

의안번호	제2023-07호
상 정 년 월 일	2023. 6. 13. [통계데이터분과위원회]

통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인 작성

제 출 자	통 계 청 장
제출년월일	2023. 6. 13.

목 차

 요약	1
 【붙임1】 통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인	 5
 【붙임2】 통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인 부록	 71

통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인 작성 보고 [요약]

1. 추진개요

- (배경) 유럽 일반개인정보보호규정(EU GDPR, '16년 제정, '18년 시행) 및 국내 관련 법규정에 대응하여 통계 작성 및 통계자료 제공 업무를 수행하는 국내 통계작성기관*을 위한 비식별화 가이드라인 제공 필요
 - * 정부부처, 지자체 등 1,306종의 통계를 작성하는 435개 관련 기관('23.5. 기준)
- 우리나라의 분산형* 통계생산 구조 및 통계 업무 특수성을 감안하여 통계작성기관이 비식별화를 편리하게 수행하고 개별 정보는 안전하게 보호하기 위한 가이드라인 제작 필요
 - * 단일 기관이 중앙집중형으로 통계를 생산하는 것이 아니라, 각각의 기관에서 관련 영역의 통계를 생산하여 제공하는 구조
- (목적) 통계작성기관이 통계법에 근거하여 국가승인통계를 작성 및 공표할 때와 자료 분석 등 이용자 요청에 따라 통계자료를 제공할 때 필요한 비식별화를 지원하기 위해 작성
 - ※ 통계작성기관 사전 의견수렴('22.12.) 및 전문가 자문('23.5.) 결과 반영

2. 주요내용 - 본문

① 개요

- 비식별화 가이드라인의 목적 및 적용대상
- 통계 작성 및 통계자료 제공 각 과정에서 비식별화 단계 도식
- 용어 정리
- 개념 정리

② 비식별화 방법

- 비식별화 평가 측도(노출위험 및 정보손실 측도, 측도의 의미 등)
- 마이크로데이터 비식별화 방법
- 매크로데이터 비식별화 방법

③ 통계 작성 단계별 비식별화

- 기획, 자료 수집, 내검·정제, 통계 작성, 공표·보관 등, 통계 작성 각 단계에서의 고려 사항

④ 통계자료 제공단계별 비식별화

- 제공용 통계자료 생성
- 이용 신청에 따른 통계자료 제공
- 통계데이터센터 분석결과 반출

⑤ 보유 정보의 안전한 관리를 위한 보호조치

3. 주요내용 - 부록

① 노출위험 평가 측도

- 자료주체의 유일성, 특이정보, k -익명성, ℓ -다양성 및 재식별위험 평가측도에 관한 설명

② 정보손실 평가 측도

- 범주형 항목 및 연속형 항목에 관한 설명

③ 비식별화 세부기법 해설

- 구조적 방법, 삭제 방법, 일반화 방법, 임의화 방법, 매크로데이터 비식별화 방법, 재현자료, 일방향 암호화에 관한 설명

4. 향후 추진계획

- 통계작성기관을 대상으로 가이드라인 본문 및 부록 책자 배포(6월)
 - 통계작성기관의 비식별화 교육 및 자료처리 업무 수행을 지원하기 위하여 「비식별화 가이드라인 프로그램」 개발('24년) 추진
 - (형식) 엑셀을 인터페이스*로 이용하고 실제 분석은 R을 이용*하여 마이크로데이터를 위한 비식별화 세부기법을 수행하는 프로그램
 - * 유럽연합은 PEFA tools 2017을 개발·보급하여 환경경제계정 작성 가이드라인 및 통계 생산에 활용, 비식별화 가이드라인 프로그램 개발에 이를 벤치마킹
 - * 통계청 통계개발원 '마이크로데이터 매스킹 기법 실무 활용 안내서('15년)' 및 '통계작성기관 마이크로데이터 제공범위 설정 가이드라인 작성('18년) 참고
 - (개발 대상) 현재 책자의 마이크로데이터 비식별화 세부 기법 위주
 - (특화 기능) 각 기법 적용 전후 노출위험 및 정보손실 측정 결과를 제공하여 담당자의 비식별화 자료처리 의사결정을 지원
- 통계작성기관에서 가이드라인을 참고하여 비식별화 방안을 마련하고, 프로그램을 활용하여 비식별화 처리 업무를 원활히 수행할 수 있도록 교육 활동 등 강화('25년~)

통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인

2023. 5. 30.



통 계 청
통계데이터기획과

목 차

I. 개요	9
1. 목적	10
2. 적용 대상	10
3. 비식별화 절차	11
4. 용어 정리	13
5. 개념 정리	18
II. 비식별화 방법	21
1. 비식별화 평가 측도	22
2. 마이크로데이터 비식별화 방법	26
3. 매크로데이터 비식별화 방법	38
III. 통계 작성 단계별 비식별화	44
1. 기획	45
2. 자료 수집	46
3. 내검 및 정제	48
4. 통계 작성	48
5. 공표 및 보관	48
IV. 통계자료 제공 단계별 비식별화	49
1. 제공용 통계자료 생성	50
2. 이용 신청에 따른 통계자료 제공	56
3. 통계데이터센터 분석결과 반출	62
V. 보유 정보의 안전한 관리를 위한 보호조치	65
1. 관리적 보호조치	66
2. 기술적 보호조치	66
3. 물리적 보호조치	67
[참고문헌]	68



I . 개 요



1 목 적

- 본 비식별화 가이드라인은 통계작성기관이 통계법에 근거하여 국가승인통계를 작성 및 공표할 때와 자료 분석 등 이용자 요청에 따라 통계자료를 제공할 때 필요한 비식별화를 지원하기 위해 작성됨

본 가이드라인에서 “비식별화”란, 통계 및 통계자료에서 특정의 개인이나 법인 또는 단체 등을 식별할 수 없도록 하는 각종 가명·익명처리를 말함

- 본 비식별화 가이드라인은 통계 작성부터 통계자료 이용 신청에 따른 제공까지 세부 단계별로 사례를 제공하고 보유한 정보의 안전한 관리 방안*을 안내하여, 통계작성기관이 비식별화를 편리하게 수행하고 정보를 안전하게 보호하는 것을 목표로 함

* 개인정보 노출위험 제어, 자료 유용성 및 이용자 편의성을 함께 고려해야 하는 통계작성기관의 특성에 적합한 방안을 지속적으로 개선·개발할 필요

2 적 용 대 상

- 본 비식별화 가이드라인*은 다음과 같은 절차에 적용됨

- ① 국가승인통계** 작성
- ② 해당 통계 작성을 위하여 수집·취득 또는 사용한 통계자료를 제공용으로 생성
- ③ 제공용 통계자료를 통계자료 이용신청자에게 제공***
- ④ 통계데이터센터 등에서 각종 통계자료를 분석한 뒤 반출

* 본 가이드라인에서 정해지지 않은 사항은 「보존용 통계자료 관리 및 이용 지침」, 「국가통계 자료제공 규정」, 「통계자료제공심의회 규정」, 「인가된 통계자료 이용 절차 등에 관한 규정」, 「통계데이터센터 운영 및 이용에 관한 규정」 등을 따름

** 국가승인통계 외에 새로운 유형의 데이터를 활용하거나 새로운 방식을 적용하여 실험적으로 작성하는 통계로, 작성 이후 신뢰성 및 타당성의 확인·점검이 필요한 통계(이하 ‘실험적 통계’)에 비식별화가 필요한 경우에 대해서도 본 가이드라인을 적용할 수 있음

*** 통계청 통계자료제공심의회 심의안건으로 상정된 통계자료 제공범위에 대해서도 본 가이드라인을 참고할 수 있음. 또한, 외부 자료와의 연계 및 결합을 위하여 통계자료를 가명처리하는 경우 다른 법령에 특별한 규정이 없으면 본 가이드라인을 우선 적용함

3 비식별화 절차

□ 예제로 이해하는 비식별화의 필요성

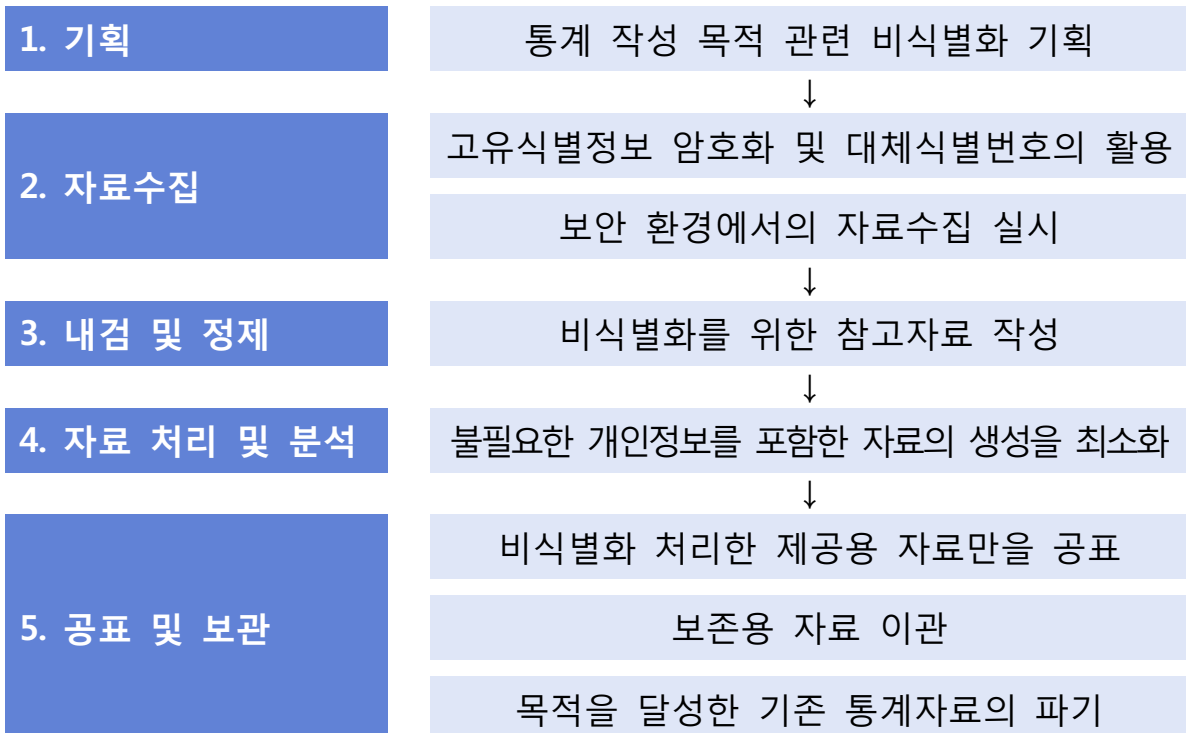
○ 마이크로데이터에서 개인정보 노출 가상 사례

ID	지역	연령	성별	직업	자산	부채	소득
1	서울	83	남	무직	28,874	1,784	570
2	제주	57	여	군인	74,475	2,547	4,785
유일한 개체 식별					속성 노출		
...	...	...	...	...	...	...	...
19,988	충남	27	남	공무원	38,147	5,787	1,879

○ 분석·집계결과(매크로데이터)에서 개인정보 노출 가상 사례

남성	50세	51~65세	50~65세	
기혼	0	5	5	- NA로 감춘 값이 1로 드러남
이혼	NA	0	NA	- 유일한 개체 식별 : 50세 남성
	NA	5	6	- 속성 노출 : 이혼

□ 통계 작성 단계별 비식별화 (구체적 내용은 III장 참조)



□ 통계자료 제공 단계별 비식별화 (구체적 내용은Ⅳ장 참조)

< 제공용 통계자료 생성 시 비식별화 >

1. 제공 범위 결정	항목 유형별 제공 범위 결정
2. 비식별화	공개 범위에 따른 제공 항목의 비식별화
3. 점검	제공용 통계자료 점검

< 이용 신청에 따른 통계자료 제공 시 비식별화 >

1. 사전 준비	신청에 따른 이용자의 자료 이용 목적 파악 및 필요한 자료의 준비
2. 유용성 및 노출위험 평가	이용자의 목적 달성에 필요한 항목에 대한 유용성 및 노출 위험성 평가
3. 제공항목 결정	노출 위험성을 고려하여 이용자의 요구에 부합하는 제공 항목의 결정
4. 통계자료제공심의회 심의	제공 항목에 대한 비식별화 검토 등을 거쳐 통계자료제공심의회 심의
5. 자료제공 및 안전한 관리	자료 제공 및 안전성 확보

< (통계데이터센터) 분석결과 반출 시 비식별화 >

1. 반출자료 비식별화	반출 요청자료의 종류·항목 확인 후 비식별화
2. 반출자료 점검 및 심의	비식별화 결과에 대한 점검 필요한 경우 반출심의위원회 심의
3. 자료 반출 및 관리	자료 반출 및 이력 관리

4 용어 정리

구분	세부구분	내용	참고 규정
식 별 정 보	자료주체	통계기초자료를 구성하는 단위를 의미하며 개인, 법인 또는 단체 등을 포함	
	항목/속성	자료주체에 대한 특성을 나타내는 변수로서 연속형 혹은 범주형 측정값으로 구성	
	식별정보	특정 개인이나 법인 또는 단체 등을 다른 개체와 구별하여 알아볼 수 있는 항목으로 고유·개체·보조식별정보 등으로 구분	
	↳ 고유식별정보	개인을 고유하게 구별하기 위하여 부여된 식별정보로서 주민등록번호, 여권번호, 운전면허번호, 외국인등록번호	개인정보보호법 제24조
	↳ 개체식별정보 (직접 식별자)	자료주체를 직접적이고 명확하게 나타내는 항목으로서 성명(사업체 대표자명), 사업자등록번호, 전화번호 등 자료주체를 직접적으로 알아볼 수 있는 정보	개체식별정보 중 개인정보보호법에서 명시한 항목을 고유 식별정보로 구분
비 식 별 화	↳ 보조식별정보 (간접 식별자)	자료주체에 대한 직접적인 정보는 없지만 다른 식별정보와 결합하여서 자료주체를 식별할 가능성을 증가시키는 정보	indirect identifier quasi identifier (Hundepool 외, 2012)
	비식별화	식별정보를 암호화하거나 삭제, 총계처리, 재현자료 생성 등을 통해 특정의 개인이나 법인 또는 단체 등을 식별할 수 없도록 하는 조치로, 가명·익명처리를 모두 포괄하는 개념	통계데이터센터 운영 및 이용 등에 관한 규정 제2조 제4호 참조
	↳ 가명처리 (가명정보)	개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가정보 없이는 특정 개인을 알아볼 수 없도록 처리하는 것(가명처리한 정보를 가명정보라 함)	개인정보보호법 제2조 제1호 다목 및 제1의2호, 신용정보법 제2조 제15호 및 제16호
	↳ 익명처리 (익명정보)	시간·비용·기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아볼 수 없도록 처리하는 것(익명처리한 정보를 익명정보라 함)	신용정보법 제2조 제17호

구분	세부구분	내용	참고 규정
각 종 통 계 자 료	통계자료	통계작성기관이 통계 작성을 위하여 수집·취득 또는 사용한 자료로, 통계기초자료와 메타데이터(공표용 보고서, 조사표, 항목 및 코드집, 가중치, 조사지침서, 파일설계서 등)로 구성	통계법 제3조 제4호
	↳ 통계기초자료 (microdata)	개인, 가구, 사업체, 법인 또는 단체에 대해 수집·취득 또는 사용된 정보로서 입력오류, 조사오류 등을 제거하여 통계 작성의 기초 자료로 사용되는 자료	통계데이터센터 운영 및 이용 등에 관한 규정 제2조 제2호, 국가통계 자료제공 규정 제2조 제2호 가목
	행정통계자료	행정자료를 주 원천자료로 하여 암호화과정, 정제, 연계 등 통계적 처리를 거쳐 생산한 자료	통계데이터센터 운영 및 이용 등에 관한 규정 제2조 제1호
	제공용 통계자료	통계별 제공범위에 따라 개인이나 사업체의 식별자(ID)와 미제공 항목 등을 삭제하여 이용자에게 제공을 결정한 자료로, 공공용 통계자료와 인가용 통계자료로 구분	보존용 통계자료 관리 및 이용 지침 제2조 제2호, 제12조
	↳ 인가용 통계자료	이용자의 통계자료 이용 신청에 따라 심사를 거쳐 제공되는 자료로, 원격접근서비스(RAS), 마이크로데이터 이용센터(RDC), 통계데이터 센터(SDC)에서 제공되는 자료	인가된 통계자료 이용절차 등에 관한 규정 제2조 제1호
	↳ 공공용 통계자료	통계자료 이용신청과 심사절차 없이 이용자가 통계자료를 이용할 수 있도록 사전에 심사를 거쳐 만들어진 자료로, 통합시스템(MDIS)을 통해 제공되는 자료	인가된 통계자료 이용절차 등에 관한 규정 제2조 제2호

구분	세부구분	내용	참고 규정
통계 기초자료 제공 서비스	마이크로데이터 통합서비스시스템 (MDIS; MicroData Integrated Service)	통계자료 제공 및 이용을 효율적으로 관리하기 위하여 통계청이 운영하는 관리시스템	인가된 통계자료 이용절차 등에 관한 규정 제2조 제3호
	원격접근서비스 (RAS; Remote Access Service)	이용자에게 온라인상에서 원격접근시스템을 통하여 인가된 통계자료를 제공하는 서비스	인가된 통계자료 이용절차 등에 관한 규정 제2조 제4호
	마이크로데이터 이용센터 (RDC; Research Data Center)	인가된 (표본)통계자료를 제공하기 위해 외부전산망을 차단하고 물리적 보안 환경을 갖춘 특정 장소 및 서비스	인가된 통계자료 이용절차 등에 관한 규정 제2조 제5호
	통계데이터센터 (SDC; Statistic Data Center)	통계청이 비식별화 처리한 (전수)자료 제공 및 자료간 연계서비스 제공을 위해 외부전산망을 차단하고 물리적 보안환경을 갖춘 특정 장소에서 운영하는 서비스 및 그 장소	인가된 통계자료 이용절차 등에 관한 규정 제2조 제6호
통계작성기관	통계작성기관	중앙행정기관, 지방자치단체 및 통계작성 지정기관	통계법 제3조, 제15조
	국가승인통계	통계청장의 승인을 받은 통계	통계법 제18조
	자료생산부서	해당 통계의 기획, 공표 등을 담당하는 부서	보존용 통계자료 관리 및 이용 지침 제2조 제3호
	자료관리부서	보존용 자료를 처리하여 이용자에게 제공하거나, 자료를 이용할 수 있도록 전산 환경을 제공하는 부서 * 통계를 작성하는 과정에서 활용 및 작성된 각종 파일로서 추후에 자료의 재집계가 가능하도록 영구 보존이 필요한 모든 자료	보존용 통계자료 관리 및 이용 지침 제2조 제5호
	자료이관기관	통계자료를 보존 및 대국민 서비스를 할 목적으로 통계청에 이관한 기관	통계법 제37조 제3항, 보존용 통계자료 관리 및 이용 지침 제2조 제6호

< 자료 특성별 항목의 종류 >

구분	정의	사례	비고
범주형 항목	가질 수 있는 값이 몇 개의 범주로 이뤄진 항목	성별, 행정구역, 직업, 산업분류 등	
준연속형 항목	가질 수 있는 값이 정수 또는 실수로 연속형이지만 실제로는 제한된 값만 가지는 항목	연령, 날짜 등	산술연산 가능, 범주형 항목 대상 비식별화 방법 적용 가능
연속형 항목	가질 수 있는 값이 정수 또는 실수인 항목	몸무게, 면적, 소득, 매출액 등 (길이, 무게, 금액 등의 단위를 갖는 항목)	산술연산 가능

< 인구 관련 범주형 항목 예시 >

성별	가구주와의관계	동거여부	혼인상태	등록장애인여부
남	가구주	같이 살고 있음	배우자있음	예
여	미혼 자녀	같이 살고 있음	미혼	아니오
남	미혼 자녀	따로 살고 있음(직장)	미혼	아니오
여	배우자	같이 살고 있음	배우자있음	아니오
여	배우자	같이 살고 있음	배우자있음	아니오
여	가구주	같이 살고 있음	배우자있음	아니오
남	가구주	같이 살고 있음	이혼	아니오
여	미혼 자녀	같이 살고 있음	미혼	아니오

< 경제활동 관련 연속형 항목 예시 >

자산금액	부채금액	순자산	경상소득	공급면적	전용면적
931,438,973	0	852,503,116	40,589,090	97	88
222,529,266	101,928,077	188,383,923	49,391,614	131	122
885,585	0	907,567	9,818,061	77	65
77,482,872	0	76,117,204	7,594,466	92	90
273,245,598	163,490,782	87,676,484	33,054,039	94	84
557,838,625	0	626,223,521	21,325,243	117	102
50,394,043	0	41,821,831	47,995,128	134	109

< 개인정보와 관련된 항목의 종류 >

구분	정의	사례	비고
개체식별 정보	자료주체를 직접적으로 나타내는 항목	주민등록번호, 여권번호, 이름, 주소 등	비식별화 대상
보조식별 정보	자료주체에 대한 직접적인 정보는 없지만 다른 보조식별정보와 결합하여 식별 가능성을 증가시키는 정보	성별, 연령, 혼인 여부, 직업, 우편번호 등	항목의 결합으로 자료주체가 유일하게 되는 경우 재식별 가능
민감정보	자료주체가 노출되는 위험성을 가진 정보는 아니지만 자료주체의 특성 중 매우 민감하여 노출되면 심각한 피해가 생기거나 사회적 논란이 야기 되는 정보	건강상태, 정치적 성향, 소득수준 등, 민감한 개인 특성과 관련된 항목	개인정보보호법 제23조의 정의보다 확장된 범위를 일컬음

< 정보노출의 종류 >

구 분	내용	사례
신분노출 (재식별)	제공용 통계자료에 포함된 자료주체의 신분을 식별하게 되는 것으로 공개된 자료에서 나이, 직업, 거주지 등을 외부 자료와 연계할 때 노출이 일어나는 경우 등을 말함	33세의 여성인 비행기 조종사가 특정 지역(읍면동)에 살고 있다는 정보가 항공사나 조종사 협회의 외부 자료와 연계되어 특정한 개인의 신분(이름, 소속, 경력 등)이 노출
속성노출	지역이나 분류 정보가 지나치게 세밀한 경우 연봉, 매출, 보험기록, 의료기록 등에 나타나는 특이 정보 또는 이상 치에 의하여 노출이 일어날 수 있음	특정 병원에 방문한 환자기록에서 산부인과를 방문한 21-25세 여성들의 상병기록이 동일한 경우에 해당 자료 주체의 질병정보 노출 (그룹속성 노출)

< 통계청 자료의 명칭 및 노출위험-자료유용성 >

자료유형	내부 자료		제공하는 자료	
마이크로데이터 (microdata)	수집된 각종 조사·행정자료 (수집된 원시자료)	통계기초자료 행정통계자료 (보관용 원시자료)	인가용 통계자료	공공용 통계자료
분석/집계 결과 (macrodata)			세분화된 집계표, 각종 기술통계 등 통계	
제어 조치	보안 조치		자료 변형 처리	
노출위험	높음		→ 낮음	
자료유용성	높음		→ 낮음	

5 개념 정리

○ 개인정보 보호법 제2조 제1항에서 용어 정의는 다음과 같음

1. **"개인정보"**란 살아 있는 개인에 관한 정보로서 다음 각 목의 어느 하나에 해당하는 정보를 말한다.

가. 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보

나. 해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 정보. 이 경우 쉽게 결합할 수 있는지 여부는 다른 정보의 입수 가능성 등 개인을 알아보는 데 소요되는 시간, 비용, 기술 등을 합리적으로 고려하여야 한다.

다. 가목 또는 나목을 제1호의2에 따라 가명처리함으로써 원래의 상태로 복원하기 위한 추가 정보의 사용·결합 없이는 특정 개인을 알아볼 수 없는 정보(이하 "가명정보"라 한다)

1의2. **"가명처리"**란 개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가 정보가 없이는 특정 개인을 알아볼 수 없도록 처리하는 것을 말한다.

○ 신용정보법 제2조 제15호~제17호에서 용어 정의는 다음과 같음

15. **"가명처리"**란 추가정보를 사용하지 아니하고는 특정 개인인 신용정보주체를 알아볼 수 없도록 개인신용정보를 처리(그 처리 결과가 다음 각 목의 어느 하나에 해당하는 경우로서 제40조의2제1항 및 제2항에 따라 그 추가정보를 분리하여 보관하는 등 특정 개인인 신용정보주체를 알아볼 수 없도록 개인신용정보를 처리한 경우를 포함한다)하는 것을 말한다.

가. 어떤 신용정보주체와 다른 신용정보주체가 구별되는 경우

나. 하나의 정보집합물(정보를 체계적으로 관리하거나 처리할 목적으로 일정한 규칙에 따라 구성되거나 배열된 둘 이상의 정보들을 말한다. 이하 같다)에서나 서로 다른 둘 이상의 정보집합물 간에서 어떤 신용정보주체에 관한 둘 이상의 정보가 연계되거나 연동되는 경우

다. 가목 및 나목과 유사한 경우로서 대통령령으로 정하는 경우

16. **"가명정보"**란 가명처리한 개인신용정보를 말한다.

17. **"익명처리"**란 더 이상 특정 개인인 신용정보주체를 알아볼 수 없도록 개인신용정보를 처리하는 것을 말한다.

- 통계를 작성 및 제공하는 환경에서 각 용어의 개념을 다음과 같은 마이크로데이터 예제를 바탕으로 이해함

예제 - 개인정보 가.		(개보법 제2조 제1항) 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보					
성명	생년월일	지역 1	지역 2	지역 3	성별	교육정도	혼인상태
		...			...		
홍길동	69.1.1	대전 서구	온천 2동	A단지 1동	남	박사	이혼
		...			...		



예제 - 개인정보 나.		해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 정보					
성명	생년월일	지역 1	지역 2	지역 3	성별	교육정도	혼인상태
		...			...		
삭제	69년 월일삭제	대전 서구	온천 2동	A단지 1동	남	박사	이혼
		...			...		

예제 - 개인정보 다. (가명정보)		가목 또는 나목을 가명처리함으로써 원래의 상태로 복원하기 위한 추가 정보의 사용·결합 없이는 특정 개인을 알아볼 수 없는 정보					
성명	생년월일	지역 1	지역 2	지역 3	성별	교육정도	혼인상태
		...			...		
☞ΦΔΔδ	69년	대전 서구	%o**φㄹ▣	ㄹ%oφ**※?	남	박사	이혼
		...			...		

예제 - 익명정보		시간·비용·기술 등을 합리적으로 고려할 때 다른 정보를 사용하여 도 더 이상 개인을 알아볼 수 없도록 처리한 정보					
성명	생년월일	지역 1	지역 2	지역 3	성별	교육정도	혼인상태
		...			...		
삭제	69년	대전 서구	삭제	삭제	남	박사	이혼
		...			...		

○ 각 마이크로데이터의 빈도표를 작성하여 각 용어의 개념을 이해

- * 아래 예제의 통계기초자료(마이크로데이터)는 자료주체의 개수(n)가 150만이며, 총 8개의 변수로 구성되어 있고, 각 변수에서 범주의 개수는 성씨 100개, 생년 85개, 지역1-2-3 2500개, 성별 2개, 교육정도 8개, 혼인상태 5개임

예제 - 개인정보 나. 관련 빈도표

빈도 1명으로 유일성을 쉽게 확인할 수 있게 되는 경우, 제공하는 마이크로데이터의 정보와 다른 정보(예. 해당 자료주체의 이웃이 가지는 정보)를 쉽게 결합할 수 있음

생년월일	지역 1	지역 2	지역 3	성별	교육정도	빈도
		...			...	
69년	대전 서구	온천 2동	A단지 1동	남	박사	1명
		...			...	

(빈도표 행의 수(m) = 변수 조합으로 생기는 경우의 수 = 340만)

예제 - 개인정보 다. (가명정보) 관련 빈도표

빈도 1명으로 유일성을 쉽게 확인할 수 있으나, 가명처리로 특정 개인을 알아볼 수 없으며, 다른 곳에 보관 중인 추가정보(암호화 키 등)가 있어야 개인정보 가. 또는 나. 유형으로 환원됨

성명	생년월일	지역 1	지역 2	지역 3	성별	교육정도	혼인상태	빈도
		...				...		
김철수	69년	대전 서구	%0x*600	m%00f**%0?	남	박사	이혼	1명
		...				...		

(빈도표 행의 수(m) = 변수 조합으로 생기는 경우의 수 = 17억)

예제 - 익명정보 관련 빈도표

빈도표로 작성할 때, 예를 들어 동일한 특성을 가지는 자료주체가 100명으로 나타난다면, 해당 자료주체를 나타내는 이 행(레코드)은 익명성 100을 확보한 익명정보가 됨
(익명정보 여부 확인을 위하여 자료에 대한 빈도표 작성이 필요)

생년월일	지역 1	성별	교육정도	혼인상태	빈도
			...		
69년	대전 서구	남	박사	이혼	100명
			...		

(빈도표 행의 수(m) = 변수 조합으로 생기는 경우의 수 = 3.4만)

○ 제공하는 모든 변수들에 대하여 집계표(빈도표)를 작성하면, 변수 범주의 세분화 수준에 따라 빈도표 행의 수(m)가 결정됨

- 자료주체의 개수(n)과 비교하였을 때, 빈도표 행의 수(m)가 큰 경우 익명성 확보가 어려울 수 있음



II. 비식별화 방법



1. 노출위험 평가 척도

비식별화 처리를 적용하는 경우 자료주체들의 개별적인 노출 위험성뿐만 아니라, 전체적인 수준에서 노출의 위험성도 측정할 필요가 있다.

개별적인 위험성을 판단할 수 있는 기준으로는 ①유일성과 ②특이정보 등이, 전체적인 위험성을 판단하는 기준은 ③ k -익명성과 ④ ℓ -다양성 등이 있다.

자료의 유용성을 최대한 보존하면서, 동시에 재식별의 위험성을 증가시키는 자료주체의 유일성을 방지하는 것이 비식별화의 가장 중요한 목표이다.

① 자료주체의 유일성

- (정의) 자료주체의 유일성이란 주어진 통계기초자료(마이크로데이터)의 모집단에서 특정 자료주체가 유일하다고 판단되는 경우*를 의미함

* 하나의 식별정보로 인하여 발생할 수도 있지만 자료에 포함된 여러 보조 식별정보의 조합으로 발생할 수도 있음

- (유의사항) 자료주체의 유일성은 자료주체의 개체식별정보로 인한 자명한 경우 이외에도, 자료주체의 중요한 특성을 포함하는 보조 식별정보들(나이, 거주지역, 성별 등)의 조합으로도 발생할 수 있어 주의하여야 함. 또한, 표본자료의 경우 가중치 등을 함께 고려하여 유일성을 판단하여야 함

② 특이정보

- (정의) 특이정보란 1개 또는 여러 개의 항목에서의 값이 고유(희소)하거나 편중된 분포를 가져서 식별 가능성이 높은 경우를 의미함
- (사례) 자료의 전체적인 패턴에서 벗어나는 극단값을 가지는 정보로 국내 최고령, 최장신, 고액체납금액, 고액급여수급자 등이 있음
- (처리방법) 특이정보를 가지고 있는 자료주체는 재식별 가능성이 매우 높으므로 해당 항목에 대한 상하단 범주화, 국소 삭제와 같은 비식별화 처리를 적용하는 등, 별도 관리가 필요

③ k -익명성

- (정의) k -익명성이란 보조식별정보들의 조합에 속하는 자료주체의 수가 최소한 k 개 이상으로 나타나도록 하는 기준을 의미함
- (처리방법) k -익명성을 구현하기 위한 비식별화 방법은 항목 삭제, 재범주화, 상하단 범주화, 국소 삭제 등이 있음

< 마이크로데이터에서의 k -익명성 판단 예시 >

번호	동거여부	혼인 상태	종사상지위	외국인 여부	표본 빈도	3-익명성	5-익명성
1	1인 가구	배우자 있음	고용원이 없는 자영업자	아니오	1	불만족	불만족
2	1인 가구	미혼	기타(무직자, 가사, 학생 등)	예	1	불만족	불만족
3	같이 살고 있음	미혼	임시·일용근로자	예	1	불만족	불만족
4	1인 가구	미혼	상용근로자	예	1	불만족	불만족
5	1인 가구	이혼	임시·일용근로자	예	2	불만족	불만족
6	1인 가구	이혼	임시·일용근로자	예	2	불만족	불만족
7	같이 살고 있음	미혼	상용근로자	예	3	만족	불만족
8	같이 살고 있음	미혼	상용근로자	예	3	만족	불만족
9	같이 살고 있음	미혼	상용근로자	예	3	만족	불만족
10	따로 살고 있음(직장)	이혼	상용근로자	아니오	7	만족	만족
11	따로 살고 있음(직장)	이혼	상용근로자	아니오	7	만족	만족
12	따로 살고 있음(직장)	이혼	상용근로자	아니오	7	만족	만족
13	따로 살고 있음(직장)	이혼	상용근로자	아니오	7	만족	만족
14	따로 살고 있음(직장)	이혼	상용근로자	아니오	7	만족	만족
15	따로 살고 있음(직장)	이혼	상용근로자	아니오	7	만족	만족
16	따로 살고 있음(직장)	이혼	상용근로자	아니오	7	만족	만족

* 1~4번 : 4개의 범주형 보조식별항목(동거여부, 혼인상태, 종사상지위, 외국인 여부)의 조합으로 각각 유일하게 나타나 3-익명성과 5-익명성을 불만족

** 5~6번 : 4개의 보조 식별항목의 조합으로 같은 특성을 가지며 표본빈도가 2개 이므로 3-익명성과 5-익명성을 불만족

*** 7~9번 : 4개의 보조 식별항목의 조합으로 같은 특성을 가지며 표본빈도가 3개이므로 3-익명성은 만족, 5-익명성은 불만족

- (사례) 위 예시는 마이크로데이터 총 4개 범주형 보조식별항목의 조합을 고려할 때 자료주체들의 3-익명성과 5-익명성 만족 여부를 보여 주는 표로, 위 예시에서 ①범주의 수가 많은 '종사상지위'를 재범주화하거나 ②4개의 보조식별항목에 대하여 국소 삭제 또는 항목 삭제 등의 비식별화 방법을 적용하면 k -익명성 위반 사례를 줄일 수 있음

④ ℓ -다양성

- (정의) ℓ -다양성이란, k -익명성을 만족하더라도 나타날 수 있는 개인정보 노출 위험성을 방지하기 위한 척도로, 보조식별정보들의 조합이 동일한 자료주체들에 대해 민감정보를 나타내는 항목의 서로 다른 값이 최소한 ℓ 개 이상 나타나야 한다는 기준임
- (배경) k -익명성을 만족하는 경우, 보조식별정보들의 조합이 같은 자료주체들의 민감정보를 나타내는 항목의 값이 모두 동일하다면 자료주체들의 민감한 정보를 쉽게 유추*할 수 있으므로, 이러한 민감한 정보의 노출을 방지하기 위하여 ℓ -다양성을 도입
 - * (예) 동거여부·혼인상태·종사자 지위 항목값이 동일한 자료주체가 3명인데 모두 암환자인 경우 암환자라는 민감정보가 노출될 수 있음
- (유의사항) ℓ -다양성은 재식별의 위험성을 낮추기보다는 특별히 민감한 항목의 정보가 노출되는 경우를 방지하기 위하여 사용되며, ℓ -다양성을 판단하는 경우 여러 개의 민감한 항목들을 동시에 고려하는 것은 어려우므로 각 항목에 대하여 분리하여 적용해야 함

◇ 자료의 유일성, 특이정보, k -익명성, ℓ -다양성에 대한 자세한 내용은 **부록 I. 노출위험 평가 척도(1~14쪽)**를 참조할 것

2. 정보손실 측도

- (배경) 자료에 대한 비식별화 처리는 자료주체의 재식별 가능성을 낮추나, 원자료의 구조나 값을 변형하는 과정에서 원자료의 유용성을 감소시키므로 정보손실을 측정하여 비교할 필요가 있음
 - (측도) 정보손실을 평가하는 측도는 범주형 항목과 연속형 항목별로 상이하며 상세한 측도는 부록에서 다룸
 - (절차) 정보 손실의 측정은 범주형과 연속형 항목에 대하여 이용자의 요구를 반영할 수 있는 다양한 측도*(예를 들어, 분포 또는 구성 비율의 변화, 평균 변화량 등)를 선택한 뒤, 이를 이용하여 비식별화 처리 전과 처리 후에 대한 유용성을 비교하는 방식으로 이뤄짐
- * 자료의 유용성 측도는 하나의 기준으로 평가할 수 없으며 여러 개의 다양한 기준을 사용하여 종합적으로 판단해야 함

◇ 정보손실 측도에 대한 자세한 내용은 **부록 II. 정보손실 평가 측도** (17~19쪽)를 참조할 것

3. 평가 측도의 의미 및 향후 과제

- 자료주체의 재식별 또는 민감정보 노출위험을 측정하는 방법은 지속적으로 개발되고 있으며, k -익명성, l -다양성 이외에도 여러 기준이 활발히 논의되고 있어 가이드라인 업데이트 지속 필요
 - 자료의 정보손실 혹은 유용성을 측정하는 방법은 매우 다양하며, 측정 방법에 따라 유용성 비교 결과가 달라져 주의할 필요가 있음
- * 각 공표자료(마이크로데이터, 매크로데이터)의 수요 및 이용현황에 대한 주기적인 분석 및 그에 따른 세부적인 정보손실 측도의 연구·개발 지속 필요
- 통계작성기관은 다양한 평가 측도를 종합적으로 고려해야 함
- * 적절한 기준 마련 및 비식별화 체계화 등을 위하여 연구개발 사업 진행 중

2 마이크로데이터 비식별화 방법

- 다양한 비식별화 방법 중 일부가 국제표준으로 정리되어 있고 각 부처의 가명·익명처리 관련 가이드라인이 이를 참고하고 있음

< 국제표준(ISO/IEC 20889)에 따른* 비식별화 방법 >

구 분	비식별화 방법		적용 항목
구조적 방법	표본추출		자료 전체
	총계처리		연속형
삭제	항목 삭제 (열 삭제)		연속형, 범주형
	레코드 삭제 (행 삭제)		연속형, 범주형
	국소 삭제 (셀 삭제)		연속형, 범주형
	기호로 표기 (masking)**		-
일반화	반올림	일반 반올림	연속형
		랜덤 반올림***	연속형, 범주형
	범주화	재범주화	연속형, 범주형
		상하단 범주화	연속형, 범주형
임의화	잡음첨가 (noise addition)***		연속형
	자료교환 (swapping)***		연속형
	부분총계 (microaggregation)***		연속형
재현자료 생성	재현자료 (synthetic data)***		자료 전체
암호화	일방향 암호화		연속형, 범주형

* ISO/IEC 20889를 따르나 실무적으로 활용하는 방법 위주로 간소화하였음

** 보통 개체식별정보의 전부 또는 일부를 기호로 표기하여 유일성을 제거

*** 자료 변조적 기법(perturbative methods)이라 부르며, 원래의 값이 무엇으로 바뀔지 결정적이지 않음

**** 재현자료, 암호화 관련 세부 설명은 부록 III. 세부 비식별화 방법 해설 참조

1. 구조적 방법

구조적 방법은 마이크로데이터를 제공하는 경우 전체 자료의 자료주체 수(행)나 구조(열)를 변형하여 제공하는 것으로, ①표본 추출과 ②총계처리가 대표적인 방법이다.

① 표본추출 (자료 전체)

○ (정의) 전체 자료에서 일부분을 추출하여 공개하는 방법

- 공개하려는 자료가 자료주체 전체를 포함하는 등록자료 또는 행정통계자료인 경우 자료의 일부분을 표본으로 추출하여 제공해야 함
- 표본으로 추출된 자료에서는 자료주체가 유일하게 구별되어도 전체 등록자료에서 유일한 것이 아닐 수 있으므로, 등록부 자료 전체를 공개하는 것보다 재식별 위험은 감소함

◇ 표본과 모집단에서의 유일성에 대한 자세한 내용은 부록 I. 노출위험 평가 측도(2~4쪽)를 참조할 것

② 총계처리 (연속형 항목)

○ (정의) 여러 개의 연관된 연속형 세부 항목들을 모두 공개하지 않고 총계만을 공개하는 비식별화 방법

* 여러 자료주체들의 특정 항목 전체를 평균값, 최빈값, 중간값 등으로 대체하여 제공하는 방법도 있으나 마이크로데이터로서의 기능이 사라짐

○ (사례1) 경제활동 항목에서 다양한 소득금액들을 모두 합하여 경상소득으로 총계처리한 예로, 6개 세부 소득금액은 미제공하고, 총계처리한 경상소득만 공개하면 재식별 위험이 감소함

< 항목 총계처리 예시 >

번호	근로소득	사업소득	재산소득	공적이전소득	사적이전소득	이전소득	경상소득(총합)
1	42,289,159	58,351,593	0	0	0	0	100,640,752
2	33,799,805	0	0	2,554,875	0	3,150,028	39,504,708
3	169,823,718	0	0	0	0	0	169,823,718
4	0	50,679,817	0	4,576,347	0	3,782,700	59,038,864
5	175,161,629	0	0	0	0	0	175,161,629
6	2,004,598	0	0	3,257,698	2,806,441	5,306,662	13,375,399
7	0	81,552,597	37,874,389	0	0	0	119,426,986
8	30,276,146	0	0	6,030,991	3,107,228	8,169,657	47,584,022

◇ 구조적 방법에 대한 자세한 내용은 부록 III-1. 구조적 방법 해설(21쪽)을 참조할 것

2. 삭제 방법

삭제 방법은 개인정보 노출의 위험성이 큰 항목 전체를 제거하는 ①항목(열) 삭제, 재식별 위험성이 큰 특정 자료주체의 모든 항목을 제거하는 ②레코드(행) 삭제가 있다. 또한 재식별의 위험성이 큰 특정 자료주체의 일부 항목의 값을 제거하여 결측값으로 표시하는 ③국소 삭제 방법이 있다. 마지막으로 개체식별정보의 전부 또는 일부를 ④기호로 표기하는 방법이 있다.

① 항목 삭제 (연속형, 범주형)

- (정의) 개인정보 노출의 위험성이 큰 항목 전체를 삭제하고 자료를 제공하는 방법으로, 개체식별정보를 항목 삭제
 - 보조식별항목 중에서 적합한 비식별화 방법을 찾아 적용하기 어렵거나 비식별화 방법을 적용해도 위험성이 큰 항목들에 적용
- (사례) 개체식별정보인 성명과 주소 항목을 삭제

< 항목 삭제 전 예시 >

번호	기준년도	성명	주소	사업체종사자	사업체매출액
1	2018	봉동석	충청남도 보령시 황율리 산 12	293	5,626,808
2	2018	권하빈	경상남도 함안군 봉촌리 길 132	3601	3,150,018
3	2018	오미경	전라남도 목포시 대의동2가 11	624	475,191
4	2018	봉성진	강원도 철원군 양지리 산 234	137	146,696
5	2018	임미나	경상북도 문경시 중평리 132-12	34	58,262
6	2018	복병지	경상북도 경산시 남하리 산 12	360	57,552
7	2018	신윤성	충청북도 청주시 상당구 종암리 꽃길 12	30	52,248
8	2018	배정숙	전라남도 여수시 나진리 192-1	148	39,805

< 항목 삭제 후 예시 >

번호	기준년도	사업체종사자	사업체매출액
1	2018	293	5,626,808
2	2018	3601	3,150,018
3	2018	624	475,191
4	2018	137	146,696
5	2018	34	58,262
6	2018	360	57,552
7	2018	30	52,248
8	2018	148	39,805

② 레코드(행) 삭제 (연속형, 범주형)

- (정의) 개인정보 노출의 위험성이 큰* 자료주체의 모든 항목 값을 삭제하고 자료를 제공하는 방법

* 특정한 자료주체가 적절한 비식별화 방법을 적용한 후에도 유일하게 구별될 가능성이 크거나 일부 항목에 특이정보를 가지고 있는 경우

- (사례) 특이정보를 가진 자료주체(자산금액이 다른 주체들보다 매우 크게 나타남)를 자료에서 제외하는 행 삭제를 적용

< 행 삭제 전 예시 >

번호	기준년도	가구주성별	가구원수	가구주_만연령	자산금액
1	2016	남자	4	57	4,619,539,958
2	2016	남자	4	45	2,108,496,937
3	2016	여자	3	53	1,623,987,294
4	2016	남자	5	46	852,757,762
5	2016	남자	4	49	774,992,697
6	2016	남자	2	36	511,537,764
7	2016	여자	1	80	57,548,536
8	2016	여자	2	69	2,479,026

< 행 삭제 후 예시 >

번호	기준년도	가구주성별	가구원수	가구주_만연령	자산금액
2	2016	남자	4	45	2,108,496,937
3	2016	여자	3	53	1,623,987,294
4	2016	남자	5	46	852,757,762
5	2016	남자	4	49	774,992,697
6	2016	남자	2	36	511,537,764
7	2016	여자	1	80	57,548,536
8	2016	여자	2	69	2,479,026

③ 국소 삭제 (연속형, 범주형)

- (정의) 재식별의 위험성이 큰 특정 자료주체의 일부 항목의 값을 제거하여 결측값으로 표시하는 방법으로, k -익명성과 같은 개인정보 기준을 충족하기 위하여 주로 사용되는 방법
 - 자료주체의 일부 항목의 특이값에 의하여 유일하게 구별될 가능성이 큰 경우에도 사용 가능
- (사례) 여러 개의 항목의 조합을 고려할 때 빈도가 1개 또는 2개로 나타나는 자료에 대하여 특정 항목을 국소 삭제
 - 다른 항목에 비하여 상대적으로 더 민감한 항목부터 또는 유용성이 낮은 항목부터 삭제하며, 아래 예시에서는 상대적으로 더 민감한 혼인상태, 종사상지위, 외국인 여부 항목을 국소 삭제함

< 국소 삭제 전 예시 >

번호	성별	가구주와의관계	동거여부	혼인상태	종사상지위	외국인여부	표본빈도
1	여	기타	같이살고있음	사별	고용원이 없는 자영업자	아니오	1
5	남	배우자	같이살고있음	배우자있음	상용근로자	예	2
6	남	배우자	같이살고있음	배우자있음	상용근로자	예	2
7	남	가구주	따로살고있음(직장)	미혼	임시.일용근로자	아니오	3
8	남	가구주	따로살고있음(직장)	미혼	임시.일용근로자	아니오	3
9	남	가구주	따로살고있음(직장)	미혼	임시.일용근로자	아니오	3

< 국소 삭제 후 예시 >

번호	성별	가구주와의관계	동거여부	혼인상태	종사상지위	외국인여부	표본빈도
1	여	기타	같이살고있음				1
5	남	배우자	같이살고있음				2
6	남	배우자	같이살고있음				2
7	남	가구주	따로살고있음(직장)	미혼	임시.일용근로자	아니오	3
8	남	가구주	따로살고있음(직장)	미혼	임시.일용근로자	아니오	3
9	남	가구주	따로살고있음(직장)	미혼	임시.일용근로자	아니오	3

◇ 삭제에 대한 자세한 내용은 부록 III-2. 삭제 방법 해설(22~26쪽)을 참조할 것

3. 일반화 방법

범주형 항목과 연속형 항목 모두에 대하여 적용할 수 있는 일반화 방법은 연속형 항목이 가진 값의 최소 단위를 크게 하거나, 범주형 항목이 가진 범주의 개수를 줄여 자료주체가 유일하게 나타나는 가능성을 줄이는 방법이다.

① 일반 반올림 (연속형)

- (정의) 연속형 항목의 값을 특정한 자리에서 반올림하는 방법으로 외부 자료와의 매칭을 어렵게 하여 노출위험을 낮추고자 사용됨
- (고려사항) 반올림을 고려하는 경우 적용하는 자릿수를 값의 크기에 따라서 적절하게 적용해야 함
 - 반올림을 적용하는 자릿수가 너무 작으면 자료의 변화가 너무 작아서 외부 자료 연계가 가능할 수도 있고 너무 크면 자료의 유용성이 심하게 감소할 수 있음
- (사례) 우측 사례는 자산금액과 부채금액에 자릿수에 따라서 반올림하는 단위를 다르게 적용한 사례임

◇ 일반화 방법에 대한 자세한 내용은 부록 III-3. 일반화 방법 해설 (26~30쪽)을 참조할 것

< 반올림 적용 전 예시 >

번호	기준년도	교육정도	혼인상태	연령코드	자산금액	부채금액
1	2016	고등학교 졸업	배우자있음	50~60세 미만	189,758,367	448,838,178
2	2016	중학교 졸업	배우자있음	60세이상	5,661,953	0
3	2016	고등학교 졸업	미혼	30~40세 미만	90,756,937	214,235
4	2016	고등학교 졸업	배우자있음	50~60세 미만	433,570,182	59,346,709
5	2016	고등학교 졸업	배우자있음	40~50세 미만	189,697,483	8,602,043
6	2016	고등학교 졸업	배우자있음	60세이상	51,829,375	0
7	2016	초등학교 졸업	사별	60세이상	406,939,721	0
8	2016	고등학교 졸업	이혼	30~40세 미만	92,874,677	18,252,622

< 반올림 적용 후 예시 >

번호	기준년도	교육정도	혼인상태	연령코드	자산금액	부채금액
1	2016	고등학교 졸업	배우자있음	50~60세 미만	190,000,000	449,000,000
2	2016	중학교 졸업	배우자있음	60세이상	5,660,000	0
3	2016	고등학교 졸업	미혼	30~40세 미만	91,000,000	210,000
4	2016	고등학교 졸업	배우자있음	50~60세 미만	434,000,000	59,000,000
5	2016	고등학교 졸업	배우자있음	40~50세 미만	190,000,000	8,600,000
6	2016	고등학교 졸업	배우자있음	60세이상	52,000,000	0
7	2016	초등학교 졸업	사별	60세이상	407,000,000	0
8	2016	고등학교 졸업	이혼	30~40세 미만	93,000,000	18,000,000

② 랜덤 반올림 (연속형, 범주형)

- (정의) 주어진 기준을 중심으로, 자료값에 따라 주어지는 확률로 올림 혹은 내림을 랜덤하게 시행하는 반올림
 - 기준값이 10이고, 랜덤 반올림하려는 값이 8이라면 0.8의 확률로 10으로 올림하고, 0.2의 확률로 0으로 내림하는 반올림 (기존의 반올림으로는 8의 경우 모두 10으로 올림하게 되어 있음)
- (사례) 연령을 10세 단위로 랜덤 반올림을 적용

< 랜덤 반올림 적용 전후 예시 >

기준년도	교육정도	혼인상태	만연령	자산금액	부채금액
2016	고등학교 졸업	배우자있음	56 → 50	189,758,367	448,838,178
2016	중학교 졸업	배우자있음	62 → 70	5,661,953	0
2016	고등학교 졸업	미혼	33 → 40	90,756,937	214,235
2016	고등학교 졸업	배우자있음	56 → 60	433,570,182	59,346,709
2016	고등학교 졸업	배우자있음	46 → 40	168,034,483	8,602,043
2016	고등학교 졸업	배우자있음	65 → 60	51,829,375	0
2016	초등학교 졸업	사별	81 → 90	406,939,721	0
2016	고등학교 졸업	이혼	35 → 30	92,874,677	18,252,622

③ 재범주화 (연속형, 범주형)

- (정의) 범주형 항목의 범주 개수를 줄이거나 연속형 항목의 값 전체를 구간화하여 범주형 항목으로 만들어 주는 방법으로, 항목이 가진 정보의 미세한 정도를 줄여서 정보의 정확성을 감소시키는 대신 개인정보 노출의 위험을 줄이는 방법

- 비식별화 작업을 적용할 때 가장 먼저 고려할 수 있는 기본적인 방법이며 다른 방법에 비해 쉽게 구현할 수 있는 매우 효과적인 방법임

- (사례) 연령을 10세 단위로 범주화하는 재범주화를 적용

< 재범주화 적용 전 예시 >

기준년도	교육정도	혼인상태	만연령	자산금액	부채금액
2016	고등학교 졸업	배우자있음	56	189,758,367	448,838,178
2016	중학교 졸업	배우자있음	62	5,661,953	0
2016	고등학교 졸업	미혼	33	90,756,937	214,235
2016	고등학교 졸업	배우자있음	56	433,570,182	59,346,709
2016	고등학교 졸업	배우자있음	46	168,034,483	8,602,043
2016	고등학교 졸업	배우자있음	65	51,829,375	0
2016	초등학교 졸업	사별	81	406,939,721	0
2016	고등학교 졸업	이혼	35	92,874,677	18,252,622

< 재범주화 적용 후 예시 >

기준년도	교육정도	혼인상태	연령코드	자산금액	부채금액
2016	고등학교 졸업	배우자있음	50~60세 미만	189,758,367	448,838,178
2016	중학교 졸업	배우자있음	60세이상	5,661,953	0
2016	고등학교 졸업	미혼	30~40세 미만	90,756,937	214,235
2016	고등학교 졸업	배우자있음	50~60세 미만	433,570,182	59,346,709
2016	고등학교 졸업	배우자있음	40~50세 미만	168,034,483	8,602,043
2016	고등학교 졸업	배우자있음	60세이상	51,829,375	0
2016	초등학교 졸업	사별	60세이상	406,939,721	0
2016	고등학교 졸업	이혼	30~40세 미만	92,874,677	18,252,622

④ 상하단 범주화 (연속형, 범주형)

- (정의) 재범주화와 동일한 방법이나 적용하는 값의 범위가 제한적인 경우로, 항목의 순서형 범주나 값이 가질 수 있는 가장 작은 부분 또는 가장 큰 부분에 재코딩을 적용하는 방법
 - 항목의 값이 가장 작은 부분 또는 가장 큰 부분에 속한 자료주체의 수가 매우 적어서 개인정보 노출의 위험이 클 때 사용하는 방법
- (사례) 자산금액이 매우 높게 나타나는 자료주체들에 대하여 자산금액의 상한을 20억 원으로 처리하여 상단 범주화

< 상단 범주화 적용 전 예시 >

기준년도	교육정도	혼인상태	만연령	연령코드	자산금액
2016	대학졸업 이상	사별	75	60세이상	3,021,381,747
2016	대학졸업 이상	배우자있음	68	60세이상	2,211,432,891
2016	고등학교 졸업	배우자있음	57	50~60세 미만	1,773,112,035
2016	대학졸업 이상	배우자있음	62	60세이상	1,305,219,958
2016	대학졸업 이상	배우자있음	53	50~60세 미만	1,158,365,179
2016	고등학교 졸업	배우자있음	62	60세이상	1,034,126,748
2016	고등학교 졸업	배우자있음	54	50~60세 미만	933,260,901
2016	고등학교 졸업	배우자있음	47	40~50세 미만	922,769,722

< 상단 범주화 적용 후 예시 >

기준년도	교육정도	혼인상태	만연령	연령코드	자산금액
2016	대학졸업 이상	사별	75	60세이상	20억원 이상
2016	대학졸업 이상	배우자있음	68	60세이상	20억원 이상
2016	고등학교 졸업	배우자있음	57	50~60세 미만	1,773,112,035
2016	대학졸업 이상	배우자있음	62	60세이상	1,305,219,958
2016	대학졸업 이상	배우자있음	53	50~60세 미만	1,158,365,179
2016	고등학교 졸업	배우자있음	62	60세이상	1,034,126,748
2016	고등학교 졸업	배우자있음	54	50~60세 미만	933,260,901
2016	고등학교 졸업	배우자있음	47	40~50세 미만	922,769,722

4. 임의화 방법

임의화 방법은 비식별화 방법 중 자료의 값을 임의적으로 변화하는 방법을 말한다. 임의화 방법을 적용하는 경우 사용되는 분포나 임의화 방법에 따라서 동일한 참값에 대하여 비식별화 처리할 때마다 그 결과값이 달라질 수 있다. (참고) 이와 달리 재범주화, 일반 반올림 등의 방법은 비식별화 처리를 실시할 때마다 참값이 동일하면 동일한 결과값을 제공한다.

① 잡음 첨가 (연속형)

○ (정의) 항목의 값에 임의의 값, 즉 잡음을 더하거나 곱하는 방법

- 잡음첨가의 주된 목적은 공개된 자료와 외부 자료와의 연계를 방지하는 것으로, 보통 잡음을 생성하는 분포의 평균은 0, 분산은 자료에 의해 결정되는 값을 가지도록 함
- 잡음첨가로 비식별화한 자료는 평균이나 중앙값 등은 원자료와 매우 유사하나, 분산이나 상관관계 등은 왜곡되는 경향이 있음. 이에 분산을 유지하는 알고리즘을 추가로 활용할 필요가 있음

○ (사례) 3개의 금액 항목에 잡음을 첨가(잡음은 평균이 0이고 표준 편차가 원자료 표준 편차의 10%를 가지는 정규분포에서 생성)

< 잡음첨가 전 예시 >

번호	년도	거주시도	연령코드	월소득액	월저축금액	월교통비용
1	2016	경기도	60세이상	3,921,000	672,000	5,148,000
2	2016	강원도	50~60세 미만	1,603,000	334,000	104,200
3	2016	충청북도	40~50세 미만	1,483,000	274,000	168,000
4	2016	경상남도	40~50세 미만	1,398,000	336,000	282,000
5	2016	경상북도	60세이상	1,226,000	104,320	100,000
6	2016	충청남도	50~60세 미만	1,209,000	656,000	105,510
7	2016	경상남도	40~50세 미만	1,143,000	250,000	120,000
8	2016	경상북도	50~60세 미만	1,025,000	181,000	150,000
9	2016	충청남도	30~40세 미만	1,015,000	147,000	128,000
10	2016	대구광역시	60세이상	755,000	113,000	125,000

< 잡음첨가 후 예시 >

번호	년도	거주시도	연령코드	월소득액	월저축금액	월교통비용
1	2016	경기도	60세이상	3,932,148	678,519	5,132,925
2	2016	강원도	50~60세 미만	1,587,325	321,384	74,246
3	2016	충청북도	40~50세 미만	1,468,231	278,774	209,722
4	2016	경상남도	40~50세 미만	1,288,881	338,227	172,920
5	2016	경상북도	60세이상	1,217,901	96,718	53,517
6	2016	충청남도	50~60세 미만	1,215,648	645,925	131,610
7	2016	경상남도	40~50세 미만	1,072,030	238,981	131,954
8	2016	경상북도	50~60세 미만	977,123	187,578	114,326
9	2016	충청남도	30~40세 미만	970,047	165,857	169,330
10	2016	대구광역시	60세이상	731,589	109,196	130,553

② 자료 교환 (연속형)

- (정의) 항목의 값의 순서를 재배열하는 방법(연속형 항목에 주로 적용)
 - 먼저 항목의 값을 순서대로 배열하고 자료주체들을 여러 개의 유사한 그룹들로 나누어 구성한 후, 각 그룹 안에서 항목들의 값들의 순서를 임의로 바꾸는 방법이 대표적(순위교환)
- (사례) 학교와 성별을 이용하여 학생들을 세 그룹으로 나누어 그룹 안에서 성적을 임의로 교환(순위교환)

< 자료교환 전과 후에 대한 예시 >

학생번호	학교	성별	그룹	자료교환 전 성적		자료교환 후 성적	
				국어	수학	국어	수학
1	A	남	1	93	80	50	80
2	A	남	1	50	70	93	60
3	A	남	1	30	60	25	70
4	A	남	1	25	50	30	50
5	A	여	2	85	90	50	90
6	A	여	2	50	70	45	70
7	A	여	2	45	50	85	60
8	A	여	2	35	60	90	50
9	B	남	3	80	90	40	90
10	B	남	3	40	75	80	40
11	B	남	3	35	40	10	75
12	B	남	3	10	20	35	20

③ 부분 총계 (연속형)

- (정의) 자료주체들을 유사성에 따라서 여러 개의 그룹으로 나누고 각 그룹에 속한 연속형 항목의 값들을 하나의 대표값(예로 해당 그룹의 평균 또는 중앙값)으로 대체하는 방법
 - 부분 총계는 연속형 항목에 주로 적용되며 매출액, 순이익, 소득, 세금 등 매우 치우친 분포를 가지는 민감한 항목에 주로 적용
- (사례) 부동산 금액 순서대로 자료주체 3개씩 한 그룹으로 묶은 후, 3개의 금액을 각 그룹의 평균으로 대체하는 부분 총계 적용

< 부분총계 적용 전 예시 >

번호	거주시도	연령코드	부동산금액	저축금액	기타자산금액
1	충청남도	50~60세 미만	24,101,000	447,000	98,000
2	충청남도	60세이상	22,742,000	6,417,000	1,711,000
3	전라남도	60세이상	22,574,000	2,658,000	655,000
4	충청북도	50~60세 미만	19,420,000	577,000	339,000
5	충청북도	60세이상	18,814,000	1,196,000	104,000
6	충청북도	60세이상	18,768,000	420,000	60,000
7	강원도	60세이상	11,692,000	3,633,000	21,000
8	전라남도	60세이상	10,030,000	182,000	1,797,000
9	경기도	60세이상	9,870,000	269,000	91,000
10	전라북도	50~60세 미만	9,698,000	554,000	86,000
11	경상남도	60세이상	9,536,000	1,928,000	1,269,000
12	경상북도	30~40세 미만	8,755,000	585,000	173,000

< 부분총계 적용 후 예시 >

번호	거주시도	연령코드	부동산금액	저축금액	기타자산금액
1	충청남도	50~60세 미만	23,139,000	3,174,000	821,333.3
2	충청남도	60세이상	23,139,000	3,174,000	821,333.3
3	전라남도	60세이상	23,139,000	3,174,000	821,333.3
4	충청북도	50~60세 미만	19,000,667	731,000	167,666.7
5	충청북도	60세이상	19,000,667	731,000	167,666.7
6	충청북도	60세이상	19,000,667	731,000	167,666.7
7	강원도	60세이상	10,530,667	1,361,333	636,333.3
8	전라남도	60세이상	10,530,667	1,361,333	636,333.3
9	경기도	60세이상	10,530,667	1,361,333	636,333.3
10	전라북도	50~60세 미만	9,329,667	1,022,333	509,333.3
11	경상남도	60세이상	9,329,667	1,022,333	509,333.3
12	경상북도	30~40세 미만	9,329,667	1,022,333	509,333.3

◇ 임의화 방법별 정보손실에 대한 자세한 내용은 부록 III-4. 임의화 방법 (31~45쪽)을 참조할 것

3 매크로데이터 비식별화 방법

- 통계기초자료 및 행정통계자료를 이용하여 다양한 종류의 통계(매크로데이터)를 작성한 이후에도 발생할 수 있는 자료주체의 재식별 혹은 민감정보의 노출위험을 검토하고 조치를 취해야 함
- 세분화된 통계를 작성할수록 자료의 유용성이 증가하나 노출위험 역시 함께 증가하므로 주의할 필요가 있음
 - 세분화에 활용된 보조식별정보들이 생성하는 경우의 수와 자료의 개수, 생성하는 통계의 가짓수를 종합적으로 고려할 필요가 있음
 - 매크로데이터 비식별화 문제 및 필요성을 아래의 사례를 통해 이해
 - * (예제) 인구 1천명 A지역에서 연령(85)*성(2)*교육정도(8)별 평균소득 비교

소득 관련 통계

특정 지역의 인구특성별 소득 통계 작성 가상 사례

case	생년	성별	교육정도	연평균소득	표준편차	(빈도)
	...	...	...	...	...	...
①	69년	남	석사	78	50	17
②	69년	남	박사	48	154	9
	...	...	...	...	...	...
③	69년	여	석사	3,110	3,785	2
④	69년	여	박사	253	-	1
	...	...	...	...	...	...

(빈도표 행의 수(m) = 변수 조합으로 생기는 경우의 수 = 1,360)

- 위의 인구특성별 연평균소득 관련 통계의 특징
 - * 다수의 조합에서 빈도가 0인 경우 발생 (예. 5세 남 박사 등)
 - * ① 및 ②의 경우, 익명성을 충분히 확보하여 재식별이나 정보노출 위험이 낮음
 - * ③의 경우, 평균과 표준편차로부터 2인 각각의 소득 57.8억 및 4.3억이 노출됨
 - * ④의 경우, 해당 개인의 소득 2.5억이 즉시 노출됨
 - * ③ 및 ④의 경우, 지역·생년·성별·교육정도에 관한 정보로 특정 개인이 식별되는 경우, 통계 제공으로 개인의 연평균소득이 드러날 수 있음.

- 통계, 통계자료 분석결과, 매크로데이터 등의 비식별화를 위하여 다음에 설명하는 셀감추기, 재범주화, 반올림, 구간값 제공, 특이 정보 처리 등의 방법을 이용할 수 있음

- * 이는 전통적인 비식별화(마스킹) 기술이며, 최근 다양한 기법 연구 진행 중
- * 현재는 다양한 서식의 집계/분석 결과 예제를 대상으로 아래의 각 방법을 설명하고, 향후 통일된 서식 이용 제안을 포함한 가이드라인 업데이트 예정

1. 셀감추기

- (정의) 집계표에서 낮은 도수를 가진 셀의 값을 공개하지 않는 방법으로, k -익명성을 기준으로 하여 적용 가능
 - 셀감추기를 적용할 때 행과 열의 합계로부터 감추어진 셀의 값을 유추할 수 있으므로 이를 방지하기 위하여 추가적인 셀감추기를 적용함
 - 셀감추기는 집계표에서 정보를 삭제하는 것이므로 정보 손실이 큼
- (사례1) 집계표에서 셀의 도수가 3 미만인 경우 셀감추기 적용
 - ① 빈도수가 3 미만인 B 분류의 남성기업수와, 해당 항목수와 대응되는 남성 매출 합계에 1차로 셀감추기 적용 마스킹
 - ② 1차 셀감추기① 적용 후, 부분 합계가 있을 경우 추론 가능한 값에 2차로 셀감추기②를 적용해야 3미만 셀의 값이 드러나지 않음

< 셀감추기 적용 전 예시 >

산업분류	남성기업수	남성매출합계	여성기업수	여성매출합계	기업수총합	매출총합
A	238	28,914	76	7,248	314	36,162
B	2	32	347	26,571	349	36,603
C	64	7,842	32	4,752	96	12,594
D	783	62,483	4	742	787	63,225
전체합계	1,087	99,271	459	39,313	1,546	138,584

< 셀감추기 적용 후 예시 >

산업분류	남성기업수	남성매출합계	여성기업수	여성매출합계	기업수총합	매출총합
A	238	28,914	76	7,248	314	36,162
B	①	①	②	②	349	36,603
C	②	②	②	②	96	12,594
D	783	62,483	4	742	787	63,225
전체합계	1,087	99,271	459	39,313	1,546	138,584

○ (사례2) 305명에 대하여 혼인상태와 교육정도의 분포를 나타내는 요약 교차표에 셀 감추기 적용

- 여러 개의 교차하는 셀들에서 도수가 5 미만의 작은 수가 나타나므로, 도수가 5 미만인 셀을 감추고① 주변합과의 관계에 의한 노출을 막기 위하여 연관된 셀까지 모두 감춤②

< 셀감추기 적용 전 예시 >

혼인상태	초등학교 졸업	중학교 졸업	고등학교 졸업	대학졸업 이상	합계
미혼	2	3	12	13	30
배우자있음	12	22	68	86	188
사별	27	10	9	3	49
이혼	8	9	18	3	38
합계	49	44	107	105	305

< 셀감추기 적용 후 예시 >

혼인상태	초등학교 졸업	중학교 졸업	고등학교 졸업	대학졸업 이상	합계
미혼	①	①	12	13	30
배우자있음	12	22	68	86	188
사별	27	②	9	①	49
이혼	②	9	18	①	38
합계	49	44	107	105	305

◇ 매크로데이터에서 셀감추기 적용 절차에 대한 자세한 내용은 부록 III-5. 매크로데이터 비식별화 방법 해설(46~54쪽)을 참조할 것

2. 재범주화

- (배경) 셀 감추기는 정보손실이 크므로 범주의 개수를 줄여 정보 노출을 방지하면서도 정보 손실을 줄일 필요가 있음
- (사례) 아래 요약 교차표는 셀감추기의 예시에 나타난 교차표에서 셀의 도수가 작은 항목끼리 행(사별과 이혼)과 열(초등학교와 중학교)의 범주를 묶어서 재범주화 처리한 것으로, 모든 셀 값 제공 가능

< 재범주화 적용 전 예시 >

혼인상태	초등학교 졸업	중학교 졸업	고등학교 졸업	대학졸업 이상	합계
미혼	2	3	12	13	30
배우자있음	12	22	68	86	188
사별	27	10	9	3	49
이혼	8	9	18	3	38
합계	49	44	107	105	305

< 재범주화 적용 후 예시 >

혼인상태	중학교 졸업 이하	고등학교 졸업	대학졸업 이상	합계
미혼	5	12	13	30
배우자있음	34	68	86	188
사별/이혼	54	27	6	87
합계	93	107	105	305

3. 반올림

- (정의) 집계표를 구성하는 자료의 값이 너무 작은 경우 해당 값에 대하여 적절한 자릿수에서 반올림을 적용
 - 반올림을 적용한 후 각 열의 총합과 행의 총합을 다시 계산해야 하며 전체 총합이 원자료의 총합과 차이가 날 수 있다는 단점이 있음

4. 구간값 제공

- (정의) 집계표에서 노출의 위험이 크거나 작은 값을 대신하여 실제 값이 아닌 구간을 제공하는 방법
 - 구간값 제공은 상하단 코딩과 유사한 방법이지만 집계표의 유용성이 크게 감소하므로 전역 코딩 등 다른 방법을 적용하는 것이 일반적임

5. 특이정보 처리

- * 향후, 연구개발 사업 성과를 반영하여 좀 더 세부적인 기준 마련 필요
- (준)연속형 변수의 경우 최대값 및 최솟값을 제공하지 않음
- 개체의 빈도수를 고려하여 분위수를 비식별화하여 반출함
 - 비식별화하지 않으면 빈도수가 적은 그룹에서 개체의 정보 노출 가능
- * 아래의 예제와 같이 사업체를 대상으로 하는 자료에서 특정 항목 분위수를 제공하는 경우, 각 그룹의 빈도수 300 미만일 때 1% 분위수를, 60 미만일 때 5% 분위수를, 30 미만일 때 10% 분위수를 기준으로 양쪽을 비식별화 필요

산업분류	빈도수	P1	P5	P10	P25	P75	P90	P95	P99
1110	2379	1	1	5	16	246	731	1264	2257
1110	149	*	3	4	14	56	1015	1275	*
1121	785	1	1	3	8	126	492	842	3751
1142	9	*	*	*	7	9	*	*	*
10302	313	1	4	11	38	640	1525	2872	8512
10730	56	*	*	8	15	578	1423	*	*

- 집계표에서 빈도수가 적은 특정 그룹에 대해서는 참값을 유추할 수 없도록 제공되는 통계의 종류를 제한할 필요가 있음
 - 그룹의 빈도수 대비 다수의 통계를 제공하는 경우, 개체별 값이 식별될 수 있음(예를 들어, 빈도수가 3인 특정 그룹의 평균·표준편차·중앙값을 모두 반출하면 각 개체의 참값을 계산할 수 있음)
- * 다만, 자료 특성상 개체 식별이 불가능한 경우에는 특이정보가 제공되더라도 개인정보 노출로 규정할 필요가 없을 것으로 판단되며, 향후 특이정보 처리를 완화시키기 위한 지속적인 논의가 필요함

< 개체수에 따른 기초통계량 비식별화 적용 전 예시 >

구분	산업분류(중)	기업체수	평균	표준편차	최솟값	중앙값	최댓값
서비스	49	373	35,344	110,259	33	12,953	244,268
	50	369	17,126	51,064	10	14,425	236,828
	51	624	12,559	10,199	1	11,857	216,427
	52	3	39,954	127,869	23	18,864	348,777
	58	541	32,454	113,499	45	10,275	215,440
	59	725	10,847	90,215	57	1,072	229,246
	62	208	31,408	57,777	3	20,162	371,500
	63	381	10,117	115,301	43	21,045	412,335
	68	267	34,966	50,686	4	18,997	327,715
	71	12	10,869	107,716	42	28,076	499,751
	72	305	40,646	119,505	16	18,209	398,603
	73	427	11,092	94,353	32	11,520	215,663
	90	988	46,948	15,019	18	12,860	207,870
	계	5,223	26,087	78,998	1	19,715	499,751
건설	41	1	19,594	-	19,594	19,594	19,594
	계	1	19,594	-	19,594	19,594	19,594

< 개체수에 따른 기초통계량 비식별화 적용 후 예시 >

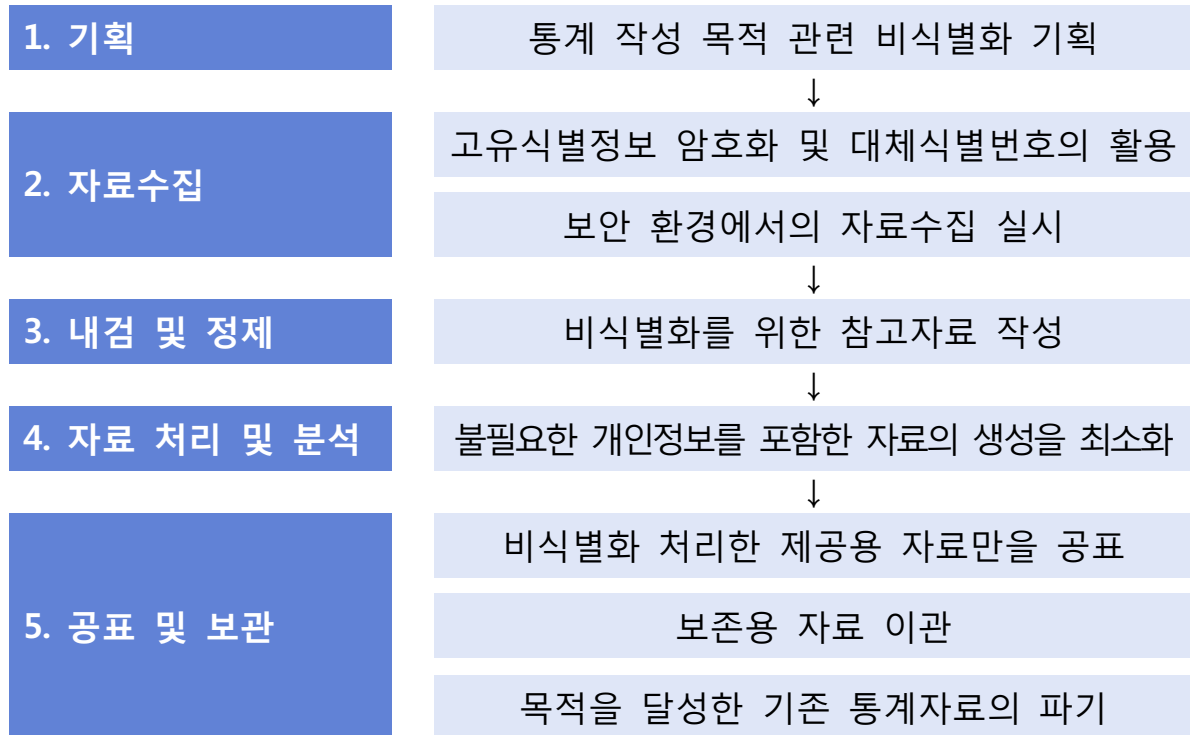
구분	산업분류(중)	기업체수	평균	표준편차	최솟값	중앙값	최댓값
서비스	49	373	35,344	110,259	*	12,953	*
	50	369	17,126	51,064	*	14,425	*
	51	624	12,559	10,199	*	11,857	*
	52	*	*	*	*	*	*
	58	541	32,454	113,499	*	10,275	*
	59	725	10,847	90,215	*	1,072	*
	62	208	31,408	57,777	*	20,162	*
	63	381	10,117	115,301	*	21,045	*
	68	267	34,966	50,686	*	18,997	*
	71	*	10,869	107,716	*	28,076	*
	72	305	40,646	119,505	*	18,209	*
	73	427	11,092	94,353	*	11,520	*
	90	988	46,948	15,019	*	12,860	*
	계	5,223	26,087	78,998	*	19,715	*
건설	41	*	*	*	*	*	*
	계	*	*	*	*	*	*



Ⅲ. 통계 작성 단계별 비식별화



< 통계 작성 단계별 비식별화 개념도 >



1 기 획

○ 통계 작성 목적 및 세부사항 검토

- 통계 작성을 위하여 수집된 개인이나 법인 또는 단체 등의 비밀에 속하는 자료는 통계작성 외의 목적으로 사용될 수 없으므로(통계법 제33조 제2항), 통계작성기관은 통계 작성목적 달성을 위한 통계 작성 사항, 대상, 방법, 자료수집 체계, 분류 또는 기준, 조사표 또는 통계표 등에서 비식별화가 필요한 부분을 검토하고 비식별화 계획을 수립하여야 함
- 특히 통계작성기관은 통계 작성 대상에 대하여 수집하려는 항목이 통계 작성에 필요한 최소한의 범위인지 검토하여야 함
- * 통계작성기관의 장은 통계의 작성을 위하여 질문을 받거나 자료제출 등의 요청을 받고 답변을 하거나 자료제출 등을 하는 개인이나 법인 또는 단체 등의 부담을 최소화하고, 비밀이 보호되도록 노력해야 한다(통계법 제4조 제3항).

- 예를 들어 통계 작성을 위해 필요한 고유식별정보는 암호화를 통해 비식별화를 수행하고, 성명, 전화번호, 상세주소, 이메일 등 민감한 항목은 통계작성에 필요한 최소한의 항목인지 검토

○ 통계 유형별 비식별화 기획

- (조사통계) 조사 항목이 통계 작성에 필요한 최소한의 범위인지 검토하고, 조사표* 설계 시 개인정보 수집을 최소화하는 방향으로 설계

* 조사표 항목 중 향후 비식별화가 필요한지 민감성을 검토

- (행정통계) 행정자료 제공기관의 고유식별정보 처리 근거*를 확인하고, 입수 대상이 되는 고유식별정보는 통계 작성 목적 달성을 위한 최소한의 범위로 제한

* 고유식별정보는 행정자료 제공기관에서 고유식별정보를 연계키로 요청한 경우나, 그 외 고유식별정보 처리가 불가피한 경우로 법령에 명확한 근거가 있는 경우에 한해서만 이용할 수 있음(통계법 시행령 제52조의3)

2 자료 수집

○ 고유식별정보 암호화

- 고유식별정보란 개인을 고유하게 구별하기 위하여 부여된 식별 정보로서 주민등록번호, 여권번호, 운전면허의 면허번호, 외국인 등록번호를 말함(개인정보 보호법 제24조, 동법 시행령 제19조)
- 통계작성기관(국가승인통계, 지정통계 해당)은 통계 작성을 위해 불가피한 경우 고유식별정보가 포함된 자료를 처리할 수 있음(통계법 시행령 제52조의3)
- 고유식별정보를 처리하는 경우에는 그 고유식별정보가 분실·도난·유출·위조·변조 또는 훼손되지 아니하도록 암호화 등 안전성 확보에 필요한 조치를 하여야 함(개인정보보호법 제24조 제3항)

- 고유식별정보 암호화를 위해서는 해시함수를 활용한 일방향 암호화*를 주로 사용함

* 일방향 암호화란 암호화한 뒤 다시 해독(복호화)할 수 없는 암호화 방법을 의미하며, 일방향 암호화를 위해 SHA256과 같은 암호화 알고리즘을 활용 (부록 참조)

- 일방향 암호화 방법에 의하여 일차암호화된 번호를 일반 상용 암호모듈(예: CubeOne)을 이용하여 64바이트의 이차암호번호를 생성하고, 최종적으로 통계작성기관의 고유의 암호화 방법을 이용하여 개인대체식별번호*를 생성할 수 있음

* 생성된 개인대체식별번호는 통계작성을 위해 수집한 자료연계를 위해 활용

◇ 일방향 암호화에 대한 자세한 내용은 부록 III-7. 일방향 암호화 해설 (58~59쪽)을 참조할 것

○ 조사통계 자료 수집 시 보호조치

- 개인정보 등 민감한 정보의 노출을 최소화하는 방법으로 조사를 수행하여야 함
- 조사자료 수집 시 부수적으로 발생하는 다양한 문서 또는 정보들 중 개인정보가 포함된 경우에는 이를 안전하게 관리하고 사용하여야 함
- 조사 결과 이동 및 처리 시 안전한 환경 하에서 수행하며, 웹구간 암호화 전송을 위한 SSL(Secure Sockets Layer)을 적용하고, 웹 방화벽 및 DB 암호화가 적용된 환경에서 수행하여야 함

○ 행정통계 자료 수집 시 보호조치

- 행정자료 입수 방법으로는 행정자료 수집 시스템(통계청), 행정정보 중계시스템(행안부), 직접 입수 방식이 있음
- 내부망과 분리된 행정자료 통합관리시스템 운영으로 보안성을 확보하며 원자료의 DB 적재와 동시에 원행정자료를 파기하고 파기공문을 발송

3 **내검·정제**

○ 비식별화 참고자료 생성

- 자료 내검·정제 시 자료주체·항목·특이정보 등을 검토하여 향후 제공용 통계자료 생성을 위한 비식별화 처리 시 참고하여야 할 사항들을 문서화하여 해당 부서로 한정하여 이관

4 **통계 작성**

○ 통계 작성 시 부수적 자료 생성 방지

- 불필요한 개인정보 등 통계적 비밀이 포함된 자료들이 부수적으로 생성되지 않도록 조치
- 통계 작성 목적의 자료연계를 위하여 생성한 대체식별번호는 연계 작업이 모두 종료된 후 파기

5 **공표·보관**

○ 공표자료 비식별화 평가

- 집계표, KOSIS 제공자료, 보도자료 등에 대한 비식별화 평가 및 추가 비식별 처리를 적용

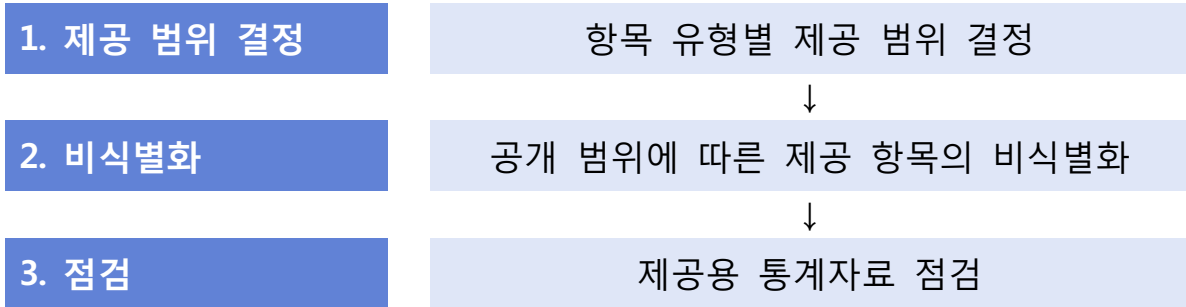
○ 자료 폐기 및 보존

- 통계 작성 목적을 달성한 행정통계자료와 파생자료는 폐기하고, 보존이 필요한 자료는 자료 보존의 절차에 따라서 보관 조치
- 통계청 통계작성부서의 경우 생산·공표한 통계조사 원자료·행정통계를 통계청(마이크로데이터과)으로 이관(보존용 전산자료 이관)
- 통계법 제27조의2(통계작성·공표과정에서의 영향력 행사, 누설 및 목적 외 사용의 금지) 내용을 준수하고, 보관은 보존용 통계자료 관리 및 이용지침 제15조, 제16조와 파기 의무를 준수

IV. 통계자료 제공단계별 비식별화

1 제공용 통계자료 생성

< 제공용 통계자료 생성단계별 비식별화 개념도 >



관련 규정
통계법 시행령 제47조(통계이용자의 통계자료 신청과 제공) 국가통계 자료제공규정 제4조(제공원칙) 국가통계 자료제공규정 제6조(자료 제공범위) 보존용 통계자료 관리 및 이용지침 제12조(제공용 자료 생성) 인가된 통계자료 이용절차 등에 관한 규정 2조(정의)

1. 제공 범위 결정

① 개체식별정보에 대한 검토

- (정의) 개체식별정보는 자료주체의 신분을 직접적이고 명확하게 나타내는 항목으로서 직접식별변수, 고유신분정보라고도 불림
- 개체식별정보는 개인의 신분 등을 직접적으로 노출하는 항목으로, 통계기초자료에 다음과 같은 식별 항목이나 응답자에 대한 정보로 나타나며 개인정보와 관련된 항목은 삭제가 원칙

식별 정보	주민등록번호, 여권번호, 가구번호, 사업체명, 사업체 대표자명, 사업자 등록번호 등
응답자 정보	응답자의 성명, 직급명, 전화번호, 이메일, 팩스번호 등

② 보조식별정보에 대한 노출의 위험성 검토

- (정의) 보조식별정보는 자료주체의 신분에 대한 직접적인 정보는 없지만 다른 보조식별정보와 결합하여 자료주체의 신분을 식별할 수 있는 위험성을 증가시킬 수 있는 정보
- (사례) 보조식별정보의 예로는 생년월일, 성별, 우편번호, 주거 지역 등이 있으며 대부분 보조식별정보 자체의 값만으로 자료주체의 신분 노출이 일어나지 않지만, 여러 개의 보조식별정보의 조합은 자료주체를 유일한 특성을 지니게 만들어 재식별 가능성을 크게 높일 수 있으므로 주의해야 함
 - 예를 들어 성별, 연령, 혼인여부, 직업은 각각 고려하면 개인정보의 노출의 위험이 매우 약하지만 정보를 서로 조합하면 노출의 위험이 증가할 수 있음 (기혼이며 35세인 여성 비행사)
 - 특히 자료주체에 대한 소재지 또는 거주지에 대한 지역 정보는 재식별의 위험이 큰 보조식별정보이므로 주의해야 함

③ 보조식별정보가 아닌 항목 중 이용목적에 부합하기 위하여 제공하는 연속형 및 준연속형항목 선정

- (고려사항) 제공용 자료의 공개 항목을 선정할 때는 다음과 같은 사항들을 고려해야 함
 - 제공용 자료의 생성에 두 개 이상의 보존용 통계자료가 필요한 경우에는 각 자료들을 구성하고 있는 자료주체와 항목들 간의 일치성과 차이점에 대한 점검을 실시
 - 표본오차, 비표본오차 등 통계적 오차가 크다고 판단되는 경우, 자료 이용 시 통계의 신뢰성이 훼손될 가능성이 있으므로 이용자의 혼란을 방지하기 위하여 공개하지 않을 수 있음
 - 원자료에 포함되어 있는 관리용 항목이나 내부검토용 항목 등은 특별한 이유가 없는 한 비공개를 원칙으로 함

* 예를 들어 자료업로드일자, 작성책임자, 조사원 정보 등은 비공개

- 조사의 응답 형식이 문장이나 긴 설명으로 수집된 경우 재식별의 위험성이 다른 범주형 또는 연속형 항목보다 크므로 공개를 하지 않을 수 있음

* 예를 들어 답변이 구체적인 문장으로 수집된 경우 응답자의 신분이나 민감한 항목이 노출될 위험이 있음

- 제공자료의 작성 시기에 따라서 항목이 다른 경우 이용자의 편의를 고려하여 시계열적 분석이 가능하도록 항목을 선정하며 폐지된 과거 항목은 공개하지 않을 수 있음

○ (사례) 다음 예시는 공개범위 선정 예시로서 2020년 출생에 대한 인구동향조사자료를 제공할 경우 공개범위를 정하고 그에 대한 해당 항목을 선정한 사례임

번호	항목명	제공여부	제공 불가 사유
1	연도	○	
2	신고날짜	○	
3	출생자 주소지(시도)	○	
4	출생자 주소지(시군구)	○	
5	성별	○	
6	혼인중 외의자	○	
7	출생년월일(년)	○	
8	출생년월일(월)	○	
9	출생장소	○	
10	부 생년월일	X	보조식별정보
11	부 교육정도	○	
12	부 연령	○	
13	모 생년월일	X	보조식별정보
14	모 교육정도	○	
15	모 연령	○	
16	부모 결혼년월일(년)	○	
17	임신주수	○	
18	다태아 출산순위	○	
19	출생아 체중	○	
20	모 총출생아수	○	
21	부모동거기간	○	
22	임신월수	X	민감정보
23	자료업로드일자(년)	X	관리용 항목
24	부국적	○	
25	모국적	○	

2. 비식별화

○ 항목 유형별 비식별화 방법 검토

- 제공하기로 결정된 항목들에 대하여, 항목 유형에 맞게 비식별화 방법*을 적절하게 선정하여 적용함

* 비식별화 방법에 대한 자세한 설명은 2장 비식별화 방법 또는 부록 참조

- 아래 표는 예시로서 개체식별정보, 보조식별정보, 연속형 항목 등과 그에 적합한 비식별화 방법을 제시한 표임

< 항목별 비식별화 방법 예시 >

항목	비식별 처리 방법 / 예시
개인정보	항목 삭제 / 주민등록번호, 이름 등
민감한 항목 또는 표본 및 비표본 오차가 많이 발생하는 항목	항목 삭제 / 소득, 부채 등 비표본오차가 많이 발생하는 항목
개별 조사대상의 개수가 매우 적은 항목	국소 삭제, 통합 / 행정구역에서 사업체수가 2개 이하인 경우
재식별 위험으로 지역별 집계 기준의 세분화 정도를 제한해야 하는 경우	재범주화 또는 항목 삭제 / 시군구 정보는 공개하지 않고 시도만 공개
항목의 범주의 값에 해당하는 정보주체의 수가 적은 경우	전역 코딩 또는 상하단 코딩 / 인구주택총조사 자료 중 85세부터는 개별 조사 대상이 적기 때문에 85세 이상 그룹으로 묶어서 제공
민감하고 특이한 정보를 가지면서 분포가 매우 치우치게 나타나는 연속형 변수	부분총계 또는 총계처리, 반올림 / 소득, 매출액 등 경제 관련 항목
민감하고 특이한 정보를 가지면서 분포가 치우치지 않고 대칭에 가까운 연속형 변수	잡음첨가 또는 자료교환 / 성적, 점수, 기록 등

- 다음 예시는 공개범위 선정과 비식별화 예시로 출생에 대한 인구 동향조사자료를 제공할 경우 2개의 공공용 자료로 공개범위를 정함
- 다양한 수준의 정보의 유용성(연령 지역 정보)을 고려하고 개인 정보 유출의 위험성을 낮추기 위하여 2개의 제공용 자료(A형과 B형) 작성
- A형은 연령에 대한 정보의 유용성을 고려하여 주소를 시도만 공개
- B형은 지역에 대한 유용성을 고려하여 주소를 시군구로 공개

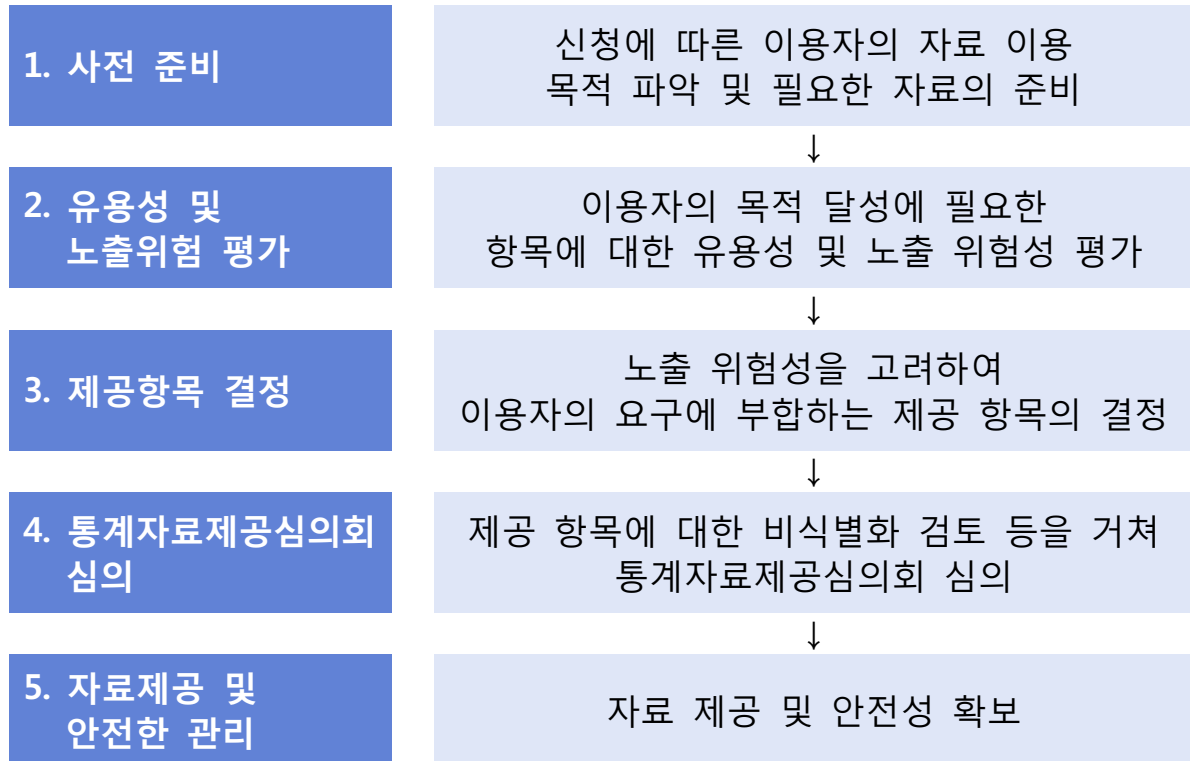
번호	항목명	제공여부	제공 불가 사유	제공시 유의사항
1	연도	○		
2	신고날짜	○		
3	출생자 주소지(시도)	○		공공용 A형만 제공
4	출생자 주소지(시군구)	○		공공용 B형만 제공
5	성별	○		
6	혼인중 외의자	○		
7	출생년월일(년)	○		
8	출생년월일(월)	○		
9	출생장소	○		
10	부 생년월일	X	보조식별정보	
11	부 교육정도	○		
12	부 연령	○		공공용 A형만 제공 B형은 5세 단위로 전역 코딩
13	모 생년월일	X	보조식별정보	
14	모 교육정도	○		
15	모 연령	○		공공용 A형만 제공 B형은 5세 단위로 전역 코딩
16	부모 결혼년월일(년)	○		
17	임신주수	○		
18	다태아 출산순위	○		
19	출생아 체중	○		소수점 셋째자리까지 제공
20	모 총출생아수	○		
21	부모동거기간	○		
22	임신월수	X	민감정보	
23	자료업로드일자(년)	X	관리용 항목	
24	부국적	○		
25	모국적	○		

3. 점검

- 보존용 자료의 메타 자료를 기반으로 비식별화가 적용된 제공용 자료에 대한 메타자료를 작성
- 제공용 자료에 대한 이용자의 편의를 위하여 항목의 이름, 항목의 형식들을 포함한 설명 자료(자료 설명 및 이용 안내서, 파일 설계서, 코드표, 분석 프로그램 등)을 마련
- 비식별화가 적용된 제공용 자료가 보존용 자료와 동일한 정보를 제공하는지에 대한 점검을 수행
 - 보존용 자료에 의하여 공표된 집계표가 비식별화가 적용된 제공용 자료로 생성된 집계표와 일치하는지 점검
 - 보존용 자료에 의하여 공표된 주요 통계량(예로 총합, 평균, 비율 등)이 비식별화가 적용된 제공용 자료로 생성된 통계량과 일치하는지 또는 유사한지 점검
 - 적용된 비식별화 방법에 의하여 보존용 자료와 제공용 자료의 집계표 또는 통계량이 일치하지 않을 수 있으며, 이러한 불일치에 대한 주요한 내용은 이용자의 편의를 위하여 제공용 자료 설명 자료에 포함

2 이용 신청에 따른 통계자료 제공

< 이용 신청에 따른 통계자료 제공단계별 비식별화 절차도 >



관련 규정
통계법 시행령 제48조 제1항 국가통계 자료제공규정 제4조(제공원칙) 통계자료제공심의회 규정 제4조(심의회 기능) 통계데이터센터 운영 및 이용규정 4장 센터이용방법과 절차 인가된 통계자료 이용절차에 대한 규정 3장 이용절차

1. 사전 준비

- 특정 이용자가 통계 작성, 정책분석, 연구 등 특정 목적으로 앞의 절에서 검토한 제공용 자료의 기본 제공 범위·방법*을 벗어나서 요청하는 경우에 이용자의 목적 및 제공 자료·방법을 검토하여야 함

* 불특정 다수를 대상으로 제공용 자료의 최초 제공 시 방법·범위의 설정도 통계자료제공심의회 심의를 거쳐 설정

- 이용 목적을 이미 제공된 자료를 통하여 달성할 수 있다고 판단 되면 이용자에게 공개된 자료 중 적합한 자료를 제공
 - 이용자의 신청 목적에 따라서 제공 자료를 만들기 위하여 필요한 원자료(통계기초자료, 행정통계자료, 명부자료, 미공표집계자료, 공간좌표자료 등)를 준비
 - 제공 자료의 생성에 필요한 보조자료가 필요한 경우 또는 두 개 이상의 원자료가 필요한 경우 자료들을 구성하고 있는 자료주체와 항목들에 대한 일치성과 차이점에 대한 점검을 실시
- 이용자의 요구는 다음과 같이 다양한 경로와 형태를 통하여 수집할 수 있음
 - (자료 형태) 관련 연구계획서, 업무추진계획, 요청 레이아웃 등으로 파악
 - (분석 목적 및 방법) 이용 신청 시 제시한 자료 사용의 방법(예를 들어 빈도분석, 관계분석, 회귀분석 및 시계열 분석 등)을 통하여 파악

< 이용 신청 시 수집 항목 사례 >

구분	세부 내용
이용지역	전국 시군구 단위
이용목적	소지역(시군구)별 가구 소득의 평균과 분포를 추정하고 지역별 소득 격차를 비교하기 위한 연구
이용자료	자료명(기준년(월)): 가계금융복지조사 (2017) 항목명: 시군구코드, 가구주정보(학력, 연령 등), 가구원수, 소득, 자산, 부채 등 경제 항목
이용기간	6개월
예상결과표 또는 반출자료형태	시군구별 가구의 경제 항목(소득, 자산, 부채 등)에 대한 요약 통계(평균, 분위수 등)

2. 유용성 및 노출위험 평가

① 항목의 유용성과 특성 파악

- 이용자가 필요한 항목들에 대하여 이용자 측면에서의 유용성을 파악
 - 항목의 유용성은 이용자의 요구와 자료의 특성에 따라서 달라질 수 있음
- 특이정보를 포함한 항목 또는 값을 파악
 - 특이정보는 정보주체를 직접 식별할 수는 없지만 특정한 정보 주체의 항목의 값이 대다수의 값과 다르거나 특이한 정보를 의미
 - * 예를 들어 각종 소득의 최댓값은 재식별 위험이 매우 큰 특이정보
 - 특이정보는 외부 자료와 쉽게 결합하여 재식별의 위험성을 크게 증가시킬 수 있으므로 주의해야 함
 - 특이정보 중에서 대중의 관심이 매우 높거나 재식별 시 사회적 파장 등 영향도가 높은 특정 정보주체 또는 항목의 값을 민감 정보라고 하며 특별한 검토가 필요함

② 항목의 노출 위험성 평가

- 항목들에 대한 노출의 위험성을 측정
 - 제공이 필요한 항목들 중에서 보조식별정보를 조합하여 나타낼 수 있는 노출의 위험성을 측정(유일성, k -익명성, ℓ -다양성 등을 이용).
 - 제공 자료를 생성하는 원자료의 범위(표본 자료 또는 센서스 자료)에 따라 노출의 위험성이 다르므로 이를 고려하여 위험성을 측정
 - 연속형 항목이라도 자료주체의 정보를 지니고 있어 재식별을 위한 항목으로 사용될 수 있는 경우를 파악
 - * 예를 들어 다자녀가구에 대한 세금 감면액은 연속형 변수이지만 금액이 0보다 크면 다자녀가 있다는 정보가 유출되어 보조 식별정보로 사용될 수 있음
- 이용자가 요구하는 항목들 중에서 외부에 존재하는 결합 가능한 정보들을 파악하고 이를 노출 위험 평가에 반영

③ 자료의 유용성 반영 지표 선택 및 이용 조건 파악

- 자료를 이용할 때 매우 중요하다고 판단되는, 즉 유용성이 큰 항목들을 다른 항목들과 구별하여 선택하고 유용성의 지표로 선정
 - 예를 들어 경제활동인구조사는 실업률, 취업자 수 등 관심 있는 주요 지표에 관련된 항목을 유용성 지표로 고려
- 자료의 이용 시 중요한 조건을 가진 항목을 파악
 - 특히 비식별화 처리 시 항목들 간의 구조적 관계*나 특성**은 훼손되지 않게 유의해야 함
 - * 항목들의 산술적 관계(항목들의 총계, 비율, 차이), 종속관계(폐업일자가 결측이면 폐업 여부도 결측) 등 항목들의 구조적 관계를 파악
 - ** 예를 들어 경제활동인구와 관련하여 성인은 15세 이상이므로 비식별화 처리 시 연령을 10-20세 구간으로 적용하지 않도록 주의

3. 제공항목 결정

- 항목들에 대하여 유용성과 재식별 위험성 평가를 종합하여 항목에 대한 평가를 수행
 - 종합적인 평가 결과 공개를 할 수 없는 항목을 선정하여 공개 항목에서 제외
- (사례) 여러 항목에 대한 노출 위험성과 유용성을 평가
 - 예를 들어 주민등록번호, 주소는 개체 식별정보로 공개하지 않으며 연령과 취업일자는 외부정보와 연계 가능성이 높으므로 비공개로 처리
 - 또한, 장애인여부와 다자녀가구_세금감면액은 특이정보인 동시에 민감한 정보이므로 공개하지 않음

< 항목에 대한 노출 위험성과 유용성 평가 사례 >

항목	항목분류	개인식별정보	보조식별정보	민감정보	노출위험성	유용성	공개여부
주민등록번호		예		예	매우높음		비공개
주소		예		예	매우높음		비공개
생년월일		예		예	매우높음		비공개
연령	준연속형		예		보통	높음	
성별	범주형		예		낮음	높음	
취업일자	준연속형		예		높음	보통	비공개
교육정도	범주형		예		낮음	높음	
직업	범주형		예		낮음	높음	
종사상지위	범주형		예		낮음	보통	
가구주여부	범주형		예		낮음	보통	
외국인여부	범주형		예		높음	보통	
장애인여부	범주형		예	예	높음	보통	비공개
혼인상태	범주형		예		낮음	높음	
주거형태	범주형		예		낮음	높음	
자산금액	연속형				낮음	높음	
부채금액	연속형			예	낮음	높음	
경상소득	연속형				낮음	높음	
근로소득금액	연속형				낮음	높음	
다자녀가구 _세금감면액	연속형		예		높음	보통	비공개

- 자료생산부서, 자료제공부서, 자료관리부서 등의 협의를 통하여 항목의 제공 범위 및 방법에 대한 의견을 수렴
- 최종적으로 이용자의 신청, 유용성과 재식별 위험성 평가, 관계 부서의 의견 등을 종합하여 제공 항목을 결정
- 항목들에 대하여 유용성과 재식별 위험성 평가에 대한 검토 내용은 추후 통계자료제공심의위원회에 심의 내용으로 사용될 수 있음

4. 통계자료제공심의회 심의

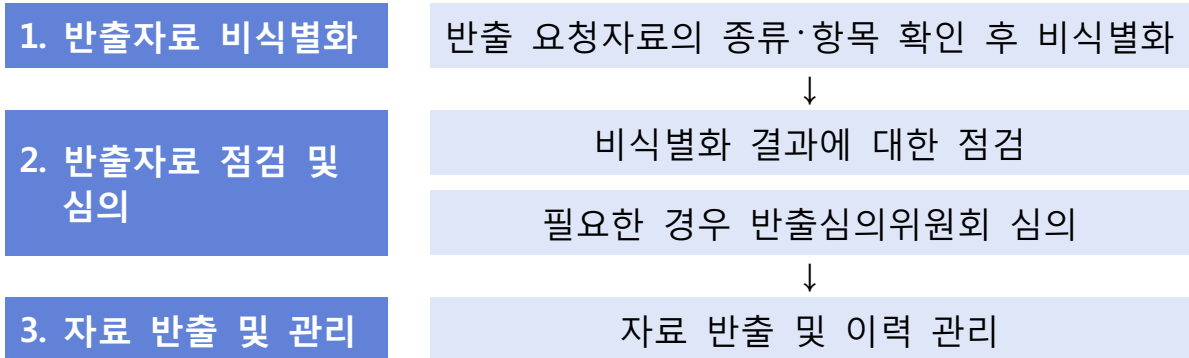
- 통계자료제공심의회 등 심의기구를 통하여 최종적인 자료 제공 방법과 항목을 결정
 - 통계자료제공심의회에서는 자료의 제공 범위(공공용, 인가용 등)에 관한 사항, 제공을 하지 않는 항목에 대한 적정성 검토, 제공하는 항목의 재식별 위험성 검토, 제공하는 항목에 적용한 비식별 처리 방법에 대한 적합성, 그 외의 민감한 항목이나 정보에 대한 검토 등을 수행할 수 있음
- 특히 제공하려는 자료가 식별할 수 없는 형태로 처리되었는지 보조 식별정보에 대한 위험성을 평가하고 제공 항목의 유용성에 대한 검토를 수행
 - 제공 항목들 중에서 보조식별 정보와 연속형 항목은 각 항목의 특성에 따라 적합한 비식별화 방법을 적용하였는지* 검토
 - * 각 항목의 특성에 적합한 방법의 선정 여부, 각 항목의 재식별 위험성과 유용성을 고려한 방법의 선정 여부, 여러 항목들의 관련성을 유지할 수 있는 적절한 방법의 선정 여부 등
 - 비식별화 전후 노출 위험성* 및 자료 유용성 비교
 - * 노출 위험성 평가는 k -익명성과 같은 형식적 기준이나 개인정보 노출 위험도와 같은 수량 측도를 주요 평가 기준으로 이용할 수 있음

5. 자료제공 및 안전한 관리

- 비식별화 검토가 완료되고 통계자료제공심의회의 승인을 받은 제공용 자료는 선정된 공개 범위와 방법에 따라서 이용자에게 제공
- 제공된 자료가 안전하게 이용될 수 있도록 관리 지침을 마련하고 시행
 - 제공되는 통계자료의 안전성 확보를 위하여 필요한 조치를 마련할 것을 요청

3 통계데이터센터 분석결과 반출

<통계데이터센터 분석결과 반출단계별 비식별화 개념도>



관련 규정
통계데이터센터 운영 및 이용규정 21조(분석결과의 반출) , (자료반출심의위원회) 인가된 통계자료 이용절차에 대한 규정 3장 이용절차

< (참고) 통계데이터센터 제공 행정통계자료(15종) >

부 문	자료명	k-익명성 기준
경제·사회	사업장기초DB(4대보험)	3 미만 값 및 이와 연관되어 추론 가능한 값
	사업장기초DB(법인)	
	기업통계등록부(대표자기준)	
	기업통계등록부(사업자등록번호기준)	
	기업통계등록부(대표자기준_분기)	
	기업통계등록부(사업자등록번호기준_분기)	
	종사자-기업체_연계DB	
인구·가구	인구가구통계등록부(인구)	5 미만 값 및 이와 연관되어 추론 가능한 값
	인구가구통계등록부(가구)	
	주택통계등록부	
	인구동태 코호트DB(1983~1993년생)	
	아동가구통계등록부	
농림어업	농업DB	
	임업DB	
	어업DB	

1. 반출자료 비식별화

① 반출 요청 자료의 종류 확인

- 인가용 제공자료 및 기타 제공자료의 항목 값은 외부로 반출이 불가능
- 개체식별정보가 포함되거나 재식별이 가능한 형태의 자료는 원칙적으로 반출이 불가능
 - ※ 예외는 통계데이터센터 운영 및 이용 등에 관한 규정 21조 참조
- 반출을 위한 자료는 원칙적으로 다음과 같은 형태로 반출하여야 함
 - 빈도표, 총계표, 기술통계, 시각화 자료(원자료값이 라벨링되지 않은 자료) 등 재분석·추가분석이 불가능한 최종 결과물 형태로만 반출이 가능
- 재식별에 이용할 수 없는 모형 적합 결과와 분석 프로그램은 반출 가능
- 프로그램을 실행할 때 발생하는 중간 결과, 로그 메시지, 경고 메시지 등은 재식별 정보를 포함할 수 있으므로 원칙적으로 반출 불가능
- 통계분석 프로그램 소스코드를 반출 시 재식별 위험을 검토
 - 통계분석 프로그램 소스코드에서 숫자, 문자 등 원자료를 유추할 수 있는 주석 등은 반드시 삭제하고 반출
 - 통계분석 시 생성된 로그파일은 반출이 원칙적으로 불가능

② 반출 요청 자료의 항목 확인 및 위험성을 검토

- 집계표에서 미리 지정한 지역과 분류의 단위보다 작은 단위에 대한 집계는 반출이 원칙적으로 불가능
- 집계표에서 k -익명성의 기준을 위반하면 원칙적으로 반출할 수 없음
 - 집계표에서 빈도에 나타날 수 있는 k 값*을 정하여 정해진 값보다 작은 경우 비식별화 처리를 적용

* 전통적으로 k -익명성 기준으로 k 값은 사업체는 3, 인구·가구·주택은 5 이용

- 특이정보가 포함된 집계표는 반출이 불가능
 - 특이정보로는 최솟값, 최댓값을 포함한 특정 분위값이 있음
 - * 모형의 결과를 시각화하여 반출할 시 라벨링이 되어 있지 않다면 반출 가능
- 반출 요청자료의 항목을 확인하고 위험성을 검토한 후 이용자에게 추가적인 비식별방법을 적용하여 반출이 가능한 형태로 다시 반출자료를 만들 수 있도록 조치

2. 반출자료 점검 및 심의

- 반출 요청 자료에 대한 검토와 비식별화를 적용한 후에 반출자료의 형식, 항목 등을 최종적으로 점검
- 분석 결과 등의 반출에 대한 승인을 위해 자료반출평가위원회*에서 심의를 진행
 - * 개인정보보호, 데이터 관리, 빅데이터 등 담당 부서 또는 전문가로 구성
- 자료반출평가위원회의 최종 심의에 따라서 다시 검토할 사항이 발생하면 반출자료를 다시 점검하여 비식별화 처리를 적용

3. 자료 반출 및 관리

- 자료반출평가위원회 최종 심의에 따라서 반출이 허가된 자료를 이용자에게 제공
- 자료반출평가위원회 최종 심의에 따라서 반출이 허가되지 않은 경우 이용자에게 심의 결과를 통보
- 비식별화 처리에 대한 이력 관리를 수행
 - 이용자의 이용 신청 이력을 기록하여, 향후 업무에 참고할 수 있도록 이용 신청에 따른 비식별화 및 제공 승인 이력을 기록
 - * 과거 반출 이력을 현재의 심의 및 승인 과정에서 고려하거나, 차등정보 보호(differential privacy) 혹은 신규 알고리즘을 도입하여 반복적인 자료 분석에 의한 개체별 노출위험을 제어할 수 있는 방안의 지속적 검토 필요

V. 보유 정보의 안전한 관리를 위한 보호조치

1 관리적 보호조치

- 제공용 자료를 관리하는 부서는 비식별화된 정보를 보호하고 자료를 안전하게 관리하기 위하여 다음과 같은 사항에 대한 내부 관리 계획을 수립·시행하여야 함
 - 비식별화된 제공용 자료에 대한 관리 담당자 지정
 - 비식별화 관련 방법 및 암호화 모듈 등 정보공유 금지
 - 제한구역 지정 및 출입자에 대한 출입 기록 관리
 - 이용목적 달성 시 파기 등
 - 비식별화 관련 정보 유출시 보호조치
- 비식별화 방법을 적용하여 제공용 자료를 생성하는 경우 다음과 같은 사항을 지속적으로 모니터링하고 보관해야 함
 - 비식별화 처리 목적
 - 비식별화를 적용한 개인정보의 항목
 - 비식별화 방법에 관한 사항

2 기술적 보호조치

- 제공용 자료를 관리하는 부서는 비식별화 정보 및 추가정보에 대하여 다음과 같은 사항에 대해 기술적 보호조치를 하여야 함
 - 비식별화 또는 암호화 관련 정보 및 추가정보의 분리 보관
 - 비식별화된 정보 및 암호화 관련 정보에 대한 접근권한 관리
- 제공용 자료 이용 시 접근 통제 및 접속기록의 보관 및 점검에 대한 기록을 모니터링할 수 있는 조치 수행
 - 행정자료는 원격분석시스템을 통해 이용 및 승인된 사용자만 접근 권한 부여

- 비식별화된 정보 처리 관련 기록 작성·보관
 - 시스템에 접속한 접속기록(접근권자, 자료명, 시간 등) 관리
 - 보안프로그램 설치·운영
- 제공용 자료를 관리하는 부서는 외부 자료 연계 시 다음과 같은 사항을 준수하여야 함
- 연계키 생성을 위해 주민등록번호를 활용하는 것은 원칙적 금지
 - 연계 과정에서 재식별 발생시 해당 정보를 즉시 파기

3 물리적 보호조치

- 비식별화된 제공용 자료를 이용하는 공간에 대하여 다음과 같은 사항을 고려하여 출입 통제 등의 절차를 수립하여야 함
- 인터넷망 차단, 업무전용망 구성
 - IC카드 등으로 접근 통제
 - 노트북, USB메모리 등 자료의 유출을 쉽게 할 수 있는 저장매체 등 반입 금지
 - 외부기관 소유 독립 서버는 시건장치가 있는 보관함에 보관



참 고 문 헌



- 관계부처 합동(2016). 개인정보 비식별화 조치 가이드라인 -비식별 조치 기준 및 지원·관리체계 안내-
 - 과학기술정보통신부, 한국정보화진흥원(2019). 개인정보 비식별 기술 가이드라인
 - 개인정보보호위원회(2020). 가명정보 처리 가이드라인
 - 개인정보보호위원회, 보건복지부(2020). 보건의료 데이터 활용 가이드라인
 - 교육부(2020). 교육분야 가명정보 처리 가이드라인
 - 금융위원회, 금융감독원(2020). 금융분야 가명·익명처리 안내서
-
- 김경미, 이의규, 정미옥 (2007). 마이크로데이터 제공에 관한 해외사례연구, 통계청 통계개발원.
 - 박민정, 권순필, 심규호 (2013). 가계금융·복지조사 마이크로데이터 제공을 위한 매스킹 방안, 통계청 통계개발원..
 - 박민정(2014). 『가계금융·복지조사 매스킹 방안 평가』, 통계청 통계개발원.
 - 박민정(2015). 『마이크로데이터 매스킹 기법 실무 활용 안내서』, 통계청 통계개발원.
 - 박민정(2016). 『통계적 노출제어 최신동향 검토 - 차등정보보호를 중심으로 -』, 통계청 통계개발원.
 - 박민정, 김정연(2017). 『재현자료 작성 방법론 검토』, 통계청 통계개발원.
 - 박민정, 성병찬, 최유성(2018). 『통계작성기관 마이크로데이터 제공범위 설정 가이드라인 작성』, 통계청 통계개발원.
 - 박민정, 이용희, 권성훈(2018). 『차등정보보호에 관한 연구』, 통계청 통계개발원.
 - 박민정, 한정임, 박노성(2020). 『통계데이터센터 DB를 활용한 재현자료 생성 방법 연구』, 통계청 통계개발원.
-
- Bethlehem, J. G., Keller, W. J. and Pannekoek, J. (1990). "Disclosure Control of Microdata", Journal of the American Statistical Association, 85(409). 38~45.
 - Differential Privacy Team, Apple (2017). Learning with Privacy at Scale, Apple Machine Learning Journal.
 - Duncan, G. T. and Lambert, D. (1989). "The Risk of Disclosure for Microdata", Journal of Business and Economic Statistics, 7, 207-217.
 - 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, 262-272. Springer-Verlag.
 - 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.
 - Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkatasubramanian, M. (2006, April). ℓ -Diversity: Privacy Beyond Kappa-Anonymity, in Proceedings of the 22nd International Conference on Data Engineering (ICDE '06). IEEE.
 - Meindl, B., Templ, M. and Kowarik, A. (2013). Guidelines for the Anonymization of Microdata Using R-package sdcMicro.
 - NCHS (2010). NCHS RDC Disclosure Manual: Rules for Researchers, NCHS.

- Park, M. and Kim, H. (2016). Statistical disclosure control for public microdata: present and future, Korean Journal of Applied Statistics, 39(6). 1041-1059.
- Reiter, J. P. (2004) New approaches to data dissemination: A glimpse into the future. Chance, 17:3 (Summer 2004), 12 – 16.
- Skinner, C. and Shlomo, N. (2008). "Assessing Identification Risk in Survey Microdata Using Log-Linear models", Journal of the American Statistical Association, 103(483). 989-1001.
- Statistics Netherlands (2007). μ -Argus version 4.1 User's manual.
- Sweeney, L. (2002). k -anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, 10, 557-570.
- Sweeney, L. (2013). Matching Known Patients to Health Records in Washington State Data. Harvard University. Data Privacy Lab. 1089-1.

통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인 부록

2023. 5. 30.



통 계 청
통계데이터기획과

목 차

목 차	
I. 노출위험 평가 척도	75
1. 자료주체의 유일성	76
2. 특이정보	79
3. k-익명성	85
4. l-다양성	87
5. 재식별위험 평가척도	88
II. 정보손실 평가 척도	91
1. 범주형 항목	92
2. 연속형 항목	92
III. 비식별화 세부기법 해설	94
1. 구조적 방법	95
2. 삭제 방법	96
3. 일반화 방법	100
4. 임의화 방법	105
5. 매크로데이터 비식별화 방법	120
6. 재현자료	129
7. 일방향 암호화	132

I . 노출위험 평가 측도

1 자료주체의 유일성

□ 개 요

- 통계기초자료(이하 ‘마이크로데이터’)를 공개할 때 재식별에 의한 개인정보 노출은 자료주체(사람, 기업 등)가 공개된 자료에서 유일할 경우 발생할 가능성이 크다.

◇ 예를 들어 2000명으로 구성된 자료에서 연령이 90세 이상인 사람이 단 한 명만 있다고 하자. 그 사람은 자료 안에서 유일하므로 식별될 수 있는 위험이 있다.

- 주어진 자료에서 유일성은 하나의 항목에 의하여 나타날 수도 있지만 여러 개의 항목들을 조합하여 발생할 수도 있다. 따라서 고려하는 식별정보의 수(예 : 나이, 성별, 학력, 지역 등)가 점점 많아질수록 유일성을 가지는 자료주체가 늘어나게 된다.

◇ 예를 들어 2000명 중 연령이 90세이면서 남자인 사람이 3명, 90세이면서 여자인 사람이 1명 있다고 한다면 90세인 사람은 4명으로 유일하지 않으나 90세이면서 여자인 사람은 1명으로 유일하다.

- 자료주체의 유일성을 판단할 때 범주형 항목과 연속형 항목에 대한 기준이 각각 다르게 나타난다.
 - (범주형 항목) 범주형 항목인 보조 식별정보들의 조합으로 유일성을 판단
 - (연속형 항목) 해당 값이 극단치 등 특이정보인지로 유일성을 판단

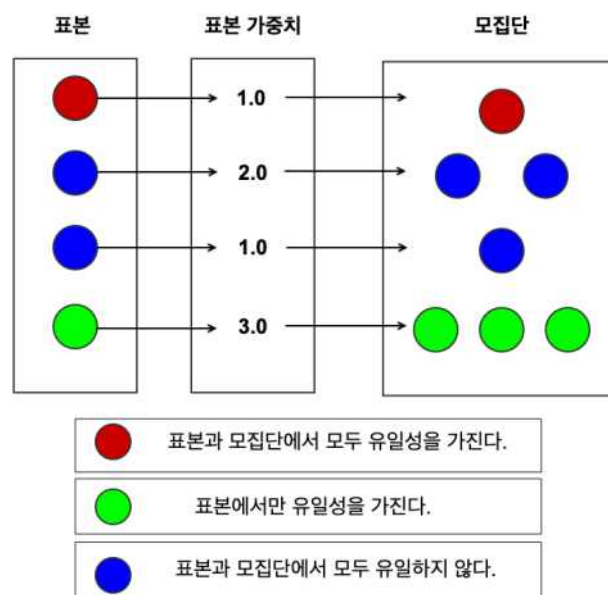
□ 표본과 모집단에서의 유일성

- 공개한 자료가 모집단 전체 자료인 경우 유일한 자료주체는 개인정보 노출의 위험성이 매우 크며 이러한 경우는 특별한 비식별 처리(행제거, 국소삭제, 자료 교환 등)가 필요하다.

- 한편, 공개한 마이크로데이터가 표본 자료일 경우 표본에서 유일한 자료주체가 모집단에서는 유일한 개체가 아닐 수 있다.
- 공개한 자료가 표본 자료일 때 자료주체가 유일한 경우 가중치의 값에 따라서 재식별 위험의 정도가 변한다.
- 표본 자료에서의 유일성을 가진 자료주체의 가중치 값이 작다면 모집단의 자료주체 개수가 적은 것이므로 재식별 위험의 정도가 증가한다.
- 아래 그림은 표본의 자료주체가 가지는 특성별로 각 가중치에 따라서 모집단에서 나타나는 특성의 변화를 설명한 것이다.
- 가중치는 표본에 속한 각 자료주체가 대표하는 모집단의 자료주체의 개수이므로, 가중치가 1보다 크면 표본에서 유일한 개체가 반드시 모집단에서 유일한 개체라고 볼 수 없다.

◇ 예를 들어 2000명으로 구성된 공개된 표본 자료에서 90세 이상의 노인이 한 명이 있고 해당 가중치가 10이라면 모집단 10명 중에 최소한 1명의 90세 이상이 존재한다는 것을 알 수 있다. 따라서 모집단에서 90세 이상인 노인은 유일하게 1명이 아닐 수 있다.

< 표본의 유일성과 가중치의 관계 >



- 아래 예시는 마이크로데이터에서 각 개체의 가중값을 이용하여 각 자료주체의 특성을 지닌 모집단의 개체의 추정값을 표시한 것이다.
- 1~4번째 자료주체는 6개의 범주형 항목조합으로 나타난 유일한 개체이며 가중값이 각각 1, 2, 2, 9 이므로 모집단에서는 같은 조합을 가진 개체가 1, 2, 2, 9개가 있다고 예측할 수 있다.
- 특히 1번째 자료주체는 가중값이 1이므로 표본과 모집단에서 모두 유일한 자료주체이며 재식별의 위험이 가장 높다.
- 5~6번째 행에 나타난 자료주체는 6개의 범주형 조합으로 나타난 같은 특성을 가진 2개의 개체이며 가중값이 각각 9과 5 이므로 모집단에서는 같은 조합을 가진 개체가 최소 14개가 있다고 예측할 수 있다.

< 표본 가중값에 의한 모집단의 개체 추정값에 대한 예시 >

번호	성별	가구주와의 관계	동거여부	혼인상태	종사상 지위	외국인 여부	표본빈도	가중값	모집단빈도 (추정)
1	여	기타	같이살고있음	사별	고용원이 없는 자영업자	아니오	1	1	1
2	남	배우자	같이살고있음	배우자있음	임시. 일용근로자	예	1	2	2
3	남	기혼 자녀 및 그 배우자	따로살고있음(직장)	배우자있음	무급 가족종사자	아니오	1	2	2
4	남	가구주	1인 가구	미혼	상용근로자	예	1	9	9
5	남	배우자	같이살고있음	배우자있음	상용근로자	예	2	9	14
6	남	배우자	같이살고있음	배우자있음	상용근로자	예	2	5	14
7	남	가구주	따로살고있음(직장)	미혼	임시. 일용근로자	아니오	3	13	19
8	남	가구주	따로살고있음(직장)	미혼	임시. 일용근로자	아니오	3	2	19
9	남	가구주	따로살고있음(직장)	미혼	임시. 일용근로자	아니오	3	4	19
10	남	가구주	따로살고있음(직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	7	5	53
11	남	가구주	따로살고있음(직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	7	9	53
12	남	가구주	따로살고있음(직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	7	13	53
13	남	가구주	따로살고있음(직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	7	15	53
14	남	가구주	따로살고있음(직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	7	3	53
15	남	가구주	따로살고있음(직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	7	5	53
16	남	가구주	따로살고있음(직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	7	3	53

2 특이정보

□ 개 요

- 특이정보는 다른 말로 이상점(outlier)이라고 하며 통계기초자료에서 노출 및 재식별 위험을 측정하는 데 중요한 요인이다.
- 일반적으로 연속형 항목에서의 특이점은 다른 개체에 비해 매우 낮은 또는 매우 높은 값을 가진 특이한 값을 의미한다.

◇ 예를 들어 소득이 다른 개체들에 비하여 월등하게 큰 자료주체는 외부정보를 이용하여 재식별할 수 있는 위험성이 높다. 또한 의학에 관련된 검사 자료에서 너무 낮은 값을 가진 개체는 재식별 위험이 높다.

□ 특이정보 검토 방법

- (백분위수) 연속형 항목에서 특이점을 검토하는 가장 기본적인 방법은 자료의 백분위수(최대값과 최소값 포함)를 이용하는 것이다. $p\%$ -백분위수는 자료를 순서대로 늘어놓았을 때 $p\%$ 의 개체들이 $p\%$ -백분위수보다 작거나 같은 값을 가지는 값이다.
- 아래의 예시는 자산금액, 부채금액, 경상소득에 대하여 평균, 중앙값, 최소값, 최대값과 함께 1%, 5%, 10%, 25%, 75%, 90%, 95%, 99% 백분위수를 계산한 것이다. 백분위수를 이용하여 연속형 항목의 퍼진 정도에 대한 기초적인 정보를 얻는 것이 특이점을 검토하는 첫 번째 단계이다.

< 요약 통계량을 이용한 특이점에 대한 검토 예시 >

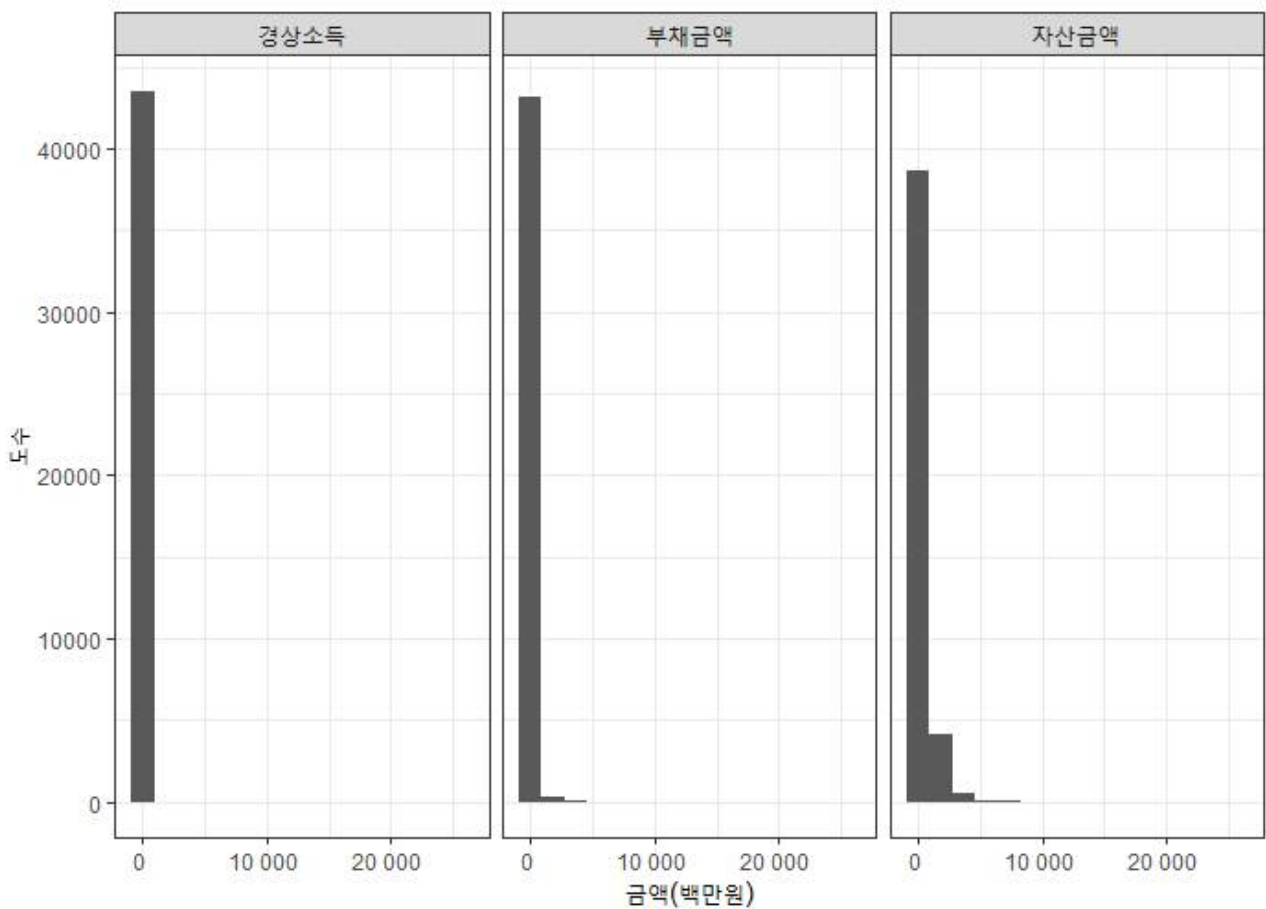
항목(원)	자산금액	부채금액	경상소득
평균	503,332,982.4	91,026,381.59	66,282,380.7
최소값	11,547	0	0
1%	787,313	0	4,151,774
5%	10,076,787	0	7,632,370
10%	31,881,932	0	15,558,944
25%	126,087,991	0	30,297,942
50%	303,795,144	21,020,080	50,609,721
75%	554,234,115	85,193,923	78,703,445
90%	966,059,984	250,397,444	147,763,238
95%	1,924,446,279	366,244,829	196,323,027
99%	3,383,472,402	842,805,049	280,132,514.2
99.5%	4,364,320,374	1,532,874,846	325,951,077
99.9%	8,928,596,067	2,856,836,250	538,998,892
최대값	25,579,485,580	6,280,539,989	1,281,557,375

- 금액을 나타내는 연속형 항목의 값의 분포는 일반적으로 오른쪽으로 치우친 경우가 많다. 즉, 다른 관측치에 비해 매우 높은 값을 가진 몇 가지 특이치가 있음을 의미하며 비식별화를 적용하는 경우 특별한 주의가 필요하다.

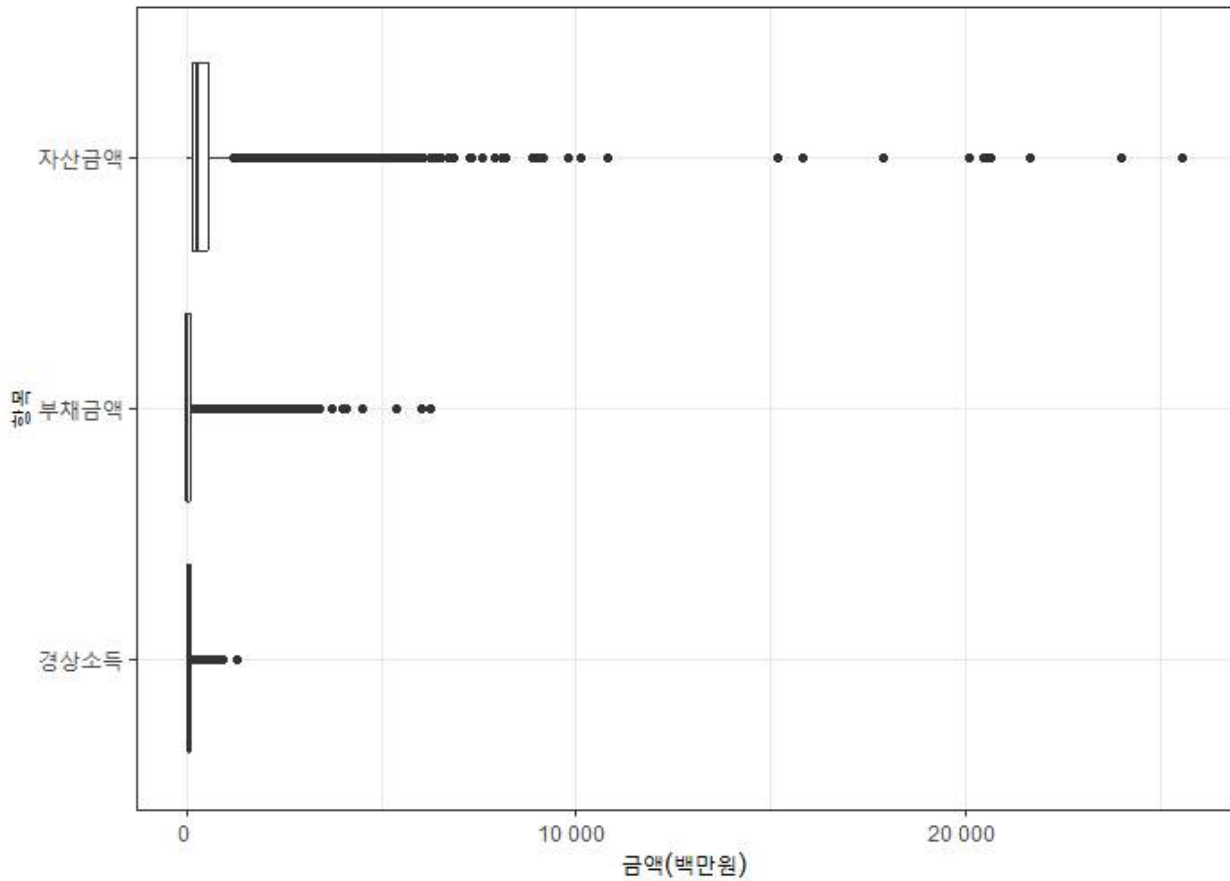
◇ 매우 높은 개인의 소득은 특이정보로 주의가 필요하다. 이와 같은 경우 매우 높은 소득과 같은 특이값이 비식별화 방법에 의하여 변형되더라도 변형된 값이 다시 가장 큰 값으로 유지되거나 순서가 바뀌지 않기 때문에 이러한 특이점에 의하여 개체가 식별될 수 있다.

- (히스토그램·상자그림) 히스토그램과 상자그림을 이용하여 오른쪽으로 치우친 분포를 나타내는 금액들의 특이정보를 검토할 수 있다.
- 일반적으로 경제활동에 관련된 소득, 수익, 부채, 자산과 같은 항목들은 매우 치우친 분포를 가지기 때문에 히스토그램으로는 상위의 개체에 대한 정보를 시각화하는 것이 어렵다. 상자그림은 자료의 퍼진 정도를 파악하는데 히스토그램보다 유용하며 로그 단위로 변환하여 보는 것이 더 유용한 경우가 많다.

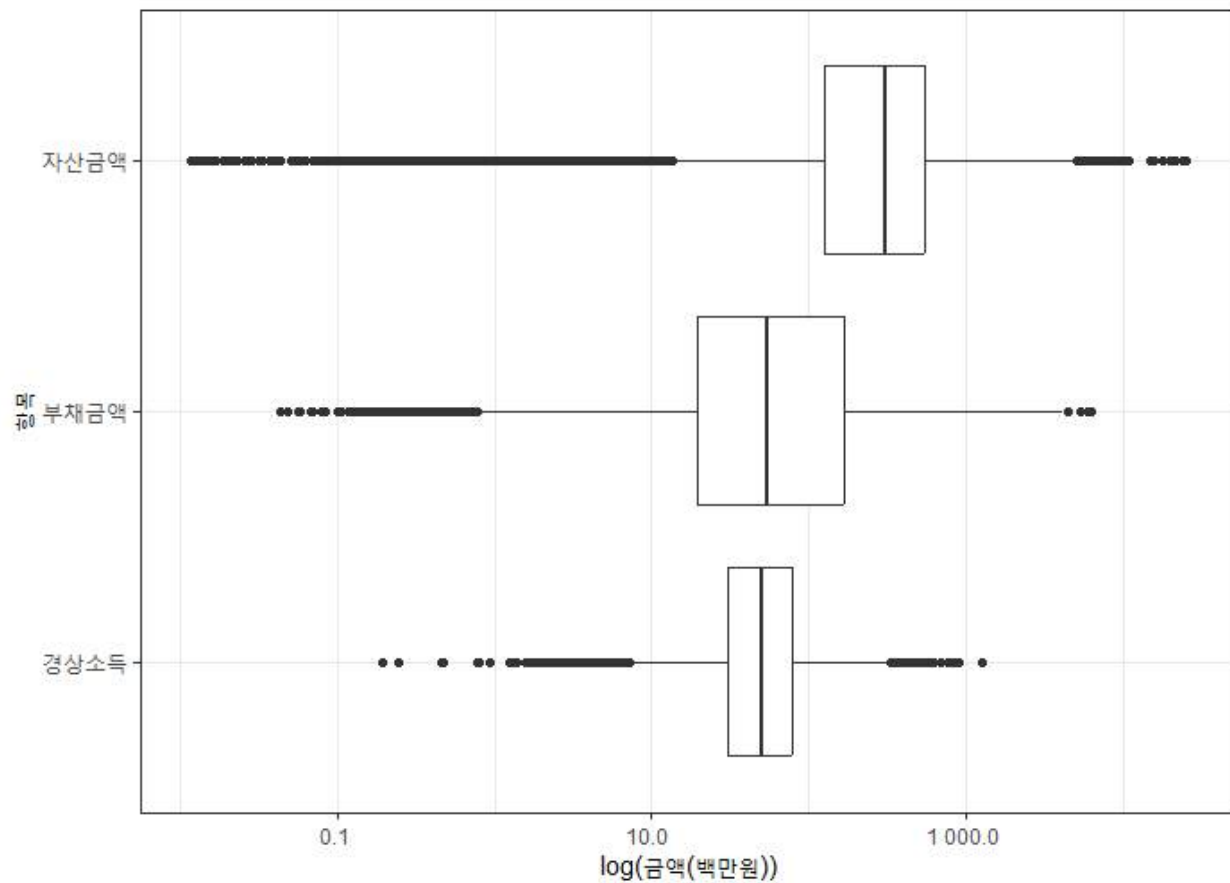
< 히스토그램을 이용한 특이점에 대한 검토 예시 >



< 상자그림을 이용한 특이점에 대한 검토 예시 >



< 상자그림(로그 단위)을 이용한 특이점에 대한 검토 예시 >



□ 특이정보를 통한 연계 가능성 검토

- 다양한 방법을 이용하여 연속형 항목의 분포를 검토한 뒤에 특정한 범위를 벗어난 극단적인 값들을 선정하여 다른 외부 정보와의 연계 가능성 등을 검토한다.
- 아래는 자산금액이 21,000백만원을 초과하는 최상위 자료주체들을 제시한 표이다. 이러한 특이정보를 가진 개체들은 연속형 항목의 특성을 파악하고 연결될 수 있는 가능한 외부 정보가 있는지에 대한 검토를 수행해야 한다. 예를 들어 아래 표의 자산금액, 부채금액은 기업 공시정보 내에 있는 대표자 등의 경제 관련 항목(소득, 세금 등)과 연계될 가능성이 있다.

< 최대값을 포함한 특이점에 대한 검토 예시 >

지역	성별	가구주와의 관계	동거여부	혼인상태	종사상지위	외국인 여부	자산금액	부채금액	경상소득	가중값
서울시 XX동	남	미혼 자녀	같이 살고있음	미혼	가정(무직자) 가사 학생 등	아니오	25,579,485,580	2,073,666,458	516,074,468	2
서울시 XX동	남	가구주	같이 살고있음	배우자 있음	상용근로자	아니오	25,579,485,580	2,073,666,458	516,074,468	2
서울시 XX동	여	배우자	같이 살고있음	배우자 있음	가정(무직자) 가사 학생 등	아니오	25,579,485,580	2,073,666,458	516,074,468	2
서울시 XX동	남	가구주	같이 살고있음	배우자 있음	고용원이 있는 자영업자	아니오	23,980,767,072	6,011,098,768	617,744,828	2
서울시 XX동	여	배우자	같이 살고있음	배우자 있음	가정(무직자) 가사 학생 등	아니오	23,980,767,072	6,011,098,768	617,744,828	2
서울시 XX동	남	가구주	같이 살고있음	배우자 있음	고용원이 있는 자영업자	아니오	21,668,546,780	787,691,456	207,046,180	2
서울시 XX동	남	미혼 자녀	같이 살고있음	미혼	상용근로자	아니오	21,668,546,780	787,691,456	207,046,180	2
서울시 XX동	여	배우자	같이 살고있음	배우자 있음	가정(무직자) 가사 학생 등	아니오	21,668,546,780	787,691,456	207,046,180	2

□ 특이정보의 비식별화 방법

- 특이정보를 포함한 연속형 항목에 대한 비식별 처리는 일반적으로 반올림, 상하단코딩, 잡음첨가, 부분총계, 국소삭제 등이 사용된다.
- 잡음첨가 방법은 자산금액과 같은 매우 치우친 분포를 가지는 항목에 적용하기 어렵다. 대칭이 아닌 분포에 잡음을 첨가하는 방법은 구현하기 어려우며 잡음을 추가하더라도 변형된 값들의 순서가 변하지 않으면 외부 자료와 연계되어 재식별 위험이 감소되지 않을 수 있기 때문이다.
- 개별적으로 한 개의 연속형 항목에 대하여 특이점을 탐색할 수도 있지만 다른 변수의 관계에 따라서 특정 조건을 만족하는 부분 집단에서 특이점인 경우도 있다. 최근 기계학습 등의 분류방법을 이용하면 하나의 항목으로는 특이점이 아니지만 여러 개의 항목을 동시에 고려하면 특정 자료주체가 특이점으로 판별되는 경우를 찾을 수 있다.
- 다음 예시는 자산금액과 부채금액이 매우 큰 특이점은 아니지만 자산금액과 부채금액이 비교적 큰 경우에 경상소득이 0으로 나타나는 특이한 자료주체를 기계학습 등의 분류방법을 이용하여 찾은 결과이다.

< 부분 집단에서의 특이점에 대한 검토 예시 >

성별	가구주와의 관계	동거여부	혼인상태	종사상 지위	외국인 여부	만연령	자산금액	부채금액	경상소득	가중값
남	가구주	같이 살고있음	배우자 있음	기타(무직자, 가사 학생 등	아니오	48	2668,786,921	473,280,991	0	15
남	미혼 자녀	같이 살고있음	미혼	기타(무직자, 가사 학생 등	아니오	16	2668,786,921	473,280,991	0	15
여	배우자	같이 살고있음	배우자 있음	기타(무직자, 가사 학생 등	아니오	44	2668,786,921	473,280,991	0	15
남	가구주	같이 살고있음	배우자 있음	기타(무직자, 가사 학생 등	아니오	59	502,945,520	128,067,413	0	3
여	배우자	같이 살고있음	배우자 있음	기타(무직자, 가사 학생 등	아니오	45	502,945,520	128,067,413	0	3
남	가구주	같이 살고있음	배우자 있음	기타(무직자, 가사 학생 등	아니오	48	380,314,339	13,610,442	0	5
여	배우자	같이 살고있음	배우자 있음	기타(무직자, 가사 학생 등	아니오	45	380,314,339	13,610,442	0	5

3 k-익명성

□ 개요

- (배경) k-익명성은 자료에 나타난 범주형 보조 식별정보들의 모든 가능한 조합을 고려했을 때 동일한 조합에 속하는 자료주체의 수가 매우 작으면 재식별 위험성이 크므로 이를 방지하기 위하여 고안된 노출위험 평가기준이다.
- (정의) 공개하려는 마이크로데이터가 k-익명성을 만족하려면 범주형 보조 식별정보들의 조합에 속한 자료주체의 개수가 최소한 k개가 되어야 한다.
 - k-익명성을 기준으로 사용할 때 일반적으로 자주 사용하는 k 값은 사업체 혹은 기업일 경우 3이며 인구, 가구, 주택은 5이다.
- (사례) 아래는 10개의 보조 식별정보의 조합을 고려할 때 각 자료주체들이 3-익명성 또는 5-익명성을 만족하는지 여부를 나타내는 표이다.

< 통계기초자료에서의 k-익명성 판단 예시 >

번호	성별	가구주와의 관계	동거여부	혼인상태	종사상 지위	외국인 여부	표본빈도	3-익명성	5-익명성
1	여	기타	같이살고있음	사별	고용원이 없는 자영업자	아니오	1	불만족	불만족
2	남	배우자	같이살고있음	배우자있음	임시. 일용근로자	예	1	불만족	불만족
3	남	기혼 자녀 및 그 배우자	따로살고있음(직장)	배우자있음	무급 가족종사자	아니오	1	불만족	불만족
4	남	가구주	1인 가구	미혼	상용근로자	예	1	불만족	불만족
5	남	배우자	같이살고있음	배우자있음	상용근로자	예	2	불만족	불만족
6	남	배우자	같이살고있음	배우자있음	상용근로자	예	2	불만족	불만족
7	남	가구주	따로살고있음(직장)	미혼	임시. 일용근로자	아니오	3	만족	불만족
8	남	가구주	따로살고있음(직장)	미혼	임시. 일용근로자	아니오	3	만족	불만족
9	남	가구주	따로살고있음(직장)	미혼	임시. 일용근로자	아니오	3	만족	불만족
10	남	가구주	따로살고있음(직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	7	만족	만족
11	남	가구주	따로살고있음(직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	7	만족	만족
12	남	가구주	따로살고있음(직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	7	만족	만족
13	남	가구주	따로살고있음(직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	7	만족	만족
14	남	가구주	따로살고있음(직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	7	만족	만족
15	남	가구주	따로살고있음(직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	7	만족	만족
16	남	가구주	따로살고있음(직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	7	만족	만족

□ k-익명성 만족을 위한 마이크로데이터 비식별화 방법

- (1차) 주어진 k-익명성을 만족하기 위해서 주로 사용하는 비식별화 방법은 **항목삭제와 전역코딩**이다. 공개하는 항목의 수를 줄이거나 세밀하게 나타나는 항목의 범주를 묶어서 전역코딩하면 항목들의 가능한 조합의 수가 줄어들기 때문에 k-익명성을 위반하는 자료주체의 수는 줄어들게 된다.
- (2차) 항목삭제와 전역코딩 후에도 k-익명성을 위반하는 자료주체들이 발생하는 경우가 있다. 이러한 경우 **행 삭제 또는 국소삭제** 방법을 이용하여 k-익명성을 위반하는 자료주체를 제거하는 것이 일반적인 방법이다.
- (평가절차) 제공용 마이크로데이터에서 k-익명성을 위반하는 자료주체의 개수를 파악하고 비식별화 처리 이후에 k-익명성을 위반하는 자료주체의 개수를 다시 검토해야 한다.
 - 예를 들어 비식별화 처리 전에 다음 예시와 같이 여러 개의 k값에 대하여 k-익명성을 위반하는 자료주체의 비율을 파악한다. (대표적으로 사업체는 3, 인구·가구·주택은 5를 기준값으로 사용함)

< k-익명성을 위반한 자료주체에 대한 검토 예시: 비식별화 처리 전 >

k 값	위반한 자료주체의 개수	위반한 자료주체의 비율(%)
3	177	0.41
5	327	0.75
7	421	0.97
10	631	1.45

- 만약 3이 기준값이라면 전역코딩, 국소 삭제, 항목 삭제 등과 같은 비식별화 방법을 적용하고 3-익명성을 위반하는 자료주체의 비율을 다시 파악하여 비식별화 기준을 만족하는지 검토한다. 다음 예시는 비식별화 처리 후의 k-익명성을 위반하는 자료주체의 비율을 파악한 결과이다.

< k-익명성을 위반한 자료주체에 대한 검토 예시: 비식별화 처리 후 >

k 값	위반한 자료주체의 개수	위반한 자료주체의 비율(%)
3	0	0
5	15	0.03
7	23	0.05
10	28	0.06

- (유의사항) k-익명성에서 적용하는 k 값의 크기가 커질수록 재식별 위험이 줄어들게 된다. 하지만 적용하는 k 값의 크기가 커지면 공개하는 항목을 줄이거나 범주형 항목들을 전역코딩해야 하므로 공개 자료의 유용성은 줄어들게 되므로 적절한 k값을 선택해야 한다.

4 1-다양성

- (정의) 1-다양성은 보조 식별정보의 조합이 동일한 집단 안에서 민감한 특성을 가지는 항목의 값이 다양하게 나타나지 않아 개인의 정보가 노출되는 것을 방지하기 위한 정보 기준이다.

< 1-다양성에 대한 검토 예시 >

번호	성별	가구주와의 관계	동거여부	혼인상태	종사상지위	외국인여부	최종학력	표본빈도	1-다양성
1	여	부모(배우자의 부모 포함)	같이살고있음	사별	무급 가족종사자	아니오	안 받음 (미취학 포함)	19	2
2	여	부모(배우자의 부모 포함)	같이살고있음	사별	무급 가족종사자	아니오	초등학교	19	2
3	여	부모(배우자의 부모 포함)	같이살고있음	사별	무급 가족종사자	아니오	안 받음 (미취학 포함)	19	2
4	여	부모(배우자의 부모 포함)	같이살고있음	사별	무급 가족종사자	아니오	안 받음 (미취학 포함)	19	2
5	여	부모(배우자의 부모 포함)	같이살고있음	사별	무급 가족종사자	아니오	초등학교	19	2
6	여	부모(배우자의 부모 포함)	같이살고있음	사별	무급 가족종사자	아니오	초등학교	19	2
7	여	부모(배우자의 부모 포함)	같이살고있음	사별	무급 가족종사자	아니오	초등학교	19	2
8	여	부모(배우자의 부모 포함)	같이살고있음	사별	무급 가족종사자	아니오	초등학교	19	2
9	여	부모(배우자의 부모 포함)	같이살고있음	사별	무급 가족종사자	아니오	초등학교	19	2
10	여	부모(배우자의 부모 포함)	같이살고있음	사별	무급 가족종사자	아니오	안 받음 (미취학 포함)	19	2
11	여	배우자	따로살고있음 (직장)	배우자 있음	무급 가족종사자	아니오	고등학교	5	4
12	여	배우자	따로살고있음 (직장)	배우자 있음	무급 가족종사자	아니오	대학교 (4년제 이상)	5	4
13	여	배우자	따로살고있음 (직장)	배우자 있음	무급 가족종사자	아니오	초등학교	5	4
14	여	배우자	따로살고있음 (직장)	배우자 있음	무급 가족종사자	아니오	중학교	5	4
15	여	배우자	따로살고있음 (직장)	배우자 있음	무급 가족종사자	아니오	고등학교	5	4

- 예시로 제시된 위의 표에서 처음 10명의 개체들은 최종학력을 제외한 다른 항목들이 동일하다. 이렇게 동일한 특성을 지닌 10명의 사람들에게 대한 최종학력을 고려할 때 2개의 항목(초등학교 또는 안 받음)밖에 없으므로 최종학력이 초등학교 이하라는 민감한 정보가 노출될 수 있다. 반대로 마지막 5명의 사람들은 최종학력의 범주가 다양하게 나타나므로 민감한 정보가 노출되지 않는다.

◇ 일반적으로 1-다양성에서 고려되는 민감한 정보는 특이정보와 같은 성격을 가지는 정보이며 사회적으로 민감한 학력, 질병, 파산 등에 관련된 것이다.

5 재식별위험 평가측도

□ 연계 확률

- (정의) 자료 연계에 의한 재식별 위험은 외부의 이용자가 가진 모집단의 정보와 공개 자료가 가진 정보를 결합하여 유일하게 분리된 개체를 모집단에 속한 개체와 정확하게 연결하는 가능성, 즉 연계 확률로 측정할 수 있다.
- (사례) 외부 공격자가 표본에서 유일하게 분리되는 자료주체를 자신이 가지고 있는 모집단 정보와 연계하여 재식별이 일어나는 가상의 시나리오를 생각해 보자.

- ◇ 노인연금에 관한 표본 자료에서 만약 연령이 90세 이상인 사람이 단 한 명만 존재하고 해당 가중값이 5라고 하자.
- ◇ 외부 이용자가 연령 정보가 포함된 병원기록에 대한 모집단 자료를 가지고 있으며, 해당 모집단 자료에는 90세 이상인 사람이 5명이고 그 중에 1명이 특별한 암환자인 것이 나타나 있다고 하자.
- ◇ 외부 이용자가 모집단의 5명 중에 1명을 임의로 선택했을 때, 선택된 개체가 표본에서 나타난 유일한 자료주체와 동일한 사람일 확률은 $1/5 = 0.2$ 즉 20% 이다.
- ◇ 20%의 가능성이 이용자의 외부 자료와 공개된 표본 자료를 연계하여 재식별이 일어날 최대 확률이다.

- 위에서 제시한 가상의 시나리오에서 고려한 재식별의 가능성은 최댓값이며 실제 재식별 가능성은 임의로 연계하는 이상적인 시나리오 가정하에서 추정한 최대 확률보다는 낮으며 정확하게 추정하기 어렵다.
- 일반적으로 외부의 공격자가 가진 정보의 양이 제한적이고 정확성이 낮을 수 있다. 또한 공개된 표본 자료에 나타난 일부 항목들은 측정오차에 의해 실제 모집단 자료와 차이가 날 수 있기 때문에 재식별 확률을 정확하게 추정하기 어렵다.

□ 초모집단 모형

- 재식별의 확률을 추정하기 위해서 사용되는 확률적 모형은 초모집단 모형(super population model)을 사용하며 포아송 분포와 이항 분포를 결합한 계층모형으로 추정할 수 있다.
- 여기서 초모집단 모형은 모집단의 개체를 생성하는 확률적 모형이 있다고 가정하는 모형이다. 이러한 모형은 조사통계 이론에서 모집단의 개체를 고정 모수(parameter)로 인식하는 전통적인 이론과는 성격이 다른 베이지안 추론(Bayesian inference)을 이용한 것이며, 모집단을 생성하는 확률적 모형이 존재한다는 가정 때문에 초모집단 모형이라고 한다.
- 아래 예시는 초모집단 모형을 이용하여 자료주체가 보조 식별 정보에 의하여 재식별될 위험을 추정한 확률을 나타난 예이다.

< 자료에서 재식별될 위험을 추정한 확률에 대한 예시 >

번호	성별	가구주와의 관계	동거여부	혼인상태	종사상지위	외국인여부	표본빈도	가중값	모집단빈도 (추정)	추정 재식별 확률
1	여	기타	같이살고있음	사별	고용원이 없는 자영업자	아니오	1	1	1	1.000000
2	남	배우자	같이살고있음	배우자있음	임시. 일용근로자	예	1	2	2	0.693147
3	남	기혼 자녀 및 그 배우자	따로살고있음 (직장)	배우자있음	무급 가족종사자	아니오	1	2	2	0.693147
4	남	가구주	1인 가구	미혼	상용근로자	예	1	9	9	0.274653
5	남	배우자	같이살고있음	배우자있음	상용근로자	예	2	9	14	0.112614
6	남	배우자	같이살고있음	배우자있음	상용근로자	예	2	5	14	0.112614
7	남	가구주	따로살고있음 (직장)	미혼	임시. 일용근로자	아니오	3	13	19	0.073171
8	남	가구주	따로살고있음 (직장)	미혼	임시. 일용근로자	아니오	3	2	19	0.073171
9	남	가구주	따로살고있음 (직장)	미혼	임시. 일용근로자	아니오	3	4	19	0.073171
10	남	가구주	따로살고있음 (직장)	배우자있음	기타(무직자 가사 학생 등)	아니오	7	5	53	0.021538
11	남	가구주	따로살고있음 (직장)	배우자있음	기타(무직자 가사 학생 등)	아니오	7	9	53	0.021538
12	남	가구주	따로살고있음 (직장)	배우자있음	기타(무직자 가사 학생 등)	아니오	7	13	53	0.021538
13	남	가구주	따로살고있음 (직장)	배우자있음	기타(무직자 가사 학생 등)	아니오	7	15	53	0.021538
14	남	가구주	따로살고있음 (직장)	배우자있음	기타(무직자 가사 학생 등)	아니오	7	3	53	0.021538
15	남	가구주	따로살고있음 (직장)	배우자있음	기타(무직자 가사 학생 등)	아니오	7	5	53	0.021538
16	남	가구주	따로살고있음 (직장)	배우자있음	기타(무직자 가사 학생 등)	아니오	7	3	53	0.021538

- 재식별 위험 확률의 추정값은 외부 이용자가 보유한 모집단에 대한 외부정보가 정확하고, 모든 개체를 연계하려고 의도하며 동시에 공개하는 자료에 측정오차가 존재하지 않는 경우를 가정한다.

◇ 재식별 위험 확률의 추정값은 “최악의 상태를 가정하고” 노출확률의 최댓값을 추정한 통계량이다. 따라서 실제 재식별이 일어나서 개인정보가 누출되는 가능성은 추정된 확률값보다 작다.

II. 정보손실 평가 측도

1 범주형 항목

- 범주형 항목을 고려할 때 비식별화 처리를 수행한 후 정보 손실의 측도로 사용될 수 있는 대표적인 방법은 다음과 같다.
- 중요한 항목에 대한 비율: 여러 개의 중요한 항목에 대하여 원자료에 의한 집계표와 비식별 처리 후 집계표에 나타난 비율의 차이 등을 비교할 수 있다.

< (사례) 비식별화 후 제공된 빈도값과 실제 빈도값의 차이 >

정보손실	0	1	2
%	93.57	4.09	2.34

- * 총 2500개의 빈도표를 제공했으며, 참값과 차이가 1인 셀이 4.09%, 2인 셀이 2.34%
- 삭제, 상하단 코딩 등으로 공개를 하지 못하는 자료의 비율도 정보 손실의 측도로 사용될 수 있다.

2 연속형 항목

- 연속형 항목을 고려할 때 비식별화 처리를 수행한 후 정보 손실의 측도로 사용될 수 있는 대표적인 방법은 다음과 같다.
- 주요 집계 통계량, 예를 들어 주요 항목에 대한 평균, 총합, 중앙값, 상관계수에 대하여 비식별 처리 전과 후의 차이를 사용할 수 있다.
- 모든 측정값의 평균 변화량을 나타내는 측도를 사용할 수 있다. 즉, 원래 자료에 나타난 항목들의 값과 비식별 처리가 된 변형된 값들의 차이를 모두 고려한 측도를 사용할 수 있다.

◇ 예를 들어 다음과 같은 IL1s가 사용될 수 있다. 아래 정의된 측도 IL1s의 값이 작을수록 정보 손실이 작다.

$$IL1s = \frac{1}{KN} \sum_{i=1}^K \sum_{j=1}^N \frac{|x_{ij} - z_{ij}|}{\sqrt{2} S_i}$$

- ◇ 위의 식에서 x_{ij} 는 i 번째 항목의 j 번째 원래 측정값, z_{ij} 는 비식별 처리를 적용한 값, S_i 는 i 번째 항목의 원자료에 대한 표준 편차이다.
- ◇ 위에서 제시된 IL1s는 연속형 항목에 대한 측도로, 범주형 항목에는 적용하기 어렵다.

- 중요한 항목들의 관계에 대한 정보 손실의 측도로서 상관계수의 변화도 고려할 수 있지만, 항목의 수가 많아지면 그만큼 고려해야 하는 상관계수의 수도 증가하므로 단일한 측도로 평가하기 힘들다. 이러한 경우 중요한 2-3개의 항목을 선정하여 상관계수의 수를 줄여서 사용할 수 있다.
- 상관계수의 변화를 정보 손실의 측도로 사용하는 경우 여러 개의 상관계수의 특성을 나타내는 하나의 측도를 선정할 수 있으며 대표적인 예가 상관계수 행렬의 고유값(eigen value)을 이용한 측도이다.

Ⅲ. 비식별화 세부기법 해설

1 구조적 방법

1. 표본추출

- 표본추출은 자료주체 전체를 포함하는 등록자료 또는 행정통계 자료인 경우 자료주체의 일부분을 추출하는 통계적 방법이다.
- 표본추출은 조사통계를 작성하는 일반적인 방법이지만 자료의 일부를 공개하는 비식별화 방법으로 사용할 수 있으며 자료 전체를 공개하지 않기 때문에 매우 효과적인 비식별화 방법이다.
- 전체 모집단 자료의 일부만 공개하면 이용자가 외부자료와 연계를 시도하는 경우에도 특정한 자료주체의 재식별 여부를 판단하기 어렵기 때문에 개인정보 유출을 효과적으로 방지할 수 있다.
- 또한 표본으로 추출된 자료에서는 자료주체가 유일하게 구별되어도 전체 등록자료에서 유일한 것이 아닐 수 있으므로 등록부 자료 전체를 공개하는 것보다 재식별 위험은 감소한다.
- 표본을 추출하는 경우 조사통계 자료를 수집할 때와 동일하게 확률적 표본추출 계획에 따라서 자료를 생성해야 하며 표본추출 계획과 관련된 정보(예를 들어 층화 방법, 가중치 등)를 공개해야 이용자가 공개 자료를 사용할 수 있다.
- 표본의 비율은 공개하는 자료의 구성과 특성 그리고 이용자의 요구 정도에 따라서 달라지지만 보통 전체 자료의 1%-10% 정도를 공개한다.

2. 총계처리

- 총계처리는 여러 개의 연관된 연속형 항목들 또는 자료주체들을 모두 공개하지 않고 총계만을 공개하는 비식별화 방법이다.

◇ 예를 들어서 사업체 자료에서 임시 및 일용근로자의 수와 상용근로자의 수를 따로 공개하지 않고 두 항목의 총계인 종사자수만 공개할 수 있다.

- 또한, 여러 자료 주체들의 특정 항목의 값을 평균값, 최댓값, 최솟값, 최빈값, 중간값 등으로 처리하여 제공할 수 있다.

2 삭제 방법

1. 항목 삭제

- 항목 삭제란 개인정보 노출의 위험성이 큰 항목의 전체를 삭제하고 자료를 제공하는 방법이다. 개체식별정보가 있는 항목은 제공 자료에서 항목삭제를 하는 것이 통상적인 절차이며, 보조식별정보 중에서 적합한 비식별화 방법을 찾아 적용하기 어렵거나 비식별화 방법을 적용해도 위험성이 큰 항목들도 항목 삭제를 적용한다.

2. 행 삭제

- 행 삭제란 개인정보 노출의 위험성이 큰 자료주체가 있는 행의 모든 항목 값을 삭제하고 자료를 제공하는 방법이다.
- 만약 자료주체의 일부 항목의 특이값에 의하여 유일하게 구별될 가능성이 크면 자료주체의 모든 항목을 삭제하는 행 삭제를 적용한다.

◇ 특정한 자료주체가 적절한 비식별화 방법을 적용한 후에도 유일하게 구별될 가능성이 크거나 일부 항목에 특이정보를 가지고 있는 경우 행 삭제를 적용한다.

- 만약 표본 추출을 비식별화 방법으로 이미 사용한 경우 특이한 자료주체를 행삭제 할 때는 같은 표본 추출층에 속하는 자료주체를 임의로 추출하여 대체한다. 이렇게 자료주체를 대체해야만 별도로 가중치를 조정할 필요가 없다.

3. 국소삭제

- 통계기초자료에서 보조식별정보의 조합을 고려할 때 보조식별정보가 많아질수록 특정한 조합에 해당하는 개체의 빈도수가 매우 작게 나타날 수 있다.
- 이러한 경우 전역코딩하여 식별 위험을 줄일 수 있으나, 비식별화 방법이 적용된 후에도 특정 자료주체의 노출 위험성이 줄어들지 않을 경우 개체의 보조 식별항목의 값을 결측값으로 만드는 방법을 국소삭제라고 부른다.

- 또한 국소삭제는 자료주체의 일부 항목의 특이값에 의하여 유일하게 구별될 가능성이 큰 경우에도 사용될 수 있다.
- 자료의 k-익명성이 위반되는 경우 이를 극복하기 위하여 개체의 보조식별정보를 국소삭제하는 경우 특수한 알고리즘을 이용할 수 있다. 알고리즘이 적용되는 경우 k-익명성을 유지하면서 국소삭제의 개수를 최소화하는 방법을 고려하는 것이 일반적이다.
- 또한 국소 삭제는 항목의 중요도를 판단하여 덜 중요한 항목부터 적용하는 것이 좋다. 그러나 중요한 항목에 대한 국소 삭제의 수가 감소하면 다른 항목의 값이 너무 많이 국소 삭제가 되는 부작용이 발생할 수 있으므로 주의해야 한다.
- 항목의 개수와 자료주체의 수가 매우 많은 큰 규모의 통계기초자료를 비식별화 하는 경우, 국소 삭제를 적용해야 하는 경우가 너무 많아질 수 있다. 이러한 경우는 다시 공개하는 보조 식별정보의 개수를 줄이거나 다시 전역코딩을 적용하는 것이 바람직하다.

◇ 국소삭제는 비식별화 조치 이후에 계획된 k-익명성이 지켜지지 않는 경우, 설정된 정보 기준을 충족하기 위하여 주로 사용되는 방법이다.

- 다음은 16개의 자료주체들로 구성된 자료에서 6개의 보조 식별항목을 조합하여 나타나는 표본빈도를 나타낸 예시 자료이다.
- 5를 k값으로 선택한 경우 아래 자료에서는 처음 9개의 개체가 k-익명성 기준을 만족하지 못한다.

< 유일성과 k-익명성에 대한 검토 예시 >

번호	성별	가구주와의 관계	동거여부	혼인상태	종사상지위	외국인여부	가중값	표본빈도
1	여	기타	같이 살고있음	사별	고용원이 없는 자영업자	아니오	1	1
2	남	배우자	같이 살고있음	배우자있음	임시. 일용근로자	예	2	1
3	남	기혼 자녀 및 그 배우자	따로 살고있음 (직장)	배우자있음	무급 가족종사자	아니오	2	1
4	남	가구주	1인 가구	미혼	상용근로자	예	9	1
5	남	배우자	같이 살고있음	배우자있음	상용근로자	예	9	2
6	남	배우자	같이 살고있음	배우자있음	상용근로자	예	5	2
7	남	가구주	따로 살고있음 (직장)	미혼	임시. 일용근로자	아니오	13	3
8	남	가구주	따로 살고있음 (직장)	미혼	임시. 일용근로자	아니오	2	3
9	남	가구주	따로 살고있음 (직장)	미혼	임시. 일용근로자	아니오	4	3
10	남	가구주	따로 살고있음 (직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	5	7
11	남	가구주	따로 살고있음 (직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	9	7
12	남	가구주	따로 살고있음 (직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	13	7
13	남	가구주	따로 살고있음 (직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	15	7
14	남	가구주	따로 살고있음 (직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	3	7
15	남	가구주	따로 살고있음 (직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	5	7
16	남	가구주	따로 살고있음 (직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	3	7

- 이제 위의 자료에서 k-익명성을 충족시키기 위하여 처음 9개의 자료주체의 항목들을 국소 삭제하는 예시는 다음과 같다.
- 국소 삭제를 적용하는 항목은 일반적으로 유용성이 작은 항목을 선택하는 것이 바람직하다.

- 하지만 아래 예시와 같이 여러 개의 자료 주체들에 대해 동시에 국소 삭제를 적용하는 경우 항목의 선택은 쉬운 작업이 아니기 때문에 일반적으로 비식별화를 위한 컴퓨터 패키지에서 제공하는 알고리즘을 사용한다.
- 패키지에서 제공하는 알고리즘을 사용하는 경우 여러 개의 보조 식별항목에 중요도를 지정하여 국소 삭제의 횟수를 항목마다 다르게 지정할 수 있다.

< 유일성과 k-익명성을 위한 국소 삭제 처리에 대한 예시 >

번호	성별	가구주와의 관계	동거여부	혼인상태	종사상지위	외국인여부	가중값	표본빈도
1				배우자있음	고용원이 없는 자영업자		1	1
2	남						2	1
3	남		따로 살고있음 (직장)	배우자있음		아니오	2	1
4	남	가구주			상용근로자		9	1
5	남			배우자있음	상용근로자		9	2
6	남			배우자있음	상용근로자		5	2
7	남	가구주	따로 살고있음 (직장)			아니오	13	3
8	남	가구주	따로 살고있음 (직장)			아니오	2	3
9	남	가구주	따로 살고있음 (직장)			아니오	4	3
10	남	가구주	따로 살고있음 (직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	5	7
11	남	가구주	따로 살고있음 (직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	9	7
12	남	가구주	따로 살고있음 (직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	13	7
13	남	가구주	따로 살고있음 (직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	15	7
14	남	가구주	따로 살고있음 (직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	3	7
15	남	가구주	따로 살고있음 (직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	5	7
16	남	가구주	따로 살고있음 (직장)	배우자있음	기타(무직자, 가사, 학생 등)	아니오	3	7

- 국소삭제를 적용한 후에 중요한 보조 식별항목의 삭제가 지나치게 많은지 검토할 수 있다. 만약 중요한 항목의 삭제가 너무 많다고 판단되면 알고리즘을 변경하여 중요한 보조 식별항목의 삭제 정도를 비교하여야 한다.

- 다음 표는 위의 예시에서 적용한 국소 삭제의 결과를 나타낸 표이다.

< 유일성과 k-익명성을 위한 국소 삭제 처리에 대한 요약 예시 >

보조식별정보	국소삭제 횟수	국소삭제 비율(%)
성별	1	6.250
가구주와의관계	5	31.250
동거여부	5	31.250
혼인상태	5	31.250
종사상지위	5	31.250
외국인여부	5	31.250

3 일반화 방법

1. 반올림

- 반올림은 연속형 항목의 값을 특정한 자리에서 반올림하는 방법으로 외부 자료와의 결합을 어렵게 하는 경우에 주로 사용된다.
- 반올림을 고려하는 경우 적용하는 자릿수를 값의 크기에 따라서 적절하게 적용해야 한다. 반올림을 적용하는 자릿수가 너무 작으면 자료의 변화가 너무 작아서 외부 자료 연계가 가능할 수도 있고 너무 크면 자료의 유용성이 심하게 감소할 수 있다.
- 금액(예를 들어 소득액, 매출액 등)으로 표시되는 항목에 대해서는 외부의 자료와 연계 가능성이 크므로 다른 비식별화 방법을 적용한 후에도 모든 값에 반올림을 적용하는 것이 좋다.
- 반올림 처리를 하는 경우 모든 자료에 대해 동일한 반올림 자릿수를 적용하면 앞에서 언급한 것과 같이 노출의 위험성이 감소하지 않거나 자료의 유용성이 심하게 감소할 수 있다.

- ◇ 변동이 큰 금액으로 나타난 자료는 반올림하는 자리를 적용하는 수의 크기에 따라서 변화시키는 것이 적절하다.
- ◇ 예를 들어 만원 미만은 그대로, 만원 이상이고 백만원 미만은 백원 자리에서, 백만원 이상은 만원 자리에서 반올림을 적용할 수 있다.

2. 전역코딩

- 전역코딩은 범주형 항목의 수를 줄이거나 연속형 항목의 값 전체를 구간화 하여 범주형 항목으로 만들어 주는 방법이다.
- 전역코딩은 항목이 가진 정보의 미세한 정도를 줄여서 정보의 정확성을 감소시키는 대신 개인정보 노출의 위험을 줄이는 방법이다.
- 전역코딩은 비식별화 작업을 적용할 때 가장 먼저 고려해야 하는 기본적인 방법이며 다른 방법에 비해 쉽게 구현할 수 있고 또한 매우 효과적인 방법이다.
- 전역코딩을 적용한 후에 보조 식별항목들의 가능한 조합의 수에 대한 변화는 유일성, k-익명성, 1-다양성 등 재식별 위험을 평가하는 전역 기준을 만족하는 수준에 영향을 준다.

◇ 전역코딩을 보조 식별항목들에 적용하는 경우 다음과 같은 두 가지 요소를 고려해야 한다.

- 1) 범주의 수가 줄어들 때 유용성이 얼마나 감소하는가?
- 2) 범주의 수가 줄어들면서 보조 식별항목들의 가능한 조합의 수가 어떻게 변하는가?

- 다음은 4개의 보조 식별항목을 고려한 원자료의 일부분에 대한 예시이다.

< 전역 코딩에 대한 예시 - 전역 코딩 적용 전 자료 >

기준년도	만 연령	가구원수	교육정도	혼인상태
2016	56	4	고등학교 졸업	배우자 있음
2016	62	2	중학교 졸업	배우자 있음
2016	33	3	고등학교 졸업	미혼
2016	56	3	고등학교 졸업	배우자 있음
2016	46	5	고등학교 졸업	배우자 있음
2016	65	2	고등학교 졸업	배우자 있음
2016	81	3	초등학교 졸업	사별
2016	35	1	고등학교 졸업	이혼
2016	35	2	대학졸업 이상	배우자 있음
2016	27	4	대학졸업 이상	미혼
2016	87	1	초등학교 졸업	사별
2016	54	2	초등학교 졸업	미혼
2016	36	4	대학졸업 이상	배우자 있음
2016	62	1	고등학교 졸업	이혼
2016	35	3	대학졸업 이상	배우자 있음

- 아래 그림은 전역 코딩을 적용한 후의 자료 일부분이다.
 - 연령을 '30세 미만', '30~40세 미만', '40~50세 미만', '50~60세 미만', '60세 이상'의 5개 범주로 전역 코딩 적용
 - 가구원의 수를 '0명', '1명', '2-3명', '4명 이상'의 4개 범주로 전역코딩
 - 교육정도를 '고등학교 졸업 이하', '대학졸업 이상' 2개 범주로 전역코딩 적용하고, 혼인상태를 '미혼', '그외의 경우' 2개 범주로 전역코딩 적용

< 전역 코딩에 대한 예시 - 전역 코딩 적용 후 자료 >

기준년도	연령(코드)	가구원수(코드)	교육정도	혼인상태
2016	50~60세 미만	4명 이상	고등학교 졸업 이하	그 외의 경우
2016	60세 이상	2-3명	고등학교 졸업 이하	그 외의 경우
2016	30~40세 미만	2-3명	고등학교 졸업 이하	미혼
2016	50~60세 미만	2-3명	고등학교 졸업 이하	그 외의 경우
2016	40~50세 미만	4명 이상	고등학교 졸업 이하	그 외의 경우
2016	60세 이상	2-3명	고등학교 졸업 이하	그 외의 경우
2016	60세 이상	2-3명	고등학교 졸업 이하	그 외의 경우
2016	30~40세 미만	1명	고등학교 졸업 이하	그 외의 경우
2016	30~40세 미만	2-3명	대학졸업 이상	그 외의 경우
2016	30세 미만	4명 이상	대학졸업 이상	미혼
2016	60세 이상	1명	고등학교 졸업 이하	그 외의 경우
2016	50~60세 미만	2-3명	고등학교 졸업 이하	미혼
2016	30~40세 미만	4명 이상	대학졸업 이상	그 외의 경우
2016	60세 이상	1명	고등학교 졸업 이하	그 외의 경우
2016	30~40세 미만	2-3명	대학졸업 이상	그 외의 경우

- 아래는 전역 코딩을 적용하기 전과 후에 대하여 4개의 항목으로 구성할 수 있는 가능한 모든 조합을 비교한 그림이다. 전역 코딩 전보다 적용 후의 조합의 수가 크게 감소함을 알 수 있다.

< 전역 코딩에 대한 예시 - 전역 코딩 적용 전과 후 가능한 조합의 변화 >

전역코딩	연령	가구원수	교육정도	혼인상태	총 조합의수
적용 전	78	10	4	4	12,480
적용 후	5	4	2	2	80

- 전역 코딩을 적용하기 전과 후 조합 수의 변화가 재식별 위험의 평가 기준인 k-익명성에 큰 영향을 미친다.
- 위의 전역코딩에 대한 예시를 고려하면 아래와 같이 전역 코딩 전과 후에 k-익명성을 위배하는 자료주체의 수가 줄어드는 것을 볼 수 있다.
- 이렇게 보조 식별항목들에 대하여 전역 코딩을 적용하는 경우는 정해진 노출 위험의 평가기준을 같이 검토해야 한다.

< 전역 코딩에 대한 예시 - 전역 코딩 적용 전 k-익명성 위반에 대한 정보 >

k 값	위반한 자료주체의 개수	위반한 자료주체의 비율(%)
3	1,230	2.83
5	2,178	5
7	2,897	6.66
10	3,882	8.92

< 전역 코딩에 대한 예시 - 전역 코딩 적용 후 k-익명성 위반에 대한 정보 >

k 값	위반한 자료주체의 개수	위반한 자료주체의 비율(%)
3	7	0.02
5	14	0.03
7	40	0.09
10	71	0.16

3. 상하단 코딩

- 상하단 코딩은 항목이 민감한 정보를 포함하고 있고 매우 작은 값과 매우 큰 값을 가지고 있는 개체들의 수가 적어서 그대로 공개 되면 개인 정보의 노출뿐 아니라 민감한 정보의 유출의 가능성이 높은 경우 사용되는 방법이다.
- 일반적으로 연속형 항목의 값이 정해진 기준 값을 초과하거나 또는 기준 값에 미달하면 값 자체를 공개하지 않고 하나의 대표값 또는 범주로 바꾸어 공개하는 방법이다.

- ◇ 예를 들어 소득에 관련된 항목을 고려하면 소득이 매우 높은 사람들이나 매우 낮은 사람들의 소득 값을 대표값이나 범주로 코딩하여 공개한다.
- ◇ 월 소득이 매우 큰 사람들의 정보를 숨기기 위하여 월 소득이 1억원 이상 인 사람들의 월 소득을 기준값(1억원)으로 모두 대체하거나 하나의 범주로 (1억원 이상) 코딩할 수 있다.

- 상하단 코딩을 적용할 때 극단값을 기준값이 아닌 범주로 대체하는 경우 산술 계산이 불가능하므로 유용성이 떨어질 수 있다.
- 상하단 코딩은 통계기초자료(마이크로데이터)와 집계표와 같은 매크로데이터에 모두 적용할 수 있다.

1. 잡음 첨가

- 잡음첨가는 주로 연속형 항목에 적용하는 비식별화 방법이며 항목의 값에 임의의 값, 즉 잡음을 더하거나 곱하는 방법을 말한다.
- 잡음첨가의 목적은 공개된 자료와 외부 자료와의 연계를 방지하는 것이다.
- 일반적으로 잡음첨가는 연속형 항목에 적용되는 방법이지만 특별하게 잡음의 형태를 조종하면 범주형 항목에도 적용이 가능하다.
- 잡음을 더하거나 곱하는 경우 편의(bias)를 피하기 위하여 자료의 값에 더해주는 잡음의 평균을 0으로 하고 곱해주는 잡음의 평균은 1이 되도록 한다. 더해주거나 곱해주는 잡음 분포의 분산이 커지면 비식별화 효과가 증가한다.
- 일반적으로 잡음을 첨가하는 방법은 적용 후에도 큰 값을 가지는 특이점의 변화가 상대적으로 크게 나타나지 않는다. 따라서 매우 큰 값이 특이점으로 나타나는 자료는 잡음첨가 후에 별도의 비식별화 방법(예를 들어 국소 삭제나 부분 총계)을 적용해야 한다.

- ◇ 잡음을 첨가하는 경우 자료의 구조적 특성에 심각한 영향을 미쳐서는 안된다.
- ◇ 예를 들어 날짜나 시간에 잡음을 첨가하게 되면 시간대별, 요일별, 월별로 나타나는 시계열적 특성이 심하게 왜곡될 수 있다.
- ◇ 잡음을 첨가한 자료는 원자료가 가지고 있는 범위를 벗어날 수 있으므로 이를 유의해야 한다. 예를 들어 잡음이 첨가된 소득이 음수가 나올 수 있다.

- 일반적으로 잡음을 첨가한 비식별화된 자료의 평균이나 중앙값은 원자료와 매우 유사하다. 하지만 잡음을 첨가하면 자료의 분산은 커지고 상관관계가 왜곡될 수 있다.
- 또한 잡음첨가는 원자료에 잡음을 더해주거나 곱해주는 방법이므로 구현하기 쉽지만 잡음을 생성할 때 대상이 되는 항목의 분포에 따라서 적절한 잡음의 분포를 결정해야 한다. 이렇게 항목의 분포와 같은 통계적인 특성을 이해해야 하는 어려움이 있다.
- 동시에 여러 개의 항목에 잡음을 첨가하는 경우 항목들의 상관관계가 크게 왜곡되지 않도록 자료에서 추정된 상관관계를 가지는 다변량 분포를 사용할 수 있다.
- 다음 두 표는 3개의 연속형 변수(월소득액, 월저축금액, 월교통비용)가 포함된 원자료와 상관계수이다.

<잡음첨가에 대한 예시 - 잡음첨가 전 자료>

번호	거주시도	연령코드	월소득액	월저축금액	월교통비용
1	충청남도	60세이상	14,975,000	132,000	415,000
2	경상북도	50~60세 미만	2,711,000	371,000	217,000
3	전라남도	30~40세 미만	2,390,000	103,060	104,460
4	경기도	50~60세 미만	2,154,000	301,000	105,310
5	경기도	50~60세 미만	1,857,000	269,000	228,000
6	충청남도	50~60세 미만	1,716,000	159,000	143,000
7	경상북도	50~60세 미만	1,541,000	145,000	119,000
8	충청북도	50~60세 미만	1,257,000	282,000	169,000
9	경기도	50~60세 미만	1,061,000	898,000	101,800
10	경기도	40~50세 미만	909,000	343,000	124,000

<잡음 첨가에 대한 예시 - 잡음첨가 전 자료의 상관계수>

상관계수	월소득액	월저축금액	월교통비용
월소득액	1	0.069	0.466
월저축금액	0.069	1	0.164
월교통비용	0.466	0.164	1

- 다음 두 표는 3개의 연속형 변수에 대하여 독립적으로 잡음을 더해주는 방법(무상관 잡음첨가)을 적용한 후 자료와 3개의 연속형 변수에 대한 상관계수이다.
- 아래 그림의 마지막 자료주체의 월소득액이 잡음첨가 후 음수로 나타난다. 이렇게 잡음첨가는 원래 항목의 값이 가질 수 있는 범위를 벗어날 수 있으므로 이에 대한 검토와 별도의 규칙(예를 들어 소득금액이 0보다 작으면 0으로 변환)이 필요하다.

< 잡음첨가에 대한 예시 - 무상관 잡음첨가 후 자료 >

번호	거주시도	연령코드	월소득액	월저축금액	월교통비용
1	충청남도	60세이상	14,745,529	120,520	424,084
2	경상북도	50~60세 미만	2,326,959	410,816	196,330
3	전라남도	30~40세 미만	2,797,447	116,712	97,359
4	경기도	50~60세 미만	1,191,178	311,136	158,585
5	경기도	50~60세 미만	2,432,188	252,129	229,484
6	충청남도	50~60세 미만	1,738,900	178,272	156,636
7	경상북도	50~60세 미만	1,547,513	80,479	81,496
8	충청북도	50~60세 미만	1,090,956	313,245	182,909
9	경기도	50~60세 미만	1,025,132	906,241	124,584
10	경기도	40~50세 미만	-200,888	397,214	147,057

< 무상관 잡음첨가 후 자료의 상관계수 >

상관계수	월소득액	월저축금액	월교통비용
월소득액	1.000	0.039	0.482
월저축금액	0.039	1.000	0.120
월교통비용	0.482	0.120	1.000

- 다음 두 표는 3개의 연속형 변수에 대하여 원자료의 상관관계를 가지는 다변량 분포를 이용하여 잡음을 더해주는 방법(다변량 잡음첨가)을 적용한 후 자료와 3개의 연속형 변수에 대한 상관계수이다.
- 잡음이 첨가된 두 표에서 볼 수 있듯이 잡음 첨가 방법은 매우 큰 값을 가진 특이점에 대한 비식별화 효과가 거의 없다.

- 무상관 잡음첨가 방법보다 다변량 잡음첨가 방법이 원자료의 상관관계에 더 가깝게 나타난다.

<잡음첨가에 대한 예시 - 다변량 잡음 첨가 후 자료>

번호	거주시도	연령코드	월소득액	월저축금액	월교통비용
1	충청남도	60세이상	14,836,620	144,838	404,691
2	경상북도	50~60세 미만	2,556,830	354,922	222,834
3	전라남도	30~40세 미만	2,490,272	106,516	88,773
4	경기도	50~60세 미만	1,847,465	274,508	108,732
5	경기도	50~60세 미만	1,748,866	253,480	228,129
6	충청남도	50~60세 미만	1,533,949	147,528	159,572
7	경상북도	50~60세 미만	1,337,458	126,720	118,235
8	충청북도	50~60세 미만	1,236,187	285,506	170,095
9	경기도	50~60세 미만	1,022,909	878,124	96,118
10	경기도	40~50세 미만	999,513	321,494	132,874

<다변량 잡음첨가 후 자료의 상관계수>

상관계수	월소득액	월저축금액	월교통비용
월소득액	1.000	0.072	0.460
월저축금액	0.072	1.000	0.169
월교통비용	0.460	0.169	1.000

2. 자료 교환

- 자료 교환은 연속형 항목에 주로 적용하는 방법으로서 항목 값의 순서를 재배열하는 방법이다.
- 자료 교환 방법은 적절하게 적용되면 단순 집계(총합, 평균, 빈도수 등)가 원자료와 동일하게 유지될 수 있도록 비식별화 처리가 가능한 방법이다.
- 자료 교환은 범주형 변수에도 적용이 가능하지만 범주형 항목들의 구조적인 관계를 왜곡시킬 수 있기 때문에(예를들어 출산경험이 있는 남자) 세심한 주의가 요구된다.

- 따라서 자료 교환을 적용하는 경우 연령, 지역 등과 같은 주요 항목(예를 들어 층화에 사용된 항목들)을 이용하여 전체 자료를 동질적인 집단으로 나누고 각 집단 안에서 자료 교환을 적용해야 한다.
- 자료 교환 방법은 다음과 같은 장점이 있다.
 - 항목들의 값을 변형하거나 왜곡하지 않고 비식별화 처리가 가능하다.
 - 보조 식별항목들에 자료 교환이 적용되면 재식별 위험성이 크게 감소한다.
 - 아주 단순하고 매우 쉽게 구현할 수 있다.
- 자료 교환 방법은 다음과 같은 단점도 있다.
 - 자료의 규모가 큰 경우 자료 교환에 긴 컴퓨터 처리시간과 많은 저장장치를 요구한다.
 - 자료 교환은 하나의 항목에 대한 분석에는 영향을 미치지 않지만 여러 항목에 적용하는 경우 부분집단이나 항목들의 다변량 관계(예를 들어 상관계수)가 왜곡될 수 있다.
- 자료 교환 방법 중 가장 대표적인 방법은 순위 교환이다. 순위 교환은 먼저 항목의 값을 순서대로 배열하고 자료주체들을 여러 개의 유사한 그룹들로 나누어 구성한 후에 각 그룹 안에서 항목들의 값들의 순서를 임의로 바꾸는 방법이다.
- 순위 교환을 적용할 때 유사 그룹의 크기는 보통 전체 자료의 개수에 대한 비율(%)로 지정해준다.
- 순위 교환을 적용하는 경우 유사 그룹의 크기가 커지면 자료의 유용성이 감소한다. 따라서 자료의 유용성에 대한 기준을 정하고 이를 유지할 수 있도록 유사 그룹의 크기를 결정하는 것이 중요하다.
- 순위 교환을 포함한 자료 교환 방법은 최대값과 같은 특이점을 그대로 유지하므로 민감한 정보(예를 들어 소득의 최대값)가 유출될 수 있다. 이러한 경우 상하단 코딩과 같은 방법을 다시 적용할 수 있다.

- 만약 연속형 항목의 값이 가질 수 있는 범위가 제한된 경우(예를 들어 연령, 발생 횟수, 시험 성적 등) 잡음첨가를 적용하면 값의 범위를 벗어 나는 왜곡이 발생할 수 있다. 또한 전역 코딩을 적용하면 연속형 자료가 범주화 되어 정보의 손실이 일어난다. 이러한 경우 순위 교환을 사용할 수 있으며 순위 교환은 다른 방법에 비하여 재식별 위험성과 유용성의 균형을 적절하게 유지할 수 있는 방법으로 알려져 있다.
- 다음 두 표는 100명의 학생들에 대한 5과목의 성적 자료에서 처음 10명의 자료와 성적들의 상관계수를 나타낸 그림이다.

< 자료 교환에 대한 예시 - 순위교환 전 자료 >

번호	국어성적	사회성적	수학성적	과학성적	영어성적
1	69	70	72	72	69
2	93	92	96	94	90
3	89	89	91	88	89
4	66	65	65	67	64
5	92	93	94	94	90
6	92	93	91	93	90
7	63	63	64	64	64
8	94	93	93	93	93
9	93	93	95	95	95
10	65	63	66	63	63

< 자료 교환에 대한 예시 - 순위 교환 전 상관계수 >

성적(상관계수)	국어성적	사회성적	수학성적	과학성적	영어성적
국어성적	1.000	0.301	0.994	0.286	0.994
사회성적	0.301	1.000	0.309	0.998	0.324
수학성적	0.994	0.309	1.000	0.295	0.994
과학성적	0.286	0.998	0.295	1.000	0.309
영어성적	0.994	0.324	0.994	0.309	1.000

- 다음 두 표는 100명의 학생들에 대한 5과목의 성적 자료에 전체 자료의 2.5%로 유사그룹 크기를 정하여 순위 교환을 적용한 후 결과이다.

< 자료 교환에 대한 예시 - 2.5 퍼센트 유사그룹으로 순위교환 적용 후 자료 >

번호	국어성적	사회성적	수학성적	과학성적	영어성적
1	68	69	89	71	88
2	93	93	91	93	90
3	93	92	96	70	89
4	88	68	42	64	68
5	89	92	93	94	91
6	94	93	94	94	90
7	65	65	66	68	66
8	92	92	94	93	92
9	91	90	91	92	95.5
10	63	41	70	44	62

< 자료 교환에 대한 예시 - 2.5 퍼센트 유사그룹으로 순위교환 적용 후 상관계수 >

성적(상관계수)	국어성적	사회성적	수학성적	과학성적	영어성적
국어성적	1.000	0.325	0.624	0.406	0.654
사회성적	0.325	1.000	0.300	0.660	0.357
수학성적	0.624	0.300	1.000	0.436	0.709
과학성적	0.406	0.660	0.436	1.000	0.348
영어성적	0.654	0.357	0.709	0.348	1.000

- 다음은 100명의 학생들에 대한 5과목의 성적 자료에 전체 자료의 5%로 유사그룹 크기를 정하여 순위 교환을 적용한 후의 결과이다.

< 자료 교환에 대한 예시 - 5 퍼센트 유사그룹으로 순위교환 적용 후 자료 >

번호	국어성적	사회성적	수학성적	과학성적	영어성적
1	89	67	89	72	89
2	93	69	90	69	67
3	69	93	93	93	90
4	89	91	64	93	68
5	93	93	70	70	91
6	67	93	94	93	67
7	65	65	89	44	65
8	90	89	96	67	70
9	93	90	92	70	95.5
10	63	61	70	70	64

<자료 교환에 대한 예시 - 5 퍼센트 유사그룹으로 순위교환 적용 후 상관계수>

성적(상관계수)	국어성적	사회성적	수학성적	과학성적	영어성적
국어성적	1.000	0.312	0.468	0.329	0.467
사회성적	0.312	1.000	0.273	0.368	0.345
수학성적	0.468	0.273	1.000	0.173	0.445
과학성적	0.329	0.368	0.173	1.000	0.277
영어성적	0.467	0.345	0.445	0.277	1.000

- 위의 순위 교환 결과를 보면 유사 그룹이 크기가 커질수록 상관계수의 왜곡이 심해짐을 알 수 있다. 하지만 순위 교환은 각 항목의 평균이나 분산이 유지되는 큰 장점이 있다.

3. 부분 총계

- 부분 총계는 자료주체들을 유사성에 따라서 여러 개의 유사그룹으로 나누고 각 그룹에 속한 연속형 항목의 값들을 하나의 대표값(예로 해당 그룹의 평균 또는 중앙값)으로 대체하는 방법이다.

- 부분 총계는 연속형 항목에 주로 적용되며 매출액, 순이익, 소득, 세금 등 매우 치우친 분포를 가지는 민감한 항목에 주로 적용한다.
- 부분 총계를 적용하는 절차는 다음과 같다.
 1. 유사한 자료주체들을 그룹화할 수 있는 기준을 정한다.
 2. 정해진 기준에 따라서 개체들의 유사그룹을 생성한다.
 3. 생성된 유사그룹 안에서 연속형 항목의 값들을 대표할 수 있는 하나의 통계량을 선정한다.
 4. 각 그룹의 연속형 항목의 값을 정해진 하나의 대표값으로 대체한다.
- 부분 총계를 적용할 때 그룹에 속하는 자료주체의 수가 증가할수록 재식별의 위험은 감소하지만 자료의 유용성도 동시에 감소한다. 따라서 적절하게 그룹의 크기를 정하는 것이 중요하다. 일반적으로 그룹의 크기는 3-5개의 개체를 포함하도록 정한다.
- 그룹의 속하는 개체의 수를 결정할 경우 미리 정해진 정보보호 기준이 있다면 이에 따라 그룹의 개체 수를 정할 수 있다. 예를 들어 k-익명성을 따라야 한다면 그룹에 속하는 개체의 수를 k 개로 정하는 것이 좋다.
- 그룹의 크기는 모두 동일하거나 서로 다를 수 있다. 그룹의 크기가 모두 동일한 경우, 특이점이 증가하는 현상이 발생할 수 있다.

◇ 예를 들어 그룹의 크기가 모두 3인 경우, 두 그룹이 (10,20,30), (30, 50, 1000) 으로 구성되어 있다고 하자. 만약 그룹의 평균으로 부분 총계를 적용하면 첫 번째 그룹은 (20,20,20) 로 원자료와 크게 차이가 나지 않지만 두 번째 그룹은 (360,360,360) 로 변하므로 특이점이 3개로 증가하는 현상이 나타난다.

- 그룹의 대표값은 일반적으로 평균이 사용된다. 그룹의 대표값으로 평균을 사용할 때의 장점은 부분 총계를 적용한 후에도 대표적인 산술적 통계량(총합, 평균, 비율 등)은 변하지 않는 점이다.

- 그룹의 대표값으로 평균을 사용할 때의 단점은 특이점이 존재하는 경우 평균을 사용하면 비식별화 후에 특이점이 증가할 수 있으며 중요한 통계량(예를 들어 소득 통계에서 지니계수)이 심하게 왜곡될 수 있다. 이러한 경우 중앙값을 사용하는 것이 안전할 수 있다.
- 부분 총계에서 그룹을 구성할 때는 다음과 같은 원칙을 유지하는 것이 중요하다.
 - 그룹 안의 개체들은 서로 유사하게 구성해야 한다.
 - 층화를 이용한 표본추출을 사용한 경우 각 층 안에서 그룹을 구성한다.
 - 가능하면 가중치가 다른 개체들을 같은 그룹으로 구성하지 않는다.
- 유사한 자료주체들을 그룹화할 수 있는 기준은 다양하며 다음과 같은 2개의 대표적인 방법이 있다.
 - 일변량 방법: 하나의 연속형 항목(또는 하나의 순서가 있는 범주형 항목)을 그 크기 순으로 정렬하여 순서대로 정해진 자료 주체의 개수로 구성된 유사그룹을 만드는 방법이다. 만약 부분 총계를 적용할 항목의 개수가 2개 이상인 경우라면, 각 항목에 대하여 각각 일변량 방법을 적용하거나, 중요한 하나의 항목을 선정하고 그 항목에 일변량 방법을 적용하여 그룹을 정할 수 있다.
 - 다변량 방법: 여러 개의 항목들에 대하여 동시에 부분 총계를 적용하는 방법을 다변량 부분 총계라고 한다. 대부분의 통계작성기관이 주로 사용하는 방법이 다변량 부분 총계 방법이다.

◇ 하나의 변수를 고려하는 경우와 마찬가지로 다변량 부분 총계 방법도 그룹을 결정하는 것이 가장 중요한 작업이다. 그룹을 정하는 방법은 다변량 거리를 이용하여 그룹 안은 최대한 유사하게, 그룹 간은 최대한 상이하게 그룹을 형성한다. 다변량 부분 총계 방법에서 가장 자주 사용되는 방법은 평균 벡터 최대거리 방법(Maximum Distance to Average Vector; MDAV)이다. 평균 벡터 최대거리 방법은 개체 간의 거리를 계산할 때 사용되는 다변량 거리의 정의에 따라서 그룹이 다르게 형성될 수 있다. 사용할 수 있는 다변량 거리들 중에서는 유클리디안 거리가 일반적으로 가장 많이 사용되고 있으며, 통계적 거리, 주성분을 이용한 거리 등도 사용될 수 있다.

1. 먼저 그룹 안에 속하는 개체의 수 k 를 결정한다.
2. 전체 자료에서 선택된 항목들의 평균값으로 구성된 벡터, 즉 평균 벡터를 구한다.
3. 평균 벡터에서 주어진 다변량 거리(예로 유클리디안 거리)의 정의를 이용하여 가장 멀리 떨어진 개체 A 를 선택한다.
4. 개체 A 에서 가장 가까운 k 개의 개체를 구하고 하나의 그룹 1을 형성한다.
5. 다음으로 개체 A 에서 그룹 1에 속한 개체를 제외한 나머지 개체 중 가장 멀리 떨어진 개체 B 를 선택한다.
6. 개체 B 에서 그룹 1에 속한 개체를 제외한 가장 가까운 k 개의 개체를 구하고 하나의 그룹 2를 형성한다.
7. 이미 형성된 그룹에 속한 개체를 제외하고 2-6의 알고리즘을 다시 실행한다.

○ 이제 예시를 통하여 서로 다른 방법을 적용한 부분 총계의 결과를 비교해보자. 먼저 다음 표와 같이 3개의 연속형 항목(부동산금액, 저축금액, 기타자산금액)이 있는 20개의 자료가 있다고 하자.

< 부분총계 처리 전 자료에 대한 예시 >

번호	거주시도	연령코드	부동산금액	저축금액	기타자산금액
1	충청남도	50~60세 미만	24,101,000	447,000	98,000
2	충청남도	60세이상	22,742,000	6,417,000	1,711,000
3	전라남도	60세이상	22,574,000	2,658,000	655,000
4	충청북도	50~60세 미만	19,420,000	577,000	339,000
5	충청북도	60세이상	18,814,000	1,196,000	104,000
6	충청북도	60세이상	18,768,000	420,000	60,000
7	강원도	60세이상	11,692,000	3,633,000	21,000
8	전라남도	60세이상	10,030,000	182,000	1,797,000
9	경기도	60세이상	9,870,000	269,000	91,000
10	전라북도	50~60세 미만	9,698,000	554,000	86,000
11	경상남도	60세이상	9,536,000	1,928,000	1,269,000
12	경상북도	30~40세 미만	8,755,000	585,000	173,000
13	전라북도	50~60세 미만	8,640,000	200,000	230,000
14	전라북도	60세이상	8,576,000	487,000	5,791,000
15	전라북도	60세이상	8,308,000	52,000	30,000
16	경기도	60세이상	7,634,000	173,000	540
17	전라북도	60세이상	7,597,000	62,000	0
18	전라북도	40~50세 미만	7,434,000	26,000	19,000
19	강원도	50~60세 미만	7,371,000	504,000	68,000
20	전라북도	50~60세 미만	7,310,000	251,000	418,000

< 부분총계(하나의 항목을 이용한 일변량 방법 적용) 처리 후 자료에 대한 예시 >

번호	거주시도	연령코드	부동산금액	저축금액	기타자산금액
1	충청남도	50~60세 미만	21,530,200	2,259,000.00	581,400.0
2	충청남도	60세이상	21,530,200	2,259,000.00	581,400.0
3	전라남도	60세이상	21,530,200	2,259,000.00	581,400.0
4	충청북도	50~60세 미만	21,530,200	2,259,000.00	581,400.0
5	충청북도	60세이상	21,530,200	2,259,000.00	581,400.0
6	충청북도	60세이상	13,496,667	1,411,666.67	626,000.0
7	강원도	60세이상	13,496,667	1,411,666.67	626,000.0
8	전라남도	60세이상	13,496,667	1,411,666.67	626,000.0
9	경기도	60세이상	9,701,333	917,000.00	482,000.0
10	전라북도	50~60세 미만	9,701,333	917,000.00	482,000.0
11	경상남도	60세이상	9,701,333	917,000.00	482,000.0
12	경상북도	30~40세 미만	8,657,000	424,000.00	2,064,666.7
13	전라북도	50~60세 미만	8,657,000	424,000.00	2,064,666.7
14	전라북도	60세이상	8,657,000	424,000.00	2,064,666.7
15	전라북도	60세이상	7,846,333	95,666.67	10,180.0
16	경기도	60세이상	7,846,333	95,666.67	10,180.0
17	전라북도	60세이상	7,846,333	95,666.67	10,180.0
18	전라북도	40~50세 미만	7,371,667	260,333.33	168,333.3
19	강원도	50~60세 미만	7,371,667	260,333.33	168,333.3
20	전라북도	50~60세 미만	7,371,667	260,333.33	168,333.3

○ 위의 표는 가장 단순한 방법인 일변량 방법을 적용한 예시이다.

- 부동산 금액에 대하여 자료주체들을 정렬하여 3개의 개체를 포함하는 유사그룹으로 나누고, 3개의 항목을 그룹 평균으로 대체한 결과이다.
- 20개의 개체를 3개의 그룹으로 나누는 경우, 모두 같은 수의 개체를 가진 유사그룹으로 나눌 수 없으므로 아래와 같이 값이 가장 큰 그룹은 5개의 개체가 포함되도록 하면 노출의 위험성을 줄일 수 있다.

< 부분총계(각 항목에 개별적으로 일변량 방법 적용) 처리 후 자료에 대한 예시 >

번호	거주시도	연령코드	부동산금액	저축금액	기타자산금액
1	충청남도	50~60세 미만	21,530,200	479,333.33	125,000.000
2	충청남도	60세이상	21,530,200	3,166,400.00	2,244,600.000
3	전라남도	60세이상	21,530,200	3,166,400.00	2,244,600.000
4	충청북도	50~60세 미만	21,530,200	572,000.00	329,000.000
5	충청북도	60세이상	21,530,200	3,166,400.00	125,000.000
6	충청북도	60세이상	13,496,667	313,333.33	37,000.000
7	강원도	60세이상	13,496,667	3,166,400.00	37,000.000
8	전라남도	60세이상	13,496,667	185,000.00	2,244,600.000
9	경기도	60세이상	9,701,333	313,333.33	81,666.667
10	전라북도	50~60세 미만	9,701,333	572,000.00	81,666.667
11	경상남도	60세이상	9,701,333	3,166,400.00	2,244,600.000
12	경상북도	30~40세 미만	8,657,000	572,000.00	125,000.000
13	전라북도	50~60세 미만	8,657,000	185,000.00	329,000.000
14	전라북도	60세이상	8,657,000	479,333.33	2,244,600.000
15	전라북도	60세이상	7,846,333	46,666.67	37,000.000
16	경기도	60세이상	7,846,333	185,000.00	6,513.333
17	전라북도	60세이상	7,846,333	46,666.67	6,513.333
18	전라북도	40~50세 미만	7,371,667	46,666.67	6,513.333
19	강원도	50~60세 미만	7,371,667	479,333.33	81,666.667
20	전라북도	50~60세 미만	7,371,667	313,333.33	329,000.000

- 위의 표는 3개의 연속형 항목에 각각 일변량 방법을 개별적으로 적용한 결과이다.
- 부동산 항목, 저축금액, 기타자산금액에 대하여 서로 개별적으로 유사그룹을 만들어 부분총계를 적용한 결과이다.
- 각 항목에 적용할 때 선택된 항목의 순서대로 3개의 개체를 포함하는 그룹으로 나누어 그룹 평균으로 대체한 결과이다.

< 부분총계(다변량 방법 적용) 처리 후 자료에 대한 예시 >

번호	거주시도	연령코드	부동산금액	저축금액	기타자산금액
1	충청남도	50~60세 미만	20,763,000	481,333.3	165,666.67
4	충청북도	50~60세 미만	20,763,000	481,333.3	165,666.67
6	충청북도	60세이상	20,763,000	481,333.3	165,666.67
2	충청남도	60세이상	19,002,667	4,236,000.0	795,666.67
3	전라남도	60세이상	19,002,667	4,236,000.0	795,666.67
7	강원도	60세이상	19,002,667	4,236,000.0	795,666.67
5	충청북도	60세이상	12,794,000	673,000.0	93,666.67
9	경기도	60세이상	12,794,000	673,000.0	93,666.67
10	전라북도	50~60세 미만	12,794,000	673,000.0	93,666.67
8	전라남도	60세이상	9,380,667	865,666.7	2,952,333.33
11	경상남도	60세이상	9,380,667	865,666.7	2,952,333.33
14	전라북도	60세이상	9,380,667	865,666.7	2,952,333.33
12	경상북도	30~40세 미만	7,945,600	179,600.0	44,508.00
15	전라북도	60세이상	7,945,600	179,600.0	44,508.00
16	경기도	60세이상	7,945,600	179,600.0	44,508.00
17	전라북도	60세이상	7,945,600	179,600.0	44,508.00
18	전라북도	40~50세 미만	7,945,600	179,600.0	44,508.00
13	전라북도	50~60세 미만	7,773,667	318,333.3	238,666.67
19	강원도	50~60세 미만	7,773,667	318,333.3	238,666.67
20	전라북도	50~60세 미만	7,773,667	318,333.3	238,666.67

- 위의 표는 평균 벡터 최대거리 방법을 이용한 다변량 방법을 적용하여 그룹을 나누는 방법을 적용한 결과이다.
- 각 유사그룹의 개체의 수를 3개로 지정한 결과로서 3개의 금액 항목에 대하여 유클리디안 거리를 이용하여 그룹 내의 개체들이 최대로 유사하게 그룹을 구성한 결과이다.
- 지금까지 살펴본 부분 총계 방법은 모두 평균을 사용하였으므로 연속형 항목의 산술적 요약 통계량(총합, 평균 등)은 비식별화 처리 후에도 변하지 않는다.

- 하지만 자료의 순서를 이용하여 작성되는 통계량들(중앙값, 백분위수, 상관계수 등)은 비식별화 처리 후 원자료의 통계량과 다르게 나타난다.
- 다음 4개의 표에 나타난 결과들은 원자료의 상관계수와, 위에서 고려한 세 가지 상이한 부분 총계 방법을 적용한 후에 항목들의 상관계수의 변화를 나타낸 것이다.
- 평균 벡터 최대거리 방법을 이용한 다변량 방법이 원자료의 상관계수와 가장 유사하게 나타나는 것을 알 수 있다.

< 원자료의 상관 계수 >

상관계수	부동산금액	저축금액	기타자산금액
부동산금액	1.0000	0.5065	-0.0416
저축금액	0.5065	1.0000	0.1444
기타자산금액	-0.0416	0.1444	1.0000

< 부분 총계 적용 후 상관 계수 (하나의 항목에 일변량 방법 적용) >

상관계수	부동산금액	저축금액	기타자산금액
부동산금액	1.0000	0.9750	-0.0244
저축금액	0.9750	1.0000	-0.0056
기타자산금액	-0.0244	-0.0056	1.0000

< 부분 총계 적용 후 상관 계수 (각 항목에 일변량 방법 개별 적용) >

상관계수	부동산금액	저축금액	기타자산금액
부동산금액	1.0000	0.5650	0.2836
저축금액	0.5650	1.0000	0.4585
기타자산금액	0.2836	0.4585	1.0000

< 부분 총계 적용 후 상관 계수 (다변량 방법 적용) >

상관계수	부동산금액	저축금액	기타자산금액
부동산금액	1.0000	0.5651	-0.1125
저축금액	0.5651	1.0000	0.1874
기타자산금액	-0.1125	0.1874	1.0000

5 매크로데이터 비식별화 방법

집계표와 같은 매크로데이터에 비식별화 처리를 하는 경우 매우 작은 수를 나타내는 셀을 감추는 셀감추기가 주로 적용된다. 이 경우 집계표의 행과 열의 소계·총계를 이용하여 감추어진 셀을 다시 추론할 수 없도록 하는 별도의 비식별화 작업이 필요하다. 이 절에서는 집계표 형태의 매크로데이터를 비식별화 하는 절차를 예시로 들어 설명하고자 한다. 다음 집계표는 산업분류와 지역에 따른 사업체의 수를 나타낸 표이다.

< 비식별화 적용 전의 집계표 >

산업분류(중)	지역	전체사업체수	개인사업체수	법인사업체수
A	a	1	1	0
	b	3	2	1
	c	5	2	3
	합계	9	5	4
B	a	3	1	2
	b	4	2	2
	c	4	2	2
	합계	11	5	6
C	a	8	7	1
	b	9	5	4
	c	9	6	3
	합계	26	18	8
D	a	1	1	0
	b	0	0	0
	c	1	0	1
	합계	2	1	1
합계	합계	48	29	19

1. 작은 수(k값 미만)에 대한 셀감추기

- 매크로데이터 비식별화 절차의 첫 번째 과정은 k-익명성에 따라 기준값으로 정해진 값*보다 작은 셀의 값을 감추는 것이다.

* 인구·가구·주택은 5, 사업체는 3을 기준값으로 정함

- 다음 표는 사업체 관련 집계표이므로 기준값이 3이며, 주황색으로 나타난 셀은 3 미만인 셀로서 셀감추기를 적용해야 하는 셀이다.

< 셀감추기를 적용한 집계표: 색깔이 있는 셀은 감추기 적용 >

산업분류(중)	지역	전체사업체수	개인사업체수	법인사업체수
A	a	1	1	0
	b	3	2	1
	c	5	2	3
	합계	9	5	4
B	a	3	1	2
	b	4	2	2
	c	4	2	2
	합계	11	5	6
C	a	8	7	1
	b	9	5	4
	c	9	6	3
	합계	26	18	8
D	a	1	1	0
	b	0	0	0
	c	1	0	1
	합계	2	1	1
합계	합계	48	29	19

2. 조직형태 합계에 의한 유추 방지 적용

- 개인사업체수와 법인사업체수를 더하여 산출된 전체 사업체수를 이용하여 감추어진 셀의 숫자를 추론할 수 없도록, 관련 셀에도 셀감추기를 적용한다.
- 다음 표에서 빨간색으로 나타난 셀들에 셀감추기를 적용해야 조직형태 합계를 이용한 숨겨진 값의 유추를 방지할 수 있다.

< 조직형태 합계에 의한 추론을 방지한 집계표: 색깔이 있는 셀은 감추기 적용 >

산업분류(중)	지역	전체사업체수	개인사업체수	법인사업체수
A	a	1	1	0
	b	3	2	1
	c	5	2	3
	합계	9	5	4
B	a	3	1	2
	b	4	2	2
	c	4	2	2
	합계	11	5	6
C	a	8	7	1
	b	9	5	4
	c	9	6	3
	합계	26	18	8
D	a	1	1	0
	b	0	0	0
	c	1	0	1
	합계	2	1	1
합계	합계	48	29	19

3. 지역별 합계에 의한 유추 방지 적용

- 산업분류별로 지역 a, b, c 의 합계를 이용하여 감추어진 셀의 숫자를 복원할 수 없도록, 관련된 셀에도 셀감추기를 적용한다.

열과 행에 의한 유추를 방지하기 위해 추가 비식별화를 적용하기 위한 일반적인 방법은 관련된 셀들 중에 최소값을 가진 셀을 감추는 것이다. 하지만, 유추를 방지하기 위한 셀감추기에 의하여 별도의 셀을 감추어야 하는 상황이 발생할 수 있다. 따라서 유추를 방지하기 위한 셀감추기를 적용하는 경우 추가로 감추는 셀의 개수가 최소한이 되도록 노력해야 한다.

- 다음 표는 각 산업분류마다 지역별 합계에 해당하는 행에 의하여 숨겨진 값을 유추할 수 없도록 관련된 셀들 중에서 최소값(파란색으로 나타난 셀)에 셀감추기를 적용한 예시이다.

< 지역별 합계에 의한 추론을 방지한 집계표(최소값 감추기 적용) >

산업분류(중)	지역	전체사업체수	개인사업체수	법인사업체수
A	a	1	1	0
	b	3	2	1
	c	5	2	3
	합계	9	5	4
B	a	3	1	2
	b	4	2	2
	c	4	2	2
	합계	11	5	6
C	a	8	7	1
	b	9	5	4
	c	9	6	3
	합계	26	18	8
D	a	1	1	0
	b	0	0	0
	c	1	0	1
	합계	2	1	1
합계	합계	48	29	19

- 하지만, 이 경우 다시 조직형태 합계로 감추어진 셀을 추론할 수 없도록 아래 예시와 같이 추가적인 셀들(녹색으로 표시)에 다시 셀감추기를 적용해야 한다.
- 예를 들어 산업분류 C에서 b지역의 개인사업체수 5는 조직형태 합계인 b지역의 전체사업자수 9를 이용하여 복원할 수 있다. 따라서 법인사업체수 4도 추가적으로 셀감추기를 적용해야 한다. 또한 같은 이유로 산업분류 C에서 지역 c의 개인사업체수 6도 추가적으로 셀감추기를 적용해야 한다.

< 조직형태 합계에 의한 추론을 방지하기 위해 추가 셀감추기 적용 >

산업분류(중)	지역	전체사업체수	개인사업체수	법인사업체수
A	a	1	1	0
	b	3	2	1
	c	5	2	3
	합계	9	5	4
B	a	3	1	2
	b	4	2	2
	c	4	2	2
	합계	11	5	6
C	a	8	7	1
	b	9	5	4
	c	9	6	3
	합계	26	18	8
D	a	1	1	0
	b	0	0	0
	c	1	0	1
	합계	2	1	1
합계	합계	48	29	19

- 최소값 감추기를 원칙적으로 적용하면 추가적인 셀들에 감추기를 적용해야 할 경우가 발생하여 정보의 손실이 커진다. 따라서 관련된 셀들에 셀감추기를 적용하는 경우 최소한의 셀들을 감출 수 있는 방향으로 셀감추기를 적용하는 것이 바람직하다.
- 다음 표는 산업분류 C에서 c지역의 법인사업체의 수 3에 셀감추기를 적용하는 대신 더 큰 값을 가진 b지역의 법인사업체의 수 4에 셀감추기를 적용한 예이다.
- 이러한 경우 앞에서의 셀감추기에 비하여 추가적인 셀을 감추기 하지 않아도 되므로 최소값을 선택하는 것보다 정보의 손실이 작다.

< C 비수도권 법인사업체수 4에 셀감추기를 적용한 집계표 >

산업분류(중)	지역	전체사업체수	개인사업체수	법인사업체수
A	a	1	1	0
	b	3	2	1
	c	5	2	3
	합계	9	5	4
B	a	3	1	2
	b	4	2	2
	c	4	2	2
	합계	11	5	6
C	a	8	7	1
	b	9	5	4
	c	9	6	3
	합계	26	18	8
D	a	1	1	0
	b	0	0	0
	c	1	0	1
	합계	2	1	1
합계	합계	48	29	19

4. 총계에 의한 유추 방지 적용

- 열과 행의 합을 이용한 복원을 방지하기 위해 셀감추기를 적용한 후 항목의 전체 합계를 이용한 복원도 방지할 수 있도록 비식별화 작업을 수행한다.
- 총계에 의한 유추 방지를 위한 비식별화 처리도 추가로 감추기가 필요한 셀들이 발생하지 않도록 최소한의 셀들을 감출 수 있는 방향으로 셀감추기를 적용하는 것이 바람직하다.
- 다음 표에서 산업분류의 총 합계(A, B, C, D의 총 합계)에 의하여 숨겨진 값을 유추할 수 없도록 보라색으로 나타난 셀들에 셀감추기를 적용해야 총계를 이용한 유추를 방지할 수 있다.

< 총계 추론을 방지한 집계표: 색깔이 있는 셀은 감추기 적용 >

산업분류(중)	전체사업체수	개인사업체수	법인사업체수
A	9	5	4
B	11	5	6
C	26	18	8
D	2	1	1
합계	48	29	19

- 추가적으로 감추어야 하는 항목 수를 최소화하기 위해 산업분류 A의 합계를 감추는 것이 가장 합리적이다.
 - 위의 표에서 만약 산업분류 B 또는 C의 전체 합계에 감추기를 적용하면 추가적으로 산업분류 B 또는 C의 전체 사업체수 구성 항목을 감춰야 함
 - 산업분류 A의 전체사업체수 9에 감추기를 적용하면 오히려 산업분류 A의 b지역에 대한 전체사업체수 3은 감추기를 적용할 필요가 없어진다는 점에서도 산업분류 A의 합계를 감춰야 함

- 위의 표에 총계를 이용한 유추를 방지하기 위한 셀감추기를 적용한 후에 최종적인 비식별화가 적용된 표는 다음과 같다. 색깔이 있는 셀들은 셀감추기를 적용해야 하는 셀들이다.

< 비식별화가 적용된 집계표: 색깔이 있는 셀은 감추기 적용 >

산업분류(중)	지역	전체사업체수	개인사업체수	법인사업체수
A	a	1	1	0
	b	3	2	1
	c	5	2	3
	합계	9	5	4
B	a	3	1	2
	b	4	2	2
	c	4	2	2
	합계	11	5	6
C	a	8	7	1
	b	9	5	4
	c	9	6	3
	합계	26	18	8
D	a	1	1	0
	b	0	0	0
	c	1	0	1
	합계	2	1	1
합계	합계	48	29	19

5. 비식별화 적용 결과

- 아래 집계표는 비식별화가 적용되어 최종적으로 반출할 수 있는 형태의 표이다. 셀감추기가 적용된 셀들은 모두 동일한 서식 및 형태로 감추기가 적용되어야 한다.

< 비식별화가 적용된 집계표: 반출 가능한 형태 >

산업분류(중)	지역	전체사업체수	개인사업체수	법인사업체수
A	a	*	*	0
	b	3	*	*
	c	5	*	*
	합계	*	*	*
B	a	3	*	*
	b	4	*	*
	c	4	*	*
	합계	11	5	6
C	a	8	*	*
	b	9	*	*
	c	9	6	3
	합계	26	18	8
D	a	*	*	0
	b	0	0	0
	c	*	0	*
	합계	*	*	*
합계	합계	48	29	19

1. 배경

- 통계작성기관이 개인정보 등 민감한 정보의 노출 방지를 위해 수행하는 다양한 비식별화 처리방법은 정보의 손실을 일으키는 단점이 있다.
- 이러한 문제를 해결하기 위해 원자료와 유사한 특성을 지니고 있어 정보의 손실을 줄이면서도 가상의 자료를 생성하여 민감한 정보의 노출을 방지하는 재현자료 기법이 주목받고 있다.
- 만약 매우 민감한 항목들이 많아서(예를 들어 세금, 상병 및 진료 내역 등) 비식별화 처리 후에도 개인정보의 유출이 우려되는 경우 재현자료를 제공하는 것이 안전하다.

2. 재현자료의 정의

- 재현자료(Synthetic data)는 실제로 수집된 통계기초자료를 기반으로 통계적 방법, 기계학습 또는 인공지능 방법을 통하여 추정된 모형에서 새롭게 생성된 모의자료를 의미한다.
- 재현자료는 비식별화 방법이 적용된 공공용 기초자료의 한계인 정보의 손실을 극복하기 위한 대안으로 사용할 수 있으며, 가상의 자료이기 때문에 비식별화가 적용된 공공용 기초자료보다 더 많은 항목과 더 세밀한 정보로 구성될 수 있다.
- 하지만 사용된 모형이 원자료가 지니고 있는 모든 특성과 항목들의 관계를 구현할 수 없으므로 원자료를 이용하여 작성한 통계와 차이가 존재할 수 있는 점을 이용자에게 충분히 고지할 필요가 있다.
- 또한, 재현자료를 생성하는 경우 사전에 이용자의 의도를 파악하여 이용자의 활용 목적에 필요한 변수의 통계량에 대해서는 원자료에 의해 생성된 통계량과 차이를 적게 하여 생성할 필요가 있다.

3. 재현자료의 목적

- 재현자료를 생성하는 경우 그 목적을 명확하게 정하는 것이 중요하다. 재현자료는 다음과 같은 목적을 위해 생성될 수 있다.
 - 통계데이터센터에서 인가된 통계기초자료 또는 행정통계자료를 이용하기 전에 사전 분석을 할 수 있도록 재현자료를 제공하여 분석 작업의 효율을 높이고 결과를 더 신속하게 얻을 수 있다.
 - 또한, 통계적 분석이나 인공 지능 개발을 위한 교육용 목적 등으로 제공할 수 있다.
- 재현자료를 생성하는 주요 목적에 따라서 자료를 생성하는 적절한 모형을 선택해야 한다.
 - * 모든 항목들 간의 관계를 분석하거나 정확한 통계량을 요구하는 경우 이를 충족하는 재현자료 생성모형 구축이 매우 어려우므로, 이러한 경우에는 통계데이터센터에서 인가용 통계기초자료 또는 행정통계자료 이용이 더욱 적절할 수 있다.

4. 재현자료의 분류

- 완전 재현자료
 - 재현자료에 실제 원자료가 하나도 포함되지 않고 모두 모형을 통하여 생성된 자료로만 이루어진 것을 완전 재현자료라고 한다.
 - 공개하려는 원자료의 모든 항목들의 값이 추정된 모형에서 생성된 자료이므로 개인정보보호의 측면에서 가장 강력한 보안성을 가지고 있는 자료이다.
 - 모든 항목들의 값을 새로 생성해야 하기 때문에 모든 변수의 관계를 설정하는 복잡한 모형이 필요하다. 또한 새로 생성하는 과정에서 원자료로부터 얻은 대부분의 실제 통계량과 차이를 보일 수 있다.

○ 부분 재현자료

- 공개하려는 항목들 중 일부만을 선택하여 재현자료로 생성한 자료를 뜻한다.
- 이러한 경우 보통 재현자료로 대치되는 변수들은 재식별에 사용되어 개인정보의 노출 위험성이 큰 민감한 변수들이다.
- 재현자료로 대치되지 않는 항목들은 일반적으로 인구사회적인 항목들(예를 들어 성별, 연령, 직업, 거주지 등)로 구성되며 재현자료 이용시 주요한 설명변수로 사용되는 경우가 많다.
- 자료의 일부분만 재현하기 때문에 완전 재현자료보다 생성에 사용되는 모형이 더 간단하고 시간과 비용이 적게 든다.

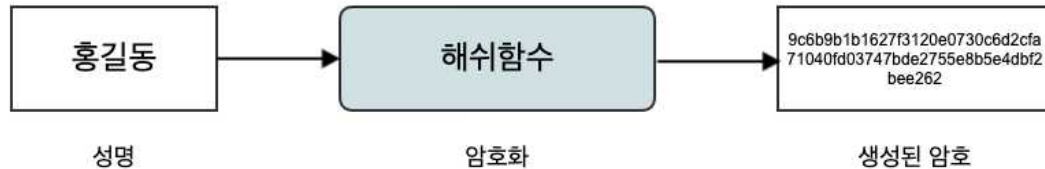
5. 재현자료의 생성 방법

- 재현자료를 생성하는 가장 쉽고 기본적인 방법은 모형을 이용하여 결측값을 대체하는 결측값 대체법을 응용하는 것이다. 재현의 대상이 되는 자료를 축차적으로 결측값으로 놓고 추정된 모형을 이용하여 대체값으로 재현자료를 만드는 방법이다.
- 최근에는 기계학습이나 인공지능을 이용하여 원자료에 대해 학습 훈련을 실행한 후에 추정된 모형을 이용하여 재현자료를 생성하는 방법이 자주 사용되고 있다.
- 특별히 베이지안을 이용한 통계적 모형, 분류나무를 이용한 기계 학습 모형과 조건부 GAN(Conditional Generative Adversarial Network) 방법이 테이블 형식의 자료에 대한 생성 방법으로 주로 사용된다.

7 일방향 암호화

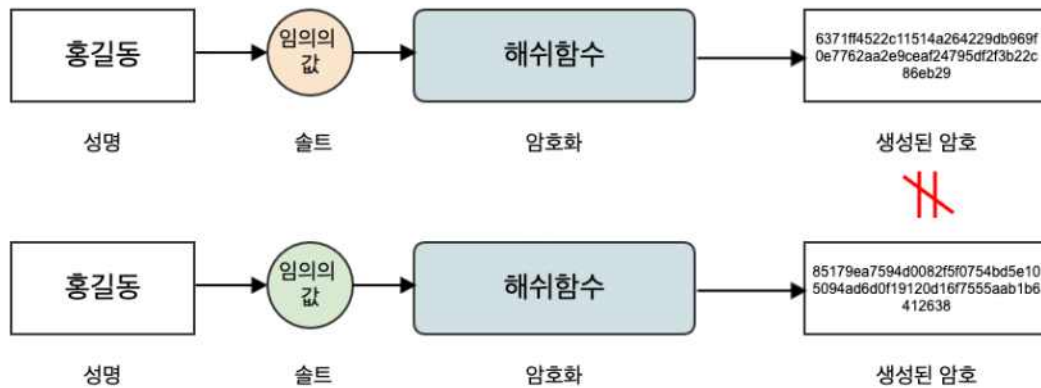
- 일방향 암호화는 개인식별정보를 암호화하여 비식별화하는 경우 주로 사용하는 방법으로, 원래의 정보에 대한 암호화의 적용만 가능하고 암호문에 대한 복호화 적용이 불가능한 암호화 기법이다.
- 일방향 암호를 생성할 때 사용되는 대표적인 방법은 해쉬(hash) 함수를 이용하는 방법으로 해쉬함수는 임의의 길이의 자료를 고정된 길이의 자료로 변환하여 매핑하는 함수이다.
- 해쉬함수로 변환된 자료는 원본 개인식별정보가 암호화되어 자료 주체의 직접적인 식별이 불가능하다. 다음은 SHA256 해쉬함수로 이름을 일방향 암호화한 예시이다.

< 해쉬함수를 이용한 일방향 암호화 >



- 암호학적 해쉬함수를 이용한 일방향 암호화는 주어진 문자열을 언제나 같은 형태의 암호로 변환하는 결정적 암호화 방법이므로, 암호화된 많은 사례를 수집하여 분석하면 다시 원래의 정보를 추론할 수 있는 위험성이 있다.
- 이러한 위험성을 제거하기 위하여 암호화할 때 원래의 정보에 임의의 잡음(솔트)를 추가하여 암호화하는 개선된 방법을 사용할 수 있다. 솔트를 추가하면 대량의 자료를 이용한 원자료의 복원을 막을 수 있으므로 결정적 암호화 방법의 단점을 보완할 수 있다.

< 솔트를 추가한 해시함수를 이용한 일방향 암호화 >



* 솔트값이 다르면 생성된 암호도 다르므로, 일방향 암호화의 위험성 방지 가능

- 해시함수를 이용한 일방향 암호화의 단점을 보완하는 다른 방법은 원래의 정보에 해시함수를 여러 번 적용하여 암호화하는 방법이 있다. 해시함수를 반복하여 적용하는 숫자를 변화시키면 대량의 자료를 이용한 원자료의 복원을 방지할 수 있다. 그러나 해시함수를 반복하는 숫자를 고정하는 경우, 반복의 숫자가 노출되면 암호를 복원할 수 있는 가능성이 높아지는 위험성이 있다.

< 해시함수를 반복 적용하는 일방향 암호화 >



< 의안 소관 부서명 >

통계청 통계데이터기획과	
연 락 처	(042) 481 - 2382