 수 없음				

	감추기		2차 감추기
조건 선택	제공 빈도수	이용자 유추	
남 + 박사 + 50~65세 + 기혼, 이혼	6		6
남 + 박사 + 50~65세 + 기혼	5		NA
남 + 박사 + 51~65세 + 기혼, 이혼	5		NA
남 + 박사 + 50세 + 이혼	NA	1명	NA

□ 또한, 위 사례의 변수가 4개인 4차원 표처럼 빈도표가 다차원인 경우, 2차로 감출 셀을 결정하는 것은 매우 복잡함 (변수가 3개 이상인 빈도표의 경우, 2차 감추기를 적용할 셀을 찾기 위한 계산이 어려움)

부록 2 물리적 노출제어 vs. 통계적 노출제어

1. 바람직한 자료 제공 요건^[7]

- ☐ 응답자의 식별 정보나 민감한 정보를 알려는 외부 시도에 대한 안전성
- ☐ 광범위한 통계적 분석을 뒷받침할 수 있는 정보의 충분성
- ☐ 보편적인 통계적 방법론들을 사용하는 이용자의 접근 편의성

		안전성	정보 충분성	접근 편의성
노출위험 축소 기법 적용		O	X	O
정보손실 축소 노력		X	O	O
이용자 접근 불편함 증가		O	O	X
위의 세 요건을 모두 충족하면서 통계자료를 제공하는 방법은 없음				

2. 이용자 유형에 따른 자료 배포 전략^[8]

종류	물리적 노출제어	통계적 노출제어	
이용자	심층 이용자	불특정 다수	
전략/기법	접근제어 및 결과통제	매스킹처리	재현자료
배포형식	▶ 인가파일 ▶ 데이터센터	▶ 가공 통계표 ▶ 원격접속	▶ 공공이용파일

- 물리적 노출제어는 노출위험이 높아도 정보손실이 적은 자료 제공 중시
 - 노출위험 축소를 위한 자료의 변형이 거의 없으나 이용자 접근이 불편
- 통계적 노출제어는 이용자 편의성을 중시하여 자료를 제공
 - 노출위험을 축소하기 위해 매스킹 등의 기법으로 자료를 변형하고 제공하므로 정보손실이 발생

3. 물리적 노출제어^[2]

- 만약 이용자들이 자료를 통계적 목적으로만 이용하였고 생산된 결과에 응답자의 정보 노출위험이 없다면, 자료는 노출에서 안전
 - (가공 통계표) 이용자 요구에 따라 통계기관이 직접 통계표 작성·제공
 - (원격접속) 접속을 원격으로 제어하고 분석결과를 검열 후 제공
 - (인가파일) 이용자 소속기관과 법적 계약 체결 후 파일 제공·감독
 - (데이터센터) 이용자의 방문 분석 및 통계기관의 반출 결과물 검열

미국 센서스국 데이터센터 (Research Data Center) 운영 사례

- 9개의 센터를 운영, 이용자는 RDC를 방문하여 필요한 작업을 완결
 - 추가 분석 시 센터 재방문, 최종 성과물은 심사위원회의 승인 후 학술지 투고 허용
 - 자료 이용 제안서 제출부터 분석결과 반출까지 총 6~12개월이 소요될 것을 예상
-

- 이용자를 규제하더라도, 분석결과를 통해 발생할 수 있는 개별정보노출 방지를 위해서는 자료제공기관이 SDC 기법을 사용해 결과물 검토 필요

4. 통계적 노출제어

- 전통적으로 공공이용파일 제공을 위해 매스킹 기법을 적용^{[4][8][9][10]}
 - (매스킹) 원래 자료를 적절히 변환하여 간접적인 식별 정보를 가리고, 변환에 관한 정보를 변환된 자료와 함께 제공하는 과정

- (마이크로데이터 마스크) suppression, aggregation, noise addition, swapping, sampling 등의 세부 기법을 활용
- (매크로데이터 마스크) suppression, rounding, swapping 등을 활용

☐ 마스크는 노출위험과 정보손실을 동시에 축소하지 못해 대안 연구 활발

☐ 최근 마스크 기능 탑재 분석서버, 재현자료, 차등정보보호 등 연구^[6]

- 분석 플랫폼을 통한 자료 제공에 적절한 SDC 방안이 연구되고 있으며, 관련 연구 결과가 나오면 데이터센터에서도 활용이 가능할 것으로 기대

부록 3

유일성과 노출위험

1. 유일성

- 마이크로데이터에서 유일성은 작성된 빈도표의 셀 값이 1일 때 발생
- 다음과 같은 4개 변수를 가지는 마이크로데이터가 있을 때, 이를 빈도표로 바꾸어 셀의 값이 1인 경우는 유일한 개체라는 의미

마이크로데이터^[11]

ID	H	A	M	S
1	20	8	1	2
2	3	5	2	1
...	...	...	...	...
8399	17	2	1	2

빈도표 (frequency table)

(H=1, A=1)	S=1	S=2	
M=1	$n_{1,1,1,1}$	$n_{1,1,1,2}$	$n_{1,1,1,\bullet}$
M=2	$n_{1,1,2,1}$	$n_{1,1,2,2}$	$n_{1,1,2,\bullet}$
	$n_{1,1,\bullet,1}$	$n_{1,1,\bullet,2}$	$n_{1,1,\bullet,\bullet}$
...			
(H=24, A=12)	S=1	S=2	
M=1	$n_{24,12,1,1}$	$n_{24,12,1,2}$	$n_{24,12,1,\bullet}$
M=2	$n_{24,12,2,1}$	$n_{24,12,2,2}$	$n_{24,12,2,\bullet}$
	$n_{24,12,\bullet,1}$	$n_{24,12,\bullet,2}$	$n_{24,12,\bullet,\bullet}$

- 4차원을 표현하기 위해 여러 개의 2차원 빈도표를 사용하는 대신, 다음과 같이 펼쳐진/편평한 형태의 빈도표 작성 가능

빈도표 (flattened frequency table)

(H,A,M,S)	(1, 1, 1, 1)	(1, 1, 1, 2)	...	(24, 12, 2, 2)
frequency	$n_{1,1,1,1}$	$n_{1,1,1,2}$	...	$n_{24,12,2,2}$

빈도표 (flattened frequency table)

H	A	M	S	frequency
1	1	1	1	$n_{1,1,1,1}$
1	1	1	2	$n_{1,1,1,2}$
...	...	...	...	...
24	12	2	2	$n_{24,12,2,2}$

- 빈도표에서 빈도수가 1인 개체를 유일하다고 하며, 노출위험이 최대
- 빈도수가 2인 경우, 같은 셀에 속하는 두 개체 중에서 한 개체가 다른 개체를 식별하므로 노출위험이 높다고 판단

2. k -익명성 및 l -다양성^[12]

- (k -익명성) 주요 변수의 각 구별 패턴이 표본 내에서 적어도 k 개 이상이면 k -익명성을 획득한다고 정의
- 다음 자료에서 두 개의 키변수가 구별 패턴을 나타낸다고 가정하면, 1~3번 3개 개체는 키변수들이 모두 (1,1)이라는 동일한 값으로 동일한 패턴을 보여 익명성 3을 획득
 - 5~6번 2개 개체는 익명성 2를 획득하고, 4번 개체는 유일성을 보임
- (l -다양성/포괄성) 같은 패턴을 갖는 표본들이 민감 변수에서 적어도 l 개의 자주

등장하는 값을 가지면 l -다양성을 획득한다고 정의

- 1~3번 개체는 민감변수(질병번호)에서 서로 다른 2개의 값을 가지므로 다양성 2를 획득
- 5~6번 개체는 익명성 2를 확보하였으나, 동일한 질병번호를 가지므로 다양성을 획득하지 못하여 민감한 정보가 노출됨
- l -다양성 미확보는 특정 그룹의 속성이 노출되는 것이기도 함

마이크로데이터				빈도표		
ID	키변수 1	키변수 2	민감변수	키변수 1	키변수 2	빈도수
1	1	1	50	1	1	3
2	1	1	50	1	2	1
3	1	1	42	2	2	2
4	1	2	42			
5	2	2	62			
6	2	2	62			

k -익명성 및 l -다양성					
ID	키변수 1	키변수 2	민감변수	k -anonymity	l -diversity
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

R패키지 sdcMicro를 이용한 k -익명성 및 l -다양성 측정

```
# microdata 작성
key1 <- c(1, 1, 1, 1, 2, 2)
key2 <- c(1, 1, 1, 2, 2, 2)
sens1 <- c(50, 50, 42, 42, 62, 62)
microdata <- cbind(key1, key2, sens1)

library(sdcMicro)
xx <- data.frame(microdata)
# microdata 자리에 행렬 형태의 마이크로데이터 변수명을 입력
anony <- freqCalc(xx, keyVars=1:2, w=NULL)
# 1:2 자리에 키변수 열 번호 입력
diverst <- ldiversity(xx, keyVars=c("key1", "key2"), ldiv_index="sens1")
# c("key1", "key2") 자리에 키변수 이름 벡터,
# "sens1" 자리에  $l$ -다양성을 알고 싶은 민감변수 이름을 입력
cbind(xx, anony$fk, diverst[,1])
# 마이크로데이터와  $k$ -익명성 및  $l$ -다양성을 출력
```

3. 노출위험

□ 현재는, 노출위험은 k -익명성을 확보하면 0인 것으로 간주

- (빈도표에서 셀 값이 k 보다 크거나 같은 경우) 해당 셀에 속하는 개체들이 k 개 이상이므로 k -익명성을 확보
- (값이 0인 셀은 공표하면서 작은 셀에는 감추기 기법을 적용하는 경우) 빈도표에서 k 미만은 NA 표기하면 해당 셀은 익명성 ($k - 1$) 을 확보
- 예를 들어 $k=5$ 일 때, 제공되는 셀 값이 NA이면 해당 셀은 1~4의 값을 가진다는 것을 의미하여 익명성 4를 확보
- (값이 0인 셀은 공표하면서 작은 셀에는 라운딩 기법을 적용하는 경우) 빈도표에

서 k 미만은 0 또는 k 로 라운딩해 해당 셀은 익명성 k 를 확보

- 예를 들어 $k=5$ 일 때 제공되는 셀 값이 0이면, 해당 셀이 0~4의 값을 가지는 것이므로 익명성 5를 확보. 또한, 제공되는 셀 값이 5이면 해당 셀이 1~5의 값을 가진다는 것을 의미하여 익명성 5를 확보

□ 분석범위의 규모가 충분히 큰 경우, 평균 등의 기술통계량이나 회귀 계수 등의 모형을 적용한 결과물은 오차를 가지는 대푯값이므로 익명성을 확보한 것으로 간주

□ (참고) 마이크로데이터의 개체마다 노출위험을 재는 측도(measure)*가 연구되어 있으나, 매크로데이터 형식인 분석결과물의 노출위험 측정에는 적절치 않아 배제

* 유일성 개념을 기초로 모형을 제안하거나 외부인(intruder)에게 의사결정론을 적용하여 개체별 노출위험을 측정하는 통계적 이론들^[13]

부록 4 정보손실

1. 파일 단위의 정보손실 측도(measure) 제공

□ (빈도표) 제공된 빈도값과 실제 빈도값의 차이

정보손실	0	1	2
%	93.57	4.09	2.34

총 2,500개의 빈도표를 제공했으며, 참값과 차이가 1인 셀이 4.09%, 2인 셀이 2.34%

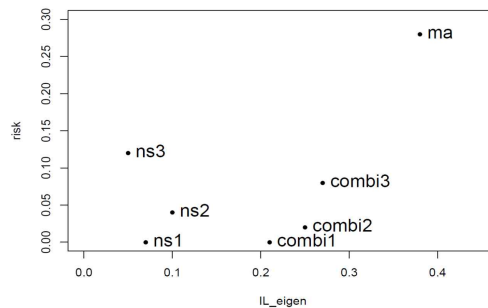
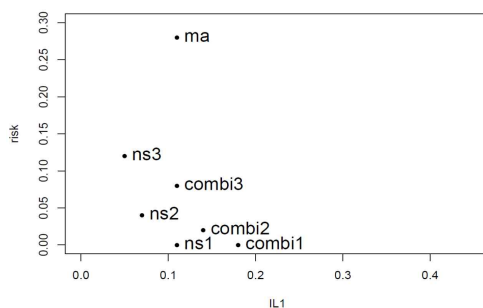
□ (총계표) 변수별 반올림 단위 제공

- 예. 매출액은 백만 원 단위 반올림 처리

2. 다양한 정보손실 측도

□ 정보손실은 ①실제 자료 값의 변화, ②자료 구조의 변화 ③ 주요 통계량의 변화 등, 다양한 측면에서 측정 가능^[13]

- 여러 방법으로 매스킹 처리하고, 각 방법의 정보손실 측정 가능
- 아래 왼쪽 그림은 매스킹 전후 자료값 사이의 거리, 오른쪽 그림은 매스킹 전후 공분산 구조를 표현하는 통계량 사이의 거리를 나타냄
 - * ma, ns1, combi1 등은 다양한 매스킹 시나리오를 나타냄
- 즉, 동일한 매스킹 처리에 대해 정보손실에 대한 판단이 다를 수 있음



부록 5

통계데이터센터 행정 통계자료 목록

□ 센터는 다음과 같은 9종의 행정 통계자료를 관리함

행정 통계자료	
경제·사회 분야 (사업체·기업체 관련)	① 기업등록부 DB(총괄)
	② 기업등록부 사업장 기초 DB(4대보험)
	③ 기업등록부 사업장 기초 DB(법인)
인구·가구 분야	⑫ 등록센서스 인구 DB
	⑬ 등록센서스 가구 DB
	⑭ 등록센서스 주택 DB
농림·어업 분야	⑮ 농업 DB
	⑯ 어업 DB
	⑰ 임업 DB

부록 6-1 자료이용 신청서식 (□ 신규, □ 수정)

※ 해당 항목에 ■ 또는 V 표시

연번	(예) 일반201706_5, 설계201706_4, 기타201706_1
제목	(예) '문화체육관광산업통계' 작성을 위한 행정자료 이용

※ 연번은 통계데이터센터 관리자가 작성

신청자 정보			
번호		이용기간	
성명		휴대전화	
소속		직위/직급	
ID(신청)			

※ 자료 분석자 인원수만큼 위의 신청자 정보 표를 작성

수정사항
(예) 자료 분석자 추가, 이용 DB 추가, 분석 변수 추가, 반출유형 추가

※ 수정 및 반출 신청 시 작성, 별도 첨부 가능

반출목적	
사용목적	☐ 일반이용 ☐ 표본설계 ☐ 기타
세부내용	(예) 행정자료 집계 결과를 활용하여 문화관광연구원에서 '문화체육관광산업통계'를 작성

※ 세부내용에는 분석결과를 활용하여 수행하고자 하는 프로젝트에 대해 상세히 작성

반출유형					
(예)	빈도표	총계표	기술통계	모형결과	기타
유형 선택	√	√	-	-	-

이용자료		
DB명	자료명	변수명

※ 최종 반출 신청 시, 반출결과의 변수명을 센터에서 제공하는 행정자료 레이아웃으로 설명하여 작성

분석결과 구조		
분석범위 (분석 대상 부분 모집단)	설명변수 (키변수)	산출변수 (민감변수, 반응변수)

◎ 반출유형별 세부 사항

빈도표/총계표 구조					
키변수1	키변수2	키변수3	키변수4	산출유형1	산출유형2
(예) SD_CD	FRGR_YN	-	-	빈도수	합계

기술통계
(예) 분석 대상 부분 모집단마다 주요 변수별 평균, 상관계수 및 왜도

모형
(예) 분석 대상 부분 모집단마다 선형회귀모형을 적용하여 얻은 회귀계수

기타
(예) SPSS Modeler 스트림 파일/ *.str 형식

부록 6-2 분석결과 반출 신청서식

연번	(예) 일반201706_5, 설계201706_4, 기타201706_1
제목	(예) '문화체육관광산업통계' 작성을 위한 행정자료 이용

※ 신규 신청 시 부여받은 연번을 기록

신청자 정보			
번호		이용기간	
성명		휴대전화	
소속		직위/직급	
ID(신청)			

※ 자료 분석자 인원수만큼 위의 신청자 정보 표를 작성

수정사항
(예) 자료 분석자 추가, 이용 DB 추가, 분석 변수 추가, 반출유형 추가

※ 수정 및 반출 신청 시 작성, 별도 첨부 가능

반출목적	
사용목적	☐ 일반이용 ☐ 표본설계 ☐ 기타
세부내용	(예) 행정자료 집계 결과를 활용하여 문화관광연구원에서 '문화체육관광산업통계'를 작성

※ 세부내용에는 분석결과를 활용하여 수행하고자 하는 프로젝트에 대해 상세히 작성

반출유형					
(예)	빈도표	총계표	기술통계	모형결과	기타
유형 선택	√	√	-	-	-

이용자료		
DB명	자료명	변수명

※ 최종 반출 신청 시, 반출결과물의 변수명을 센터에서 제공하는 행정자료 레이아웃으로 설명하여 작성

분석결과 구조		
분석범위 (분석 대상 부분 모집단)	설명변수 (키변수, 설명변수)	산출변수 (민감변수, 반응변수)

추론에 의한 노출 가능성 검토
(예) 반출을 요청하는 빈도표는 k-익명성을 확보하도록 처리되었고, 이는 실제 참값을 추론하는 경우 언제나 참값 후보가 k개 이상이 됨을 의미. 즉, 추론에 의해 정확한 참값이 노출될 가능성은 $1/k$ 이하로 낮음 (예) 반출을 요청한 회귀 계수는 오차를 가지는 추정값이며, 시도 단위로 계산되어 모집단의 크기가 충분히 큼. 따라서 분석을 통해 도출된 모형으로 예측한 특정 개체의 소득이 정확히 참값이라 할 수 없음 (예) 반출을 요청하는 잔차 산점도는 회귀분석 결과 타당성을 입증하기 위해 논문 에 수록할 것이나, 10,000개 자료에서 20% 표본 추출하여 5%의 잡음을 추가한 것으로 어떤 개체의 정보를 정확히 노출시킬 수 없음

최종 성과물 입증 (개별정보 비식별 입증)
(예) 반출을 요청하는 회귀 분석결과를 정책 보고서에 사용될 수치이며, 개별정보 식별 등에 사용할 수 있는 중간 산출물 형식의 데이터가 아님 (예) 반출을 요청하는 빈도표는 '문화체육관광산업통계'를 공표하는데 사용할 최종 결과물이며, 익명성을 확보하여 개별정보 노출에 사용할 중간 산출물이 아님

※ 반출 결과물이 공표를 위한 최종 성과물이며, 다른 분석을 위해 사용하는 중간 산출물이 아님을 설명

이용 목적 이외 사용 금지 서약
신청 목적 이외의 용도로 반출물을 사용하는 경우, 향후 데이터센터 이용 금지 등의 조치를 감수할 것을 서약

◎ 반출유형별 세부 사항

빈도표/총계표 구조					
키변수1	키변수2	키변수3	키변수4	산출유형1	산출유형2
(예) SD_CD	FRGR_YN	-	-	빈도수	합계

※ 키변수(설명변수)는 4개까지 허용하며, 4개 이상의 변수는 분석범위로 처리

기술통계
(예) 분석 대상 부분 모집단마다 주요 변수별 평균, 상관계수 및 왜도

모형
(예) 분석 대상 부분 모집단마다 선형회귀모형을 적용하여 얻은 회귀계수

기타
(예) SPSS Modeler 스트림 파일/ *.str 형식

◎ 다차원 빈도표 반출 요청을 위한 펼쳐진/편평한 빈도표 작성 방법

※ 반출 테이블을 분석범위별로 펼쳐진/편평한 테이블(flattened table) 형식으로 기재하여 요청하나, 실제 반출 시 형태 변경 가능. (예) 아래 두 표는 같은 형식으로 간주

※ 관리자 및 이용자는 펼쳐진 테이블의 빈도수를 확인하면 익명성 확보를 한눈에 파악

반출 신청

반출 결과물

변수 1	변수 2	변수 3	산출 유형
성별	혼인 상태	교육 수준	빈도 수



성별 = 남성		
	혼인1	혼인2
교육1	100	100
교육2	100	100

성별 = 여성		
	혼인1	혼인2
교육1	100	100
교육2	100	100

※ R 및 SAS 프로그램

R을 이용해 flattened table을 작성하는 방법

```
fable(microdata[, c(1,2,3)]
# 빈도표를 작성하려는 마이크로데이터, microdata에서 설명변수 열 번호를
  선택
# ?fable 라고 입력하여 도움말을 참조
# aggregate 명령어도 사용 가능. ?aggregate 라고 입력하여 도움말 참조
```

SAS을 이용해 flattened table을 작성하는 방법

```
proc freq data = microdata;
  table a*b*c/ out=newresult;
run;
# 빈도표를 작성하려는 마이크로데이터, microdata로 freq 프로시저 실행
# 빈도표를 작성하려는 설명변수 이름들을 *로 연결해 입력
# 빈도표 결과를 저장할 변수를 newresult에 입력
```

부록 7 익명성 확보를 위한 빈도표 SDC 기법

1. 빈도표 재디자인(Re-design)^[4]

- 작은 셀들이 많아 노출위험이 높은 경우, 설명변수 범주를 병합하여 빈도표의 행과 열을 다시 구성하는 방법
- 마이크로데이터 매스킹 기법 중 전반적 재코딩(global recoding)과 유사한 개념
 - 마이크로데이터 재코딩은 특정 범주의 빈도가 유난히 낮지 않도록 사용
 - 이는 빈도수가 낮은 범주가 있는 경우, 그 범주에 속하는 개체들을 다양한 방법들로 더 분석할 때 유일성이 높아지는 현상을 완화하기 위함
 - 반면, 빈도표 재디자인은 작성된 빈도표의 노출위험을 줄이기 위함
 - (예) 직업 변수에서 의사 및 간호사 범주를 의료계 종사자로 병합 등

2. 셀 감추기(suppression)^[4]

- 작은 셀들을 감추고 정보를 제공하지 않는 방법, 1차(primary) 감추기
 - 노출위험 기준이 k -익명성 확보일 때, k 보다 작은 값을 가지는 셀들을 NA 처리하여 해당 셀의 익명성을 확보하는 방법
 - (값이 0인 셀을 NA 처리) NA 처리된 작은 셀은 $0 \sim (k - 1)$ 중에 참값을 특정할 수 없으므로 k -익명성을 확보
 - (값이 0인 셀을 공표) NA처리된 작은 셀은 $1 \sim (k - 1)$ 중에서 참값을 특정할 수 없으므로 $(k - 1)$ -익명성을 확보.
 - 이를 1차 감추기(primary cell suppression)라고 부름
- 주변합에서 안전한 셀들을 빼서 작은 셀 값을 계산해 낼 수 있으므로, 안전한 셀들을 대상으로 추가로 셀 감추기를 시행할 필요

- 2차 셀 감추기(secondary/complementary cell suppression)라고 부름

* [부록 1]의 가상사례를 참조

3. 라운딩(rounding)^[4]

- 주어진 값(base)를 기준으로 빈도값을 올림/내림/반올림 처리하는 방법

- 기준값이 5이면 작은 셀은 1~4가 되며, 모든 빈도를 0, 5, 10, ... 등

- 확률적으로 라운딩 값을 결정하면 랜덤 라운딩. 총합의 기댓값을 유지

원래 빈도값	랜덤 라운딩	
1	2/3 확률로	0
	1/3 확률로	3
2	1/3 확률로	0
	2/3 확률로	3

- 주변합이 셀들을 직접 더한 계산 결과와 일치하도록 선형계획법(linear programing)을 사용하면 제어된 라운딩(controlled rounding)

- 계산이 복잡하고 해가 항상 존재하는 것은 아님

라운딩 적용 예제								
원래 빈도표			라운딩					
1	0	1	1 → 0	0	1 → 0			
3	3	6	3 → 5	3 → 5	6 → 5			
12	20	32	12 → 10	20	32 → 30			
16	23	39	16 → 15	23 → 25	39 → 40			
랜덤 라운딩			제어된 라운딩					
1 → 0	0	1 → 0	1 → 5	0	1 → 5			
✓ 3 → 5	✓ 3 → 0	6 → 5	3 → 0	3 → 5	6 → 5			
12 → 10	20	32 → 35	12 → 10	20	32 → 30			
16 → 15	23 → 20	39 → 40	16 → 15	23 → 25	39 → 40			

* B = 5일 때 3의 random rounding : 2/5 확률로 0으로 변환, 3/5 확률로 5로 변환

4. 작은 셀 조정(small cell adjustments, SCA)^{[4][5]}

- 빈도표 총합의 기댓값을 유지하도록 작은 셀들에 대해 랜덤 라운딩 처리하고, 주변합을 계산하여 전체 빈도표를 완성하는 방법
 - 영국 ONS에서 2001년 센서스 빈도표 공표에 다른 기법들과 함께 적용함
- SCA를 적용한 빈도표는 작은 셀들이 모두 k -익명성을 획득하여 안전하고, 총합의 정보손실은 k 이하로 기대되어 효율적인 노출제어 방안임
- SCA를 적용한 빈도표를 다시 집계하여 상위 수준의 빈도표를 만들 때, 정보손실이 커질 수 있음
 - (예) 100개 지역 빈도표를 합쳐서 상위 지역의 빈도표를 만들면, 정보손실은 확률적으로 0부터 $100 \times k$ 까지 발생할 수 있음
 - * 실제로 SCA를 사용하는 프로그램을 작성하면 $100 \times k$ 와 같은 극단적인 정보손실을 어느 정도 완화하는 방법이 있으나 (그럼에도 상위 지역 총합에서 정보손실이 여전히 수용하기 어려운 수준으로 존재), 여기서는 이해를 위해 원리를 설명함
- SCA 적용 예로 총 100개 구에서 다음과 같은 빈도표 작성을 가정
 - 100개 시군구에서 빈도표가 동일하여 4개 셀은 0, 1, 2, 3의 값을 나머지 셀은 모두 4의 값을 가져서 각 시군구의 빈도표 총합은 182이고, 익명성 3을 기준으로 노출위험을 측정한다고 가정
 - 총 5개의 시도가 각각 20개 시군구를 가진다고 가정
 - 총 100개 시군구 빈도표 각각에 SCA 처리하여 작은 셀에 3 기준 랜덤 라운딩을 적용하면, 셀 값이 1 또는 2인 경우 랜덤하게 0 또는 3으로 변환되며 각 빈도표 총합은 179~185의 값을 가짐
 - 시군구 빈도표마다 총합의 정보손실은 3 이하로 정보손실이 적음
 - 반면, 각 시도별 빈도표 총합의 정보손실은 0에서 60(3×20)까지 발생

SCA 적용 예제											
1번 구						i 번 구					
번호	A	M	S	빈도 수 참값	SCA 처리	번호	A	M	S	빈도 수 참값	SCA 처리
1	1	1	1	0	0	1	1	1	1	0	0
2	1	1	2	1	0	2	1	1	2	1	0
3	1	2	1	2	3	3	1	2	1	2	0
4	1	2	2	3	3	4	1	2	2	3	3
...	...	...	...	...	...	...	...	...	...	...	...
48	12	2	2	4	4	48	12	2	2	4	4
합				182	182	합				182	179
						...					
j 번 구						100번 구					
번호	A	M	S	빈도 수 참값	SCA 처리	번호	A	M	S	빈도 수 참값	SCA 처리
1	1	1	1	0	0	1	1	1	1	0	0
2	1	1	2	1	0	2	1	1	2	1	3
3	1	2	1	2	3	3	1	2	1	2	3
4	1	2	2	3	3	4	1	2	2	3	3
...	...	...	...	...	...	...	...	...	...	...	...
48	12	2	2	4	4	48	12	2	2	4	4
합				182	182	합				182	185

5. 정보손실이 제한된 작은 셀 조정(Bounded SCA)

- ☐ 영국 ONS의 작은 셀 조정은 상위 지역 수준에서 정보손실을 많이 발생시키는 경향이 있으므로 이를 해결할 필요
- ☐ 하나의 마이크로데이터에서 매우 많은 빈도표를 생성해야 하는 경우, 노출위험 제어를 위해 k -익명성을 확보하고 정보손실을 항상 작게 유지하기 위한 알고리즘을 개발 (2016년 통계개발원)
 - 현재 학계의 의견을 수렴하는 과정을 진행 중
- ☐ 필요한 경우 통계데이터센터에서 현재 사용 가능

부록 8 용어 설명

기술통계	영어로는 descriptive statistics로 표현되며, 특정 모집단의 특징을 나타내는 통계들을 일컫는 용어이다. 평균, 총합, 분산 등의 통계를 예로 들 수 있다.
마이크로 데이터	데이터를 구성하는 각 개별 자료들을 쌓아놓은 집합으로 행렬 형식으로 표현하면 열에는 각 변수들이, 행에는 응답 개체들이 나열된 것으로 생각할 수 있다.
물리적 노출제어	SDC는 주로 자료를 변형하여 노출위험을 제거하므로 정확한 정보를 필요로 하는 심층 이용자가 SDC 처리된 자료를 이용하는 것은 적절하지 않다. 원자료 이용을 허락하고 이용자를 규제하는 전략을 물리적 노출제어라 부르도록 한다.
매스킹	대부분 통계기관에서 매스킹은 자료에서 개체가 식별되는 것을 방지하기 위한 여러 기법과 그러한 처리과정에 관련된 정보제공 행위를 통칭하는 용어로 사용된다. 쉬운 사례로, 연령 변수의 각 세 단위를 5세 단위로 조정하여 마이크로데이터를 작성하고, 자료의 연령은 실제 연령이 아니라 5세 단위로 조정되었음을 공지하는 것까지 매스킹 처리라 한다.
매크로 데이터	보통 마이크로데이터를 집계하여 도출된 통계표 등을 매크로데이터라고 한다.
분석범위	특정 분석을 시행할 때 대상이 되는 부분 모집단을 정의할 수 있는데, 본 가이드라인에서는 편의상 이를 분석범위 혹은 분석대상 부분 모집단이라 표현하였다. 예를 들어, 시도별로 회귀추정식을 계산하는 경우, 분석범위는 시도가 된다.
산출변수	빈도표/총계표에서 셀의 값을 계산할 때 사용하는 변수, 다차원 표에서 빈도 혹은 총합을 구할 때 사용하는 변수, 모형에서 반응변수에 해당하는 변수를 산출변수라 표현하였다.

설명변수	빈도표/총계표에서 행과 열을 구성하는 변수, 다차원 빈도표에서 키변수를 구성하는 변수, 모형에서 설명변수로 사용하는 변수들을 설명변수라 표현하였다.
속성	어떤 변수 혹은 개체의 특징을 속성이라 표현한다.
총계	영어로 magnitude를 총계라 표현하였으며, 규모 등으로 번역되어 사용되기도 한다. 평균, 총합 등 자료의 규모를 나타내는 집계결과를 총계라 표현하였다.
키변수	키변수는 데이터베이스에서 사용하는 용어로, 데이터베이스에서 각 레코드(개체)의 특징을 나타내는 변수들을 지칭한다.
통 계 적 노출제어	영어로는 Statistical Disclosure Control 또는 Limitation으로 표현된다. 국내에는 통계적 노출제어, 노출조절, 노출관리 등으로 번역되어 사용되고 있다. 비식별화 조치 역시 통계적 노출제어의 한 종류로 볼 수 있다. 영어로 Confidentiality라고 표현되는 비밀보호라는 용어도 통계적 노출제어의 개념을 표현하기 위해 사용되고 있다. 참고로 노출제어 등의 낱선 용어 대신, 통계적 비밀보호, 비식별화 등의 표현이 사용될 수 있다고 판단된다. 통계적 노출제어는 전통적인 매스킹 기법을 포함하여 개별정보 노출을 억제하기 위해 배포하는 자료에 통계적인 조치를 취하는 행위들을 통칭하는 개념으로 이해된다. 이는 자료에서 개별정보 노출위험을 검토하고, 노출위험과 정보손실을 축소시키기 위한 모든 노력을 포함한다. 자료 자체에 행해지는 기법들을 포괄한다는 점에서 보안 강화 등의 공학적 조치와 다르다. 통계를 공표하면서 자료를 통한 개별정보 노출을 방지하기 위한 과정을 보통 통계적 노출제어(SDC)라 부르도록 하겠다.
NA	Not Available의 약자
output	본문에서 분석결과, 반출 결과물, 반출물 등으로 표현되었다.
RDC	Research Data Center의 약자
RAS	Remote Access System/Service의 약자

집필진

- 박민정 (통계청 사무관)
- 박영옥 (통계청 통계서비스정책관 과장)
- 황영자 (통계청 통계교육원 교육운영과 서기관)

연구보고서 2020-19

통계데이터센터 반출 결과물의 통계적 노출제어 가이드라인(안)

인 쇄	2021년 4월 12일
발 행	2021년 4월 13일
발 행 인	통계개발원장 전영일
발 행 처	통계청 통계개발원 35220 대전광역시 서구 한밭대로 713 Tel.(042)366-7100 Fax.(042)366-7123
홈페이지	http://sri.kostat.go.kr
ISSN(Online)	2733-4120





통계청
통계개발원