최종보고서

매스킹 교육과정 및 교재 개발

2011. 10

책임연구원: 안형진 (고려대학교 의학통계학교실)
공동연구원: 이용희 (서울시립대학교 통계학과)
공동연구원: 송주원 (고려대학교 통계학과)
공동연구원: 김규영 (통계청)

제 출 문

통계개발원장 귀하
매스킹 교육과정 및 교재 개발 결과보고서 제출

본 보고서를 “매스킹 교육과정 및 교재 개발” 결과 보고서로 제출합니다.

2010년 10월

책임연구자: 안형진 (고려대학교 의학통계학교실)
공동연구자: 이용희 (서울시립대학교 통계학과)
공동연구자: 송주원 (고려대학교 통계학과)
공동연구자: 김규영 (통계청)

차 례

<표 차례>

<그림 차례>

제 1장. 통계적 정보보호의 개념

1.1 정보의 노출(disclosure) 개념

정보의 노출이란 주제는 매우 어렵고 하나의 정해진 개념은 없다. 하지만 이 절에서는 자료를 공공에게 제공하는 과정에 있어 비밀정보 노출에 한정하여 Duncan 등(1993)이 소개한 세 가지 종류의 정보 노출에 관하여 소개한다.

신원노출 (identity disclosure) : 제3자가 제공된 자료를 통하여 개인의 신원을 알아내는 경우이다. 개인의 신원을 알아낸 것만으로는 기밀성(confidentiality)의 필요조건을 위배한 것은 아닐 수도 있다. 매크로 자료인 경우 일반적으로 알려진 개인의 기밀자료가 누설되지 않는 한 신원을 알아내는 것이 신원노출(identity disclosure)이라고 할 수 없다. 하지만 마이크로 자료인 경우는 그 자료가 포함하고 있는 정보가 매우 세부적이라 개인의 기밀자료가 누설되었다고 볼 수 있으므로 개인의 신원을 알아내는 것이 신원노출이라고 할 수 있다.

속성노출 (attribute disclosure) : 한 개인의 기밀정보가 드러나고 그 정보가 특정한 개인의 속성으로 돌려질 때 이를 속성노출이라 한다. 즉, 정보의 누설과 그 정보가 특정한 개인과 매칭되는 두 가지 조건을 만족하여야 한다. 속성노출은 기밀정보가 정확히 누설되거나 매우 근접하게 추정되는 경우에 발생한다.

추론노출 (inferential disclosure) : 제공된 자료의 통계적 성질로부터 정보가 신뢰성 있게 추정되는 경우에 발생하는 노출을 추론노출이라고 한다. 예를 들어,

자료가 주택 구매 가격과 소득은 아마도 높은 상관을 보이고 주택 구매 가격이 공공정보라면 제3자는 주택 구매 가격으로부터 소득을 유추할 수 있다. 정보제공기관은 이 추론노출에 관해서는 큰 신경을 쓰지 않는다. 만일 이 추론노출을 너무 걱정한 나머지 제공된 자료 사이의 변수들을 왜곡시킨다면 자료제공의 의미가 없으며 일반적으로 추론은 모집단의 전체적 특성을 예측하기 위함이지 개개인의 정보를 정확히 예측하기 위함은 아니다. 따라서 추론노출로 알게 된 개인의 정보는 정확하지 않은 경우가 대부분이다.

1.2 노출위험의 개념적 모형

정보제공기관은 특정한 개인의 특징에 관한 정보 또는 신원 노출을 차단하기 위한 정책과 절차를 개발한다. 노출위험을 평가하기 위하여 정보제공기관은 위험요인과 노출의 결과를 잘 알고 있어야 한다. 정보를 제공한 개인은 정보제공기관이 개인의 신원을 노출시키지 않을 것이라는 믿음이 있으며 이를 지켜주어야 한다. 정보보호 보장에 관한 사항은 설문지 첫 장에 기술되어야 하며 면접조사인 경우에는 면접시작 시점에서 이를 제공하여야 한다. 따라서 자료를 수집하고 가공하여 제공하는 모든 과정에서 정보보호를 보장할 수 있어야 한다. 비밀노출은 정보제공기관의 임무에 중대한 해를 입히는 재난이 될 수 있다. 단지 몇 명의 정보가 노출되어도 정보제공기관과 정보제공자 사이의 믿음은 깨질 수 있으며 이는 결국 많은 응답거부로 이어질 수 있다. 하지만 이러한 이유로 정보제공기관이 노출의 위험을 전혀 감수하지 않는 것은 높은 유용성을 가진 자료를 공공에게 제공해야 하는 의무와 상충될 수 있다. 이는 정보제공기관이 좀 더 실용적이어야 한다. 이러한 실용성은 인식적인 (perceived) 위험과 객관적인 (objective) 위험이 상호작용하고 있다. 만일 위험이 매우 작다고 인식되는 경우에는 자료유출 시도자(제공된 자

료로부터 개인정보를 얻고자 하는 자 또는 기관, data snooper)는 다른 방법을 통하여 목적을 달성하려고 할 것이며, 따라서 자료유출 시도자의 노출위험 인식이 감소하면 사실상 객관적인 위험도 감소하게 된다. Marsh 등 (1991)은 이러한 상호작용을 다음과 같이 조건부 확률을 이용하여 식으로 표현하였다.

P(신원확인 성공) = P(신원확인 성공 | 확인 시도) × P(확인 시도)
P(신원확인 성공 | 확인 시도하지 않음) = 0

신원확인을 시도할 확률과 신원확인을 시도한 경우 신원확인에 성공할 확률을 통제함으로써 실제 노출의 위험을 제한할 수 있다.

1.2.1 노출위험의 요소

한 명 또는 그 이상의 모집단 개체에 관한 부분적 정보를 가진 자료유출 시도자는 그 개체에 관하여 좀 더 많은 정보에 관하여 추론을 하고 싶어 한다. 이런 경우 자료유출 시도자는 부분적 정보를 목적 데이터 셋에 있는 항목들과 연결하고자 할 것이다.

1.2.1.1 마이크로 자료

마이크로 자료는 정보제공기관이 가지고 있는 원시자료이다. 즉, 개인별(또는 가구별)로 습득된 자료이다. 마이크로 자료는 일반적으로 개인, 가구, 또는 기관 등의 자료를 표본조사나 센서스를 통하여 습득한다. 하지만 마이크로 자료는 다양한 방

법으로 수집할 수 있다. 예를 들어, 국가수준에서 습득된 경제 마이크로 자료는 국민계정의 보고로부터 수집하며 병원의 입원환자 자료는 병원의 의무기록으로부터 수집한다. 마이크로 자료에 관한 중요한 점은 데이터 셋에 있는 각 모집단 개체에 대하여 다수의 속성 (attributes) 값의 레코드가 있다는 것이다.

전체 모집단에 대하여 수집된 모든 변수를 포함하고 있는 마이크로 자료 셋을 완전한 표(full table)이라고 한다. 이는 마이크로 자료가 빈도표로 전환될 수 있음을 시사한다. 정보제공기관은 완전한 표를 제공하지 않는다. 대신, 완전한 표의 주변표(marginal table)인 부분자료 또는 마이크로자료의 표본을 제공한다. 만일 마이크로자료의 표본을 제공하는 경우에 일반적으로 변수의 수 및 그 변수의 정보의 수준을 줄인다. 예를 들면, 시간이나 공간에 관한 자세한 정보를 제한한다. 제공하는 자료를 제한하는 것은 전형적으로 통계적 노출 제한의 초기 단계이다.

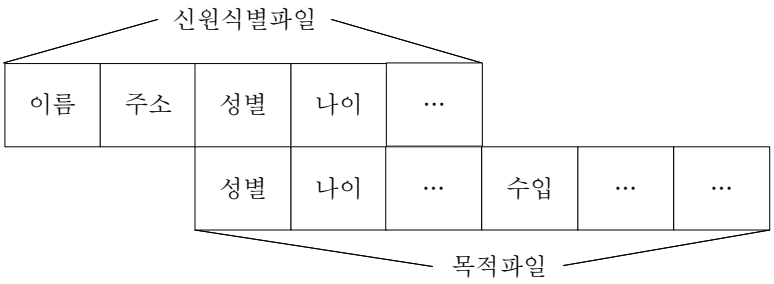
마이크로 자료의 경우 노출과 관련되어 신원파악(identification)과 속성파악(attribution)이라는 두 개의 개념이 있다. 신원파악은 알려진 모집단 개체와 특정한 마이크로자료 레코드와의 관계를 의미한다. 속성파악은 마이크로자료 셋과 특정한 모집단 개체와의 관계를 의미한다. 자료유출 시도자는 모집단 개체에 관한 속성을 알기 위해서 그 개체의 신원을 알아야 한다. 속성파악은 그 자체가 노출을 의미한다고 할 수 있다. 하지만 신원파악과 속성파악은 독립적으로 일어날 수도 있다. 만일 자료유출 시도자가 하나의 모집단 개체에 대하여 마이크로 자료 셋에 포함된 모든 정보를 아는 경우에는 속성파악 없이 신원파악이 가능할 수 있다. 반대로 주어진 변수 값들에서 다른 속성을 갖는 둘 이상의 모집단 개체들에게서 각 모집단 개체의 추가적인 정보의 속성파악은 직접적인 신원파악 없이도 가능할 수 있다. 예를 들면, 하나의 데이터 셋이 수입과 직업에 관한 정보를 포함하고 있는 경우 같은 직업을 갖는 모든 개인들이 개별적인 수입을 갖는 경우에 직접적인 신

원파악 없이도 속성파악이 가능하다. 위와 같은 예외의 경우도 있으나 일반적으로 신원파악은 자동적으로 속성파악이 가능하다.

1.2.1.2 고의적 결합 (Deliberate Linkage)

고의적 결합(deliberate linkage) 또는 매칭(matching)는 전형적인 노출의 형식이다. 고의결합의 전제조건은 자료유출 시도자가 모집단 개체의 이름과 주소와 같은 직접적인 신원 식별요인들을 포함하고 있는 데이터 셋 및 목표 데이터 셋에 포함되어 있는 식별변수들 (key variables)에 접근할 수 있어야 한다. 이 식별변수들은 식별요인과 목표 데이터 셋과 연결시키는 데 사용된다<그림 1.1>.

<그림 1.1> 식별변수의 매칭을 통한 신원노출의 예



1.2.1.3 합계자료 (Aggregate Data)

합계자료는 일반적으로 빈도 및 상대빈도를 포함한 표로 제공된다. 물론, 평균과 합 등의 통계량을 포함하는 교차표의 형태도 합계자료이다. 이러한 통계표의 특징은 주변합을 제공하므로 통계표 내의 값과 주변합과 결합할 수 있다. 예를 들면,

성별의 변수에서 남자의 빈도와 여자의 빈도를 합치면 주변합을 구할 수 있다. 값들 간의 이러한 결합관계는 마이크로 자료와 달리 합계 데이터를 보호하는 문제에 어려움이 있을 수 있다.

마이크로 데이터의 경우 노출과 관련되어 제거(subtraction)와 속성파악(attribution)이라는 두 개의 개념이 있다. 합계 데이터에서의 속성파악 개념은 마이크로 데이터에서의 속성파악 개념과 비슷하다. 즉, 제공된 합계 데이터에서의 정보와 특정한 모집단 개체와의 관계이다. 하지만, 마이크로 데이터에서는 신원파악이 되는 경우 자동적으로 속성파악이 되지만 합계표 데이터에서는 합계표 내의 비구조적인 0이 존재하는 경우 조건적으로 속성파악이 된다. 제거방법은 자료유출 시도자에게 알려진 합계표 데이터 내의 값 또는 변수들을 합계표로부터 제거하는 방법이다. 먼저 속성파악을 알아보기 위해 <표 1.1>과 같은 합계표 자료를 고려해 보자.

<표 1.1> 워크샵에 참석한 두 직업군과 소득수준의 교차표 1

	높음	중간	낮음	합계
교수	0	100	50	150
변호사	100	50	5	155
합계	100	150	55	305

<표 1.1>의 모집단은 한 워크샵에 참석한 모든 개인이다. 저녁만찬을 하는 동안 한 사람이 지난 1년 동안 2억 이상을 벌었다고 이야기하는 것을 자료유출 시도자가 들었다고 하자. 만일 자료유출 시도자가 <표 1.1>에 제시된 정보를 알고 있는 경우, 그 사람은 변호사임을 추론할 수 있을 것이다. 이는 양의 속성노출(positive attribution)의 예이다. 양의 속성노출이란 특정한 값과 특정한 모집단 개체와의 연

관성(association)이다. 반대로, 만일 자료유출 시도자가 몇 명의 사람들과 이야기를 나눈 후 이 들이 학계에서 참석한 사람들이란 사실을 알았다고 하면 그 들의 수입은 높지 않음을 알 수 있다. 이는 음의 속성노출(negative attribution)의 예이다. 음의 속성노출이란 특정한 모집단 개체로부터 한 변수의 특정한 값을 분리(disassociation)하는 것이다. 연관(association)과 분리(disassociation)는 같은 과정의 다른 형태이다. 합계표의 셀에 비구조적 0이 존재하면 이는 잠재적으로 노출의 위험이 있다. 이제 <표 1.2>를 고려해 보자.

<표 1.2> 워크샵에 참석한 두 직업군과 소득수준의 교차표 2

	높음	중간	낮음	합계
교수	1	100	50	151
변호사	100	50	5	155
합계	101	150	55	305

이 표의 모집단은 <표 1.1>의 모집단과는 다르다. 즉, 학계에서 참석한 한 명은 높은 소득을 올리고 있다. 이 경우 이전의 표처럼 추론을 할 수는 없다. 하지만 자료유출 시도자가 학회의 참석자이고 높은 소득을 올리며 학계에 있다면 자료유출 시도자는 아래 표의 높은 소득이며 학계의 셀 빈도인 1을 제거할 수 있다. 이렇게 1을 제거하고 나면 <표 1.2>는 이전의 <표 1.1>과 동일해 지며 따라서 이전의 노출 상황과 같아진다.

이제 논의를 좀 더 확장해 보자. 만일 자료유출 시도자가 좀 더 많은 사람의 정보를 알고 있다면 그러한 사람들의 빈도를 합계표에서 제거하고 잔여 합계표를 구성할 수 있다. 예를 들면 <표 1.1>의 자료에서 자료유출 시도자가 <표 1.3>과 같은 부분 정보를 알고 있다고 한다면 잔여 합계표는 <표 1.4>가 된다. 이렇듯 잔

여 합계표는 좀 더 많은 0을 가진 셀을 만들어 낼 것이다.

<표 1.3> 이전의 <표 1.2>에서 자료유출 시도자가 알고 있는 정보를 표시한 교차표

	높음	중간	낮음	합계
교수	0	80	25	105
변호사	70	20	5	95
합계	70	100	30	200

<표 1.4> <표 1.2>에서 <표 1.3>을 제거한 후의 잔여 빈도를 표시한 교차표

	높음	중간	낮음	합계
교수	1	20	25	46
변호사	30	30	0	65
합계	31	50	30	111

사실 셀의 빈도가 0이 아니더라도 일반적으로 셀의 빈도가 작으면 노출의 위험 또한 높아진다. 빈도가 높은 경우 보다 작은 경우 외부로부터 정보를 얻어 셀의 빈도를 제거하여 제로 셀을 만들기 쉽다. 0셀로 못 만들더라도 셀의 빈도가 낮은 경우 높은 수준의 신뢰정도를 가지고 정보를 노출시킬 수 있다.

이러한 제거방법이 합, 평균 등을 제시하는 요약표에 미치는 영향을 이해하는 것도 중요하다. 예를 들어 <표 1.5>는 서울, 경기, 그 외 지역의 연간 대학병원 매출액 총계를 정리한 요약표라고 하자. 물론 이 총액은 설명의 목적으로 만들어낸 가상의 값이다.

<표 1.5> 서울, 경기, 그 외 지역의 연간 대학병원 매출액 총계를 정리한 요약표

(가상자료)

	연매출액 (단위: 천원)
서울 지역	10,000,000
경기 지역	70,500,000
그 외 지역	80,000,000

이제 서울 소재 고려대학교 병원의 연매출액이 90억원이고 이 사실이 알려져 있다면 90억원을 제거함으로써 고대병원을 제외한 서울소재 대학병원의 연매출액의 합은 10억원임을 알 수 있다. 이는 지배원리(principle of dominance)의 한 예이다. 고려대학교 병원이 서울지역의 셀에서 우세하기 때문에 고려대학교 병원의 연매출액을 아는 누구도 서울지역 다른 대학병원들의 연매출액에 관한 제한된 추론이 가능하다.

1.2.1.4 합계표의 결합

합계 자료에서 노출의 위험은 합계표를 결합할 때 일어날 수 있다. 만일 두 개 이상의 표에 공통으로 들어 있는 변수들이 있는 경우에 합계표를 결합할 수 있다. 모든 합계표는 마이크로 데이터에서 작성된 완전한 표(full table)의 주변표(marginal table)이다. 여기서 완전한 표는 데이터 내의 모든 변수를 이용하여 작성한 합계표를 말한다. 만일 충분한 외부의 정보를 보유하고 있다면 합계자료를 이용하는 이용자는 완전한 표를 재구성할 수도 있다. 물론 완전한 표의 재구성은 거의 불가능하지만 자료유출 시도자는 하위 합계표(subtables)를 결합하여 좀 더 큰 합계표를 재구성할 수 있고 이런 좀 더 큰 합계표는 노출의 위험정도가 더 높

아질 수 있다. <표 1.6>-<표1.9>을 이용하여 예를 들어보자. <표 1.6>-<표1.8>의 합계표는 모두 같은 가상의 모집단에서 추출한 자료를 이용하여 합계한 빈도표이다.

<표 2.6>-<표 2.9> 합계표의 결합과 노출의 위험 정도를 보기위한 가상의 합계표

<표 2.6>

	변수1	
변수2	A	B
C	3	9
D	2	2

<표 2.7>

	변수1	
변수2	A	B
C	1	10
D	4	1

<표 2.8>

	변수1	
변수2	A	B
C	8	3
D	4	1

<표 2.9>

	변수1 과 변수2			
변수3	A,C	A,D	B,C	B,D
E	0	1	8	2
F	3	1	1	0

<표 1.6>, <표 1.7>, <표 1.8> 세 개 모두 그 자체로 노출의 위험을 가지고 있지는 않다고 할 수 있다. 물론 위의 표 세 개에서 몇 몇 셀의 빈도는 작아서 위험하다고 할 수 있다. 이 세 표는 서로 오버랩되며 삼원 빈도표 (three-way table)의 주변 빈도표이기 때문에 여러 가지 방법을 이용하여 주변 빈도표로부터 가능한 모든 삼원 빈도표를 파악할 수 있다. 위의 예에서 세 개의 이원 주변 빈도표를 이용하여 오직 하나의 삼원 빈도표(<표 1.9>)를 재구성할 수 있다. 비록 세 개의 이원 주변 빈도표는 0셀을 포함하지 않더라도 재구성된 삼원 빈도표는 2 개의 0셀

을 포함하고 이는 노출의 위험을 높인다.

1.2.1.5 계층적 합계표 (Hierarchial Tables)

한 변수에 속한 값들이 다시 다른 값들로 구성된 경우 이 변수는 계층적 변수라고 한다. 이러한 계층적 변수를 이용하여 구성한 합계표를 계층적 합계표라고 한다. 계층적 변수의 예로 “지리적 위치”가 있다. 이러한 지리적 위치는 국가, 시도, 동.읍, 통.리 등으로 계층적으로 구성되어 있다. 이러한 합계표 안의 계층적 변수는 부수적인 결합을 발생시켜 자료의 정보보호를 좀 더 어렵게 할 수 있다.

1.2.1.6 익명 자료의 결합

내부의 자료를 결합할 수도 있지만 자료를 외부의 자료와 결합하여 양질의 자료를 얻고 통계적 추론을 향상시킬 수 있다. 예를 들면 국민건강영양조사의 자료와 보험심사평가원의 자료를 결합하여 좀 더 세부적인 분석을 통하여 양질의 연구결과를 도출할 수 있다. 하지만 이러한 외부자료와의 결합은 노출 위험에 영향을 미칠 수 있다.

1.2.1.7 무의식적 인식 (spontaneous recognition)

만일 어떤 개인 X의 속성 값이 일반적인 값이 아님을 알고 있다. 그런데 자료에서 그러한 속성 값을 갖는 레코드를 발견하였다고 하면 그 레코드는 X의 것이라

고 추론할 것이다. 이러한 신원파악이 무의식적으로 발생하면 이를 무의식적 인식이라고 한다. 만일 의식적으로 개인의 신원을 파악한다고 하면 이는 고의적 결합(deliberate linkage)이다. 이러한 무의식적 인식으로 추론한 결과는 틀릴 수도 있다. 즉, 같은 속성 값을 갖는 레코드가 X의 것이 아닐 수도 있다. 자료에서 그 값이 유일한 값이라고 하더라도 모집단에서는 유일한 값이 아닐 수도 있다.

1.2.2 인식된 위험과 실제 위험 (perceived and actual risk)

지금까지는 제공된 데이터와 자료유출 시도자의 정보와의 관계를 바탕으로 한 위험의 "객관적" 요소에 관하여 논의하였다. 하지만 전체적 위험에 영향을 미칠 수 있는 "주관적" 요소도 있다. 몇 가지 예를 들어보자. 목적한 자료의 인식된 민감성은 자료유출 시도자의 동기에 영향을 미칠 수도 있지만 또한 무응답의 가능성에 영향을 줄 수도 있다. 인식된 정보노출의 성공 가능성은 자료유출 시도자의 동기에 영향을 줄 수 있다. 합계자료에서 낮은 빈도의 셀에 있게 되면 낮은 빈도의 셀이 정보노출의 위험을 실제로 높이든 그렇지 않든 간에 응답자는 자료가 위험하다고 생각할 수 있다.

이러한 주관적으로 인식된 위험과 실제 위험과의 관계는 복잡하기 때문에 통계적 정보보호 방법은 객관적인 요소에 초점을 맞추어 개발된다. 하지만, 적어도 정보제공기관은 실제 위험을 통제하는 데 있어 주관적 요소의 영향을 조심스럽게 평가하여야 한다.

1.2.3 정보 노출의 시나리오

정보 노출은 어떠한 경로로 일어나는 지 이해하는 것이 정보 보호에서 매우 중요하다. 예를 들면 자료유출 시도자는 누구인지, 정보노출을 통하여 무엇을 얻고자 하는 지 등을 파악하는 것이 중요하다. 정보보호 단계에서 이러한 문제를 먼저 인식하고 대답하여야 한다. 위험의 노출을 평가하는데 있어서 고려해야 할 정보노출의 시나리오는 정보노출의 동기 (motivation), 수단(means), 기회(opportunity), 공격의 종류(types of attack), 매칭 및 식별변수 (matching/key variable), 자료변이의 영향 (effect of data divergence), 노출시도의 성공 가능성 (likelihood of success), 노출시도의 영향 (consequence of attempt), 노출시도의 가능성 (likelihood of attempt), 데이터베이스 구조에서 변이의 영향 (effect of variations in database structure) 등이 있다. 이런 시나리오들을 간략하게 살펴보자.

1.2.3.1 동기

먼저 정보노출의 동기(motivation)에 관하여 살펴보자. 이러한 정보노출의 동기는 일반적으로 근본적 이유와 목적으로 나눌 수 있다. 근본적 이유는 예를 들어 정보제공기관의 평판을 나쁘게 하기 위해서와 같이 동기에 관한 기술이다. 목적은 자료유출 시도자가 이루고자 하는 상황에 관한 자세한 기술이다. 예를 들면 개인을 매치하여 공공에게 배포하는 것이 목적이 될 수 있다.

1.2.3.2 수단

수단은 어떻게 정보노출을 할 것인가에 관한 것으로 기술(skill), 지식(knowledge), 계산능력(computational power) 등이 있다. 기술이란 정보노출을 위한 적절한 매칭 방법을 선택하고 그 결과를 해석할 수 있는 통계적/계산적 능력을 말한다. 지식은 정보노출을 위하여 자료유출 시도자가 사용할 수 있는 실제 정보를 말한다. 자료유출 시도자의 목적을 이루기 위하여 필요한 분석을 수행하기 위하여 충분한 계산능력이 필요하다. 이 계산능력은 하드웨어와 소프트웨어 모두 포함한다.

1.2.3.3 기회

자료유출 시도자가 데이터에 접근할 수 없다면 정보를 노출시킬 기회가 작아질 것이다. 만일 자료유출 시도자가 목표한 데이터가 오직 법적으로 허가받은 사람(예: 정보제공보호 동의서에 서명한 사람)에게만 선택적으로 제공된다면 자료유출 시도자가 데이터에 접근할 기회가 매우 제한될 것이다. 일반적으로 자료유출 시도자는 자료가 공공에게 제공하거나, 허가된 자료 이용자와 공모하거나, 자료를 해킹하거나 훔침으로써 목표한 데이터를 습득할 수 있다. 불법적으로 자료를 습득하고 정보를 노출시키는 것은 법적문제를 야기할 수 있다. 따라서 자료유출 시도자가 자료를 습득할 기회의 확률은 실제로는 낮을 수 있으나 만일 자료의 허가된 이용자가 많은 경우 자료제공기관은 이 확률을 거의 1로 가정하여야 한다. 즉, 어떤 개인이나 기관이 목표한 자료를 습득하고자 시도하면 그 자료에 접근할 수 있다고 가정하여야 한다.

1.2.3.4 정보노출 시도의 종류

정보노출 시도의 종류는 자료유출 시도자의 목적을 이루기 위한 방법이다. 먼저 마이크로자료에 대한 시도의 종류로는 데이터 셋을 상호 매치하는 방법, 특정한 개인을 매치하는 방법, 임의의 개인을 매치하는 방법, 특정한 개인들로 구성된 집단을 매치하는 방법, 피싱(fishing)이 있다. 매크로데이터는 제거(subtraction)를 통하여 정보노출 시도를 할 수 있다.

1.2.3.5 식별변수

모든 정보노출 시도에서 식별변수는 신원노출을 위하여 필수적이다. 식별변수란 자료유출 시도자가 목적으로 하는 데이터에 포함되어 있으면서 자료유출 시도자 또한 알고 있는 변수를 의미한다. 따라서 식별변수를 통하여 개인을 매치할 수 있다. 이상적으로 식별변수의 코딩은 매치하고자 하는 두 개의 데이터 셋에서 동일하여야 한다.

1.2.3.6 목적변수 (target variable)

자료유출 시도자는 식별변수를 이용하여 목적으로 하는 변수의 값을 알기를 원한다. 만일 자료유출 시도자가 정보를 얻고자 한다면, 목적변수의 내용은 직접적으로 관련이 있다. 하지만 단지 신원노출만이 목적이라면 목적변수의 내용은 그리 중요하지 않을 수도 있다. 하지만 많은 경우 자료제공기관은 노출의 영향에 있어 목적변수가 담고 있는 정보는 매우 중요하게 생각한다.

1.2.3.7 자료 변이의 영향

응답자가 항상 정확한 정보를 제공하는 것은 아니며 자료를 코딩하는 데 오류가 발생할 수 있고 무응답이 발생할 수 있으므로 모든 데이터 셋은 오류와 부정확성을 포함하고 있다. 또한, 자료제공기관은 자료를 습득한 날짜로부터 오랜 시간 후에 자료를 공공에게 제공하기도 한다. 따라서 분석시점에서의 개인이나 가구의 특성은 수집된 당시의 특성과 다를 수 있다. 이러한 오류 및 부정확성은 자료유출 시도자가 가지고 있는 자료에도 있을 수 있다.

데이터 내의 이러한 "잡음(noise)"을 자료 변이(data divergence)라고 한다. 데이터-데이터 변이는 두 개의 데이터 간의 차이를 의미하고 데이터-실제 변이는 데이터와 실제와의 차이를 의미한다. 이러한 오류는 일반적으로 자료 매칭을 어렵게 한다. 하지만 만일 두 자료 모두 동일하게 실제와 다른 경우, 즉 평행한 변이(parallel divergence)를 갖는 경우에는 자료 매칭에 영향을 주지 않는다. 예를 들어 많은 변수들이 건강보험공단의 자료와 보험심사평가원의 자료에 공통으로 들어 있다. 이 두 기관은 한 개인의 같은 정보를 보유하고 있을 수 있는데 만일 그 정보가 틀리다면 두 기관 다 동일하게 틀린 정보를 가지고 있을 것이다.

1.2.3.8 노출 시도의 성공 가능성

노출 시도의 성공 가능성은 노출 시도를 하여 신원을 파악할 가능성과는 다르다. 노출 시도의 성공 가능성은 자료유출 시도자가 자신의 목적을 성취하는 가능성을 말한다. 자료유출 시도자의 목적에는 신원파악만이 있는 것은 아니다. 이제 이 가능성을 노출 시도의 영향과 노출 시도의 가능성으로 나누어 보도록 하자.

※ 노출 시도의 영향

노출 시도의 영향은 자료유출 시도자의 목적과 시도의 성공/실패 여부에 달려 있다. 노출 시도의 영향은 정보노출이 일어났는지 또한 이러한 시도가 공공에게 어떠한 영향을 주는 지를 평가하여야 한다.

만일 적어도 하나의 매치가 발생하게 되면 정보는 노출될 수 있으며 따라서 목적 데이터 내의 레코드가 파악될 수 있다. 만일 매치가 되지 않는 경우 자료유출 시도자의 시도는 정보를 노출 시키지 않게 된다. 어떤 경우에 자료유출 시도자는 그의 목적이 정보를 노출하는 것이 아니라 단지 정보가 노출될 수 있다는 위험을 보이기 위하여 정보노출을 시도할 수도 있다.

개인의 정보를 노출하고자 하는 시도가 있다는 사실을 공공이 아는 경우에 이러한 시도가 성공을 하던 실패를 하던 자료정보기관에 대한 공공의 믿음에 영향을 주므로 자료정보기관에 이러한 시도는 위험한 영향을 줄 수 있다. 만일 이 시도가 성공하면 이는 명백히 자료정보기관에 부정적인 영향을 줄 것이다. 하지만 만일 시도가 실패한다면 이는 두 가지 다른 영향이 있다. 시도가 실패하였더라도 시도가 있었다는 자체가 자료정보기관에 약점이 있다는 것을 보여 주는 것이다. 다른 측면에서는 자료정보기관이 시도를 무산시켰기 때문에 이는 그 기관의 정보보호는 안전하다는 반증이 될 수도 있다.

시도가 성공하여 노출된 개인의 신원정보가 공공에게 배포된 경우는 자료제공기관에 가장 큰 부정적 영향을 주게 되며 앞으로 공공은 이 기관의 조사에 협조하지 않게 될 가능성이 높아진다.

※ 노출 시도의 가능성

정보의 보호를 위하여 식별변수가 주어진 경우 노출 시도의 가능성을 아는 것이 중요하다. 이러한 노출 시도의 확률을 수치적으로 계산하는 것은 어렵다. 이러한 확률을 계산하기 위해서는 여러 가지 요소를 고려하고 가정하여야 한다.

1.3 노출 위험의 평가

정보 노출의 위험의 평가는 자료제공기관에게 매우 중요한 일이다. 정보노출의 평가에 있어 (특히, 마이크로데이터) 유일성(uniqueness)는 매우 중요한 개념이다. 유일성 가운데 모집단 유일성이란 만일 한 개인이 모집단에서 식별변수들의 집합에서 유일한 값을 가진다면 그 개인은 모집단 유일이라고 한다. 주어진 모집단에서 모집단 유일인 개인의 비율은 모집단 유일성의 수준이라고 한다. 모집단 유일성은 단순하며 노출위험과 직관적인 관련을 가지고 있다는 장점이 있다. 만일 한 개인이 모집단 유일이고 제공된 자료에서 레코드 매칭이 된다면 매우 높은 신원노출이 일어날 확률이 있다. 하지만 모집단 유일성 방법은 모집단의 정보에 접근할 수 있어야 한다. 이러한 단점을 극복하는 대안으로 가지고 있는 정보를 이용하여 그 정보 안에서의 유일성을 이용하여 모집단 유일성에 관한 추론하는 방법을 이용한다. 이러한 방법들은 5장에서 다루도록 한다.

1.3.1 매칭/재신원 파악 실험 (matching/reidentification experiments)

자료유출 시도자가 사용한 같은 방법을 이용한 모의실험을 통하여 매칭 또는 신

원 파악의 노출위험을 평가할 수 있다. 이 방법은 유일성 통계량을 통한 이론적 값을 이용하는 것이 아니라 실제 자료를 이용하여 모의실험을 진행한다는 장점이 있다. 하지만 적용되는 자료가 달라지면 그 결과도 달라지는 단점이 있으며 모의실험의 진행이 매우 복잡하며 시간이 많이 걸린다.

1.3.2 합계 자료에서의 노출 위험 평가

마이크로데이터에서의 노출위험 평가방법은 많이 개발되었으나 합계자료에서의 노출위험 평가 방법은 아직 많지 않다. 이러한 이유 중 하나는 마이크로데이터에서 공격의 개념적 구조(신원-속성 파악)는 잘 설정되어 있으나 합계 자료에서의 그러한 개념적 구조 (제거-속성 파악)는 잘 설정되어 있지 않기 때문이다. 일반적으로 합계표 자료에서의 위험 평가는 임기응변의 대리측정값을 이용한다. 빈도표에서 많이 사용되는 노출위험 평가 측정값은 작은 빈도를 갖는 셀의 수에 바탕을 둔다. 합, 평균 등을 제시하는 요약표의 경우 p/q 규칙을 이용하여 우세한 응답자를 갖는 셀을 파악하는 방법을 이용하여 노출위험을 측정한다. 이러한 방법들은 5장에서 좀 더 자세하게 설명한다.

1.4 노출위험의 통제 (controlling the risk)

노출위험을 평가한 후 자료제공기관은 노출위험에 따라 자료의 제한 정도를 달리해야 한다. 이 절에서는 몇 가지 노출위험 통제방법에 관하여 기술한다.

1.4.1 메타데이터 통제

메타데이터 수준에서의 노출위험 통제는 제공될 자료의 전체적인 구조에서 이루어져야 한다. 이러한 통제의 주요한 요소는 표본추출의 비율, 변수의 선택, 선택된 변수의 세부적인 수준이 있다.

1.4.1.1 표본추출의 비율

표본조사에서 표본추출 비율은 표본설계에서 결정된다. 일반적으로 표본추출 비율을 선택할 때 노출의 위험을 고려하지는 않는다. 그럼에도 불구하고 표본추출의 비율은 마이크로데이터의 노출 위험을 측정하는 데 중요한 요소이다.

1.4.1.2 변수의 선택

노출의 위험을 통제 또는 제한하는 데 가장 많이 사용하는 방법은 특정 변수를 제공되는 데이터에서 제외하는 것이다. 자료제공기관은 잠재적인 자료유출 시도자가 접근할 수 있을 것 같은 변수들을 제거하거나 자료유출 시도자가 목적으로 하는 변수들을 제거할 수 있다. 이전에 설명한 시나리오 분석을 통하여 제거할 변수를 선택할 수 있다.

1.4.1.3 선택된 변수의 세부적인 수준

변수내의 정보를 어느 정도의 수준으로 제공해야 하는 것을 정하는 것은 변수를 선택하는 것과 비슷한 과정이다. 예를 들어, 매크로데이터의 노출통제를 위하여 자료제공기관은 작은 빈도를 갖는 셀들을 조사하여 다른 셀과 병합할 수 있다. 이 때 병합과정은 정보의 유용성은 보존하면서 위험노출의 가능성은 줄일 수 있도록 하여야 한다. 데이터 이용자들은 제공된 데이터 내의 정보는 최대한 자세하기를 원한다. 하지만 자료제공기관은 정보보호의 차원에서 어떤 변수내의 정보를 통제할 수밖에 없다.

1.4.2 데이터 바꾸기 (Distorting the Data)

메타데이터 통제방법의 대안으로 데이터를 바꾸는 방법이 있다. 이를 변조(perturbation)라고 한다. 위험노출 통제를 위하여 사용하는 변조방법으로 자료의 교환(data swapping), 반올림(rounding), 셀 변조(cell perturbation), 셀 감추기(cell suppression) 등의 방법이 있다. 이는 4장과 5장에서 자세하게 설명한다.

1.4.3 접근제한 (Controlling Access)

자료제공기관은 누가 또 어떤 방법으로 자료에 접근 할 수 있는 지 통제할 수 있다. 일반적으로 공공에게는 매크로데이터가 제공된다. 마이크로데이터의 접근은 일반적으로 접근허가권이 필요하다. 노출의 위험을 줄이기 위하여 자료접근 권한, 목적, 방법들은 노출제한기법과 함께 사용된다. 예를 들어, 센서스 자료를 제공할

때 CD를 이용하여 배포하는 경우에는 제공되는 자료의 세부적인 수준이 낮고 자료제공기관 내에서 허가받은 사람만이 이용할 수 있는 자료의 세부적인 수준은 매우 높을 것이다.

1.5 자료의 유용성

정보노출에 초점을 맞추어 자료를 제한하게 되면 제공되는 자료의 유용성이 낮아 잘못된 분석결과를 초래할 수 있다. 따라서 노출의 위험 평가와 더불어 유용성도 함께 평가하여야 하는데 이 유용성은 정보의 손실을 수치화하여 측정한다. 이렇게 수치를 이용하여 유용성을 평가하게 되면 데이터 간의 유용성 정도를 객관적으로 비교할 수 있다. 하지만 이 방법은 제공된 자료를 이용하여 여러 가지 방법으로 분석을 하는 경우에 실제적인 유용성을 평가하기 어렵다는 단점이 있다. 만일 자료 이용자가 그가 하는 분석에서 유효한(valid) 결과를 얻을 수만 있다면 자료제공기관이 어떤 노출제한 기법을 이용하여 자료를 통제하는 지 관심이 없을 것이다. 반대로 만일 자료의 이용자가 하려는 분석을 할 수 없는 경우에는 제공된 자료의 정보손실이 낮다고 하더라도 그 이용자에게는 쓸모없는 자료일 수 있다.

노출제한기법이 자료의 유용성에 미치는 영향은 “분석적 완비성의 감소(reduction of analytical completeness)”와 “분석적 유용성의 손실 (loss of analytical validity)”로 구분할 수 있다. 노출제한기법을 이용하여 자료를 통제하는 경우 하고자 하는 분석을 못하게 될 수 있다. 이를 분석적 완비성의 감소라고 한다. 예를 들어 마이크로 자료에서 지역적 스레시홀드를 사용하여 작은 지역들을 병합하면 이러한 작은 지역들의 사회적 영향에 관심이 있는 자료 이용자들은 제공된 자료를 효율적으로 사용할 수 없다.

만일 자료의 이용자가 같은 분석방법을 제공된 자료와 원 자료에 적용했을 때 다른 결과를 도출하는 경우 분석적 유용성의 손실이 있다고 한다. 데이터에서 하나의 변수를 제거하는 경우 자료의 질(정보의 정도)에는 영향을 줄 수 있으나 만일 자료 이용자가 제거된 변수에 관심이 없는 경우에는 (즉, 분석에 사용하지 않는 경우) 분석적 완비성에는 영향을 주지 않을 수 있다. 반대로 자료에서 아주 작은 부분을 변조하는 경우 자료의 질(정보의 정도)에는 큰 영향이 없을 수 있으나 변조가 중요한 변수를 불균형하게 만드는 경우 분석적 유용성에 큰 영향을 줄 수 있다. 6장에서 자료의 유용성에 관하여 좀 더 자세하게 살펴보도록 한다.

제 2장. 노출위험의 평가

자료를 공공에게 제공하기 이전에, 정보제공기관은 자료유출 시도자가 개인정보를 누출시키는 위험 정도를 평가하여야 한다. 원 자료는 받아들이지 못할 정도의 노출위험을 가진다. 따라서 제공되는 자료는 받아들일만한 정도의 노출위험이 되도록 변형되어야 한다. 노출위험을 낮추는 방법은 3장과 4장에서 소개하고 이 장에서는 노출의 위험을 이해하고 노출의 위험을 평가하는 적절한 방법에 관하여 소개한다.

노출의 위험을 평가하기 위해서, 정보제공기관은 다음과 같은 세 가지 요소를 인지하여야 한다. (1) 자료유출 시도자의 행동에 따른 여러 가지 가능한 결과 (2) 그러한 결과의 비효율성 (3) 그러한 결과가 나올 가능성

여러 가지 복잡함으로 인하여 정보제공기관은 위의 세 가지 요소를 완벽하게 알 수는 없다. 따라서 정보제공기관은 노출과 관련된 비효율성은 높다고 가정하는 동시에 자료유출 시도자의 부분은 단지 중간정도의 정보를 가지고 있다고 가정한다. 많은 정보제공기관들은 어디서 공격이 일어나며 자료유출 시도자가 가지고 있는 수단은 무엇인 지에 관한 개념적 모형을 알아보기 위하여 시나리오 분석방법을 이용한다. 이 시나리오 분석에서 정보제공기관은 자료유출 시도자가 자료를 노출시키는데 이용할만한 식별변수들 (key variables)을 지정한다.

자료유출 시도자의 공격이 있는 경우 그 공격이 성공할 가능성을 측정할 수 있는 측정값이 있으면 노출조절에 매우 유용할 것이다. 시나리오 분석은 성공할 만한 공격과 성공하지 못할 공격을 추려내는 과정이다. Marsh 등 (1991)은 다음과 같은 공식을 제시하였다.

$$P(\text{신원노출}) = P(\text{신원노출} \mid \text{공격 시도}) \times P(\text{공격 시도})$$

여기서 공격 시도는 성공할 것 같은 경우는 1, 성공하지 못할 것 같은 경우는 0인 이분형 변수로 간주하고 위험을 평가하는 몇 가지 방법에 관하여 살펴본다.

2.1 임계값과 다른 대리 측정값 (Thresholds and Other Proxies)

노출위험의 측정은 종종 모집단의 임계값을 이용한다. 임계값 검정을 실시하여 위험이 받아들이지 못할 정도로 높은지 아니면 받아들일 수 있을 정도로 낮은지 판단할 수 있다. 예를 들면, 노출위험 측정값의 종류와 제공되는 자료의 보호정도에 따라 하나의 임계값(τ)를 정한다. 만일 위험 측정값이 τ보다 작으면 자료는 제공되고 만일 위험 측정값이 τ보다 크면 좀 더 노출제한을 하여야 한다. 받아들일 만한 위험의 수준은 어떻게 또 누구에게 자료를 제공하느냐에 따라 달라질 수 있다. 이러한 임계값 규칙은 이해하기 쉽고 또 실제로 실행하기 쉽다는 장점이 있다. 하지만 이 방법은 자료유출 시도자가 실제로 자료를 노출시키는 방법과 상관없을 수 있는 단점이 있다. 따라서 좀 더 정교한 측정값이 필요할 수 있다. 다음은 마이크로 데이터에서의 위험 평가에 관하여 살펴보도록 한다.

2.2 마이크로 데이터에서의 위험평가

이전에 언급한대로 마이크로 자료에서는 공공에게 제공하는 파일의 레코드와 자료유출 시도자가 가지고 있는 파일의 레코드를 자료유출 시도자가 매칭하는 과정

을 통하여 노출의 위험이 발생할 수 있다. 따라서 마이크로 자료의 노출위험을 측정하는 기준은 어느 정도까지 매칭이 가능한가에 달려있다. 또한 노출위험의 측정은 개인수준과 전체적인 수준에서 할 수 있다.

2.2.1 파일 수준의 노출위험 측정 (File-level risk metrics)

파일 수준의 노출위험 측정은 전체 데이터 파일의 평균 위험을 측정하는 것이다. 이 방법은 식별변수를 범주화하는 것이 좋은 지 결정하는 데 도움이 되는 방법이다. 대부분의 파일 수준의 노출위험 측정은 모집단 유일성의 개념에 바탕을 두고 있다.

유일성은 마이크로 데이터에서 노출위험을 이해하는데 중요한 개념이다. 유일성 가운데 모집단 유일성이란 만일 한 개인이 모집단에서 식별변수들의 집합에서 유일한 값을 가진다면 그 개인은 모집단 유일이라고 한다. 주어진 모집단에서 모집단 유일인 개인의 비율은 모집단 유일성의 수준이라고 한다. 모집단 유일성은 단순하며 노출위험과 직관적인 관련을 가지고 있다는 장점이 있다. 만일 한 개인이 모집단 유일이고 제공된 자료에서 레코드 매칭이 된다면 매우 높은 신원노출이 일어날 확률이 있다. 하지만 모집단 유일성 방법은 모집단의 정보에 접근할 수 있어야 한다. 이러한 단점을 극복하는 대안으로 가지고 있는 정보를 이용하여 그 정보 안에서의 유일성을 이용하여 모집단 유일성에 관한 추론하는 방법을 이용한다.

Skinner와 Elliot (2002)은 유일한 매치가 주어진 경우 정확한 매치의 확률에 초점을 맞추는 방법을 제시하였다. 마이크로 데이터의 경우 자료유출 시도자의 공격은 다음 세단계의 과정을 통해 모방할 수 있다 (Elliot, 2000). 먼저 하나의 데이터 셋

에서 레코드를 제거한다. 원래의 표본 비율(sampling fraction)과 동일한 확률로 제거된 레코드를 조건부로 대체한다. 제거된 레코드를 식별변수를 이용하여 그 데이터 셋에 대하여 매칭한다. 이 과정은 6가지의 가능한 결과가 발생할 수 있다 <표 2.1>.

<표 2.1> 마이크로데이터에서 하나의 레코드를 제거한 후 다시 재표본 추출하는 경우에 가능한 결과

레코드	재표본 추출 됨	재표본 추출되지 않음
표본 유일	**올바른 유일 매치**	매치되지 않음
두 개의 표본짝 중 하나	올바른 매치를 포함한 다중 매치	**잘못된 유일 매치**
두 개 이상의 표본으로 구성 된 동등집단 중 하나	올바른 매치를 포함한 다중 매치	잘못된 다중 매치

<표 2.1>에서 6가지 경우 중에 굵은 글씨로 표시된 부분이 중요하다. 첫 번째는 제거된 레코드가 식별변수의 셋에 대하여 표본 유일이며 그 레코드가 데이터 셋으로 다시 표본 추출되는 경우이며 이는 올바르게 유일 매치가 된 경우이다. 두 번째는 식별변수 셋에 대하여 같은 정보를 갖는 표본이 둘이며 제거된 레코드가 데이터 셋으로 다시 표본 추출되지 않는 경우이며 이는 잘못된 유일 매치이다. 이를 이용하여 유일 매치인 경우의 올바른 매치일 조건부 확률을 아래와 같이 구할 수 있다.

$$P(\text{올바른 매치} \mid \text{유일 매치}) = \frac{\sum_j I(f_j = 1)\pi}{\sum_j I(f_j = 1)\pi + \sum_j I(f_j = 2)(1-\pi)}$$

여기서 f_j는 j번째 레코드에 대하여 식별변수 셋에 대하여 같은 정보를 갖는 표본의 수로 표본 유일인 경우는 1이고 식별변수 셋에 대하여 같은 정보를 갖는 표본이 둘이면 2이다. 유일매치가 되기 위해서는 f_j값은 1과 2만 가질 수 있다. I(.)는 지시변수이고 π는 표본 비율이다.

이 방법은 모집단 데이터를 필요로 하지 않으며 자료유출 시도자가 할 수 있는 공격을 모방한다는 장점이 있다. 하지만 이 방법은 자료유출 시도자의 파일과 목적 파일 간에 차이가 있을 수 있다는 점을 고려하지 않았기 때문에 위의 조건부 확률은 항상 과대 추정된다. 또한, 자료유출 시도자가 한 개체를 모집단으로부터 무작위로 표본 추출한다는 강한 가정을 가지고 있다.

2.2.2 레코드 수준의 노출위험 측정 (Record-level Risk Metrics)

식별변수들의 값이 모집단에서 드문 경우를 갖는 레코드는 노출위험이 높다. 하지만 표본에서 식별변수들의 값이 드물거나 유일한 경우 항상 노출위험이 높은 것은 아니다. 이를 보기 위해 식별변수는 범주형이라고 가정하고 이 변수들의 교차표를 고려해 보자. 교차표 안의 각 셀은 식별변수들의 범주의 교차곱(cross-product)이다. F_k를 모집단에서 셀 k에 속하는 개체 수라고 하고 f_k는 셀 k에 속하는 표본수라고 하자. 이제 $1/F_k$는 셀 k에 속하는 한 개체가 다시 신원노출이 되는 확률이라고 정의하자. 이 때 모집단 빈도인 F_k는 표본 빈도인 f_k로부

터 추정해야 한다. 이러한 세팅에서 두 가지 방법으로 노출위험을 평가할 수 있다. 하나는 $F_k|f_k$는 포아송 분포를 따른다고 가정하는 포아송 모형을 이용하는 방법이고 (Skinner와 Holmes, 1998; Elamir과 Skinner, 2006), 다른 하나는 $F_k|f_k$는 기하분포(negative binomial distribution)를 따른다고 가정하는 Argus 모형을 이용하는 방법이다 (Benedetti 등, 2003; Polettini와 Stander, 2004). 두 가지 방법 모두 개인의 노출위험을 측정하고 이를 합쳐서 전체 데이터에 대한 포괄적인 노출위험 측정값을 구할 수 있다. 포괄적 노출 위험은 $E[1/F_k|f_k]$로 정의된다.

자료유출 시도자가 정보를 노출시키고자 하는 특정한 모집단 개체들의 셋을 “목표 집단”이라고 한다. 만일 자료유출 시도자가 목표 집단의 식별변수에 관한 사전정보를 가지고 있다면 주요 속성에 관한 값을 이용하여 사전정보와 공개된 마이크로 데이터를 연결을 시도할 것이다. 자료유출 시도자가 범주형인 식별변수 X에 관하여 마이크로 데이터의 레코드 r을 목표와 매치한다고 가정하자. F_i는 $X=i$에 속하는 모집단 개체 수이고 $i(r)$은 레코드 r에 대한 X의 값이라고 하자. 만일 $F_{i(r)}$이 알려져 있다면 정확한 연결의 확률을 $1/F_{i(r)}$로 추정할 수 있고 만일 $F_{i(r)}=1$이면 연결의 확률은 1이다. 하지만 일반적으로 마이크로 데이터 셋은 표본이기 때문에 자료유출 시도자는 $F_{i(r)}$의 실제값 을 모른다. 하지만 초 모집단 모형 (super-population model)을 고려함으로써 자료유출 시도자는 확률 $P(F_i=j)$를 셀 빈도에 적용할 수 있다.

2.3 합계자료에서의 노출위험 측정

마이크로 자료에 비하여 합계 자료에서의 노출 위험 측정 방법은 많이 개발되어

있지 않다. 가장 간단한 방법은 문제가 될 만한 낮은 빈도를 고려하여 임계값을 정하는 것이다. 좀 더 이론적인 방법은 노출 위험의 필요충분조건으로 모집단의 교차표 안에 하나 이상의 0셀이 있어야 한다는 주장에 근거한다. 만일 한 개인이 어떤 모집단에 있고 변수들의 셀에 근거하여 구성한 그 모집단의 교차표에서 하나의 0셀이 있는 경우 그 모집단 개인은 0셀에 관련한 정보는 가지고 있지 않다. 일반적으로 모집단 교차표들의 집합을 이용하여 개별 교차표에 관한 경계를 결정할 수 있다. 이는 제공된 교차표에 있는 모든 변수로 구성된 완전한 교차표 (이를 "기본" 교차표라고 한다.)를 고려할 수 있다. 기본 교차표 내 변수의 상위집합으로 이루어진 모든 교차표가 0셀을 포함하는 필요충분조건은 기본 교차표가 0셀을 포함하는 경우이다.

Smith와 Elliot (2008)은 모집단의 완전한 교차표에 0셀이 존재한다는 가정 하에 위험노출을 측정하는 방법을 제안하였다. 이 측정치는 "무작위로 추출된 n개의 모집단 개체를 제거한 경우 완전한 교차표 내에 하나 이상의 0셀을 포함하는 확률"로 정의된다. Smith와 Elliot (2008)은 이 확률을 제거 속성노출 확률(subtraction attribution probability: SAP)이라고 하였다. 이는 다시 말해 교차표 내의 변수 값이 알려진 모든 신원노출이 된 개체를 제거한 후 모집단 교차표 안에 0셀을 포함할 확률이다. 셀 빈도가 $c_i, i = 1,...,m$ 인 기본 교차표를 고려해 보자. 각 셀이 독립적으로 변형된 주변 빈도표가 제공된 경우 셀 안의 빈도를 x라고 하면 실제 셀 빈도 c는 하한 값 l과 상한 값 u 사이에 존재하게 된다. 즉, $l \leq c \leq u$이다. 이런 경우 모집단의 알려진 표본을 제거함으로써 빈도표가 0셀을 포함하게 될 필요충분조건은 알려진 표본이 $s_i = c_i = u_i'$인 경우이다. 여기서 s_i는 알려진 표본의 빈도이고 u'은 u_i의 집합이다. 예를 들어, 실제 빈도가 0-2인 경우 제공되는 빈도는 0이고 실제 빈도가 3-7인 경우에 실제 빈도는 5이고 실제 빈도가 8-12인 경우에 제공되는 빈도는 10인 규칙을 가진다고 하자. 만일 자료유출 시도자가 제공된 빈

도표의 한 셀에서 5라는 값을 관찰하게 되면 그 셀의 실제 빈도는 3-7인 것을 알 것이다. 그 셀에서의 최대값은 7이므로 제거에 의하여 0셀을 만들기 위해서는 이 셀의 실제 빈도가 7이어야 한다. 즉, 제거를 이용하여 이 셀의 빈도를 0으로 만들기 위해서는 $7(s_i) = 7(c_i) = 7(u_i)$이다. 이를 일반화하면 값이 알려진 개체의 빈도를 안다고 할 때 자료유출 시도자가 적어도 하나의 0셀을 만들어 낼 확률은 다음과 같다.

이 식에서 S는 모든 가능한 표본 빈도표의 집단이고 p는 모집단의 빈도표이다. 또, n은 비복원 단순임의 추출된 표본의 수이다. 이 방법의 장점은 제공된 표가 변형이 되었든 되지 않았든, 하나의 표가 제공되는 여러 개의 표가 제공되든 같은 유효성을 가지고 적용할 수 있다. 하지만 만일 자료가 방대한 경우에 이 방법을 이용하여 노출위험을 측정하려면 계산이 복잡하고 오래 거리는 단점이 있다 (Smith와 Elliot, 2005).

2.4 민감도

지금까지 논의한 방법들은 한 모집단 개체의 속성이 노출되었다면 노출된 정보는 매우 중요하다는 가정을 함축적으로 하였다. 하지만 노출위험의 측정 방법에서 자료유출 시도자가 노리는 정보의 민감도 또한 노축위험 측정에서 중요하다. 노출된 정보의 중요성에 영향을 미치는 요소는 다음과 같다. (1) 얼마나 쉽게 노출된 정보를 다른 방법으로 획득할 수 있나? (2) 노출된 정보가 얼마나 큰 손실을 노출된 개체에게 입힐 수 있나? (3) 또한 노출된 정보가 공공에게 입히는 영향은 무엇인가?

물론 이 세 가지는 서로 관련이 되어있다. 예를 들어 어떤 특정한 사람의 자료가 데이터 셀에 있고 그 사람이 서울의 아파트에 산다는 정보가 노출되었다고 가정하면 이 노출은 개인에게나 공공에게 그리 큰 영향을 주지 않을 것이다. 노출된 개인은 다른 사람들이 저자가 아파트에 사는 것을 아는 것에 대해 무관심할 가능성이 높으며 (실제로 무관심하다) 다른 사람들(공공)도 노출된 개인이 아파트에 사는 것에 대한 관심이 없을 것이다. 이러한 노출정보는 민감하지 않다고 할 수 있다. 하지만 이러한 정보의 민감성은 모집단에 따라 달라질 수 있다. 만일 특정한 사람을 스토킹 하고자 하는 스토커가 있다면 이 스토커에게는 그 특정한 사람이 아파트에 사는지 주택에 사는지의 정보가 중요할 수도 있다. 다른 예로는, 어떤 사람들은 자신의 수입이 노출되는 것에 둔감할 수 있고 다른 사람들은 자신의 수입이 노출되는 것에 매우 민감할 수 있다. 일반적으로 어떤 변수의 민감성은 다른 변수에 달려 있는 경우가 많다. 물론 국내와는 다를 수 있겠지만 영국국민에게서 조사한 결과에 따르면 다음과 같은 정보들이 일반적으로 민감하다: 상세한 개인의 접촉정보, 재정정보, 인종에 관한 정보, 범죄기록, 의무정보, 정치성향, 정치집단의 회원여부, 본인이 이용한 웹사이트 경로정보, 종교, 유전적 정보, 성생활 정보, 학력, 고용기록, 노동조합 가입여부 (McCullagh, 2007). 이러한 개인에게 민감한 정보는 연구자들에게 관심 있는 정보이기도 한다.

McCullagh의 분석에 포함되지 않은 점은 민감성이 변수에 따라 다를 수 있지만 또한 그 변수에 대해 조사된 값에 따라 달라질 수 도 있다는 것이다. 예를 들어, 에이즈에 감염되지 않은 사람에게는 에이즈 감염 여부라는 변수가 민감하지 않겠지만 에이즈에 감염된 사람에게는 이 변수가 매우 민감하다. 결국, 민감성이란 사회적, 심리학적, 정치적, 윤리적, 법적인 문제들이 얽힌 매우 복잡한 문제이다. 정보제공 기관에게 중요한 문제는 모집단에서 어느 정도의 비율이 제공되는 정보가 민감하지 않도록 느끼도록 정보를 보호해야 하는지 정하는 것이다. 물론 이는 매

우 어려운 주제이다. 이에 관한 논의는 본 교재의 범위를 넘으므로 건너 띄도록 하겠다.

제 3장. 매크로자료 정보보호 기법

<학습목표>

(1) 매크로 자료의 정보보호 개념에 관하여 설명한다.

(2) 매크로 자료의 정보보호와 실시하는 방법을 단계별로 설명하고 관련 용어들을 정의한다.

(3) 매크로 자료의 정보보호 기법들을 이해한다.

3.1 매크로자료의 정보보호 개념

매크로 자료의 정보보호는 마이크로 자료가 표 (table) 형태로 합계되어 발간된 경우 합계 자료(aggregated data)에서 발생하는 정보유출(information disclosure)의 문제점을 논의하고 이를 막기 위한 정보보호 개념을 다룬다. 이 방법은 표 형태 자료의 정보보호 개념을 다룬다는 의미로 “표 자료(tabular data)의 정보보호”라고도 부른다(Duncan et. al., 2011). 자료로부터 표를 구성할 때 범주 각 셀(cell)의 값 뿐 아니라 주변합(marginal sum)도 보고하는 것이 일반적인데 주변합 정보를 이용하여 정보가 유출되는 경우가 발생하므로 이를 막기 위한 노력을 경주해야 한다는 점에서 마이크로 자료의 정보보호보다 복잡한 문제를 안게 된다.

매크로 자료의 정보 유출은 표의 일부 셀의 값이 비구조적으로 0(nonstructural zero)인 경우 발생하게 된다. 예를 들어 <표 3.1>은 사회경제수준(socioeconomic status)을 “상”, “중”, 그리고 “하”로 구분한 후 A, B 두 지역의 사회경제수준의 분포를 표로 나타낸 가장 자료이다. B 지역의 경우 사회경제수준이 “상”인 가구가 존재하지 않으므로 만약 특정 가구가 사회경제수준이 “상”이라고 응답한 경우 이

가구가 A 지역에 속한다는 것을 알 수 있게 되고 이는 이 가구의 지역정보 유출이 발생하는 결과가 된다. 마찬가지로, 만약 우리가 B 지역에 속한 한 가구원을 안다면 이 가구의 사회경제수준은 “상”이 아니라는 것을 추론할 수 있으므로 이 가구의 사회경제수준 정보가 유출되는 것이다.

<표 3.1> 두 지역의 사회경제 수준별 가구 숫자

		사회경제수준			
		상	중	하	합계
지역	A	65	30	5	100
	B	0	40	60	100
	합계	65	70	65	200

문제는 표의 셀 값이 0이 아닌 경우에도 정보의 유출이 발생할 수 있다는 것이다. 예를 들어, <표 3.1>에 의하면 지역 A에는 사회경제수준이 “하”인 5개 가구가 존재한다. 연구자 K가 A 지역에서 조사를 실시하였고 사회경제수준이 “하”인 5개 가구에 관한 자료를 가지고 있다면 <표 3.1>에서 본인의 자료 정보를 빼고 <표 3.>을 재구성할 수 있다. 이렇게 재구성된 자료 <표 3.2>에는 지역 A의 사회경제수준인 “하”인 가구가 존재하지 않게 되어 나머지 가구의 사회경제수준은 “하”가 아니라는 정보가 유출되게 된다. 이와 같이 자료유출 시도자의 자료에 근거하여 원 자료의 정보 일부를 제외시키는 경우를 자료빼기(subtraction)라 부른다. 여러 자료유출 시도자들은 본인들이 가지고 있는 정보가 서로 다르므로 각각 가지고 있는 자료를 바탕으로 자료빼기를 시행하게 되고 이는 다양한 형태로 감해진 자료들을 형성하게 된다. 이 때, 자료빼기로 인한 정보 유출은 이와 같이 재형성된 자료 중 한 개의 자료에서라도 비구조적 0이 존재한다면 발생하게 된다는 점에서 복잡해진다.

정보유출은 비구조적 0이 존재한다면 100% 발생하게 되며 비구조적 0은 존재하지 않더라고 0에 가까운 작은 값을 지닌 셀이 존재한다면 발생할 확률이 높아지게 된다. 이는 0에 가까운 작은 값을 가진 셀은 정보 유출 시도자의 자료빼기를 통해 비구조적 0의 값을 지닌 셀로 바뀔 수 있기 때문이다.

<표 3.2> 연구자 K의 정보를 <표 3.1>에서 제외시킨 후 두 지역의 사회경제수준별 가구 숫자

		사회경제수준			
		상	중	하	합계
지역	A	65	30	0	95
	B	0	40	60	100
	합계	65	70	60	195

더 심하게는 0에 가까운 셀이 아닌 다른 셀의 정보에서도 발생할 수 있다는 점이다. 예를 들어, 연구자가 지역 A의 한 구획에서 60가구에 관한 연구를 진행하였고 조사된 60가구 모두에서 사회경제 수준이 “상”이었다면 이 지역의 나머지 구획의 경우 사회경제수준이 “상”인 가구가 5가구밖에 없어 상대적으로 적을 것이라는 정보가 유출된다. 이와 같이 특정 집단이 특정 셀의 정보 대부분을 지배하는 경우를 지배원리(principle of dominance)라 하고 지배원리가 존재하는 경우 정보유출의 가능성은 높아지게 된다.

매크로 자료의 정보유출 가능성은 표를 연결하는 게 가능하다면 증가하게 된다 (Chowdhury, et. al., 1999). <표 3.3>부터 <표 3.5>는 성별, 지역에 따른 소득수준을 이원분할표를 사용하여 나타낸 결과이다. 비구조적 0을 지니거나 0에 매우 가까운 셀은 없으므로 정보유출이 발생할 위험성이 적어 보인다. 하지만 이 세 개의 표의 정보를 재정리하면 <표 3.6>을 구성할 수 있고 이 표에서는 비구조적 0

의 값을 지니는 셀이 생기게 된다. 즉, “A 지역 여성” 중 소득이 “상”인 경우가 존재하지 않고 “B 지역 남성” 중에는 소득이 “하”인 경우가 존재하지 않아 정보 유출이 발생하게 된다.

<표 3.3> 성별과 지역 분포의 이원분할표

		성별		
		남	여	합계
지역	A	10	8	18
	B	3	9	12
	합계	13	17	30

<표 3.4> 성별과 소득의 이원분할표

		성별		
		남	여	합계
소득	상	7	12	19
	하	6	5	11
	합계	13	17	30

<표 3.5> 지역과 소득의 이원분할표

		지역		
		A	B	합계
소득	상	15	4	19
	하	3	8	11
	합계	18	12	30

<표 3.6> 성별, 지역 조합과 소득의 이원분할표

		성별 및 지역				
		A지역 남성	A지역 여성	B지역 남성	B지역 여성	합계
소득	상	7	0	8	4	19
	하	3	3	0	5	11
	합계	10	3	8	9	30

지금까지는 빈도표(frequency table)에서 발생하는 정보유출에 관하여 논하였는데 연속형 자료의 합이나 평균 등으로 구성된 표에 대하여 적용도 가능하다. 예를 들어, <표 3.7>은 X, Y, Z 세 가지 제품에 대한 A, B, C 세 지역에서의 판매총액을 나타낸 가상자료이다.

<표 3.7> 세 지역에서 판매된 세 가지 제품의 제품별 판매총액 (단위: 천만원)

		지역			
		A	B	C	합계
제품	X	20	50	10	80
	Y	8	19	22	49
	Z	17	32	12	61
	합계	45	101	44	190

각 셀의 값은 0보다 크지만 특정 회사 S가 제품 Y를 지역 C에서 2억 어치를 판매한 사실을 알고 있는 자료유출 시도자는 지역 C에서 다른 회사들의 제품 판매량이 상대적으로 매우 적다는 사실을 알 수 있게 되고 지배원리가 적용되어 이 셀의 값은 정보유출 위험이 높아지게 된다.

3.2 매크로자료 정보보호의 단계

표에 0의 값을 지니는 셀이 존재한다면 정보유출은 발생하게 되지만 0의 값을 지니는 값이 존재하지 않더라도 0에 가까운 값을 지니는 셀이 존재하면 정보유출 위험이 높아지고 0과 크게 다른 값을 지니는 셀에서조차 자료빼기에 의해 정보유출이 발생할 수 있으므로 우선 어느 셀이 정보유출 발생 가능성이 높은 위험한 셀(risky cell)인지 확인해야 하며 이와 같은 셀이 발견된다면 정보유출을 막기 위해 적절한 조치가 취해져야 한다. 이를 논의하기 위하여 우선 표 자료의 구조적 특성을 논의하고 이를 선형연립방적식으로 표현하는 방법을 설명한다.

3.2.1 표 자료의 구조

표 자료에는 마이크로자료 자료와 달리 행에 대한 합(row sum) 및 열에 대한 합(column sum), 그리고 총합(total sum)의 주변합 정보가 같이 제공되는 게 일반적이다. 따라서 표 자료는 각 셀들의 행에 대한 합의 정보, 열에 대한 합의 정보들로 구성된 선형연합방정식(a linear system of equations)을 만족해야 한다. 이를 위하여 표 자료의 각 셀의 값, 행의 합, 열의 합, 그리고 총합을 원소로 지니는 배열 y를 고려하자. 즉, K개의 범주를 가진 변수와 L개의 범주를 가진 변수의 이원분할표의 배열은 $n=(K+1)\times(L+1)$개의 원소로 구성된다. 예를 들어 <표 3.8>에 나타난 가상의 자료에 대한 배열은 $y=[y_1,y_2,\cdots,y_{16}]$으로 표현할 수 있고 이때 $n=(3+1)\times(3+1)=16$이 된다.

<표 3.8> 가상의 자료

	A	B	C	합계
X	y_1	y_2	y_3	y_4
Y	y_5	y_6	y_7	y_8
Z	y_9	y_{10}	y_{11}	y_{12}
합계	y_{13}	y_{14}	y_{15}	y_{16}

이 원소들의 일부는 행의 합이므로 이 합들은 각 셀들의 합으로 표현될 수 있으므로 다음의 식들을 만족해야 한다.

$$y_1 + y_2 + y_3 - y_4 = 0;$$
$$y_5 + y_6 + y_7 - y_8 = 0;$$
$$y_9 + y_{10} + y_{11} - y_{12} = 0.$$

마찬가지로 일부는 열의 합이므로 다음의 식들을 만족해야 한다.

$$y_1 + y_5 + y_9 - y_{13} = 0;$$
$$y_2 + y_6 + y_{10} - y_{14} = 0;$$
$$y_3 + y_7 + y_{11} - y_{15} = 0.$$

또한, 행의 합들과 열의 합들의 합은 총합이므로 다음의 식들을 만족한다.

$$y_4 + y_8 + y_{12} - y_{16} = 0;$$
$$y_{13} + y_{14} + y_{15} - y_{16} = 0.$$

표는 위의 식들을 모두 만족해야 하므로 이를 다음의 8개의 선형연립방정식으로 나타낼 수 있다.

$$y_1 + y_2 + y_3 - y_4 = 0;$$

$$y_5 + y_6 + y_7 - y_8 = 0;$$

$$y_9 + y_{10} + y_{11} - y_{12} = 0;$$

$$y_1 + y_5 + y_9 - y_{13} = 0;$$

$$y_2 + y_6 + y_{10} - y_{14} = 0;$$

$$y_3 + y_7 + y_{11} - y_{15} = 0;$$

$$y_4 + y_8 + y_{12} - y_{16} = 0;$$

$$y_{13} + y_{14} + y_{15} - y_{16} = 0.$$

$I = \{1,2,\cdots,n\}$, $J = \{1,2,\cdots,m\}$에 대하여 위의 연립방정식을 식으로 나타내면

$$\sum_{i \in I} m_{ij} y_i = b_j \quad \text{for } j \in J$$

와 같이 표현할 수 있다. 위 자료에서

$n = 16,$

$m = 8,$

$$M = \{m_{ij}\} = \begin{Bmatrix} 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & -1 \end{Bmatrix},$$

$b = \{b_j\} = \{0,0,\cdots,0\}$

이 된다. 이 때, 행렬 M의 계수(rank)는 7이므로 위 선형연립방정식에서 임의로

한 개의 식을 제외할 수 있다.

3.2.2 위험한 셀의 정의

Willenborg and de Waal (2001)은 정보보호를 위한 단계를 제 1차 문제(primary problem)와 제 2차 문제(secondary problem)으로 나누어 정의한다. 제 1차 문제는 표의 각 셀이 위험한 셀(risky cell)인지 여부를 판단하는 것을 의미한다. 위험한 셀로 판단된다면 정보보호 기법을 사용하여 정보유출을 막아야 하므로 이 결정은 매우 중요한 단계이다. 이를 위해 흔히 사용되는 세 가지 기법은 다음과 같다.

3.2.2.1 $n-$룰

빈도표(frequency table)의 위험한 셀을 정의하기 위하여 흔히 사용되는 $n-$룰은 임의의 작은 숫자 n에 대하여(예를 들어 $n \leq 1$) 셀의 도수(frequency)가 n보다 작은 셀을 위험한 셀로 정의한다. 이 때, n을 작게 할수록 자료유출 시도자는 0의 셀을 얻기 위하여 필요한 정보가 줄어들게 되고 n을 크게 할수록 위험한 셀로 정의되는 셀들이 늘어나는 부담이 생긴다.

3.2.2.2 지배룰 또는 (n,k)-룰

연속형 자료의 합으로 구성된 표에서 사용되는 지배룰(dominance rule) (또는 (n,k)-룰로 불림)은 두 개의 모수인 (n,k)로 구성되는데 이 때 n은 양의 정수를

나타내고 k는 비율(percentage)를 나타낸다. 각 셀의 값을 구성하는 응답값들 중 가장 큰 n개의 응답들의 합이 전체 합인 해당 셀의 값의 $k\%$ 이상이 되는 경우 해당 셀은 위험한 셀로 정의한다 (Willenborg and de Waal, 2001). 예를 들어, <표 3.7> 자료에서 제품 Y를 지역 C에서 판매한 5개의 기업이 존재하고 이들의 매출은 각각 10, 8, 2, 1, 1이라 가정하자. $n=3, k=70\%$으로 정의한 지배룰에 의하면 가장 매출이 큰 세 기업의 매출액의 합은 20이 되고 제품 Y의 C 지역 전체 매출액 22의 90.9%나 차지하므로 지배기준 70%를 넘게 되고 이 셀은 위험한 셀로 정의된다.

3.2.2.3 사전/사후 불명확성 룰

자료유출 시도자가 자료빼기를 통해 정보 유출을 시도하려고 할 때 자료빼기 후 가장 높은 기여자(contributor)의 값을 뺀 나머지 값들의 합이 가장 높은 기여자의 원자료 값의 $p\%$ 미만이 되면 그 셀을 위험한 셀로 정의하는 방법을 p-룰이라 부르는데 사전/사후 불명확성 룰(prior/posterior ambiguity rule) (또는 p/q 룰이라고도 부름)은 이 방법을 사후 불확성까지 고려하여 확장한 정의이다. 이 방법은 p와 q 두 개의 모수로 구성되어 있다. 표 자료를 발표하기 전에 한 기여자(contributor)가 기여하도록 허락하는 비율을 $p\%$로 한정하고 표 자료를 발표한 후에 한 기여자(contributor)가 기여하도록 허락하는 비율을 $q\%$로 한정하는 경우 사전/사후 불명확성 룰은 자료빼기 후 남아있는 가장 높은 기여자의 값을 뺀 나머지 값들의 합이 가장 높은 기여자의 원자료 값의 p/q 미만이면 위험한 셀로 정의한다. <표 3.7> 자료에서 제품 Y를 지역 C에서 판매한 5개의 기업이 존재하고 이들의 매출은 각각 10, 8, 2, 1, 1이고 두 번째로 기여가 높은 기업에 대한 정보를 지닌 자료유출 시도자가 존재한다고 가정하자. $p=25\%, q=50\%$로 정의하는 경우

사전/사후 불명확성 룰에 의하면 가장 매출액이 큰 기업의 총 판매액이 자료10으로 자료빼기 후 남아있는 가장 높은 기여자의 값을 뺀 나머지 값들의 합인 $2+1+1=4$가 $10\times\frac{25}{50}=5$보다 작아 이 셀은 위험한 셀이 된다.

3.2.2 위험한 셀에 대한 정보보호

Willenborg and de Waal (2001)의 정보보호를 위한 제 2차 문제는 위험한 셀에 대하여 정보보호 기법을 적용하는 것을 의미한다. 정보보호 기법을 적용하기 위하여 우선적으로 고려할 것은 자료 유출 시도자의 정보를 파악하는 것이 필요하다. 자료 유출 시도자는 정확한 정보값을 가지고 있을 수도 있지만 특정 구간으로 표현할 수 있는 구간 정보를 지니고 있을 가능성이 크다. 예를 들어, <표 3.7>의 자료유출 시도자는 직관적으로 판매량은 음수일 수 없으므로 [0, ∞)의 구간에 있을 것이라고 예측할 수 있다. 또는, 자료유출 시도자가 알고 있는 일부 기업들의 판매량이 5라면 이 기업들을 포함한 전체 기업들의 총 판매량에 대한 정보는 [0, ∞) 구간 대신 [5, ∞)의 구간으로 예측이 가능하다. 이를 일반화시켜 수식으로 표현하면 다음과 같다. 유출시도자 k가 i번째 셀의 값에 대한 사전정보(priori knowledge)에 근거하고 이 사전정보의 상한과 하한은 각각 Ub_i^k와 Lb_i^k로 정의한다면 유출시도자의 구간정보는 $\left[lb_i^k, ub_i^k\right]$로 표현할 수 있다. 즉,

$$lb_i^k \le y_i \le ub_i^k$$

의 구간 정보로 표현할 수 있다.

자료유출 시도자의 구간정보는 원자료의 구조에 근거한 연립방적식을 풀 때 함께 고려하여야 하는 제약조건(constraint)이 된다. 이를 식으로 표현하면 정보보호 기법을 적용한 후 자료는 다음의 조건을 만족해야 한다.

$$\sum_{i \in I} m_{ij} y_i = b_j \quad \text{for } j \in J;$$

$$lb_i^k \leq y_i \leq ub_i^k \quad \text{for } i \in I.$$

이 제약식을 만족하는 표를 일치성이 있는 표(congruent table)라 부른다. 물론, 정보보호 기법이 적용되기 전 원 자료도 일치성이 있는 표가 되지만 정보보호 기법의 원리는 일치성이 있는 다른 자료들도 존재하도록 하여 자료 셀의 참값을 알 수 없도록 만듦으로써 정보보호가 가능하게 하는데 있다.

정보보호의 난제는 일치성이 있는 자료가 많아지는 경우 제공한 자료의 정확도가 떨어져 활용도가 떨어지게 된다는 점이다. 가장 활용도가 많은 자료는 원자료이겠지만 원 자료의 문제는 정보유출이 발생한다는 점이다. 즉, 자료의 활용도를 높이려고 하면 정보유출이 발생하기 쉽고 정보유출 가능성을 낮추기 위하여 일치성이 있는 자료가 많이 존재하게 한다면 자료의 활용도가 낮아지는 것이다. 따라서 정보의 유출이 생기지 않으면서 자료의 활용도를 가능한 한 높이는 정보보호 기법을 적용하는 게 바람직하다. 즉, 정보보호 기법은 자료유출이 생기지 않도록 하는 것을 가장 중요하게 두고 자료유출이 생기지 않는 한에서 정보의 손실(loss of information)을 최소화하여 자료의 활용도를 높이는 방법을 의미한다.

3.3 매크로자료 정보보호 기법

표 자료의 자료 유출을 막기 위해 흔히 사용되는 정보보호 기법들을 소개한다.

3.3.1 표 재설계 (table redesign)

표 자료의 정보보호를 위해 가장 손쉽게 시행할 수 있는 방법은 위험한 셀을 연관된 셀과 통합하여 표를 재설계하는 방법이다. 예를 들어, <표 3.1>에 의하면 B 지역에 사회경제 수준이 “상”인 가구가 존재하지 않아 자료유출이 발생하므로 이 경우 사회경제 “상”과 “중”을 통합하여 한 개의 범주로 만든 후 재설계된 자료 <표 3.9>를 제공하는 것이다.

<표 3.9> 재설계된 두 지역의 사회경제 수준별 가구 숫자

		사회경제수준		
		상, 중	하	합계
지역	A	95	5	100
	B	40	60	100
	합계	135	65	200

이 방법은 손쉽게 시행할 수 있는 장점을 지니는 데 반하여 사회경제 수준 “상”인 가구와 “중”인 가구들의 비교 분석이 더 이상 가능하지 않아 활용도가 떨어진다는 데 있다. 또한 2×2 이차원 분할표에 대하여 적용한다면 표는 더 이상 이차원 분할표가 아니고 변수간 연관성을 파악하는 것이 불가능하게 된다, 따라서, 정보의 심각한 손실이 발생할 가능성이 크므로 어느 범주를 통합하여 표를 재설계해야 하는지 조심해서 결정해야 한다.

3.3.2 셀 감추기 (Cell Suppression)

셀 감추기 기법은 위험한 셀의 값을 감추는 정보보호 기법을 의미한다. 감추어진 셀(suppressed cell)의 값은 특수 문자(*나 문자 s가 흔히 사용됨)로 나타나게 된다. 셀 감추기는 두 가지 단계로 이루어지는데 제 1차 감추기 (primary suppression) 단계에서는 위험한 셀의 값을 감추게 된다. <표 3.7>에 나타난 가상 자료에서 지역 C에서 판매된 제품 Y의 판매총액이 위험한 셀로 나타난다면 이 셀에 해당한 값을 감추게 되며 그 결과는 <표 3.10>에 나타난다.

<표 3.10> 세 지역에서 판매된 세 가지 제품의 제품별 판매총액 자료에 대한 1차 단계 셀 감추기 (단위: 천만원)

		지역			
		A	B	C	합계
제품	X	20	50	10	80
	Y	8	19	*	49
	Z	17	32	12	61
	합계	45	101	44	190

* 는 해당 셀의 값이 감추어졌다는 것을 의미함.

마이크로 자료와 달리 표 자료에서는 각 행 및 열의 합계가 제공되기 때문에 제 1차 감추기 단계를 실행한다고 하더라도 지역 C에서 판매된 제품 Y의 판매총액 정보는 보호되지 않는다. 즉, 제품 Y의 행의 합계가 49로 나타나므로 49에서 A와 B 지역의 판매총액인 $8+19=27$을 제외한 22가 C 지역의 판매액임을 단번에 알 수 있는 것이다. 마찬가지로 열의 합계와 총합을 이용하여도 지역 C에서 판

매된 제품 Y의 판매총액을 알 수 있어 이 정보에 대한 보호가 되지 않는 것이다. 따라서 제 2차 감추기 단계(secondary suppression) (또는 complementary suppression이라고도 부름)가 실행되어야 한다.

제 2차 감추기 단계에서는 위험한 셀의 정보보호를 위하여 다른 셀의 값들이 추가적으로 감추어진다. 이 때, 너무 많은 셀의 정보를 감추면 정보의 손실이 많아지고 가능한 한 작은 셀의 정보를 감추면서 가능한 한 많은 정보를 유지해야 한다. 이를 위하여 우선 행의 합에서부터 추론된 정보 유출을 방지하기 위해 다른 셀의 값이 추가로 감추어져야 한다. 즉, 제품 Y의 지역 C의 정보 유출을 방지하기 위하여 제품 Y에 대한 다른 지역(A 또는 B)의 판매량 정보도 감추어져야 하는 것이다. 마찬가지로 열의 합에서부터 추론된 정보 유출을 방지하기 위해 지역 C에서 판매된 다른 제품(X나 Y)의 판매량 정보도 감추어져야 하는 것이다. 추가로 총합으로 인한 정보 유출을 막기 위한 감추기도 실행해야 한다. 문제는 어느 행과 열의 값을 감추어야 정보유출이 최소화되는지 파악해야 하므로 이에 대한 해결을 위해 최적화 문제(optimization problem)이 적용된다.

<표 3.11>은 제 1차와 2차 셀 감추기가 모두 실행된 후 결과 표를 나타낸다. 원래 위험한 셀인 제품 Y의 지역 C의 판매량 뿐 아니라 제품 Y의 지역 A의 판매량 정보 및 제품 Z의 지역 A과 C의 판매량 정보도 감추어진 것으로 나타난다. 하지만 정확한 행의 합계, 열의 합계, 그리고 총계는 표에 유지되고 있다.

<표 3.11> 세 지역에서 판매된 세 가지 제품의 제품별 판매총액 자료에 대한 2차 단계 셀 감추기를 시행한 결과표 (단위: 천만원)

		지역			
		A	B	C	합계
제품	X	20	50	10	80
	Y	*	19	*	49
	Z	*	32	*	61
	합계	45	101	44	190

최적화된 셀 감추기는 수학적 프로그래밍 문제로 표현할 수 있다. 자료유출 시도자 k에 대항하여 원 자료 y에 대한 셀 감추기를 시행하기 위하여 다음의 방정식

$$\sum_{i \in I} m_{ij} y_i = b_j \quad \text{for } j \in J,$$

$$y_i = y_i \qquad \text{for all } i \notin \text{supressed cell},$$

$$lb_i^k \le y_i \le ub_i^k \quad \text{for all } i \in \text{supressed cell}.$$

을 만족하는 해답을 찾아야 한다.

<표 3.7> 자료에 대하여 셀 감추기가 시행되었을 때 감춰진 셀의 값이 음수가 아니라는 정보($lb_i^k = 0$ and $ub_i^k = \infty$)만을 가진 한 명의 자료유출 시도자가 존재하는 경우 제품 Y의 지역 C의 판매량 $y_{Y,C}$의 최소값(minimum value) $\underline{y_{Y,C}}$는 다음의 조건들

$y_{Y,A} + y_{Y,C} = 30,$

$y_{Z,A} + y_{Z,C} = 29,$

$y_{Y,A} + y_{Z,A} = 25,$

$y_{Y,C} + y_{Z,C} = 34,$

$y_{Y,A} \geq 0,\ y_{Z,A} \geq 0, y_{Y,C} \geq 0, y_{Z,C} \geq 0$

을 만족하는 최적 y값으로 계산된다.

마찬가지로 제품 Y의 지역 C의 판매량 $y_{Y,C}$의 최대값(maximum value) $\overline{y_{Y,C}}$를 계산할 수 있고 이 자료에 대하여 구해진 최소값과 최대값은 $\underline{y_{Y,C}} = 5,\ \overline{y_{Y,C}} = 30$으로 나타나 정확한 값으로 추측할 수 있는 구간이 충분히 넓으므로 제품 Y의 지역 C의 판매량 $y_{Y,C}$에 대한 정보보호가 잘 이루어졌음을 알 수 있다. 이 때 최소값과 최대값의 구간인 [5,30]을 $y_{Y,C}$에 대한 정보보호구간(disclosure limitation interval)이라 부른다.

3.3.2.1 셀 감추기로 인하여 손실된 정보량

표 자료의 정보보호를 위하여 셀 감추기가 실시된 경우 일부 셀의 값이 감추어져 정보의 손실(loss of information)이 발생하므로 이렇게 손실된 정보의 량을 추정하는 것이 바람직하다. 셀 i에 대한 정보 손실량을 ω_i라 하면 이 값은 자료 y_i, 셀 i의 자료값 y_i 를 계산하는데 기여한 마이크로자료의 응답값들의 숫자들의 함수로 계산된다 (Willen borg and de Waal, 2001). 표 자료의 손실된 정보량은

$$\sum_{i \in suppressed} w_i$$

로 표현된다. 이 정보 손실량의 계산 방법은 Salazar(2010)에서 상세히 다루고 있다.

3.3.2.2 구간값 제공 (Interval Publication)

셀 감추기가 정보 유출의 위험이 있는 셀의 값을 감추어 발표하는 기법이라면 구간값 제공(interval publication)은 셀의 정확한 값 대신 구간으로 표현하여 제공하는 정보보호 기법을 의미한다. 셀 감추기는 정보유출 위험이 있는 셀의 값을 전혀 제공하지 않지만 구간값 제공 방식은 정확한 값이 무엇인지는 감추어져 있으나 구간 정보를 제공하므로 조금 더 많은 정보를 제공하는 셀 감추기 기법의 확장이라 할 수 있다.

<표 3.7>에 나타난 가상자료에서 지역 C에서 판매된 제품 Y의 판매총액이 위험한 셀로 나타난다면 이 셀에 해당한 값을 구간값으로 제공한 결과는 <표 3.12>에 나타난다. 지역 C에서 판매된 제품 Y의 정확한 값 대신 이 값의 범위를 20 - 26 사이의 구간으로 제공하는 게 이 정보보호 기법의 특징이다. 셀 감추기와 마찬가지로 정보 유출 위험이 있는 셀의 정확한 값을 추측하지 못하도록 주변 다른 셀들에 대한 값도 구간으로 표현되어야 한다. 즉, 지역 A에서 판매된 제품 X와 Y의 판매량, 그리고 지역 C에서 판매된 제품 X의 판매량도 더불어 구간값으로 나타나고 있다.

<표 3.12> 세 지역에서 판매된 세 가지 제품의 제품별 판매총액 자료에 대한 구간값 제공 (단위: 천만원)

		지역			
		A	B	C	합계
제품	X	[18 - 24]	50	[6 - 12]	80
	Y	[4 - 10]	19	[20 - 26]	49
	Z	17	32	12	61
	합계	45	101	44	190

구간값 제공에서 어느 셀을 원자료 대신 구간값으로 제공해야 하는지 수학적 프로그래밍 문제로 표현할 수 있다. 자료유출 시도자 k에 대항하여 원자료 y에 대한 구간값 제공을 시행하는 경우 각 셀 i에 대하여 제공된 구간값 y_i는 $y_i \in [y_i^-, y_i^+]$으로 표현하게 된다. 이 때 구간값을 결정하는 문제는 다음의 방정식

$$\sum_{i \in I} m_{ij} y_i = b_j \quad \text{for } j \in J,$$

$$y_i^- \le y_i \le y_i^+ \quad \text{for all } i \in I,$$

$$lb_i^k \le y_i \le ub_i^k \quad \text{for all } i \in I.$$

을 만족하는 해답을 찾는다.

이 방법의 정보손실은 y_i 대신 $[y_i^-, y_i^+]$을 제공하므로 발생하게 되고 정보 손실량은 $y_i - y_i^-$와 $y_i^+ - y_i$의 비율로 측정된다. 셀 감추기 방법에서 정의한 정보 손실량 개념을 확장하면 i번째 셀의 정확한 값 대신 구간값을 제공함으로써 얻어지는 정보 손실은 셀 감추기의 ω_i의 개념을 확장하여 사용할 수 있다. 구간의 최대값과 최소값을 제공하여 생기는 정보 손실 ω_i^+와 ω_i^-의 함수로 나타내면 구간값 제공

방법을 사용한 표 자료의 정보손실량은

$$\sum_{i \in I} \left[\omega_i^+ \left(y_i^+ - y_i \right) + \omega_i^- \left(y_i - y_i^- \right) \right]$$

으로 추정된다 (Fischetti and Salazar, 2003).

구간값 제공은 셀 감추기보다 더 많은 정보를 활용하게 되므로 상대적으로 적은 정보가 손실되고 계산도 쉬운 장점을 지닌다. 한편, 이 방법은 셀 감추기에 비해서 상대적으로 더 많은 셀을 구간으로 표현하는 경향이 있다는 점이 단점이라 할 수 있다.

3.3.3 반올림 (Rounding)

셀 감추기나 구간값 제공 방법은 표의 일부 셀의 값이 감춰져 나타나지 않거나 구간으로 표현되므로 분석에 어려움이 발생한다. 따라서 표의 모든 셀을 한 개의 값으로 표현하는 방식이 선호되는 데 이 때 사용되는 방법으로 반올림(rounding)을 고려할 수 있다.

i번째 셀의 기저값을 r_i로 나타내자. 기저값 r_i는 모든 셀에 동일하게 설정되는 것이 일반적이나 셀에 따라 다르게 나타낼 수도 있다. 예를 들면, 모든 i에 대하여 $r_i = 1$로 정한다면 모든 셀의 값을 정수값으로 반올림하는 것을 의미하고 $r_i = 5$인 경우 모든 셀의 값은 5의 배수값으로 반올림하여 표현하게 된다. <표 3.13>은 모든 i에 대하여 $r_i = 5$로 반올림한 표를 보여준다.

<표 3.13> 세 지역에서 판매된 세 가지 제품의 제품별 판매총액 자료에 대한 반올림 결과표 (단위: 천만원)

		지역			
		A	B	C	합계
제품	X	20	50	10	80
	Y	10	20	20	50
	Z	15	30	15	60
	합계	45	100	45	190

각 셀의 값을 어떤 기저값으로 반올림해야 하는지는 제약식을 만족하는 최적 해를 구하여 결정한다 (Bacharach, 1966). $\lfloor y_i \rfloor$ 를 r_i를 기저로 하는 반올림에서 버림한 값을 나타내고 $\lceil y_i \rceil$ 를 r_i를 기저로 하는 반올림에서 올림한 값을 나타내자. 각 셀의 값은 버림하거나 올림하여 기저값으로 표현할 수 있으므로 반올림 결과표는 $v = [v_i : i \in I]$로 표현되는데 이 때 $v_i \in \{ \lfloor y_i \rfloor , \lceil y_i \rceil \}$으로 표현할 수 있다. 일반적으로 정보 유출 시도자는 기저값 r_i를 손쉽게 알 수 있고 이 때 반올림 문제는 제약을 지닌 방정식

$$\sum_{i \in I} m_{ij} y_i = b_j \quad \text{for } j \in J,$$

$$v_i - r_i \le y_i \le v_i + r_i \quad \text{for } all\ i \in I,$$

$$lb_i^k \le y_i \le ub_i^k \quad \text{for } all\ i \in I.$$

을 풀어 얻는다.

3.3.3.1 반올림으로 인하여 손실된 정보량

표 자료의 정보보호를 위하여 반올림이 실시된 경우 대부분의 셀의 값이 반올림되므로 정보의 손실이 일어난다. 이 때 손실된 정보량은 원래 값과 반올림하여 제공한 값 사이의 절대값 차이로 표현할 수 있으며 표 자료의 손실된 정보량은 각 셀에서 발생한 손실된 정보량의 합계로 나타난다.

이 방법을 적용하는데 가장 큰 어려움은 위 방정식을 만족하는 해가 항상 존재하는 것은 아니라는 것이다. 모든 셀에 동일한 기저값을 사용하는 경우 이차원분할표에서는 해를 발견할 수 있으나 다중분할표에서는 해가 존재하지 않을 수도 있다. 다중분할표에서 해를 발견하기 위해서 branch-and-bound 방법(Kelly, et. al., 1990)이 제안되었고 발견적 방법(heuristic method)로 반올림 문제를 해결하기 위하여 branch-and-bound 방법이 사용되어 왔다 (Kelly, et. al., 1990, 1993; Salazar, et. al, 2004; Salazar, 2006).

3.3.4 셀 변조 (Perturbation)

표 자료의 정보보호를 위한 반올림 기법의 주요 문제점은 기저값에 근거하여 반올림을 실시해야 하므로 제약을 만족하는 연립방정식의 해를 구할 수 없는 경우가 발생할 수 있다는 점이다. 셀 변조 (cell perturbation) 방법은 반올림 기법의 제약을 완화하여 해를 구하는 방법으로 널리 사용되고 있다. 이 방법은 반올림 기법의 제약 $v=[v_i : i \in I]$에서 $v_i \in \{ \lfloor y_i \rfloor , \lceil y_i \rceil \}$에서 $\lfloor y_i \rfloor$를 r_i를 기저로 하는 반올림에서 버림한 값 대신 $\lfloor y_i \rfloor = y_i - t_i$, 그리고 $\lceil y_i \rceil$를 r_i를 기저로 하는 반올림에서 올림한 값 대신 $\lceil y_i \rceil = y_i + t_i$로 정의한다. 이 때, t_i는 임의의 기저

값이 된다. <표 3.14>은 <표 3.7> 자료에 대한 정보보호를 위해 셀 변조를 실시한 후 제공한 표를 나타낸다.

<표 3.14> 세 지역에서 판매된 세 가지 제품의 제품별 판매총액 자료에 대한 셀 변조 결과표 (단위: 천만원)

		지역			
		A	B	C	합계
제품	X	20	50	10	80
	Y	7	16	26	49
	Z	18	35	8	61
	합계	45	101	44	190

3.3.4.1 셀 변조로 인하여 손실된 정보량

표 자료의 정보보호를 위하여 셀 변조가 실시된 경우에도 많은 셀의 값이 바뀌게 되므로 원 자료의 정보가 손실된다. i번째 셀에서 손실된 정보는 $|v_i - y_i|$ 에 비례하므로 정보보호된 자료의 손실된 정보량은

$$\sum_{i \in I} |v_i - y_i|$$

로 표현할 수 있다. 문제는 이 기준 하에서 가장 작은 손실된 정보량을 보이는 경우는 $v_i = y_i \text{ for } all\ i \in I$가 되므로 정보보호가 되지 않은 자료를 해로 선택하게 되는 점이다. 따라서 대신 변조된 표 자료 $v = \{v_i\}$와 적절하게 선택된 배열 y' 사이의 거리를 최소화하는 해를 찾는다.

3.3.5 자료 교환 (Data Swapping)

표 자료의 정보보호를 위한 자료 교환(data swapping)은 Griffin, et. al. (1989)가 처음 제안한 방법으로서 다음과 같은 단계로 진행된다.

단계 1: 마이크로 자료로부터 표본을 추출한다.
단계 2: 중요한 변수들을 사용하여 다른 지역의 자료 중 짝(match)이 되는 자료를 찾는다.
단계 3: 짝 자료의 지역 코드를 교환한다.

마이크로 데이터에 대하여 자료 교환을 실시한 후 표 자료는 더 이상의 수정 없이 제공된 수 있다. 이 경우 자료 교환으로 인하여 정보보호가 실제로 이루어졌는지 그리고 자료의 유용성에 어떤 영향을 주는지는 경험적으로 평가가 진행되어야 한다. Navarro 등(1988)이 진행한 모의실험은 자료 교환이 작은 규모의 모집단을 제외하고 적절한 정보보호를 제공하는 것으로 나타난다.

3.3.6 그 외의 방법

3.3.4절은 셀 변조 기법에 관하여 설명하였는데 Duncan and Roehrig(2007)은 순환 셀 변조(cyclic cell perturbation) 기법을 제안하였다. 순환 셀 변조기법은 셀 변조 기법을 연속적으로 시행하는데 각 변조 단계에서 4 - 5개의 셀을 변조한다. 이 때 일부 셀의 값은 1을 증가시키고 일부 셀의 값은 1을 감소시켜 행과 열의 합이 변하지 않도록 유지한다. 또한 변조 기법을 적용할 때 각 셀의 변조 횟수를 전체 연속에서 동일하게 유지하는 순환 구조를 지닌다. 이 방법은 셀들의 변조 순서에

따라 결과가 달라지는 특징을 지닌다.

Cox 등(2004)는 셀 변조 기법의 변형인 표 보정 관리(controlled tabular adjustment) 기법을 제안하였는데 이 방법에 따르면 위험한 셀의 값을 가장 가까운 안전한 값으로 바꾼 후 다른 셀들을 보정하여 행의 합과 열의 합을 유지해 준다. 셀 변조 기법과 표 보정 관리 기법은 모두 가능한 한 원래의 표와 가까운 정보보호된 표를 찾는다는 점에서 유사성을 지난다. 한편, 셀 변조 기법은 모든 셀에서 정보보호 유지를 위한 요구조건을 만족하는 데 반하여 표 보정 관리 기법에서는 정보보호를 위한 제약이 포함되지 않아 이 조건이 만족되지 않는다.

3.3.3절은 반올림 기법을 설명하였는데 반올림할 때 행의 합과 열의 합이 고정되어 있어 관리된 반올림(controlled rounding) 기법이라고도 부른다. 이에 반하여 임의 반올림(random rounding) 기법은 행의 합과 열의 합이 셀의 반올림된 값에 따라 변하는 반올림 기법을 의미한다. 이 방법의 장점은 단순함에 있으며 이렇게 생성된 표는 원래 표에 대한 불편성(unbiasedness)을 지니므로 바람직한 통계적인 특성을 지닌다. 한편, 이 방법의 단점은 행의 합, 열의 합, 또는 총합이 변동 가능하므로 이 값이 원래 값과 크게 달라질 수 있다는 점이다. 따라서 정보보호 자료의 적절성에 대한 감사(audit)이 필요하다.

미국 센서스 국에서 고려한 다른 방법으로는 마이크로 자료에 잡음(noise)를 추가하는 정보보호 기법을 들 수 있다. 이 방법은 마이크로 자료의 데이터 값에 잡음을 더하여 준 후 표를 만들어 제공하는 기법이다. 이 때 충분한 잡음을 더해준다면 자료의 정보가 유출될 가능성이 줄어들게 되는 반면에 자료의 유용성은 줄어들게 된다. Evans et. al. (1998)은 합계자료의 값을 보존하기 위하여 잡음첨가방법 전의 자료에 근거한 통계량과 잡음첨가방법 후의 자료에 근거한 통계량이 동

일하게 되도록 보정을 하는 방법을 제안하였다.

표 자료의 정보보호는 꼭 위에 고려한 방법들 중 한 가지를 선택하여 사용해야 하는 것이 아니다. 한 개의 표에서도 여러 가지 정보보호 기법을 사용할 수 있다. 연결된 표(linked table)는 한 개 이상의 이원분할표가 한 개의 표로 연결된 표를 나타내는데 이렇게 연결된 각각의 표에 대하여 다른 정보보호 기법을 적용하는 것이 가능하다.

제 4장. 마이크로자료 보호 기법

4.1 서론

마이크로자료는 단위자료들이 모여서 만들어진 집합이다. 각 단위자료는 다양한 형태의 변수들의 값으로 구성되어 있다. 예를 들면 자료의 구성단위가 사람이라면 각 단위자료는 사람에 대한 특성을 나타내는 변수인 나이, 성별, 직업 등을 포함한다. 마이크로자료는 이용자들에게 매우 유용한 정보를 제공하지만 동시에 개인의 자료가 유출되어 사생활이 침해당할 가능성도 동시에 존재한다. 따라서 마이크로자료에 포함되어있는 자료 중 민감한 정보가 노출 될 수 있는 가능성을 줄이는 노력이 필요하다. 이러한 정보 보호를 위한 노력은 기본적으로 정보의 유용성을 저하시키므로 정보보호와 유용성유지를 동시에 고려해야한다.

1980년대 IT기술의 혁명이 시작되고 소규모 기업 및 개인에게 개인용 컴퓨터의 접근성이 커지면소 마이크로자료에 대한 요구가 점점 커졌다. 마이크로자료는 연구자, 기업, 정책결정자가 중요한 연구 및 의사결정을 신속하고 저렴하게 그리고 효율적으로 수행할 수 있도록 할 수 있는 정보를 제공한다. 따라서 사용자들은 크고 매우 자세한 자료를 요구한다. 그러나 불행하게도, 통계기관이 더 많은 정보를 제공할수록 사용자가 일부 응답자에 대한 정보를 확인할 수 있는 위험성이 더 커진다. 이러한 이유로 통계기관은 모든 응답자의 기밀성을 보장하기 위한 필수적인 조치를 할 것을 요구한다. 오늘날 사용자는 과거보다 더 많은 식별정보를 가지고 있으며 더 정교한 방법을 이용하므로 응답자의 기밀성을 보호할 수 있는 방법도 이에 따라 더 정교해야만 하는 것이 최근의 현실이다.

4.2 자료 공개의 필요성과 노출 위험

통계자료를 공표하는 기관은 이용자 또는 연구자의 요구에 따라서 마이크로 자료를 이용하는 방법을 제공한다. 최근 마이크로 자료를 공개하는 이유는 꼭 이용자의 요구에 부응하는 측면도 있지만, 더욱 중요한 이유는 여러 가지 사회적 이슈나 정책결정에 필요한 정보를 제공하는 경우 증거에 기반을 둔 분석(evidence-based analysis)이 제공되어야 하며 이를 위해서는 마이크로자료의 이용이 필수적이기 때문이다. 자료를 요약한 집계표 또는 기본적인 요약 통계량만 공개하지 않고 더 나아가 마이크로 자료를 공개하는 경우 아래와 같은 이점이 있다.

(1) 연구자나 정책결정자들이 중요한 이슈에 대하여 증거에 기반을 둔 분석을 통하여 폭넓은 정보를 얻을 수 있고 이를 통하여 건전한 논쟁과 합리적인 의사결정이 가능하다.

(2) 마이크로 자료의 분석을 통하여 모집단에 대한 새로운 정보나 경향을 발견할 수 있다. 이를 통하여 통계작성기관은 새로운 통계를 개발하고 이동자의 요구를 적절히 반영할 수 있다. 또한, 통계기관이 작성한 집계표나 요약통계의 신뢰성을 높이는데 마이크로 자료가 큰 역할을 한다.

(3) 통계작성기관이 공표하는 집계통계보다 좀 더 세밀하고 자세한 통계 (예를 들어 소지역통계)를 생산할 수 있는 마이크로 자료를 공개하는 기회를 제공함으로써 이용자들의 다양한 요구에 적절하게 부응할 수 있다. 또한, 연구자에게는 연구주제에 대한 다양한 통계적인 모형이나 가설을 검증할 기회를 제공한다.

역사적으로 미국이 1960년 인구 총 조사 자료를 일반인에게 공개한 것을 시점으

로 선진국의 통계작성기관은 마이크로자료를 공개하기 시작하였다. 캐나다가 1971년 총 조사 자료를 공개하기 시작하였고, 오스트레일리아는 1981년부터 영국은 1989년부터 여러 종류의 마이크로 자료를 공개하기 시작하였다.

마이크로자료를 공개하는 경우에는 응답자 또는 조사단위의 신분이 노출될 위험성이 매우 크다. 마이크로자료를 공개하는 경우 조사단위의 신분을 확인할 수 있는 직접적인 인식정보(direct identifier)를 공개하는 경우에는 노출을 제한할 방법이 없다. 예를 들어, 조사단위의 이름, 주소, 주민번호, 등록자 번호 등이 직접적인 인식정보이며, 이러한 정보는 마이크로자료를 공개할 때 특별한 이유가 없는 한 공개를 하지 않는다. 하지만 직접적인 인식정보를 공개하지 않더라도 마이크로 자료에 포함된 간접적인 인식정보(indirect identifier)의 조합을 통하여 조사단위의 신분을 확인할 수 있는 위험성이 항상 존재한다.

간접적인 인식정보는 일반적으로 조사단위의 지역정보, 생일, 설립연도, 규모, 크기, 유일한 특이 값 등이며 이러한 간접정보를 적절히 조합하여 이용자가 보유하고 있는 외부정보와 결합하면 조사단위의 신분노출 위험성은 크게 높아진다. 최근에는 IT의 발전과 여러 가지 레코드연결 프로그램의 사용화로(Record linkage) 인하여 마이크로 자료를 공개하는 경우 직접적인 인식정보와 간접인식정보를 공개하지 않더라도 외부이용자가 공개된 자료와 외부자료의 분석을 통하여 조사단위의 신분을 쉽게 파악하는 방법이 많다. 이러한 매칭(matching)이나 레코드 연결방법에 의한 신분노출은 흔히 일어나며 다음과 같은 예가 있다.

(1) 다양한 사설정보제공기관은 여러 가지 산업에서 매출액에 의한 상위 업체들의 여러 가지 정보를 판단한다. 사업체의 매출액 규모, 종사자 수, 신용정보 등을 판단하는데 이러한 관련정보와 공개된 마이크로자료를 결합하면 쉽게 사업체의 신

분이 노출되고 민감한 정보가 유출될 가능성이 크다.

(2) 2009년 교육과학기술부는 학생들의 수능성적을 A 국회의원에게 공개하였다. 이때 여러 가지 직접적인 신용정보를 제외하고 학교의 지역 정보화 학생 수능점수만을 공개하였지만, 지역정보, 학생의 수, 남녀비율 등 여러 가지 정보를 조합하여 학교의 이름을 알아낼 수 있었다. (연합뉴스, 2009년 10월 15일)

위와 같은 이유 때문에 통계작성기관은 마이크로자료를 공개할 때 직접적인 인식정보와 간접적인 인식정보를 적절하게 제거하고 변형하여 제공해야 한다. 더 나아가서 외부이용자가 보유한 외부자료를 통하여 신분이 노출될 수 있는 여러 가지 가능성을 고려하여 마이크로자료에 적절한 노출제한기법을 적용한 후에 공개해야 한다. 특별하게 통계작성기관이 마이크로자료를 공개할 때 신중하게 고려해야 할 사항은 다음과 같다.

(가) 외부이용자의 정보: 외부이용자가 공개하려는 마이크로자료에 대한 정보를 어떤 형태로 얼마나 자세하게 보유하고 있는지를 정확하게 인식하고 이에 대응하여 노출제한기법을 적용해야 한다.

(나) 마이크로자료를 공개하는 경우 지역정보를 어떤 범위로 어떤 형태의 정보를 제공할지 세심하게 검토하여 공개해야 한다. 지역정보는 조사단위의 신분을 쉽게 파악할 수 있는 매우 유용한 정보이기 때문에 외부이용자가 신분노출을 위하여 자주 이용하는 정보이다. 공개 자료에서 지역정보의 범위를 자세하게 공개하면 직접적인 인식정보가 없더라도 조사단위의 위치를 쉽게 파악할 수 있고 이를 이용하여 다른 정보와 연결을 통해서 신분노출의 위험성이 증가한다. 예를 들어, 학교의 지역정보를 시/군/구 단위로 공개할 때 직접적인 인식정보가 없더라도 학교의

여러 가지 특성들(학교의 종류, 학생의 수, 남녀비율, 교사의 수 등)을 조합하여 외부공개정보와 결합하여 쉽게 학교의 이름을 확인할 수 있다. 이럴 때 노출의 위험성을 줄이려면 학교의 지역정보를 더 큰 단위 예를 들어, 광역시/도 단위를 공개할 수 있다.

(다) 시계열 정보: 마이크로자료를 (longitudinal structural) 시간에 따라서 연속적으로 공개하는 경우 조사단위 특성들의 시간에 따른 변화를 파악할 수 있기 때문에 노출의 위험성이 증대된다.

(라) 특이 값(Outlier): 마이크로자료 중 여러 가지 변수들의 특성을 고려할 때 조사 단위가 전체적인 경향과 매우 다른 경향을 보이는 특이한 단위라면 노출의 위험성이 증대한다. 예를 들어 사업체의 매출액이 매우 크거나, 가구의 구성원이 매우 많으면 다른 정보와 결합하여 신분의 노출이 쉽게 이루어질 수 있다. 대한보건통계는 특이한 질병을 보유하고 있거나 매우 크거나 작은 단위의 검사자료를 가진 조사단위로 쉽게 신분노출이 될 수 있다. 이러한 특이 값을 가진 조사단위의 신분이 노출될 위험성을 줄일 수 있는 적절한 노출제한기법이 마이크로자료에 적용되어야 한다.

통계작성기관은 마이크로자료를 공개하는 경우 조사단위의 신분노출이나 민감한 정보의 유출의 위험성을 최소화하면서 동시에 자료의 유용성을 어느 정도 유지해야 하는 두 가지 상충한 목표를 달성해야 한다. 신분이나 민감한 정보의 노출 위험성을 줄이기 위해서는 마이크로자료에 대한 접근을 제한하거나 마이크로자료에 노출제한기법을 적용하여 안전한 자료로 변형시켜서 공개하는 방법을 적절하게 적용해야 한다.

4.3 마이크로자료 노출제한 방법

4.3.1 접근의 제한

마이크로자료에 노출제한방법을 적용하여 대중이 제한 없이 이용할 수 있도록 공개를 할 때에는 자료의 유용성이 상대적으로 감소하여 이용자들이 다양하고 세밀한 분석이 불가능한 경우가 많다. 이러한 경우를 위하여 통계작성 기관에서는 노출제한 방법을 최소한으로 적용한 원래자료에 매우 가까운 마이크로자료를 이용할 수 있는 환경을 제공하는 경우가 필요하다. 이렇게 마이크로자료에 대한 제한적인 접근을 허용하는 경우에는 다음과 같은 사항을 고려하여 그에 대한 공개 정책을 수립하고 시행해야 한다.

○ 마이크로자료를 이용할 수 있는 자격요건의 선정
○ 마이크로자료를 제공하는 수단과 방법
○ 마이크로자료를 분석할 수 있는 범위

마이크로자료를 이용할 수 있는 자격을 가진 사람들은 자료에 포함된 정보를 누출할 위험성이 매우 작은 신뢰받는 개인이나 조직이어야 한다. 또한, 마이크로 자료를 이용하려는 사람들에 대해서는 자료유출에 대한 도덕적 또는 법적인 책임에 대하여 이용 전에 구속력 있는 서약서를 받는 것이 중요하다. 이렇게 신뢰할 수 있는 이용자에게 서약서를 받고 자료의 접근을 허용하는 경우에는 언제나 자료의 이용이 공익에 맞는지 또는 사회적, 정치적인 문제 해결에 이바지할 수 있는지 또는 통계작성기관의 목적에 맞는지에 대한 요건을 미리 심사하여 접근의 허용을 결정하는 것이 바람직하다.

마이크로자료에 접근을 허용하는 기준으로서 다음과 같은 사항을 주로 고려해야 한다.

(1) 제안된 자료이용의 목적이 꼭 마이크로자료를 이용해야 가능한가?

(2) 마이크로자료를 이용한 연구가 보고서나 논문의 형태로 출간될 만큼 중요한 주제를 다루고 있는가?

(3) 자료를 이용하려는 사람이 사회적으로 권위가 있고 신뢰할 수 있는 기관에 속해 있는가?

(4) 마이크로자료의 이용목적이 통계생산기관의 목적에 맞는가?

(5) 마이크로자료를 이용하려는 사람의 자격이 통계작성기관이 정해놓은 보안수칙에 부합되는가?

마이크로자료에 접근이 허용된 사람들에게 자료를 어떤 형태로 제공하고 어디서 사용을 할 수 있게 하는지에 대한 정책도 세심하게 고려돼야 한다. 공개 자료의 접근방법과 형태는 다음과 같은 방법이 있다.

(1) 기본적으로 물리적인 저장매체 (USB, CD등)나 데이터베이스의 직접적인 접근을 통해서 마이크로자료를 제공할 수 있는데 이러한 방법은 자료 이용자의 편의는 증대되지만 자료유출의 가능성이 높아지는 단점이 있다.

(2) 인터넷이나 포탈시스템(portal system)에 자료 분석을 위한 시스템을 구성하고 마이크로자료의 직접적인 접근을 하지 않고도 필요한 분석을 수행하여 이용자가 원하는 분석결과만 얻을 수 있도록 하는 방법이 있다. 이러한 방법은 최근 IT의 발달에 의해 많은 통계생산기관에서 제공하는 방법이며 자료의 직접접근이 불가능하기 때문에 자료의 유출가능성이 낮아지며 통계작성기관이 이용자를 통제할 수 있는 여러 가지 방법을 사용할 수 있는 장점이 있다.

(3) 자료를 접근할 수 있는 공간과 도구를 제한적으로 지정된 곳으로 지정할 수 있지만 (예를 들어 자료 분석센터를 세우고 제한된 컴퓨터를 이용하는 경우) 이러

한 제한방법은 이용자가 지리적으로 이동해야하고 정해진 시간에만 사용할 수 있기 때문에 자료의 활용도가 매우 떨어질 수 있다.

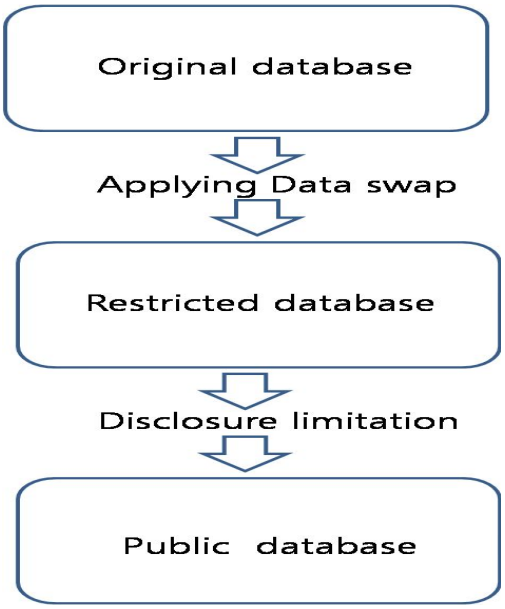
이용자가 접근한 마이크로자료를 이용하여 분석할 수 있는 통계적인 분석범위 또한 통계작성 기관을 세심하게 고려하여 제한을 주어야 한다. 보통 일반대중에게 제한 없이 공개되는 마이크로자료보다 접근의 제한이 있는 마이크로자료는 상대적으로 노출의 위험성이 높다. 따라서 마이크로자료를 분석한 결과물 (집계표, 요약 통계량, 등)에 조사단위의 신분이 노출될 정보를 포함한 결과가 있을 수가 있기 때문에 이러한 위험성을 포함할 수 있는 분석은 제한돼야 한다. 일반적일 때에 자료의 접근이 허가된 사람은 서약서를 통하여 노출에 대한 위험성과 법적 책임을 인지하고 있지만, 통계분석의 특성상 의도하지 않은 노출의 결과로 발생할 수 있다. 예를 들어서 분석의 결과도 요약 집계표를 생성할 때 특이한 또는 유일한 조사 단위가 공개된다거나 또는 회귀분석 시 잔차의 크기가 매우 큰 단위가 공개되면 노출의 위험성이 커진다. 이러한 분석결과에 포함된 노출의 위험성을 낮추기 위하여 분석결과를 통계작성기관이 검토하거나 또는 자료 분석시스템 자체에 매크로자료에 대한 노출제한방법을 구현할 수 있다.

미국의 국가교육통계센터(The National Center for Education Statistics; NCES)는 1988년, 1994년, 2002년의 일련의 법 개정에 의하여 교육 자료에 있어서 강력한 사생활 보호 장치가 필요하게 되었다. 2002년에 NCES는 개인정보 보호를 위한 표준을 제정하였으며 동시에 Disclosure Review Board에 의하여 통계적 노출제한 방법에 대한 재검토 및 개선 작업이 시작되었다 (Report on Statistical Disclosure Limitation Methodology, 2005). NCES는 학교, 기관, 교사, 학생, 학부모에 다양한 자료를 조사하고 수집하고 있으며 이를 대중이 이용할 수 있는 공개자료 형태 또는 자료 분석 시스템의 형태로 일반 사용자에게 제공하고 있는데 제공되는 모든

자료는 Disclosure Review Board에서 공인하는 노출제한 방법을 적용한 뒤에 공개하고 있다. 최근 2011년에는 National Institute of Statistical Sciences Data Confidentiality Technical Panel에서 자료의 노출제한 정책 및 통계적 방법에 대한 진단을 받고 보고서를 발간하였다 (Krenzke 2006; Karr 2011)

NCES에서 채택한 자료 공개의 절차는 <그림 4.1>과 같다. NCES가 획득한 표본조사 자료를 저장하고 있는 원래 데이터베이스(original database)에 자료교환 등 통계적 노출관리 기법을 적용한 제한적 데이터베이스(restricted database)를 만든다. 제한적 데이터베이스에 접근하려면 NCES에 사용 신청을 하여 이용할 수 있는 허가를 받아야 하며 제한적 데이터베이스의 이용은 Data Analysis System으로 명명된 인터넷기반 시스템에 접속으로 가능하다. 이용자는 Data Analysis System에서 기초통계분석과 회귀분석을 이용하여 자료를 분석할 수 있다. 또한 일반 이용자가 제한 없이 사용할 수 있는 공개자료(public database)는 제한적 데이터베이스에 좀 더 강력한 통계적 노출관리 기법이 적용되어 공개된다. 공개 자료는 인터넷 상에서 누구나 다운로드 받을 수 있다.

<그림 4.1> NCES가 채택한 마이크로 교육자료 공개의 절차



4.3.2 자료감추기 (Suppression)

자료감추기(suppression)는 자료의 일부를 공개하지 않고 숨기는 방법이다. 자료의 일부가 공개되었을 때 노출 위험이 상당하게 높다고 판단되면 자료의 일부를 제외하고 공개하는 비법을 말한다. 자료 감추기에는 변수 감추기(Variable or Attribute Suppression)와 레코드 감추기(Record Suppression)가 있다.

(1)레코드 감추기 : 마이크로자료에서 특정한 성질을 가진 단위를 공개하지 않는 방법이다. 예를 들어서 가구조사에서 월 소득이 1000만 원을 넘는가구를 공개 자료에서 제외할 수 있다.

(2)변수 감추기 : 마이크로자료를 구성하고 있는 변수 중에 공개되면 노출의 위험이 크게 증대하는 변수의 값을 모든 조사단위에 대하여 공개하지 않는 것이다. 예

를 들면 출생지(birth place)나 특정질병의 유무 등은 마이크로자료 공개할 때 공개하지 않는다.

자료감추기는 다른 노출제한기법과 달리 자료의 일부를 공개하지 않는 것이므로 자료를 공개했을 때 노출의 위험성이 매우 커지는 경우에 주로 적용한다. 일반적으로 마이크로자료에 시간적 변화에 대한 정보가 포함된 자료(longitudinal data)는 조사단위의 노출위험성이 크다. 이러한 경우 각 조사단위의 연속적인 자료를 연결하는 변수를 공개하지 않음으로서 노출의 위험성을 줄일 수 있다.

또한, 많은 경우에 조사단위의 지역적인 정보는 노출의 위험성을 증가시킨다. 이럴 때에 조사 단위가 속한 작은 지역에 대한 정보를 공개하지 않고 큰 지역에 대한 소속정보만을 공개하는 것이 일반적이다. 예를 들어 조사 단위가 기업체일 때 읍/면/동 단위로 지역정보를 공개하면 작은 지역에 소재한 기업의 개수가 줄어들어서 노출의 위험성이 커진다. 이럴 때에는 더 큰 지역 단위 즉, 시/군/구 단위에 대한 정보만 공개하여 노출의 위험성을 낮춘다.

조사단위의 자료에 계층적인 구조 (hierarchical structure)를 포함하는 경우 노출의 위험성이 증대한다. 예를 들어서 학생, 학교, 교사에 대한 조사 자료에는 학생이 속한 학급, 학급의 주임교사, 학생과 교사가 속한 학교에 대한 정보가 계층적으로 연결되어 있어서 조사단위의 신분이나 정보를 역 추적하여 알아내기 쉽다. 이러한 계층적인 구조로 가지는 자료는 각 조사단위를 연결하는 변수를 (예를 들어 소속 학교) 공개하지 않으면 노출의 위험성이 많이 줄어든다.

여러 가지 변수들의 조합으로 특정한 성질을 지닌 레코드 유일하거나 또는 드문 경우 해당하는 조사단위의 일부 변수 값을 숨길 수 있다. 이러한 방법을 국소적

자료 감추기(local suppression)이라고 한다. 만약 어떤 레코드가 지역=“강원도”, 직업=“건설기술자”, 성별=“여”, 라고 한다면 이러한 조사단위는 노출의 위험이 크다. 이럴 때 해당하는 조사단위에 대해서 직업을 숨길 수도 있지만 이러한 경우 실제 고용통계에 큰 영향을 미치므로 직업=“결 측”, 지역을 결 측 치로 처리할 수 있다. (지역=“결 측“)

4.3.3 범주화 (Recoding)

범주화(recoding)는 변수가 가질 수 있는 가능한 값들을 합쳐서 몇 개의 구간으로 범주화하거나 (categorization)이나 이미 범주화되어 있는 자료의 값들을 더 넓은 범위의 범주로 합치는 방법이다. 범주화는 변수의 모든 값에 일률적으로 적용하는 전체 범주화(global recoding) 방법이 있고 또는 자료의 일부분에, 특히 매우 작은 값 또는 매우 큰 값과 같은 극단값, 적용하는 극단값 범주화 (top coding or bottom coding) 방법이 있다.

전체범주화의 일반적인 방법은 변수가 가질 수 있는 특성들을 관련된 것으로 묶어서 범주의 포함 범위를 늘리고 범주의 개수를 줄이는 것이다. 예를 들어 사업체에 대한 자료에서 보통 사업체의 업종은 소분류, 중분류, 대분류로 분류된다. 예를 들어서 농가에 대한 자료를 공개하는 경우 소분류를 사용한다면 두 개의 농가가 각각 “소 사육업”과 “양돈업”으로 서로 다른 특성을 가진다. 만약에 이러한 세부 정보가 작은 지역 정보와 결합되어 농가의 노출 위험성이 커진다고 판단되면 소업종보다 범위가 큰 중소업종인 “축산업”으로 공개하면 노출의 위험성이 감소하게 된다. 다른 예는 개인 자료의 공개 시 “토목기술자”와 “화공기술자”와 같이 세부적으로 나뉘어져 있는 직업의 범주를 합쳐서 더 큰 범주인 “기술자”로 공개하는

것도 전체범주화에 위한 노출제한방법이다. 전체범주화를 적용하는 경우 적용할 변수의 수와 범주화의 정도는 공개하고자 하는 마이크로 자료의 특성(노출의 위험성이 있는 중요한 변수들의 개수와 관계)과 외부이용자가 보유하고 있는 매칭을 위한 외부변수들을 종합적으로 고려하여 결정해야 한다.

연속적인 값을 가지는 변수를 범주화하는 방법도 전체범주화 방법이다. 예를 들어 개인의 자료에서 연령(age)이 그대로 공개되면 노출의 위험이 커진다고 판단되었을 때 이를 범주화하여 (예를 들어 19세 미만, 20세에서 29세 미만, ... , 60세 이상) 노출의 위험성을 줄일 수 있다. 연속형 변수를 범주화 하는 경우 범주의 개수가 너무 작으면 정보의 유용성이 감소되어 이용자의 요구에 부응할 수 없게 되며 범주의 개수가 너무 많으면 자료의 유용성은 어느 정도 유지 되지만 노출의 위험성이 크게 줄어들지 않을 수 있으므로 범주화를 하는 경우에는 자료의 특성과 이용자의 요구를 고려하여 적절하게 적용해야 한다.

극단 값 범주화는 변수의 특성이 민감한 정보를 포함하고 있고 또한 매우 작은 값과 매우 큰 값을 가지고 있는 단위들의 수가 적어서 그대로 공개되면 신분의 노출뿐 아니라 민감한 정보의 유출의 가능성이 높은 경우 사용되는 방법이다. 변수의 값이 정해진 기준 값을 초과하거나 또는 기준 값에 미달하면 값 자체를 공개하지 않고 하나의 대표 값 또는 범주로 바꾸어 공개하는 방법이다. 예를 들어 개인 자료를 공개하려는 경우 수득에 관련된 변수를 고려하면 소득이 매우 높은 사람들이나 매우 낮은 사람들의 소득 값을 대표 값이나 범주로 코딩하여 공개한다. 예를 들면 아래 표에서 100명의 사람에 대하여 월 소득 자료가 있다고 가정하자. 월 소득이 매우 큰 사람들의 정보를 숨기기 위하여 월 소득이 1000만 원 이상인 사람들은 월 소득을 기준값(1000만원)으로 모두 대체하거나 하나의 범주로 (1000만 원 이상) 대체할 수 있다 (top coding). 또한 소득이 매우 낮은 사람들에

게도 유사한 방법을 적용하여 노출의 위험을 줄일 수 있다 (bottom coding)

<표 4.1> 소득 자료에 대한 극단값 범주화의 예

ID	월소득(만원) (원자료)	극단값범주화	
		기준값으로 대체	범주로 대체
1	1500	**1000**	**1000 이상**
2	1150	**1000**	**1000 이상**
3	950	950	950
4	870	870	870
5	750	750	750
6	550	550	550
7	450	450	450
8	430	430	430
9	440	440	440
.....			
90	100	100	100
91	95	95	95
92	90	90	90
93	86	86	86
94	85	85	85
95	80	80	80
96	74	74	74
97	50	**50**	**50 이하**
98	45	**50**	**50 이하**
99	30	**50**	**50 이하**
100	25	**50**	**50 이하**

4.3.4 잡음첨가방법 (Noise Addition)

잡음첨가방법(Noise Addition)은 자료의 값이 잡음(noise)을 더하고나 곱하여 원래 자료에 약간의 변형을 가하여 공개하는 방법이다. 중요한 개인 정보를 포함한 변수에 잡음을 첨가하여 변형된 자료를 공개하면 외부이용자가 가지고 있는 외부정보와 차이가 있기 때문에 외부 변수와 결합하여 조사단위의 신분을 알아낼 수 있는 가능성이 적어진다. 더 나아가서 개인의 신분이 노출되어도 만약에 민감한 정보에 대한 변수들에 잡음첨가방법이 적용되었다면 외부이용자가 알아낸 개인의 정보가 실제 값과 차이가 있음으로 직접적인 정보의 누출 피해를 줄일 수 있다.

잡음을 더하거나 곱하는 경우 편이(Bias)를 피하기 위하여 자료의 값에 더해주는 잡음의 평균을 0으로 하고 곱해주는 잡음의 평균은 1이 되도록 한다. 일반적으로 잡음첨가방법은 변수가 연속적인 값을 가지는 변수에 적용되는 방법이지만 특별하게 잡음의 형태를 조종하면 범주형 변수에도 적용이 가능하다.

잡음첨가방법은 실제 자료에 잡음을 더해주거나 곱해주는 방법이므로 구현하기 쉽지만 잡음을 생성할 때 변수의 분포를 고려해야 하기 때문에 변수의 통계적인 특성을 이해해야 하는 어려움이 있다. 노출첨가방법은 개인의 신상정보를 포함하고 있는 식별변수의 노출제한에는 잘 쓰이지 않지만 민감한 정보를 포함하고 있는 변수(소득, 성적 등)에는 적용이 가능하다. 하지만 잡음을 원래 값이 더해주거나 곱해준 후에 변형된 자료의 평균은 크게 변하지 않지만 분산이 증가하는 중요한 단점이 있다. 하나 이상의 변수에 잡음을 첨가하는 경우에는 여러 개의 변수들의 상관관계를 왜곡시킬 위험이 있다. 각 변수에 독립적인 잡음을 첨가하면 원래의 상관관계보다 변형된 자료의 상관관계가 작아진다. 이러한 단점을 보완하기 위하여 서로 상관이 있는 잡음을 여러 개의 변수에 첨가해주면 원래 자료의 상관관

계를 유지할 수 있다. 하지만 무상관 잡음첨가 (uncorrelated noise addition)나 상관 잡음첨가(correlated noise additon) 모두 분산이 증가하는 것을 방지하지는 못한다. 변수의 가질 수 있는 값이 양수인 경우 잡음을 원래 자료에 더해주는 방법은 변형된 자료가 음수가 나올 수 있는 경우가 있기 때문에 이러한 것을 방지할 수 있는 방법은 잡음이 평균이 1인 양수의 값만 가능한 분포에서 생성하여 곱해준다. 이러한 방법을 승법잡음(noise multiplication)이라고 한다.

4.3.4.1 무상관 가법 잡음 (Uncorrelated Noise Addition)

무상관 가법 잡음 (Uncorrelated Noise Addition)은 여러 개의 변수에 각각 서로 독립인 잡음을 첨가하는 방법이다. 잡음이 가지는 분포는 여러 가지 분포가 가능하지만 설명을 위하여 노출제한을 하려는 변수들과 잡음 모두 정규분포를 따른다고 가정한다.

무상관 가법 첨가방법은 원래의 데이터에 평균이0, 분산이$\sigma_{\epsilon_j}^2$인 정규분포를 따르고 공분산이 0인 난수를 추출하여 더한다. 그리고 일반적으로 잡음 ϵ_j의 분산 $\sigma_{\epsilon_j}^2$은 원래변수의 분산σ_j^2에 비례한다고 가정한다. ($\sigma_{\epsilon_j}^2 := \alpha\sigma_j^2$). 원래의 데이터가 p 차원 자료이며 정규분포를 따른다고 가정한다. : $X \sim (\mu, \Sigma)$. 또한 잡음은 평균이 0인 정규분포를 따르며 $\epsilon \sim N(0, \Sigma_\epsilon)$ 주어진 $\alpha > 0$ 에 대하여 잡음의 공분산행렬은 다음과 같다고 가정한다.

$$\Sigma_\epsilon = \alpha \bullet \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p^2 \end{bmatrix}$$

가법 잡음 방법을 행렬로 표시하면 다음과 같이 나타낼 수 있다.

$$Z = X + \epsilon$$

예를 들어 변수가 3개일 때, 무상관 가법 잡음는 잡음의 평균 $E(\epsilon_j)$가 0이기 때문에 원래데이터와 변형된 데이터의 평균이 유지되고

$$E(Z_j) = E(X_j) + E(\epsilon_j) = E(X_j) + 0 = E(X_j) = \mu \quad (j=1,2) \ ,$$

공분산도 잡음의 공분산 $Cov(\epsilon_1, \epsilon_2)$이 0이기 때문에 원래데이터의 공분산과 마스킹 된 데이터의 공분산이 유지 됩니다.

$$Cov(Z_1, Z_2) = Cov(X_1, X_2) + Cov(\epsilon_1, \epsilon_2) = Cov(X_1, X_2) + 0 = Cov(X_1, X_2) \quad (1 \neq 2)$$

그러나 잡음의 분산 $V(\epsilon_j)$이 원래데이터의 분산 $V(X_j)$에 α를 곱한 것과 같기 때문에 변형된 자료의 분산이 원래 자료의 분산보다 크게 된다.

$$V(Z_j) = V(X_j) + V(\epsilon_j) = V(X_j) + \alpha V(X_j) = (1+\alpha) V(X_j) \ ,$$

공분산은 유지되지만 분산이 $V(Z_j) = (1+\alpha) V(X_j)$으로 증가하기 때문에 변형된 자료의 상관계수는 원래 자료의 상관계수보다 작게 나타나다.

$$\rho_{Z_1, Z_2} = \frac{Cov(Z_1, Z_2)}{\sqrt{V(Z_1) V(Z_2)}} = \frac{Cov(X_1, X_2)}{\sqrt{(1+\alpha) V(X_1)(1+\alpha) V(X_2)}}$$

$$= \frac{Cov(X_1, X_2)}{(1+\alpha)\sqrt{V(X_1) V(X_2)}} = \frac{1}{1+\alpha} \rho_{X_1, X_2}$$

4.3.4.2 상관 가법 잡음 (Correlated Noise Addition)

상관 가법 잡음(Correlated Noise Addition)은 여러 개의 변수에 상관관계가 있는 잡음들을 첨가하는 방법이다. 정규분포를 가정하면 원래의 자료의 평균이 0이고 분산은 $\sigma_{\epsilon_j}^2$인 정규분포를 따르고 상관관계가 있는 난수를 생성하여 원래의 자료에 더해주는 방법이다. 무상관 가법잡음 방법과의 차이는 잡음이 정규분포를 따르며 $\epsilon \sim N(0, \Sigma_\epsilon)$ 잡은 ϵ_j의 공분산 행렬 Σ_ϵ은 원래 자료의 공분산행렬 Σ에 비례하게 한다. 이차원 자료의 예를 들면

$$\Sigma_\epsilon = \alpha \bullet \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

따라서 상관 가법 잡음이 더해진 변형된 자료의 공분산행렬은 다음과 같다.

$$\Sigma_Z = \Sigma + \alpha\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} + \alpha \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} = (1+\alpha) \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

상관 가법 잡음는 무상관 가법 잡음와 마찬가지로 잡음의 평균 $E(\epsilon_j)$가 0이기 때문에 원래 자료와 변형된 자료의 평균은 같다.

$$E(Z) = E(X) + E(\epsilon) = E(X) + 0 = E(X) = \mu \ ,$$

분산은 무상관 가법 잡음와 마찬가지로 잡음의 분산 $V(\epsilon_j)$이 원래 자료의 분산 $V(X_j)$ 에 $1+\alpha$를 곱한 것과 같기 때문에 변형된 자료의 분산이 증가한다.

$$V(Z_j) = V(X_j) + V(\epsilon_j) = V(X_j) + \alpha V(X_j) = (1+\alpha) V(X_j) \ .$$

공분산은 잡음의 공분산 $Cov(\epsilon_1,\epsilon_2)$이 원래 자료의 공분산 $Cov(X_1,X_2)$에 $1+\alpha$을 곱한 것과 같기 때문에 원래데이터의 공분산과 변형된 자료의 공분산이 유지되지 않는다.

$$Cov(Z_1,Z_2) = Cov(X_1,X_2) + Cov(\epsilon_1,\epsilon_2) = Cov(X_1,X_2) + \alpha\, Cov(X_1,X_2)$$
$$= (1+\alpha)Cov(X_1,X_2), \quad (1 \neq 2 \text{일 때})$$

상관계수는 공분산이 $Cov(Z_1,Z_2) = (1+\alpha)Cov(X_1,X_2)$으로 유지가 안 되고 분산도 $V(Z_j) = (1+\alpha)V(X_j)$으로 유지가 안 되지만 분자와 분모의 $(1+\alpha)$가 상쇄되므로 원래 자료의 상관계수와 변형된 자료의 상관계수는 같다.

$$\rho_{Z_1,Z_2} = \frac{Cov(Z_1,Z_2)}{\sqrt{V(Z_1)V(Z_2)}}$$
$$= \frac{(1+\alpha)Cov(X_1,X_2)}{\sqrt{(1+\alpha)V(X_1)(1+\alpha)V(X_2)}} = \frac{1+\alpha}{1+\alpha}\frac{Cov(X_1,X_2)}{\sqrt{V(X_1)V(X_2)}} = \rho_{X_1,X_2}$$

4.3.4.3 가법 잡음 예제

위에서 논의한 여러 가지 통계적 성질은 이론적인 성질이기 때문에 실제 잡음을 첨가하는 경우에 자료가 실제 어떻게 변하는지 알아보기 위하여 모의실험을 실시하였다. 원래자료는 초등학생들의 언어(X1),사회(X2),수학(X3) 학업성취도 데이터로부터 20개의 표본으로 뽑고 100점 만점으로 변환한 자료를 사용하였다<표 4.2>.

원 자료의 변수 X1은 평균이 73.65이고 분산이 161.71, 변수 X2는 평균이 73.8이고 분산이 142.59, 변수 X3는 평균이 69.9이고 분산이 142.41이다. 원 자료의 공분산 행렬은 다음과 같다.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix} = \begin{bmatrix} 161.71 & 136.45 & 108.38 \\ 136.45 & 142.59 & 76.98 \\ 108.38 & 76.98 & 142.41 \end{bmatrix}$$

무상관 잡음 첨가를 적용하여 변현된 자료 Z의 공분산은 아래와 같으며

$$\Sigma_Z = \Sigma + \alpha\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix} + \alpha \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix} = \begin{bmatrix} (1+\alpha)\sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & (1+\alpha)\sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & (1+\alpha)\sigma_3^2 \end{bmatrix}$$

상관 잡음 첨가를 적용하여 변현된 자료 Z의 공분산은 아래와 같다.

$$\Sigma_Z = \Sigma + \alpha\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix} + \alpha \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix} = (1+\alpha) \begin{bmatrix} 161.71 & 136.45 & 108.38 \\ 136.45 & 142.59 & 76.98 \\ 108.38 & 76.98 & 142.41 \end{bmatrix}$$

이러한 잡음을 더해주는 방법을 1000번 반복수행하여 변형된 자료의 상관계수와 분산에 어떤 변화가 있는지 알아보았다.

<표 4.3>은 무상관 잡음이 적용된 변형 자료의 한 가지 예이며 (α=0.0609) <표 4.4>는 상관 잡음이 적용된 변형 자료 (α=0.0609)의 예이다.

<표 4.2> 원래 자료

X1 (언어)	X2 (사회)	X3 (수학)
71	67	81
83	79	79
96	85	77
84	89	63
66	62	61
56	62	64
74	75	82
65	71	62
66	68	59
70	66	67
66	68	58
68	77	58
56	50	58
100	93	95
79	76	87
81	78	85
85	89	63
85	90	79
54	54	55
68	77	65

<표 4.3> 무상관 잡음이 적용 된 변형 자료(α=0.0609)

X1 (언어)	X2 (사회)	X3 (수학)
74	70	82
85	80	80
97	83	82
83	88	61
67	63	63
52	58	63
71	73	80
67	74	63
58	61	49
67	64	65
64	68	59
75	83	65
55	51	55
97	90	94
80	78	88
80	75	86
77	80	60
88	90	83
62	60	61
68	75	68

<표 4.4> 상관 잡음이 적용 된 변형 자료(α=0.0609)

X1 (언어)	X2 (사회)	X3 (수학)
74	71	81
77	73	77
94	82	74
84	89	65
67	64	59
56	62	65
73	76	80
64	70	62
70	70	62
69	65	65
66	68	60
68	76	56
52	47	60
96	87	90
80	75	85
79	76	83
85	91	58
86	92	82
51	52	52
69	77	65

(1) 무상관가법잡음의 모의실험 결과

무상관 가법잡음은 원래에 공분산이 0인 난수를 추출하여 α값이 각각 0.0100, 0.0201, 0.0609, 0.1205일 때 상관계수와 분산에 변화가 있는지 알아보았다. 상관계

수는 α값에 따라 원래 자료의 $1/(1+\alpha)$배이고 분산은 원래 자료보다 α값에 따라 $1+\alpha$배 높은 것을 알 수 있습니다. 따라서 표준편차는 $\sqrt{1+\alpha}$ 배 증가한다.

<표 4.5> 무상관가법잡음의 모의실험 결과 - 변형된 자료의 상관계수와 표준편차는 1000개의 모의실험에서 생성된 통계량의 평균값. ()안의 값은 원래 자료의 통계량과의 비율.

상관계수		언어 & 사회	사회 & 수학	언어 & 수학
원래 자료	표준편차 증가	0.8986	0.5402	0.7142
α=0.0100	0.5%	0.8896(0.9900)	0.5349(0.9903)	0.7073(0.9904)
α=0.0201	1%	0.8823(0.9819)	0.5295(0.9802)	0.6993(0.9792)
α=0.0609	3%	0.8488(0.9446)	0.5101(0.9443)	0.6726(0.9418)
α=0.1205	5%	0.8026(0.8931)	0.4843(0.8966)	0.6388(0.8944)
표준편차		언어	사회	수학
원래 자료	표준편차 증가	12.72	11.94	11.93
α=0.0100	0.5%	12.79(1.0054)	12.00(1.0050)	12.01(1.0061)
α=0.0201	1%	12.82(1.0081)	12.05(1.0090)	12.03(1.0082)
α=0.0609	3%	13.10(1.0301)	12.29(1.0289)	12.30(1.0304)
α=0.1205	5%	13.39(1.0532)	12.62(1.0568)	12.63(1.0582)

(2) 상관가법잡음의 모의실험 결과

상관 가법잡음은 원 자료의 공분산을 가지는 난수를 추출하여 α값이 0.0100, 0.0201, 0.0609, 0.1205일때 상관계수와 분산에 변화가 있는지 알아보았다. 상관계수는 α값에 상관없이 항상 유지가 되지만 분산은 무상관가법잡음과 마찬가지로

원래 자료보다 α값에 따라 $1+\alpha$배 높은 것을 알 수 있습니다. 따라서 표준편차는 $\sqrt{1+\alpha}$ 배 증가한다.

<표 4.6> 상관가법잡음의 모의실험 결과 - 변형된 자료의 상관계수와 표준편차는 1000개의 모의실험에서 생성된 통계량의 평균값. ()안의 값은 원래 자료의 통계량과의 비율.

상관계수		언어 & 사회	사회 & 수학	언어 & 수학
원래 자료	표준편차 증가	0.8986	0.5402	0.7142
α=0.0100	0.5%	0.8983(0.9996)	0.5405(1.0006)	0.7143(1.0001)
α=0.0201	1%	0.8983(0.9997)	0.5421(1.0035)	0.7153(1.0015)
α=0.0609	3%	0.8978(0.9991)	0.5389(0.9977)	0.7117(0.9965)
α=0.1205	5%	0.8969(0.9981)	0.5339(0.9883)	0.7099(0.9939)
표준편차		언어	사회	수학
원래 자료	표준편차 증가	12.72	11.94	11.93
α=0.0100	0.5%	12.79(1.0060)	12.02(1.0065)	12.01(1.0063)
α=0.0201	1%	12.85(1.0107)	12.06(1.0097)	12.05(1.0094)
α=0.0609	3%	13.07(1.0280)	12.26(1.0267)	12.27(1.0281)
α=0.1205	5%	13.40(1.0539)	12.58(1.0532)	12.64(1.0595)

4.3.4.4 승법잡음 (Noise Multiplication)

승법잡음은 가법잡음과는 달리 원래 자료 X에 잡음 ϵ 을 곱하여 데이터를 변형하는 방법이다. 원래 자료를 X_j, 잡음을 ϵ_j, 변형된 자료를 Z_j라고 할 때 행렬표기는 다음과 같다.

$$Z_j = X_j \bullet \epsilon_j, \ j = 1, \cdots, n$$

이때, 잡음 ϵ_j는 1이 아니면서 1에 가까운 값을 가지는 것이 효과적이다. 왜냐하면, 잡음 ϵ_j가 1이면 원래 자료와 변형된 자료가 같아지기 때문에 변환의 의미가 없기 때문이다. 하지만 잡음 ϵ_j은 1에서 멀리 떨어진 값을 가질수록 정보의 손실이 커진다. 잡음 ϵ_j과 원래 자료 X_j와 독립이기 때문에 Z_j의 평균은

$$E(Z_j) = E(X_j) \bullet E(\epsilon_j)$$

이고, 분산은

$$V(Z_j) = V(X_j)V(\epsilon_j) + [E(\epsilon_j)]^2 V(X_j) + [E(X_j)]^2 V(\epsilon_j)$$

이다.

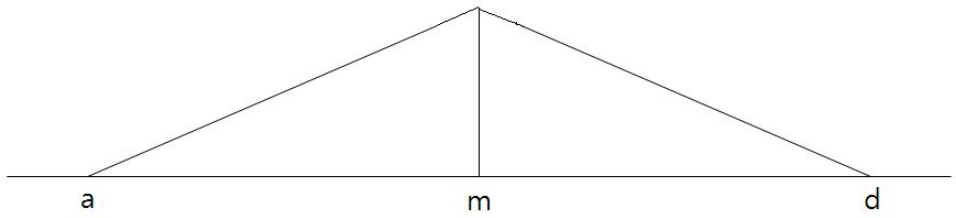
일반적으로 잡음 ϵ에 대한 적절한 분포를 가정하고 난수 ϵ의 값을 추출하는데 승법잡음 방법에서는 삼각분포, 평균μ를 중심으로 좌우가 절단된 정규분포, 절단된 삼각분포, 절단된 삼각분포를 변형한 다양한 형태의 분포 등을 사용한다.

(1) 삼각분포

삼각분포는 자료의 최빈값 m을 중심으로 좌우로 일정부분 절단시킨 분포

를 말하고 좌우가 비대칭인 경우도 있습니다.

<그림 4.2> 삼각분포



잡음의 최소값을 a, 최대값을 d라고 할 때 대칭일 경우의 평균과 분산은 다음과 같다.

$$E(\epsilon) = m, \quad V(\epsilon) = \frac{(d-m)^2}{6}$$

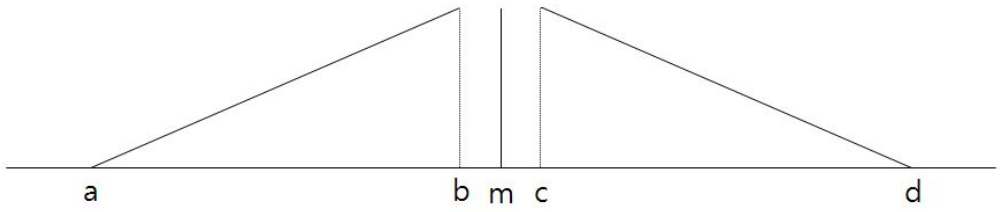
비대칭일경우의 평균과 분산은 다음과 같다.

$$E(\epsilon) = \frac{(m+a+d)}{3}, \quad V(\epsilon) = \frac{[d^2+a^2+m^2-m(a+d)-ad]}{18}$$

(2) 절단된 삼각분포

절단된 삼각분포는 자료의 최빈값 m을 중심으로 좌우로 일정부분 절단시킨 분포를 말하고 좌우가 비대칭인 경우도 있다.

<그림 4.3> 절단된 삼각분포



잡음의 최소값을 a, 최대값을 d, b와 c(>b)를 절사점이라고 할 때 대칭일 경우의 평균과 분산은 다음과 같다.

$$E(\epsilon)=m,\quad V(\epsilon)=\frac{(d+c)^2+2\left[(2m-c)^2-m(m+2d)\right]}{12}$$

비대칭일 경우의 평균과분산은 다음과 같다.

$$E(\epsilon)=\frac{\left\{(d-m)(b-a)^2(2a+a)+(m-a)(d-c)^2(2c+d)\right\}}{3\left\{(b-a)^2(d-m)+(d-c)^2(m-a)\right\}}$$

$$V(\epsilon)=\frac{1}{18\left\{(b-a)^2(d-m)+(d-c)^2(m-a)\right\}}\times\left\{\begin{matrix}(b-a)^6(d-m)^2+(d-c)^6(m-a)^2+(b-a)^2(d-m)(d-c)^2\\(m-a)\left[(3b+a)^2+(3c+d)^2+2(a^2+d^2)-4(2b+a)(2c+d)\right]\end{matrix}\right\}$$

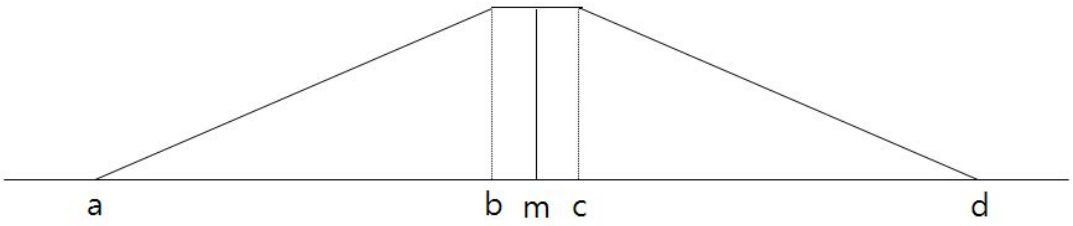
(3) 변형된 삼각분포

(a) 사다리꼴분포

사다리꼴분포는 자료의 최빈값 m을 중심으로 좌우로 일정부분 변형시킨 분포를

말하고 좌우가 비대칭인 경우도 있다.

<그림 4.4> 사다리꼴분포



잡음의 최소값을 a, 최대값을 d, b와 c(>b)를 절사점이라고 할 때 대칭일 경우의 평균과 분산은 다음과 같다.

$$E(\epsilon)=m, \qquad V(\epsilon)=\frac{[d^3+c^3-4m^3+(6m^2+cd)(d+c)-4m(d^2+c^2+cd)]}{6(d+c-2m)}$$

비대칭일 경우의 평균과 분산은 다음과 같다.

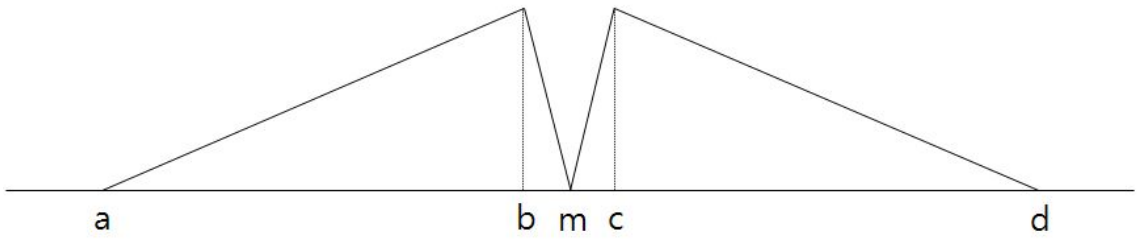
$$E(\epsilon)=\frac{(d^2+c^2-b^2-a^2+cd-ab)}{3(d+c-b-a)}$$

$$V(\epsilon)=\frac{1}{18(d+c-b-a)}3(d^3+c^3-b^3-a^3-a^2b-ab^2+c^2d+cd^2) - \frac{2[(d+c-b-a)(d+c+b+a)+ab-cd]^2}{(d+c-b-a)}$$

(b) 이중삼각분포

이중삼각분포는 자료의 최빈값 m을 중심으로 좌우로 일정부분 변형시킨 분포를 말하고 좌우가 비대칭인 경우도 있다.

<그림 4.5> 이중삼각분포



잡음의 최소값을 a, 최대값을 d, b와 c(>b)를 절사점이라고 할 때 대칭일 경우의 평균과 분산은 다음과 같다.

$$E(\epsilon) = m, \qquad V(\epsilon) = \frac{\left[(d-m)^2 + (c-m)^2 - m(d+c-m) + cd\right]}{6}$$

비대칭일 경우의 평균과 분산은 다음과 같다.

$$E(\epsilon) = \frac{(a+b+c+d+2m)}{6}$$

$$V(\epsilon) = \frac{1}{36}\left\{2a^2 + 2b^2 + 2c^2 + 2d^2 + 2m^2 - m(a+b+c+d) - 2(a+b)(c+d) + ab + cd\right\}$$

4.3.5 자료교환 (Data Swapping)

자료교환(data swapping)은 어떤 영역에 속한 자료를 같은 영역에 속한 다른 자료로 교환하는 방법을 말한다.

데이터 교환은 첫 번째로 Dalenius and Reiss(1978)에 의해 처음 소개되었고, Dalenius and Reiss(1982)에 의해 발전되었다. Dalenius과 Reiss는 범주형 자료에서 도수 분포 (frequency distribution)가 변하지 않도록 하는 자료교환의 절차를 제안하였다. 제안된 절차는 다음과 같은 바람직한 특성이 있다.

(1) 자료와 응답자 사이의 관계를 제거하여 개인 식별이 불가능하다.
(2) 민감하지 않은 변수를 방해하지 않고 민감한 변수에 적용할 수 있다.
(3) 가장 보호가 필요로 하는 부분에 선택적으로 적용할 수 있다

초기의 연구는 대부분 범주형 자료에 대한 방법이었으나 Reiss (1984) and Dalenius (1988)가 연속형 자료에 대하여 그 방법을 확장하였다. 자료교환 방법에 대한 자세한 개관은 Fienberg and McIntyre(2004)에 잘 설명되어 있다.

자료교환은 자료의 값 자체를 변화사키지 않고 그 위치를 바꾸는 방법으로 원래의 자료와 교환된 자료에 의해 계산된 통계량이 유사하게 얻어지도록 구현된다. 자료교환의 방법을 잘 설계하면 자료의 교환 후에도 주어진 일부의 통계량들이 원래 자료의 통계량과 같아지게 또는 매우 유사하게 할 수 있다. 자료가 교환되는 정도와 범위가 커지면 민감한 정보의 유출 가능성이 줄어들지만 자료 교환 전과 후에 대한 통계량의 차이가 커지기 때문에 자료의 유용성은 감소하게 된다. 따

라서 자료교환의 방법을 구현하는 경우 정보유출의 위험성과 자료의 유용성에 대한 상호적인 고려를 하여 주어진 목적을 달성도록 교환방법을 설계한다.

일반적으로 자료를 공개할 때 이용자가 가지고 있는 이부의 정보와 공개된 자료의 정보가 결합되어 응답자의 신분이 노출되고 민감한 정보가 유출된다. 자료교환을 적용한 후 공개를 하면 노출된 응답자의 신분에 대한 불확실성이 커지기 때문에 외부 이용자들의 비밀누출 행위에 대한 시도를 차단하는 효과가 있다.

자료교환은 다른 정보보호방법들보다 비용이 많이 드는 단점이 있다. 자료에 대한 전반적인 구성이나 변수의 특성 및 분포를 알고 이를 이용하여 교환을 하기 때문에 자료의 일부 또는 전체를 간단한 규칙에 따라 변형시키는 방법보다 시간이나 비용이 많이 든다. 또한 구현을 하는 경우 컴퓨터 작업을 하기 위하여 공간과 시간이 필요하다. 하지만 이러한 비용의 지출에 비하여 원래 자료의 유용성에 어느 정도 가까운 공개 자료를 제공할 수 있기 때문에 많은 통계기관들이 자주 사용하고 있다.

4.3.5.1 자료교환의 기본방법

4.3.5.1.1 Dalenius와 reiss의 방법

범주형 자료에 대해 Dalenius와 Reiss가 제안한 기본적인 자료교환의 방법의 예를 들어보자. 현재 어느 기업에 다음과 같이 6명의 직원에 대한 자료가 있다고 가정하자

<표 4.7> 자료교환 전의 원 자료

원 자료

#	직급	성별	징계이력
1	대리	남	없음
2	대리	여	없음
3	과장	남	있음
4	과장	여	없음
5	대리	남	있음
6	과장	여	없음

<표 4.2>에 나타난 자료에서 성별과 직급에 대한 결합 분포를 살펴보면 다음과 같다.

<표 4.8> 원 자료에서 직급과 성별의 결합 분포

		성별		
		남	여	합
직급	대리	2	1	3
	과장	1	2	3
합		3	3	6

외부의 이용자가 다음과 같이 개인 식별에 대한 외부 정보를 가지고 있다면 에서 <표 4.2> 에 나타난 6명의 자료와 결합하여 3번째 사람의 신분을 알아내고 그 사람이 징계를 받은 이력이 있음을 알아낼 수 있다.

<표 4.9> 외부 이용자가 소유한 정보

홍길동	과장	남자

이러한 외부의 식별정보와 공개 자료의 연결을 통하여 민감한 정보가 노출되는 위험성을 줄이기 위하여 자료교환을 실시한다. Dalenius와 Reiss의 자료교환의 방법은 위의 자료에서 직급과 성별의 결합 분포를 유지하며 자료를 다음과 같이 교환한다.

(1) 6명의 직원을 성별이 남자인 집단과 여자인 집단으로 나눈다. 이렇게 나눈 집단의 성별에 대하여 동등집단(equivalent class)이다. 직원 1, 3, 5는 남자집단이고 직원 2, 4, 8은 여자집단이다.
(2) 남자집단에서 2명을 선택하여 직급의 값을 교환한다. 예를 들어 1번과 3번의 남자직원들의 직급을 교환한다. 또한 여자집단에서 2명을 선택하여 직급의 값을 교환한다. 예를 들어 2번과 4번의 여자직원들의 직급을 교환한다.
(3) (2)번에서 얻어진 자료에서 이제는 6명의 직원을 직급에 대한 동등집단으로 나눈다. 직원 1, 2, 6는 과장집단이고 직원 3, 4, 5은 대리집단이다.
(4) 과장집단에서 2명을 선택하여 성별의 값을 교환한다. 예를 들어 1번과 2번의 과장직원들의 성별을 교환한다. 또한 대리집단에서 2명을 선택하여 직급의 값을 교환한다. 예를 들어 3번과 4번의 대리직원들의 성별을 교환한다.

<표 4.10> 직급, 성별에 대하여 자료교환 후의 자료

자료교환 후의 자료

#	직급	성별	징계이력
1	과장	여	없음
2	과장	남	없음
3	대리	여	있음
4	대리	남	없음
5	대리	남	있음
6	과장	여	없음

위와 같이 자료교환을 실행한 후의 자료는 <표 4.5>와 같다. 자료교환을 적용한 자료를 공개하면 외부 이용자가 소유한 식별 자료를 이용하여 그 직원의 징계의 이력은 알 수 있지만 외부 이용자가 알아낸 정보가 실제 값이 아니기 때문에 민감한 자료의 유출은 일어나지 않는다.

<표 4.5>에 주어진 자료교환을 적용한 자료에 대하여 직급과 성별의 결합 분포를 살펴보면 <표 4.3>와 같음을 알 수 있다. 이렇게 자료 교환한 후에도 직급과 성별에 대한 결합 분포는 변하지 않음을 알 수 있다. 따라서 위의 동등집단을 설정하여 축차적으로 실행하는 자료교환 방법은 교환에 쓰인 변수들의 분포는 유지할 수 있다. 하지만 자료교환을 하지 않은 징계이력에 대해서는 주변 분포는 유지되지만 (예로 징계를 받은 사람은 2명) 특정한 직급 또는 성별에 대한 징계이력의 조건부 분포는 유지되지 않는다. 예를 들어 자료교환 전에 여자는 모두 징계이력이 없지만 자료교환 후에는 3명 중에 1명이 징계를 받은 것으로 나타난다. 이렇게 자료교환은 민감한 정보의 유출을 막을 수 있지만 자료의 조건부 분포가 바뀌어 자료의 유용성이 감소하는 결과를 가져올 수 있다.

Dalenius와 reiss의 방법 자료교환의 절차를 t개의 변수에 대하여 적용하는 일반적 방법을 t-차수 도수 계산법(t-order frequency counts) 이라고 부르며 그 절차를 설명하면 다음과 같다.

우선 $X_1, X_2, \dots, X_t$ 를 자료교환을 하기 위한 미리 지정된 t개의 변수라고 하자. t개의 변수
$X_1, X_2, \dots, X_t$ 가 가질 수 있는 값들의 모든 조합을 생각해보자. 예를 들어 X_1 이 가질 수 있는 범주의 개수를 M_1, X_2 가 가질 수 있는 범주의 개수를 M_2, X_t 이 가질 수 있는 범주의 개수를 M_t 라고 한다면 $X_1, X_2, \dots, X_t$ 의 값들이 가질

수 있는 모든 조합의 수를 M 은 다음과 같다.

$$M = M_1 \times M_2 \times ... \times M_t$$

전체의 자료의 개수를 N이라 하면 N개의 자료를 $X_1, X_2, \cdots, X_t$ 이 가지는 값에 따라 M 개의 그룹으로 나눌 수 있다.

예를 들어 앞의 예제에서 t=2개의 변수, 즉 성별(X_1)과 직급(X_2)를 자료교환을 위한 변수로 선택하였으며 (2-차수 도수 계산법) 성별은 2개의 범주(남, 여; $M_1 = 2$) 그리고 직급도 두 개의 범주(대리, 과장; $M_2 = 2$)을 가질 수 있으므로 가능한 범주의 조합은 총 4개이다 ($M = 4 = M_1 M_2$).

이제 하나의 변수를 고정하고 나머지 변수들의 값들이 모두 같은 동등조건을 정의한다. 만약 두 개의 레코드가 X_1을 제외한 $t-1$개의 나머지 변수의 값이 모두 같다면 두 레코드를 X_1 을 제외하고 동등하다고 말한다. 이러한 X_1을 제외한 동등한 레코드들을 모아서 동등집단을 구성할 수 있다. 마찬가지로 모든 t 개의 변수에 대하여 X_i 을 제외한 동등집단을 구성할 수 있다.

예를 들어 예제에서 성별(X_1)을 제외한 동등한 집단은 직급(X_2)의 값이 같은 레코드를 모아 놓은 집단이며 직급의 범주가 대리와 과장이므로 두 개의 집단을 구성할 수 있다. 즉, 직급이 대리인 1,2,5와 직급이 과장인 3,4,6이 X_1을 제외하면 동등하다.

<표 4.11> 성별(X_1)을 제외한 동등한 2개의 집단

#	직급 X_2	성별 X_1	징계이력
1	대리	남	없음
2	대리	여	없음
3	과장	남	있음
4	과장	여	없음
5	대리	남	있음
6	과장	여	없음

t-차수 도수 계산법을 이용한 자료교환은 다음과 같이 실시한다.

(1단계) X_1 을 제외한 동등집단내에서 X_1 의 값들을 임의로 교환한다.

(2단계) X_2 을 제외한 동등집단내에서 X_2 의 값들을 임의로 교환한다.

.....

(t단계) X_t 을 제외한 동등집단내에서 X_t 의 값들을 임의로 교환한다.

4.3.5.1.2 미국 통계청의 방법

미국 통계청(US Census Bureau)는 2000년 전국인구총조사의 자료에 대하여 자료교환을 적용하였다 (Moore 2005). 자료교환의 적용 이유는 총 조사 자료를 이용하여 각종 집계표 또는 교차표를 만들어 사용할 때 표의 셀에 나타나는 도수가 매우 작은 값이어서 신분의 노출 위험성이 있고 또한 총 조사 자료의 일부를 공개할 때 개인의 신분의 노출이 우려되어 이를 방지하기 위함이다. 자료교환의 방법은 다음과 같다.

우선 전체 자료를 교환에 대한 목적 또는 특성(swapping attribute)에 따라 여러 개의 교환집단으로 나눈다. 집단을 나누는 단위는 지역이 될 수도 있고 (예로 미국 총 조사의 경우 County, 우리나라의 경우 읍/면/동) 자료의 성격에 따라 레코드가 모여 있는 같은 특성을 지닌 집단(예를 들어 학교, 병원 등)이 된다. 다음으로 각 교환집단에서 자료교환을 적용할 레코드(record, 예를 들어 가구)의 비율을 결정한다. 이때 자료교환의 비율은 공개하지 않는다. 또한 교환을 하기위한 변수들도 먼저 정하며 교환을 하는 변수들은 민감한 값을 가지고 있는 변수들이다.

교환을 할 레코드의 비율을 정해지면 먼저 몇 개의 변수를 선택하고 그 변수들을 식별변수(key variable)로 정한다. 식별변수는 서로 다른 교환집단에서 개체들를 연결하는 수단이 된다. 교환 집단 내에서 교환할 개체(R1)가 정해지면 다른 집단에서 선택된 개체와 식별변수의 값이 동일한 개체(R2)를 찾는다. 이 두 개체(R1과 R2)에서 교환을 하기로 정한 변수의 값을 교환한다. 이러한 교환 작업을 미리 정해진 레코드의 수만큼 실시한다.

위에서 설명한 미국 통계청 방법에 대한 간단한 예제를 살펴보자. 어느 지역에서 다음과 같은 병원과 의사에 대한 자료가 있다고 가정하자.

<표 4.12> 원래의 자료

원 자료

	교환집단	식별변수	교환하는 변수
#	병원	연령	월소득(만원)
1	A	34	400
2	A	56	700
3	A	45	450
4	B	34	600
5	B	55	1000
6	B	60	1500
7	C	43	460
8	C	56	600

만약에 자료의 이용자가 의사의 이름과 나이, 근무하는 병원에 대한 정보를 가지고 있다면 쉽게 의사의 신분을 알아내고 그 의사의 월 소득을 알 수 있을 것이다. 이러한 정보의 유출을 막기 위하여 다른 교환집단(병원)에 있는 의사 중에 식별변수(나이)의 값이 같은 의사를 찾아 월 소득을 교환 한다 <표 4.7>. 이렇게 자료교환을 하면 자료의 이용자가 의사의 신분을 확인하려는 시도를 막는 효과가 있다. 또한 설사 외부 이용자가 신분을 확인하더라도 유출된 민감한 정보의 신빙성이 떨어져 자료의 유출에 대한 시도를 줄일 수 있다.

<표 4.13> 자료교환에 적용된 자료

자료교환에 적용된 자료

	교환집단	식별변수	교환하는 변수
#	병원	연령	월소득(만원)
1	A	34	400
2	A	56	600
3	A	45	450
4	B	34	600
5	B	55	1000
6	B	60	1500
7	C	43	460
8	C	56	700

4.3.5.2 순위자료교환방법 (Rank-Based Proximity Swapping)

1987년 Brian Greenberg는 발표되지 않은 논문에는 새로운 자료교환 방법을 소개했다. 이 방법을 순위 자료 교환 법 (A Rank-Based Proximity Swapping)이라고 부르며 연속변수에 대해 적용할 수 있는 방법이다 (Moore 2005). Delenius와 Reiss의 방법과는 다르게 연속형 자료에 적용하는 방법이며 선택된 변수의 값을 기준으로 개체의 자료를 순서대로 정렬하여 선택된 변수의 값을 정해진 범위 내에서 교환하는 방법이다. 교환하는 자료의 범위에 제약을 두면 자료의 왜곡을 조절할 수 있으며 자료교환 후에 얻은 통계는 원래 자료로 부터 얻은 통계의 추정값과 유사하지만 교환의 범위에 영향을 받는다.

<표 4.14> 연령 순서로 정렬된 원 자료

연령 순서로 정렬된 원 자료

	교환하는 변수		
#	병원	연령	월소득(만원)
1	A	34	400
2	B	34	600
3	C	43	460
4	A	45	450
5	B	55	1000
6	A	56	700
7	C	56	600
8	B	60	1500

순위자료교환방법에 대한 간단한 예제를 살펴보자. <표 4.6>에 있는 자료를 개체의 연령 순서대로 정렬하면 <표 4.8>과 같은 형태를 얻게 된다. 연령에 범위가 2인 순위자료교환방법을 적용하면 먼저 가장 어린 의사 1번을 연령을 다음으로 큰 값을 가진 두 의사 (2번과 3번)중 하나를 임의로 선택하여 교환한다. 예를 들어 의사 3번이 선택되었다면 의사 1번의 연령 34와 의사 3번의 연령 43을 교환한다. 이러한 절차를 교환되지 않은 개체에 축차적으로 적용하고 교환할 자료가 없으면 중단 한다 <표 4.9>

<표 4.15> 순위자료교환방법이 적용된 자료

순위자료교환방법이 적용된 자료

		원 자료	교환된 자료	
#	병원	연령	연령(원래 #)	월소득(만원)
1	A	34	43 (3)	400
2	B	34	45 (4)	600
3	C	43	34 (1)	460
4	A	45	34 (2)	450
5	B	55	56 (7)	1000
6	A	56	60 (8)	700
7	C	56	55 (5)	600
8	B	60	56 (6)	1500

순위교환이 적용되면 변수에 대한 주변분포는 유지되지만 조건부 분포나 다른 변수와의 결합분포는 변하게 된다. 결합분포가 변하는 정도는 교환을 할 때 선택하는 값들의 범위에 따라 달라진다. 교환의 범위가 커지면 다른 변수와의 상관관계는 크게 변화된다. 예를 들어 <표 4.9>를 보면 원 자료에서 연령과 소득의 상관계수는 0.6900 이지만 순위자료교환이 적용된 연령과 소득과의 상관계수는 0.6369으로 순위자료교환이 적용되면 상관계수가 감소함을 알 수 있다. 만약에 2보다 큰 범위를 사용한다면 상관계수는 더 작게 나타날 것이다. 이렇게 순위자료교환을 적용하는 경우 교환의 범위가 클수록 노출의 위험성이 작아지게 되지만 자료의 유용성이 감소하므로 미리 정해진 노출관리방법의 목적에 맞게 두 요소를 적절히 고려해야 한다.

아래절차는 Greenberg가 제안한 일반적인 순위자료교환방법을 설명하는 것이다.

(1) 레코드의 수가 N인 자료에서 하나의 변수 X를 오름차순으로 정렬한다. 즉,

만약 $i < j$ 이면 $X_i < X_j$ 이다. 모든 정렬된 자료에 “비 교환”이라는 표시를 한다. 이때 극단 값 범주화 (top-coding, bottom-coding) 결 측 값(missing value), 대체 값 (imputed value)는 “교환”이라고 표시한다.

(2) $0 \leq P(X) \leq 100$ 인 $P(X)$ 값 을 결정한다. $P(X)$값은 비율로서 정렬된 X의 값을 교환하는 자료의 범위를 결정한다. 교환을 할 수 있는 자료의 X_i 와 X_j의 범위의 차이는 $j-i$ 이며 이 범위의 차이 $j-i$ 는 $P(X)*N/100$ 보다 작아야 한다. 예를 들어 200개의 레코드가 있다고 가정하고 ($N=200$) $P(X)$ 의 값을 5%라고 하면 자료를 교환할 때 교환을 할 수 있는 자료의 X_i 와 X_j의 범위의 차이는 5*200/100=10 미만이어야 한다. 만약 정렬된 자료 중 첫 번째의 값 X_1은 범위의 차이가 10 미만인 X_2 과 X_{10}중 사이의 하나를 임의로 선택하여 교환한다.

(3) 정렬된 X의 값 X_1를 (2)에 기술된 방법으로 교환할 자료 X_k를 선택하여 교환하고 교환된 두 개의 값에 “교환”이라는 표시를 한다.

(4) 교환되지 않은 가장 낮은 순위 j를 찾는다. X_j 에 대하여 (2)에서 구한 교환의 범위를 계산하고 범위 내에 있는 교환되지 않는 값 중 하나를 선택하여 교환하고 교환된 두 개의 값에 “교환”이라는 표시를 한다.

(5) (4)의 절차를 교환할 값이 없어질 때까지 계속한다.

(6) 교환을 원하는 다른 변수들 $Y, Z, ..$에 대하여 (1)-(5)의 절차를 차례로 적용하여 순위자료교환을 실시한다.

Greenberg의 순위교환방법에서 교환의 범위를 지정하는 비율 $P(\bullet)$의 값이 작으

면 교환이 이루어진 변수들의 상관계수 또는 다른 다변량 통계 값이 원래 자료의 값과 크게 차이나지 않는다. 하지만 $P(\bullet)$ 값이 너무 크면 상관계수 또는 다른 다변량 통계 값이 크게 왜곡되어 자료의 유용성이 현저히 떨어지므로 이를 유의해야 한다. Greenberg의 논문에서는 $P(\bullet)$값들을 변화시키면서 자료교환을 하여 적절한 통계량이 얻어지는 $P(\bullet)$값을 선택하라고 제안하였다.

[순위교환방법의 예제] 점포수와 매출액 교환

사업체의 자료에서 매출액과 점포수가 나타나있는데 점포수와 매출액에 순위교환방법을 적용하여 외부 이용자가 매출액과 점포수의 정보를 조합하여 사업체의 신분을 알아내는 위험성을 줄이려고 한다.

(1) 다음은 점포수에 순위교환방법을 적용하는 절차이다.

Step1. 자료를 점포수에 대하여 정렬한다.

Step2. 자료의 수(N)=25, P(a)=20으로 정하기로 한다.

Step3. j=2일 때 , M=min[25, 2+20*(25/100)]=7 이므로 구간 [3, 7]에서 교환되지 않은 임의의 자료인 k=6을 선택한다.

Step4. a_2=(16), a_6=(21) 을 교환한다.

Step5. j=3일 때 M=min[25, 3+20*(25/100)]=8 이므로 구간 [4, 8]에서 교환되지 않

은 임의의 자료인 k=5을 선택한다.

Step6. a_3=(18), a_5=(20) 을 교환한다.

Step7. 모든 순위들이 교환이라고 표시될 때까지 위의 4, 5의 과정을 반복함

점포수의 원 자료에 대한 순위교환을 통해 최종결과 파일은 각 응답자에 대한 정보를 충분히 정보보호하게 된다. 즉 원 자료에서 자회사의 수와 매출액의 상관관계는 0.42인데 교환 후의 상관관계는 0.41로 교환 후에도 상관관계는 거의 변하지 않음을 나타내고 있다.

<표 4.16> 매출액에 순위교환방법을 적용

원 자료		
기업체 번호	점포수	매출액
1	5	15
2	18	1600
3	24	2100
4	19	1100
5	16	20100
6	20	1600
7	119	15100
8	24	3100
9	34	1900
10	35	1200
11	26	1000
12	26	2100
13	32	2000
14	61	2600
15	72	3100
16	28	21000
17	31	2500
18	229	16000
19	41	4000
20	65	2800
21	53	2100
22	34	1900
23	21	3000
24	42	2900
25	81	3500

점포수 교환과정											
기업체 번호	점포수	매출액	순위	교환전	1단계	점포수	2단계	점포수	…….	마지막 단계	점포수 최종교환
1	3	15	1	coding	교환	3	교환	3		교환	3
5	16	20100	2		교환	21	교환	21		교환	21
2	18	1600	3			18	교환	20		교환	20
4	19	1100	4			19		19		교환	24
6	20	1600	5			20	교환	18		교환	18
23	21	3000	6		교환	16	교환	16		교환	16
3	24	2100	7			24		24		교환	24
8	24	3100	8			24		24		교환	19
11	26	1000	9			26		26		교환	31
12	26	2100	10			26		26		교환	34
16	28	21000	11			28		28		교환	28
17	31	2500	12			31		31		교환	26
13	32	2000	13			32		32		교환	41
22	34	1900	14			34		34		교환	26
10	35	1200	15			35		35		교환	40
9	40	1900	16			40		40		교환	35
19	41	4000	17			41		41		교환	32
24	42	2900	18			42		42		교환	72
21	53	2100	19			53		53		교환	61
14	61	2600	20			61		61		교환	53
20	65	2800	21			65		65		교환	81
15	72	3100	22			72		72		교환	42
25	81	3500	23			81		81		교환	65
7	119	15100	24	coding	교환	119	교환	119		교환	119
18	229	16000	25	coding	교환	229	교환	229		교환	229

(2) 다음은 매출액에 순위교환방법을 적용하는 절차이다.

Step1. 점포수에 순위교환방법을 적용한 자료를 매출액에 대하여 정렬한다.

Step2. 자료의 수(N)=25, P(a)=30으로 정하기로 한다.

Step3. j=2일 때 , M=min[25, 2+30*(25/100)]=10 이므로 구간 [3, 9]에서 교환되지 않은 임의의 자료인 k=8을 선택한다.

Step4. a_2=(1000), a_8=(1900) 을 교환한다.

Step5. j=3일 때 M=min[25, 3+30*(25/100)]=7 이므로 구간 [4, 10]에서 교환되지 않은 임의의 자료인 k=10을 선택한다.

Step6. a_3=(1100), a_{10}=(2100) 을 교환한다.

Step7. 모든 순위들이 교환이라고 표시될 때까지 위의 4, 5의 과정을 반복함

<표 4.17> 매출액에 순위교환방법을 적용

원자료		
기업체 번호	점포수	매출액
1	5	50
2	18	1600
3	24	2100
4	19	1100
5	16	20100
6	20	1600
7	119	15100
8	24	3100
9	34	1900
10	35	1200
11	26	1000
12	26	2100
13	32	2000
14	61	2600
15	72	3100
16	28	21000
17	31	2500
18	229	16000
19	41	4000
20	65	2800
21	53	2100
22	34	1900
23	21	3000
24	42	2900
25	81	3500

매출액 교환과정												
기업체 번호	점포수	매출액	순위	교환전	1단계	매출액	2단계	매출액	…….	마지막 단계	매출액 최종교환	점포수 최종교환
1	5	50	1	coding	교환	50		50		교환	50	3
11	26	1000	2		교환	1900		1900		교환	1900	21
4	19	1100	3			1100	교환	2100		교환	2100	20
10	35	1200	4			1200		1200		교환	1600	24
2	18	1600	5			1600		1600		교환	1200	18
6	20	1600	6			1600		1600		교환	2100	16
9	34	1900	7			1900		1900		교환	2000	24
22	34	1900	8		교환	1000		1000		교환	1000	19
13	32	2000	9			2000		2000		교환	1900	31
3	24	2100	10			2100	교환	1100		교환	1100	34
12	26	2100	11			2100		2100		교환	3000	28
21	53	2100	12			2100		2100		교환	1600	26
17	31	2500	13			2500		2500		교환	2800	41
14	61	2600	14			2600		2600		교환	3500	26
20	65	2800	15			2800		2800		교환	2500	40
24	42	2900	16			2900		2900		교환	3100	35
23	21	3000	17			3000		3000		교환	2100	32
8	24	3100	18			3100		3100		교환	20100	72
15	72	3100	19			3100		3100		교환	2600	61
25	81	3500	20			3500		3500		교환	2900	53
19	41	4000	21			4000		4000		교환	4000	81
7	119	15100	22			15100		15100		교환	15100	42
18	229	16000	23			16000		16000		교환	21000	65
5	16	20100	24			20100		20100		교환	3100	119
16	28	21000	25			21000		21000		교환	16000	229

매출액의 원 자료는 순위교환을 통해 각 응답자에 대한 정보를 충분히 정보보호하게 된다. 즉 원 자료에서 매출액과 점포수의 상관관계는 0.42인데 교환 후의 상관관계는 0.59로 교환 후에도 상관관계는 점포수에 비해 조금 더 변한 것을 나타내고 있다. 또한 최종 변환된 점포수와 매출액의 상관관계는 0.54로 처음 0.42에 비해 달라진 것을 볼 수 있다.

이러한 상관관계를 유지하기 위해서 위에서 설명한 바와 같이 p(a) 값을 줄여주는 것이 좋다. 그래서 현재 순위교환에 관한 최근 연구는 파일의 왜곡이 적도록 교환을 관리하는 방법에 중점을 두고 있으며 현재 p(a)을 어느 기준으로 두어야 가장 효율적인지에 대한 연구방법에 중점을 두고 있다.

4.3.5.3 자료교환의 장단점

모든 자료교환의 절차는 다음과 같은 장점을 가진다.

(1) 자료교환은 각 응답자에 대해서 정확히 정보를 보호한다. 즉 개인적인 자료를 서로 교환하여 원 자료와 교환된 자료가 일치하지 않음으로 개인정보노출을 보호한다.

(2) 만약 모든 잠재적인 식별변수에 자료교환이 실행되면 (변수 값은 응답자자료의 연결에 기여할 것이다.), 교환은 자료와 응답자 간의 모든 관계를 제거한다. 즉 변수 값과 응답자간의 연결고기를 제거함으로서 개인정보노출을 보호할 수 있다.

(3) 절차가 매우 간단하다. 즉 마이크로데이터와 랜덤 난수 생성이외에는 아무것

도 필요하지 않음으로 programming이 매우 간단하다.

(4) 다른 변수에 대한 응답을 방해하지 않고 민감한 변수에 적용할 수 있다. 즉 민감하지 않다고 생각하는 변수에 대해서는 아무런 교환을 하지 않고, 민감하다고 생각하는 변수에 대해서만 자료교환을 해도 된다.

(5) 연속형 변수의 교환은 그것이 가장 보호가 필요로 하는 부분에 선택적으로 적용할 수 있다. 희귀하고 독특한 응답은 일반적으로 응답자를 식별하는데 사용될 것임으로 이 값은 변경될 가능성이 매우 높다. 하지만 자주 반복되는 응답은 침입자에게 가치가 작을 것이고 교환에 의해서 바뀔 가능성이 적을 것이다.

(6) 절차는 연속변수, 범주형 변수(인종, 성별, 직업)에 제한되지 않고 둘 다 적용이 가능하다.

자료교환 절차는 다음과 같은 단점을 가진다.

(1) 임의의 범주형 자료교환은 색다른 조합과 함께 수많은 조합을 생산할 수 있다. 범주형의 경우 이상한 조합을 만들어 파일의 유용성을 감소시킬 수 있다. 예를 들어 남자 출산과 같은 조합을 뜻한다. 또한 연속형 변수에서도 발생 가능하다. 예를 들어 점원의 낮은 소득이 뇌 외과의사의 높은 소득과 서로 교환될 수 있다.

(2) 파일에서 교환의 대상이 되는 자료 수와 변수의 수이다. 즉 오리지널 파일과 교환된 파일을 저장하는 상당한 시간과 교환하고 난 후에 원 자료와 교환 후 자료에 대한 컴퓨터 저장 공간을 상당히 요구한다.

(3) 자료교환은 마이크로데이터의 분석 가치를 상당히 약화시킬 수 있다.
비록 교환이 전제 모집단에서 일변량 분석에 영향을 미치지 않을 지라도 모든 하위 도메인에서 분석에 영향을 미칠 것이다. 예를 들어 관리인의 수입에 분산과 평균인 조건부 분포가 변한다. 또한 서로간의 다변량 관계를 깨버릴 수 있다. 예를 들어 두개 또는 그 이상의 변수의 회귀분석과 상관분석에 대한 유용성이 약화된다.

4.3.6 표본추출

전체자료에서 자료의 일부분을 표본 추출하여 일부분만 공개하는 것이 표본추출에 의한 노출제한방법이다. 일반적으로 전수자료(예를 들어 센서스 자료)를 공개하는 경우 자료의 일부분만 표본추출을 통해서 공개하고 더 나아가 원 자료가 표본추출로 얻어진 자료라 하더라도 다시 표본자료에서 일부만을 추출하여 부분자료만 공개하는 경우도 있다.

일단 표본으로 추출된 자료는 전체자료 일부분이기 때문에 추출한 것만으로도 자료의 노출 위험성이 감소한다. 하지만 표본으로 추출된 자료라 하더라도 언제나 노출의 위험성이 충분히 낮아서 다른 노출제한방법을 적용할 필요성이 없는 것은 아니다. 따라서 표본으로 추출되어 공개되는 일부의 자료에 대해서도 노출에 대항 위험성을 다시 확인하여 필요한 경우 적절한 노출제한방법을 다시 적용해야 한다.

예를 들어 캐나다 통계청에서는 센서스 자료의 약 3%를 일반에게 공개한다. 일반인이 사용할 수 있는 자료를 공개하는 경우에 규모가 작은 지역에 대한 자료는

공개하지 않는다. 공개 자료는 선택된 센서스 구역과부구역, 도시 지역 자료만 주로 공개되며 민감한 변수에 대해서도 노출제한방법을 적용한 후에 공개한다. 외부 이용자가 개인이나 가구의 신분을 알아내기 위하여 사용할 수 있는 중요한 변수들 중에 일부만 공개하며 변수의 값 중 극단 값으로 판단되는 것들은 값을 감추고 공개한다. 또한, 가구의 소득 같은 민감한 자료는 최소값이나 최대값이 나타나지 않도록 감추거나 구간화(coding)하여 공개한다.

표본추출에 의한 노출제한방법을 적용하면 외부 이용자가 외부변수와 공개변수의 조합 연결로 찾을 수 있는 유일한 조사 단위들이 실제로 모집단에서의 유일한 단위인지를 확신할 수 없다. 따라서 표본추출은 이러한 정보 유출을 위한 외부 공격의 의지를 약화시키는 기능도 있다.

4.3.7 인위자료 (Synthetic data)

인위자료(synthetic data)는 앞에서 살펴본 통계적 노출 기법과 매우 다르다. 주로 전통적인 노출 제한 기법은 원래 자료를 바꾸거나 변형시켜서 노출의 위험성를 줄인다. 인위자료는 원자료를 생성하는 가상의 통계적 모형(statistical model)을 가정하고 원래 자료를 가상의 통계적 모형에서 발생한 모의자료로 대체하는 방법을 말한다. 모의자료는 원 자료와 직접적인 관련이 없는 자료이지만 원자료를 생성하는 가상의 모형에서 추출된 자료이기 때문에 원자료가 지닌 통계적인 특성들을 가진다. 따라서 이러한 인위자료에 기반한 분석은 원자료에 기반한 분석과 매우 유사할 것이다. 이러한 인위자료의 공개는 노출의 위험성을 크게 줄일 수 있다. 이러한 장점이 있음에도 불구하고 자료를 생성하는 적절한 통계적인 모형을 찾는 것은 매우 어려운 작업이므로 아직까지 널리 쓰이지 못하는 상황이나 근래

에는 많은 연구가 활발히 이루어지고 있다 (Ducan et al 2011).

[인위자료 예제]

원래자료는 중학생들의 언어(X1),사회(X2),수학(X3) 학업성취도 데이터로부터 340개의 표본으로 뽑고 100점 만점으로 변환한 자료를 사용하였다 <표 4.18>.

원 자료의 변수 X1은 평균이 70.97이고 분산이 101.88, 변수 X2는 평균이 68.87이고 분산이 126.41, 변수 X3는 평균이 67.54이고 분산이 123.58이다. 원자료의 공분산 행렬은 다음과 같다.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix} = \begin{bmatrix} 101.88 & 85.90 & 80.36 \\ 85.90 & 126.41 & 93.47 \\ 80.36 & 93.47 & 123.58 \end{bmatrix}$$

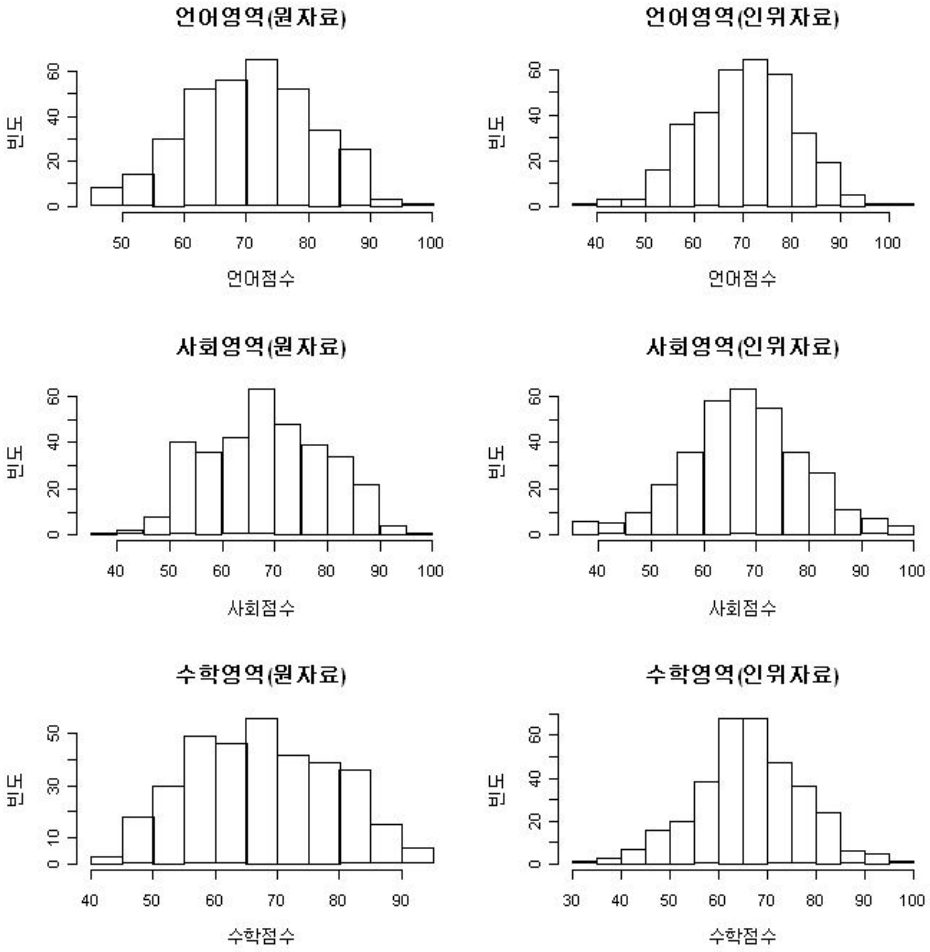
인위자료는 원래 자료와 평균과 분산이 같은 삼 변량 정규분포에서 340개의 난수를 발생시켜서 만들 수 있다. 이때 다변량 정규분포에서 생성된 난수는 실수값 이므로 소수 첫째자리에서 반올림하여 자연수로 만들어 주어야한다 <표 4.19>.

인위자료의 변수 X1은 평균이 70.94이고 분산이 106.35, 변수 X2는 평균이 68.20이고 분산이 131.11, 변수 X3은 평균이 66.83이고 분산이 117.93이다. 인위자료의 공분산 행렬은 다음과 같다.

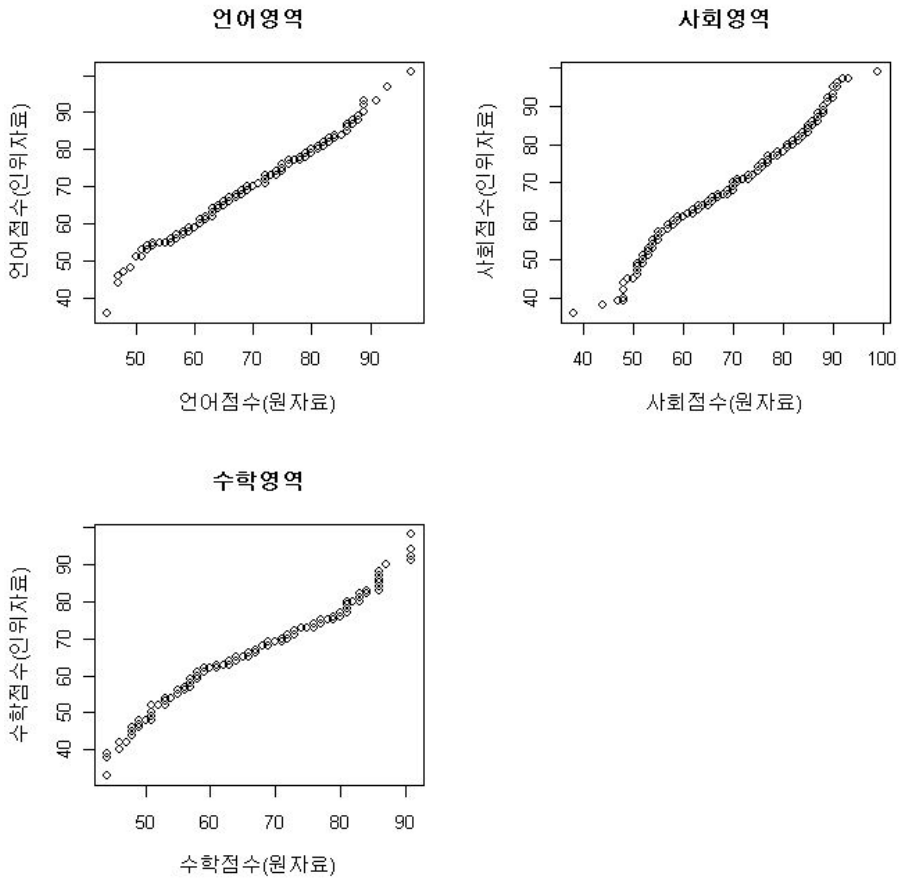
$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix} = \begin{bmatrix} 106.35 & 88.04 & 77.74 \\ 88.04 & 131.11 & 93.24 \\ 77.74 & 93.24 & 117.93 \end{bmatrix}$$

언어영역의 히스토그램을 보면 원래 자료와 인위자료 모두 대칭인 종 모양(bell shape)이지만 원래자료보다 인위자료가 꼬리가 더 얇다. 이는 원래자료의 분포가 정규분포에 가깝지만 꼬리가 더 두터운 분포를 가짐을 알 수 있다. 사회영역은 언어영역과 비슷한 경향을 보인다. 수학영역은 원래 자료가 두터운 꼬리를 기진 분포로 정규분포를 이용하여 모형화하는 것이 부족할 수 있다 <그림 4.6>. 언어영역과 사회영역에서 원래자료와 인위자료의 Q-Q 그림을 보면 45도 직선상에 점들이 분포되어 있으므로 원래자료의 인위자료의 분포가 같다. 하지만 수학영역은 원래 자료와 인위자료의 Q-Q 그림을 보면 꼬리가 정규분포보다 두터운 것을 보여준다<그림 4.7>. 이렇게 원래자료를 모형화하여 인위자료를 만드는 방법은 유용하고 간편하지만 원래 자료의 특성을 잘 반영하는 통계적 모형을 찾는 것은 쉬운 일이 아니다.

<그림 4.6> 원래 자료와 인위 자료의 히스토그램



<그림 4.7> 원래자료와 인위자료의 Q-Q 그림



<표 4.18> 원래 자료

X1 (언어)	X2 (사회)	X3 (수학)
71	67	81
83	79	79
96	85	77
84	89	63
66	62	61
56	62	64
74	75	82
65	71	62
66	68	59
70	66	67
66	68	58
68	77	58
56	50	58
100	93	95
79	76	87
81	78	85
85	89	63
85	90	79
54	54	55
68	77	65

<표 4.19> 인위자료

X1 (언어)	X2 (사회)	X3 (수학)
74	70	82
85	80	80
97	83	82
83	88	61
67	63	63
52	58	63
71	73	80
67	74	63
58	61	49
67	64	65
64	68	59
75	83	65
55	51	55
97	90	94
80	78	88
80	75	86
77	80	60
88	90	83
62	60	61
68	75	68

제 5장. 위험노출과 자료의 유용성

개인의 정보보호와 자료의 유용성은 타협이 있다. 즉, 개인의 정보보호 정도를 높이면 자료의 유용성은 낮아지고 자료의 유용성을 높이면 개인의 정보보호 정도는 낮아진다. 따라서 정보제공기관은 안전한 자료를 제공함으로써 노출의 위험(R)을 낮추어 개인의 정보를 보호하는 동시에 분석적으로 유용한 자료를 제공함으로써 자료의 유용성(U)을 높여야 한다. 따라서, 자료를 보호하는 문제는 두가지 기준을 모두 고려하여야 한다. 여러 가지 정보보호 수준에서 (R, U)를 계산함으로써 어떻게 R과 U의 상호 득실이 발생하는 지 평가하여야 한다. 이 득실 관계는 R-U 정보보호 지도 (R-U confidentiality map)을 이용하여 조사할 수 있다 (Duncan 등, 2001). 또한, 여러 가지 다른 정보보호 방법에 관한 R-U 지도를 그리고 비교함으로써 (예를 들어 잡음첨가방법(noise addition)방법에 관한 R-U 지도와 자료교환(data swapping)에 관한 R-U 지도를 비교함으로써) 좀 더 적절한 방법을 선택할 수 있다.

5.1 위험노출과 자료 유용성의 기본 개념

노출은 자료유출 시도자가 자료안의 개인이 누구인 지 알아내는 경우에 발생할 수 있다. 이를 신원노출이라고 한다. 또한 자료유출 시도자가 자료 안의 개인 정보를 제공된 것 보다 더 많이 알아내는 경우 노출이 발생한다. 이를 속성노출이라고 한다(Duncan 과 Lambert, 1989). 자료유출 시도자의 공격으로부터 안전한 데이터를 제공하기 위해서는 이전 장들에서 배운 통계적 노출 제한 기법들을 적용하여 데이터를 조작하여야 한다. 이런 방법들은 노출 위험 R을 낮추는데 효과적이다. 하지만 R을 낮추게 되면 자료의 유용성 U 또한 낮아지게 된다. 원 자료를

매스킹된 자료로 바꾸는 경우의 영향을 보도록 하자. 먼저 자료 안 속성값의 범위에 따라 셀을 구성하여 자료를 범주화시키고 각 셀의 빈도를 조사한다. 예를 들어 월수입을 100만원 단위로 하여 빈도표를 만들어 자료를 제공할 수 있다(예: 월수입이 200만원에서 300만원 사이의 인원수). 식별변수를 범주화하게 되면 노출위험을 낮출 수 있다. 하지만 자료 이용자에게는 이런 범주화가 분석을 어렵게 만들 수 있다. 예를 들어 교육정도가 월수입에 어느 정도 영향을 미치는지 파악하기 위하여 범주화된 월수입을 종속변수로 교육정도를 독립변수로 하여 회귀분석을 하는 경우 연속형으로 측정된 월수입을 종속변수로 하는 경우보다 분석이 어려워지며 정보의 손실이 발생하게 된다. 만일 범주형 자료를 좀 더 작은 범주로 재구성하게 되는 경우는 문제가 더 심각해 질 수 있다. 예를 들어, 남자와 여자를 합쳐서 하나의 범주로 구성하게 되면 성별이라는 정보는 완전히 없어지게 된다. 제공된 자료를 이용하여 아주 중요한 연구를 하거나 정책을 결정해야 하는 경우 이렇게 범주화된 자료의 유용성 U가 낮기 때문에 제공된 자료에 만족하지 못할 수 있다. 현재 다양한 통계적 노출제한 기법이 있으며 각 방법은 노출위험과 유용성에 고유한 영향을 미친다.

5.1.1 통계적 노출제한 기법에서 모수의 선택

일반적으로 노출의 위험과 유용성을 평가하기 위하여 R과 U는 수치값으로 제시한다. 하나의 통계적 노출제한 기법에서 노출의 위험 R과 자료의 유용성 U는 그 기법의 모수들에 의존하게 된다. 예를 들어, swapping 기법에서는 swapping rate이 모수가 된다. 일반적으로 “가장 좋은” 모수값을 선택하기 위한 두 가지 방법이 있다.

(1) 고정된 임계값 ρ보다 작은 노출 위험을 갖는 유용성 중에 유용성을 최대화

하는 모수를 구한다. 즉, $\max U : R \leq \rho$.

⑵ R과 U의 가중 평균값을 최대화하는 모수를 찾는다. 즉, $\max \lambda R + (1-\lambda) U$ for some λ.

R에 제약을 준 상태에서 U를 최대화하는 첫 번째 방법이 좀 더 많이 사용된다. 이는 가설검정의 과정에서 제1종의 오류를 범할 확률 α를 고정된 어떤 수준에 둔 상태에서 검정력을 최대화하는 전략을 사용하는 방법과 유사하여 통계학자에게 좀 더 친숙한 방법이라 할 수 있다. 방법 1에 비해 방법 2는 λ를 선택하여야 하고 R과 U의 척도가 다른 경우를 고려해야 하는 문제가 있다. 방법 1은 정보노출의 위험이 임계값 아래에 있는 경우 매스킹된 자료가 제공되므로 정보보호를 좀 더 강조한다. 물론 방법 1도 제약점이 있다. 임계값 ρ는 실제로는 다중 속성일 것이다. 예를 들어, 합계표 자료를 보호하는 방법에서 ρ는 모든 민감한 셀과 모든 자료유출 시도자에 대하여 정보제공기관이 요구하는 모든 노출제한 수준의 집합이 된다. 또한 이 방법은 극단적 기준(extreme criteria)을 초래할 수 도 있는 점에서 비판받기도 한다. 예를 들어, R과 U모두 0과 100사이의 값을 갖으며 정보보호를 위한 임계값 ρ는 9라고 하자. 이제 공공에게 제공될 데이터가 2개 (D1과 D2) 있다고 하자. 이 때 D1의 R과 U는 각각 8.999와 10이고 D2의 R과 U는 각각 9.001과 90이다. 1번 방법을 적용하면 제공될 데이터로 D1을 선택한다. 하지만 이 경우 D2가 제공하기 좀 더 좋은 자료라고 할 수 있을 것이다. 사실 여러 가지 요소를 고려하여 하나의 모형을 선택하는 경우 여러 가지 요소를 모두 만족하는 모형을 찾는 것이 불가능한 경우가 많으므로 이러한 선택의 문제에 빠질 수밖에 없다.

5.2 자료의 유용성 측정

지금까지 위험노출을 측정하는 방법에 관하여 논의하였다. 이제 자료의 유용성을 측정하는 몇 가지 방법을 살펴보자. 자료의 유용성은 통계적 추론에 있어 자료의 값과 관련이 있으므로 통계학자에게 좀 더 친숙하다. 일반적으로 자료의 유용성을 측정하는 척도는 통계적 검정에서 검정력, 평균제곱오차, 신뢰구간으로 표현될 수 있다. 예를 들어 Agrawal과 Srikant (2000)은 95%신뢰구간의 넓이로 자료의 유용성 척도를 제시하였다. 통계적 노출 제한기법을 적용하여 원 자료를 매스킹 자료로 변환하게 되면 자료의 유용성은 낮아지는데 자료의 양이 작아지는 것처럼 보이지 않기 때문에 이 유용성의 손실이 모든 자료 이용자에게 명백한 것은 아닐 수 있다. 따라서 자료제공기관은 매스킹된 자료를 제공하는 경우 매스킹으로 인한 자료의 손실이 있음을 자료 이용자에게 고지해야 하며 이 손실정도의 측정값을 제공하여야 한다. 자료의 손실은 원 자료를 이용한 추론의 정확성에 비하여 얼마나 매스킹 자료를 이용한 추론이 정확한지를 평가하여 측정할 수 있다. 하지만 자료 이용자가 어떤 방법을 이용하여 통계적 추론을 할지 모르기 때문에 자료제공기관이 이러한 정확성을 비교하기에는 어려움이 있다. 대신에 자료제공기관은 원 자료의 일반적 정보구조와 크게 다르지 않은 매스킹 자료를 제공할 수 있다. 이런 경우 자료의 유용성은 정보손실척도(information loss metrics)에 근거하여 측정할 수 있다. 정보의 손실을 최소화하여 제공되는 데이터가 원 자료의 정보를 충분히 보존하기를 원한다.

정보의 손실을 평가하기 위한 두 가지 보완적인 방법은 다음과 같다.

(1) 원 자료와 매스킹된 자료를 비교한다.
(2) 원 자료에서 계산한 어떠한 통계량과 매스킹된 자료에서 계산한 그 통계량을

비교한다.

첫 번째 방법을 이용하여 측정한 정보의 손실을 “뒤틀림 측정값 (distortion measure)”라고 한다 (Gomatam and Karr, 2005). 통계량의 비교를 이용하는 두 번째 방법을 이용하여 측정한 정보의 손실은 “대리 측정값 (proxy measure)”라고 한다.

예제를 통하여 정보손실척도가 어떻게 작용하는 지 보자. 자료유출 시도자가 레코드 결합방법을 통하여 정보를 습득하고자 하는 경우 일반적으로 사용하는 식별변수는 범주형 인구학적 지리학적 식별자(identifier)일 것이다. 회귀분석을 실시하는 경우 이러한 식별자들은 독립변수로 사용될 것이다. 이런 경우 정보제공기관이 범주를 합하는 방법으로 매스킹을 한 경우에 다음과 같이 두 가지 방법으로 정보손실 정도를 측정할 수 있다.

(1) 회귀모형의 예측력의 손실은 R2의 함수로 표현될 수 있다.
(2) 변수나 범주가 합쳐지는 경우 정보의 손실은 엔트로피를 이용하여 측정할 수 있다.

5.3 유용성의 직접 측정

위에서 언급한 몇 가지 방법들의 단점은 유용성을 직접 측정하지 않는다는 것이다. 이는 데이터 이용자가 실시하고자 하는 특정한 분석과 자료의 유용성과의 관계를 고려하지 않았다는 것이다. 이러한 측정값들은 노출제한기법이 분석적 검정력에 미치는 영향에 대한 대리자이다. 여기서 분석적 검정력이란 하나의 주어진

데이터로부터 자료 이용자가 올바른 결론을 내릴 능력을 의미한다. 분석적 검정력은 통계적 검정력과 관련은 있으나 별개의 개념이다. 분석적 검정력은 분석적 완비성 (analytical completeness)과 분석적 유효성(analytical validity)이라는 두 가지 요소로 구성된다. 분석적 완비성은 자료의 이용자가 시행하고 싶은 분석을 시행할 수 있는지를 측정하는 개념적 척도이다.

분석적 유효성은 만일 자료 이용자가 완벽한 데이터를 소유하고 있다면 같은 분석을 통하여 같은 결론을 도출할 수 있는지를 측정하는 척도이다. 이 두 가지 모두 노출제한기법 외의 요소에 영향을 받는다. 예를 들면, 완비성은 처음에 수집된 자료와 모든 형태의 데이터 오류에 영향을 받는다. 하지만 완비성 및 유효성은 노출제한기법에 영향을 많이 받게 된다. 범주를 합치게 되면 작은 빈도를 가진 범주를 고려해서 실시해야 하는 분석은 할 수가 없게 되어 완비성이 낮아진다. 예를 들어 마이크로 데이터에서 지리학적 임계값을 사용하여 작은 단위의 지역을 합치게 되면 그러한 작은 단위 지역에 관한 추론을 할 수 없게 된다. 반면에 자료를 조작하는 경우에는 같은 통계적 방법을 적용하더라도 원 자료를 이용한 결과와 조작된 자료를 이용한 결과와 상이해질 수 있다. 일반적으로 재코딩을 이용한 정보보호방법은 완비성을 저하시키고 자료를 조작하는 정보보호방법은 유효성을 저하시킨다.

5.4 R-U 정보보호 지도

많은 자료제공기관들은 경험적 방법으로 정보노출 위험과 자료의 유용성 간 득과 실을 평가하고 있다. 하지만 자료제공기관은 이러한 득실을 좀 더 체계적으로 평가할 수 있는 분석적 기법을 사용하여야한다. 이 장에서는 그러한 방법으로 R-U

정보보호 지도를 소개한다.

이 기법은 Duncan과 Fienberg (1999)가 합계표 자료에서 정보노출 위험과 자료의 유용성 간 득과 실을 평가하는 방법으로 처음 소개하였다. 이 장에서는 Duncan 등 (2001)이 제안한 연속형 마이크로데이터의 맥락에서의 R-U 정보보호 지도 방법을 소개한다.

R-U 정보보호지도 방법은 통계적 노출제한 기법의 모수값을 달리할 때 노출위험 R과 정보의 유용성 U에 결합적으로 미치는 영향을 추적한다. 가장 기본적인 R-U 정보보호지도는 데이터를 제공하기 위한 하나의 방법에서 모수를 달리하여 R과 U를 평가한 후 그 값의 집합으로 표현하여 R과 U의 득실을 보여준다. 예를 들어, 잡음첨가방법으로 데이터를 매스킹하는 경우 오차의 분산인 σ^2의 값을 증가시키면서 R과 U를 측정하고 R을 Y축에 U를 X축에 표시함으로써 지도를 완성한다. 이 때 σ^2의 값을 증가시키는 것은 좀 더 높은 정보보호를 실시한다는 의미이다. R-U 정보보호지도 방법은 매우 일반적이며 원칙적으로는 어떠한 통계노출제한 기법에도 다 사용할 수 있다는 장점이 있다. 이제 두 개의 다른 맥락에서 R-U 정보보호 지도를 설명한다.

5.4.1 다변량 잡음첨가방법를 이용한 통계적 노출제한 방법에서의 R-U 정보보호 지도 작성방법

이제 다변량 잡음첨가방법을 이용한 통계적 노출제한 방법에서 어떻게 R-U 정보보호 지도를 작성하는 지 보도록 하자. 먼저 자료제공기관이 자료 이용자와 자료 유출 시도자에 관하여 다음과 같은 두 가지 특성을 따른다고 가정하자.

(1) 자료 이용자. 자료의 유용성 U는 모집단 평균 벡터 μ의 임의의 선형결합인 $c'\mu$의 추정에서 발생한 자료 이용자의 평균제곱오차(mean squared error)의 역수라고 한다.

(2) 자료유출 시도자. 노출위험 R은 자료유출 시도자가 모집단의 한 개인에게서 목표하는 속성값인 τ를 추론하는데서 발생하는 평균제곱오차의 역수라고 한다. 여기서 자료제공기관은 신원노출이 아닌 속성노출 제한에 더 중점을 둔다. 즉, 개인의 신원이 파악되더라도 그 개인의 어떤 속성이 노출되지 않기를 바라는 것이다. 잡은첨가방법은 자료유출 시도자가 자료의 레코드가 어떤 특정한 개인의 것이라는 것을 알아도 실제 속성값은 모르게 하는 노출제한기법이다.

위의 두 특성을 고려하여 이제 자료유출 시도자가 목표로 하는 속성값인 τ에 대하여 알고 있는 상태가 다음과 같이 두 가지로 나뉜다고 가정하자. 먼저 자료유출 시도자는 단지 τ가 일반 모집단 값들 중에 하나라는 사실은 알고 있다. 이를 모집단 인식 상태(population knowledge state)라고 한다. 이 상태에서 자료유출 시도자가 알고 있는 개인의 속성값 τ의 확률분포는 모집단에서의 속성값 τ의 분포와 같다. 두 번째 상태는 레코드 인식 상태(record knowledge state)라고 한다. 이 상태에서는 자료유출 시도자가 목적으로 하는 레코드가 있는 변수 X에서 이 자료유출 시도자가 특정한 레코드를 파악하기 위한 충분한 외부의 정보를 알고 있다.

만일 데이터가 모집단에서 추출될 추출비가 작은 경우와 자료유출 시도자가 목적으로 하는 개인이 표본에 속해 있을 것이라고 확신하기 어려운 경우에 이 모집단 인식 상태를 가정하는 것이 적절하다. 이에 비해, 자료유출 시도자가 충분한 외부의 정보를 이용하여 목적으로 하는 레코드를 연결할 수 있다고 생각되는 경우에

는 레코드 인식 상태를 가정하는 것이 적절하다.

이제, n개의 레코드가 p개의 속성을 가지고 있는 원 자료를 생각해보자. 이 데이터를 행렬의 형태로 표현하면 다음과 같다. $X=[X_{ij}]=[X_1,...,X_n]^T$. 여기서 레코드는 평균벡터 μ와 분산-공분산 행렬 Σ를 갖는 모집단으로부터 무작위 추출된 표본이라고 하자. 따라서 $X_i \overset{iid}{\sim} (\mu, \Sigma)$ 라고 표현할 수 있으며 잡음첨가방법을 이용하여 매스킹한 자료는 다음과 같이 표현할 수 있다.

$$Y=X+\epsilon \text{ where } \epsilon \sim (0, \lambda^2, \Sigma)$$

자료의 유용성: 데이터 사용자는 모집단의 평균 벡터 μ를 매스킹된 자료의 표본 평균 벡터 $\hat{\mu}=\overline{X}$를 이용하여 추정한다. 이 경우 $E(\hat{\mu})=\mu$이고 $Var(\hat{\mu})=\frac{1+\lambda^2}{n}\Sigma$이다. 이전에 설명한대로 자료의 유용성 U는 모집단 평균 벡터 μ의 임의의 선형결합인 $c'\mu$를 추정할 때 발생하는 평균제곱오차의 역수로 표현된다. 이를 구하여 보면 다음과 같다.

$$U=\frac{n}{1+\lambda^2}(c^T\Sigma c)^{-1}$$

노출 위험: 자료유출 시도자의 목적은 특정한 개체의 신원을 노출하는 것이고 모집단 인식 상태에 있다고 가정하자. 자료유출 시도자는 속성 j에서 특정한 속성값 τ를 목적으로 하고 $\hat{\tau}=\overline{Y_j}$를 이용하면 노출의 위험은 다음과 같이 표현된다.

$$R = \frac{1}{E(\tau - \hat{\tau})^2} = \frac{n}{(1+\lambda^2)\sigma_j^2 + n(\mu_j - \tau)^2}$$

이제 자료유출 시도자가 목적 속성값 τ와 같은 매스킹된 레코드를 평가할 수 있는 상태인 레코드 인식 상태를 고려하여 보자. 이 경우 노출의 위험은 자료제공기관에 최악이 된다. 여기서 만일 자료유출 시도자가 $\hat{\tau} = Y_{ij} = \tau + \epsilon_{ij}$를 이용하면 노출 위험 R은 다음과 같다.

$$R = \frac{1}{E(\tau - \hat{\tau})^2} = \frac{1}{E(\epsilon_{ij})^2} = \frac{1}{\lambda^2 \sigma_j^2}$$

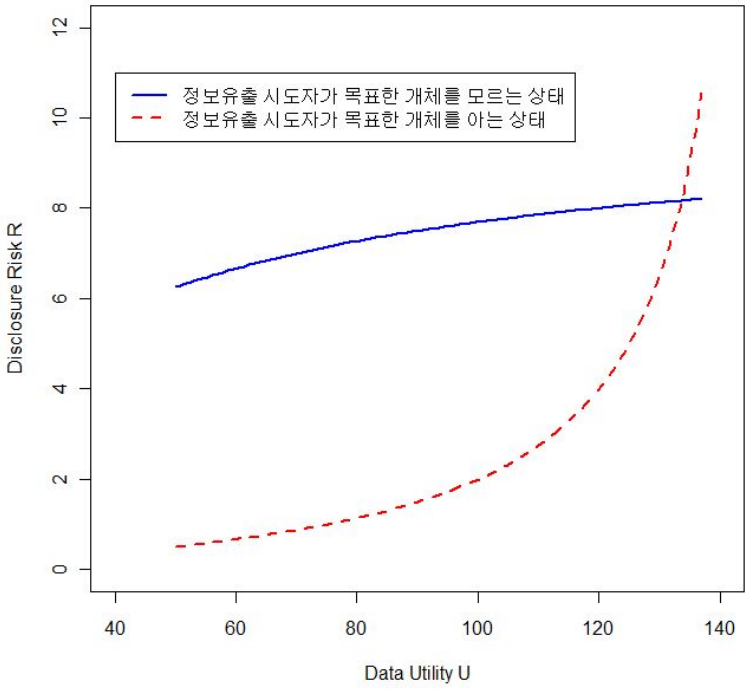
여기서 데이터 매스킹을 하지 않는 경우의 노출위험은 $\sigma_j^2 = 0$일 때의 R이며 $\sigma_j^2 = 0$이면 R은 무한대이다. 자료유출 시도자가 레코드 연결(record linkage)을 할 수 있는 경우, 원 자료를 제공하는 것은 명백히 정보보호에 위협이 된다.

목표하는 개체를 알고 있는 경우, 자료유출 시도자는 목적 속성값의 추정값으로 항상 $\hat{\tau} = Y_i^j$를 사용하여야 자료유출 시도자에게 좋은가? 위의 두 개의 R에 관한 식을 비교하면 자료유출 시도자는 $\lambda^2 > \left(\frac{n}{n-1}\right)\left(\frac{\tau - \mu_j}{\sigma_j}\right)^2 + \frac{1}{n-1}$인 경우 $\hat{\tau} = \overline{Y_j}$를 이용하는 것이 자료유출 시도자에게 좋다. 따라서 정보의 노출위험을 줄이기 위해서는 충분한 잡음을 더하는 것이 좋다.

위의 두 개의 노출위험 R을 고려한 R-U 정보보호 지도를 그려보면 아래와 같다. 이 지도를 그리기 위하여 $n = 50$, $\sigma_j^2 = 1$, $(c^T \Sigma c)^{-1} = 3$, $(\mu_j - \tau)^2 = 0.1$로 가정하였다.

이 R-U 지도는 자료유출 시도자가 목표한 개체를 아는 경우 (푸른 실선)와 목표한 개체를 모르는 경우 (붉은 점선) 노출 제한 모수 λ^2가 증가함에 따라 자료의 유용성과 노출 위험의 변화를 (즉, 영향정도) 보여준다. 이 지도에서 알 수 있는 점은 잡음이 충분히 큰 경우 목표한 개체를 아는 것이 자료유출 시도자에게 도움이 되지 않는다. 만일 대략 U < 133.6인 경우 자료유출 시도자는 목표한 개체를 모르는 것이 낫다.

<그림 5.1> R-U 정보보호 지도



이러한 R-U 지도 방법은 두 개의 다른 자료유출 시도자의 인식 상태를 비교하는데 유용하게 사용할 수 있다.

< 용어집 >

결합 (linking)

제공된 데이터 내의 개체와 외부자료에서 이미 식별된 개체와 매칭하는 과정

계층적 데이터 (hierarchical data)

한 변수에 속한 값들이 다시 다른 값들로 구성된 데이터 (예: 지리적 위치)

구간값 제공 (interval publication)

매크로데이터에서 특정한 셀의 값이 구간으로 대체되는 정보보호 방법. 각 구간은 원 값을 포함하고 있어야 한다.

극단값 범주화 (top coding)

어떤 임계값보다 높은 값들은 개인자료 대신 합계만 제공하여 정보를 보호하는 방법

노출 위험 측정값 (risk metrics)

주어진 가정 하에 정보노출이 발생할 가능성을 측정하기 위한 양적 지표

다변량 잡음 첨가 (multivariate additive noise)

원 자료에 잡음의 확률벡터를 첨가하여 제공될 데이터를 구성하는 정보보호 방법. 확률벡터의 다변량 분포는 적절한 데이터의 유용성을 유지하면서 노출의 위험을 줄일 수 있도록 정한다.

레코드 감추기 (record suppression)

마이크로자료에서 특정한 성질을 가진 단위를 공개하지 않는 정보보호 방법

레코드 결합 (record linkage)

다른 데이터 셋에서 레코드를 결합하는 과정.

마이크로 그룹화 (microaggregation)

레코드를 그룹화하는 정보보호 방법. 개인 레코드의 실제값을 제공하는 대신 일반적으로 그룹의 평균을 제공한다.

마이크로데이터 (microdata)

개별 레코드로 구성된 데이터. 각 레코드는 개인, 가구, 사업체 등이 될 수 있다.

매스킹 (masking)

원 자료의 확률적(stochastic) 또는 결정적(deterministic) 변환을 통한 모든 정보보호 방법

매크로데이터 (macrodata)

요약된 자료. 대부분의 경우 범주형의 속성을 가지며 분할표의 형태로 제공된다.

모집단 유일성 (population uniqueness)

한 모집단에서 주어진 변수의 집합에 대하여 값들의 조합이 유일한 개체의 비율

목적 변수 (target variable)

데이터 안에서 자료노출 시도자가 관심있어 하는 변수. 이 목적 변수의 값은 일반적으로 노출 시도자에게 알려져 있지 않으며 민감하다.

민감성 (sensitivity)

주어진 모집단 개체에 관하여 정보가 노출되는 경우 발생한 손상의 질적/주관적 척도

반올림 (rounding)

매크로데이터에서 값이 특정한 기저로 반올림되는 정보보호 방법

분석적 완비성 (analytical completeness)

노출제한기법이 적용되기 전의 자료와 적용된 후의 자료에서 같은 자료 분석법을 적용할 수 있는 지의 정도

분석적 유효성 (analytical validity)

노출제한기법이 적용되기 전의 자료와 적용된 후의 자료에서 같은 자료 통계적 추론 결과를 도출할 수 있는 지의 정도

변수 감추기 (attribute suppression)

하나 이상의 변수의 모든 값 또는 몇 몇 값을 제거하는 통계적 노출제한 기법

변조 (perturbation)

한 개체의 값을 임의 또는 규칙적으로 바꾸는 모든 정보보호 방법

사전정보 (priori knowledge)

자료유출 시도자가 모집단 개인에 관한 정보를 노출을 촉진하는 정보. 사전정보는 모집단 개인에 관하여 이미 알고 있는 정보일 수도 있고 데이터에 적용되는 노출제한과정에 관한 정보일 수도 있다.

셀 감추기 (cell suppression)

매크로데이터에서 특정한 셀의 빈도를 제공하지 않는 정보보호 기법

셀 변조 (cell perturbation)

매크로데이터에서 모든 또는 몇 몇 셀의 빈도를 어떤 구간 내의 값으로 대체하는 정보보호 기법

속성 (attribution)

데이터의 정보와 특정한 모집단 개체와의 관계

속성 노출 (attribute disclosure)

데이터 내에서 어떤 모집단 개체의 신원파악 없이 그 개체의 정보를 노출하는 것. 전형적인 속성노출은 합계 데이터에서 개인에 관한 추론이 가능한 경우이다.

수학적 프로그래밍 (mathematical programming)

최적화(optimization) 문제를 푸는 수학적 모형과 알고리즘을 다루는 모든 수학 분야.

순환 셀 변조 (cyclic cell perturbation)

매크로 데이터에서 모든 셀의 빈도는 다른 빈도값으로 대체되는 정보보호 기법. 이 방법은 자료 순환이라고 불리우는 비슷한 양식의 4개 이상의 셀의 빈도를 임의로 몇 셀은 1씩 더하고 다른 셀은 1씩 빼서 행과 열의 합은 유지한다.

식별변수 (key variable)

노출 위험을 높이는 속성변수들의 조합

식별자 제거 (deidentification)

데이터에서 직접적인 신원 식별자를 제거

신원파악 (identification)

알려진 모집단 개체의 신원과 특정한 마이크로데이터의 레코드와의 관계

위험한 셀 (risky cell)

표에서 정보 노출의 위험이 높아 보이는 셀, 민감한 셀 (sensitive cell)이라고도 함.

익명화 (anonymization)

데이터베이스에서 개체들의 신원파악을 불가능하게 만드는 과정. 즉, 신원파악의 위험을 매우 작게 만드는 과정.

인위 자료 (synthetic data)

모형을 만들고 원 자료로부터 모형을 추정하여 확률적으로 생성한 자료로 원 자료를 대체하는 정보보호 방법

임의 반올림 (random rounding)

매크로데이터에서 행의 합과 열의 합이 셀의 반올림된 값에 따라 변하는 반올림 기법. 이 방법에서 반올림의 방향은 확률을 이용하여 결정한다. 이 방법은 쉽고 생성된 표가 불편성(unbiasedness)을 가진다. 하지만 주변 셀(marginal cell)이 같은 방법으로 수정되면 생성된 표에서 셀의 합과 주변이 다를 수 있다.

자료 감추기 (local suppression)

마이크로데이터 내의 특정한 값을 결측으로 코딩하는 정보보호 기법.

자료 교환 (data swapping)

원 자료의 어떤 영역에 속한 속성값을 같은 영역에 속한 다른 속성값과 교환하는 정보보호 기법.

자료 변이 (data divergence)

두 개의 데이터 셋 간의 기록된 정보의 차이 (이를 데이터-실제 변이라고 함) 또는 하나의 데이터 셋과 실제 간의 기록된 정보의 차이 (이를 데이터-데이터 변이라고 함). 자료변이는 입력의 차이, 응답오차, 조사 시점의 차이 등으로 발생할 수 있다.

자료유출 시도자 (data snooper)

데이터 내에서 모집단 개체의 신원을 파악하고자 하는 개인, 그룹, 또는 기관. 일반적으로 자료유출 시도자는 이미 알고 있는 정보와 데이터에 포함된 정보를 통계적 결합방법을 이용하여 신원노출을 시도한다.

자료 제공 (data dissemination)

데이터가 자료 이용자에게 제공되는 과정

잡음 첨가 (noise addition)

원 자료의 몇 몇 값 또는 모든 값에 확률적으로 잡음을 첨가하여 정보를 보호하는 방법

재신원 파악 (reidentification)

정보보호 방법을 적용하여 제공한 데이터에서 개인의 신원을 다시 파악해 내는 과정

전체 범주화 (global recoding)

전체 자료에 일률적으로 적용하는 범주화

정보노출 감사 (disclosure auditing)

정보보호 기법이 적용된 데이터에서 정보가 잘 보호되었는 지 검사하는 과정. 합계 자료인 경우 위험하다고 생각되는 모든 셀의 최소값과 최대값을 계산하여 감사를 할 수 있다.

정보제공기관 (data stewardship organization)

모집단에 관한 자료를 수집하고 가공하고 제공하는 기관. 정보제공기관은 개인의 정보를 보호하고 또한 높은 유용성을 가진 정보를 자료 이용자에게 제공하는 법적, 도덕적 의무를 갖는다.

제거 (subtraction)

셀의 구성원들 중에서 이미 알려진 모집단 개체를 제거한 후 남은 구성원들로 셀의 빈도를 만들어 정보노출을 시도하는 방법.

직접적인 인식정보 (direct identifier)

이름, 주민등록번호와 같이 자료 내 개체를 직접적으로 인식할 수 있는 정보. 이는 두 개 이상의 변수의 조합일 수도 있다 (예: 주소와 나이)

통계적 노출 (statistical disclosure)

모집단 개인에 관하여 정보가 추론되는 과정

통계적 노출 제한 (statistical disclosure limitation)

통계 분석의 목적으로 제공되는 데이터에서 정보노출 위험을 낮추는 과정, 통계적 노출 통제 (statistical limitation control)이라고도 함.

통제된 반올림 (controlled rounding)

매크로데이터에서 모든 셀의 값을 한정 집단 내의 값으로 대체하는 정보보호 기법. 일반적으로 주어진 기저값(예: 5또는 10)으로 반올림하여 대체한다. 이 방법은 주변 합을 유지할 수 있다.

평행한 변이 (parallel divergence)

두 개의 데이터 셋 내의 값은 같으나 이 두 값 모두 실제 값과 동일하게 다른 변이.

표 보정 관리 (controlled tabular adjustment)

매크로데이터에서 모든 셀의 빈도는 다른 빈도값으로 대체되는 정보보호 기법. 이 방법은 셀 변조 방법 중 하나이다. 이 방법은 위험에 노출된 셀의 보호범위를 지정하고 범위의 하한값이나 상한값 중 하나로 그 셀의 빈도를 대체한다. 가법성을 보장하기 위하여 위험하지 않은 셀은 작은 값을 더하거나 빼서 보정한다.

표본 유일성 (sample uniqueness)

한 표본에서 주어진 변수의 집합에 대하여 값들의 조합이 유일한 개체의 비율

합계 데이터 (aggregated data)

모집단 또는 표본 전체의 정보를 요약한 자료. 요약 통계값, 빈도표 등이 합계 데이터의 예이다.

n 규칙 (n rule)

표에서 위험한 셀을 찾는 과정. 임계값 n이 주어진 경우, 한 셀의 빈도가 n이하인 경우 그 셀은 위험하다고 분류한다.

p/q 규칙 (p/q rule)

표에서 어떤 셀이 위험한 지 탐지하는 규칙. n 규칙보다 좀 더 정교하다. 각 셀에 속하는 개인의 숫자 (즉, 빈도)와 더불어 주된 영향을 미치는 개체의 값들을 같이 고려한 규칙이다.

R-U 정보보호 지도 (R-U confidentiality map)

주어진 정보보호 방법에서 정보보호의 정도가 높아짐에 따라 노출 위험과 데

이터 유용성의 관계를 평면에 나타낸 그래프

< 참고문헌 >

Agrawal, R., Srikant, R. (2000) Privacy-preserving data mining. Proceedings of the 2000 ACMSIGMOD on Management of Data, Dallas, TX, 15 - 18.

Bacharach, M. (1966) Matrix rounding problem, Management Science, 9, 732 - 742.

Benedetti, R., Franconi, L., Capobianchi, A. (2003) Individual risk of disclosure using sampling design information. Istat Contributi n. 14/2003. Available at http://www.istat.it/dati/pubbsci/contributi/Contr_anno2003.htm.

Chowdhurry, S. D., Duncan, G. T., Krishnan, R., Roehrig, S. F., Mukherjee, S. (1999) Disclosure detection in multivariate categorical database: Auditing Confidentiality Protection through Two New Matrix Operatiors, Management Science, 45, 1710-1723.

Cox, L. H., Kelly, J. P., and Patil, R. (2004) Balancing quality and confidentiality for multivariate tabular data, Privacy in Statistical Databases, Lecture Notes in Computer Science, Domingo-Ferrer, J., Torra, V. (eds.), vol. 3050, Springer: New York.

Duncan, G. T., Elliot, M., and Salazar-Gonzalez, J-J. (2011) Statistical Confidentiality, Springer: New York.

Duncan, G.T., Fienberg, S.E. (1999) Obtaining information while preserving privacy: aMarkov perturbation method for tabular data. Eurostat. Proceedings of Statistical Data Protection 98,

Lisbon, pp. 351 - 362.

Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R., Roehrig, S.F. (2001) Disclosure limitation methods and information loss for tabular data. In: Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (eds.) Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 135 - 166. North-Holland, Amsterdam.

Duncan, G.T., Jabine, T.B., de Wolf, V.A. (eds.) (1993) Panel on Confidentiality and Data Access, Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, National Research Council and the Social Science Research Council, Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics. National Academy of Sciences, Washington, DC.

Duncan, G.T., Lambert, D. (1989) The risk of disclosure for microdata. J. Bus. Econ. Stat. 7, 207 - 217.

Duncan, G. T., Roehrig, S. F. (2007) Reconciling information privacy and information access in a globalized technology society, Database Technologies: Concepts, Methodologies, Tools, and Applications, Erickson, J. (ed.), 1823 - 1843.

Elliot, M.J. (2000) DIS: a new approach to the measurement of statistical disclosure risk. Risk Manage. Int. J. 2(4), 39 - 48.

Elliot, M.J., Dale, A. (1998) Disclosure risk for microdata. End of Framework IV project report to the European Union.

Elamir, E., Skinner, C.J. (2006) Record level measures of disclosure

risk for survey microdata. J. Official Stat. 22, 525 – 539.

Evans, T., Zayatz, L., and Slanta, J. (1998) Using noise for disclosure limitation of establishment tabular data, Journal of Official Statistics, 14, 537 – 551.

Fienberg, S.E., McIntyre, J. (2004) Data swapping: variations on a theme by Dalenius and Reiss. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. Lecture Notes in Computer Science, vol. 3050, pp. 14 – –29. Springer, Berlin, Heidelberg.

Fischetti, M. and Salazar, J. J. (2003) Partial cell suppression: a new methodology for statistical disclosure control, Statistics and Computing, 13, 13 – 21.

Gomatam, S., Karr, A.F., Sanil, A.P. (2005) Data Swapping as a Decision Problem. J. Off. Stat. 21(4), 635 – 655.

Griffin, R. A., Navarro, A., and Flores-Baez, L. (1989) Disclosure avoidance for the 1990 Census, Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, 516 – 521.

Kelly, J. P., Golden, B. L., and Assad, A. A.(1990) Using simulated annealing to solve controlled rounding problems, ORSA Journal on Computing, 2, 174 – 185.

Kelly, J. P., Golden, B. L., and Assad, A. A. (1993) Large-scale controlled rounding using tabu search with strategic oscillation, Annals of Operations Research, 41, 69 – 84.

Krenzke, T., Roey, S., Dohrmann, S, Mohadjer, L., Huang, W.-C., Kaufman, S., Seastrom, M. (2006) Tactics for Reducing the Risk

of Disclosure Using the NCES DataSwap Software, 1650 Research Blvd., Rockville, MD 20850 National Center for Education Statistics, 1990K St., N.W., Washington, D.C.

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., Walford, N. (1991) The case for samples of anonymized records from the 1991 census. J. R. Stat. Soc. Ser. A 154, 305 - 340.

McCullagh, K. (2007) Data sensitivity: proposals for resolving the conundrum. J. Int. Commer. Law Technol. 2(4), 190 - 201.

Navarro, A., Flores-Baez, L., and Thompson, J. (1988) Results of data switching simulation, Presentation at the Spring meeting of the American Statistical Association and Population Statistics Census Advisory Committees, Washington, DC.

Polettini, S., Stander, J. (2004) A Bayesian hierarchical model approach to risk estimation in statistical disclosure limitation. In Domingo-Ferrer, J., Torra, V. (eds.) Privacy in Statistical Databases, pp. 247 - 261. Springer, Berlin.

Reiss, S.P. (1984) Practical data-swapping: the first steps. ACM Trans. Database Syst. 9, 20 - -37.

Richard A. Moore, Jr. (2005) CONTROLLED DATA-SWAPPING TECHNIQUES FOR MASKING PUBLIC USE MICRODATA SETS, Statistical Research Division, US Bureau of the Census

Salazar, J. J. (2006) Controlled rounding and cell perturbation: statistical disclosure limitation methods for tabular data, Mathematical Programming. 105, 251 - 274.

Salazar, J. J. (2010) Branch-and-cut versus cut-and-branch algorithms for cell suppression, Privacy in Statistical Database, Lecture Notes in Computer Science, Domingo-Ferrer, J., Magkos, E. (eds), Springer: Berlin.

Salazar, J. J., Lowthian, P., Young, C., Merola, G., Bond, S., and Brown, D. (2004) Getting the best results in controlled rounding with the least effort, Privacy in Statistical Databases, Lecture Notes in Computer Science, Domingo-Ferrer, J. (ed.), Springer: New York.

Skinner, C.J., Holmes, D.J. (1998) Estimating the re-identification risk per record in microdata. J. Off. Stat. 14, 361 - 372.

Smith, D., Elliot, M.J (2005) An experiment in Naive Bayesian record linkage. Proceedings of Conference of the International Statistical Institute, Sydney.

Smith, D., Elliot, M.J. (2008) A measure of disclosure risk for tables of counts. Trans. Data Priv. 1(1), 34 - 52 With Smith, D.

Stephen E. Fienberg and Julie McIntyr (2004) Data Swapping: Variations on a Theme by Dalenius and Reiss, in Privacy in Statistical Databases, Lecture Notes in Computer Science, springer, 2004

Willenborg, L.C.R., and de Waal, T. (2001) Elements of Statistical Disclosure Control. Lecture Notes in Statistics, vol. 155. Springer; New York.

STATISTICAL POLICY WORKING PAPER 22 (Second version, 2005)

Report on Statistical Disclosure Limitation Methodology (2005), Federal Committee on Statistical Methodology, Statistical and Science Policy-Office of Information and Regulatory Affairs-Office of Management and Budget

NISS Data Confidentiality Technical Panel : Final Report (2011) Alan Karr, Education Statistics Services Institute American Institutes for Research

Statistical Confidentiality Principles and Practice (2011). George T. Duncan in carnegie Mellon University Santa Fe, NM87505, USA, Mark Elliot Juan-Jose Salazar-Gonzalez IN University of Manchester, UK

〈부록 1〉

통계적 정보보호 방법 기초과정 (안)

	1일차	2일차	3일차
09:10 ~ 10:00 (1교시)	교육안내 -교육운영-	마이크로 데이터 정보보호 - 노출제한 방법 1	위험노출과 자료의 유용성 1
10:10 ~ 11:00 (2교시)	개인 정보보호의 개념 1	마이크로 데이터 정보보호 - 노출제한 방법 2	위험노출과 자료의 유용성 2
11:10 ~ 12:00 (3교시)	개인 정보보호의 개념 2	마이크로 데이터 정보보호 - 노출제한 방법 3	예제 실습 1
12:00 ~ 13:10	중 식		
13:10 ~ 14:00 (4교시)	기본 통계 I	매크로 데이터 정보보호 - 노출위험 평가 방법 1	예제 실습 2
14:10 ~ 15:00 (5교시)	기본 통계 II	매크로 데이터 정보보호 - 노출위험 평가 방법 2	예제 실습 2
15:10 ~ 16:00 (6교시)	마이크로 데이터 정보보호 - 노출위험 평가 방법 1	매크로 데이터 정보보호 - 노출제한 방법 1	수료식
16:10 ~ 17:00 (7교시)	마이크로 데이터 정보보호 - 노출위험 평가 방법 2	매크로 데이터 정보보호 - 노출제한 방법 2	♣
17:10 ~ 18:00 (8교시)	♣	매크로 데이터 정보보호 - 노출제한 방법 3	♣

▣ 교재명 : 통계적 정보보호 방법 (정책연구용역 결과물 활용)

〈부록 2〉

통계적 정보보호 방법 전문가 과정 (안)

Ⅰ 목적

❑ 통계적 정보보호의 개념, 마이크로 및 매크로데이터에서의 정보보호 방법들, 정보보호와 데이터의 유용성을 상호 평가하는 방법을 학습하여 통계작성기관 담당자의 통계역량을 강화함.

Ⅱ 활용 업무

❑ 통계작성에 필요한 자료를 수집하고 제공하는 분야로 조사관리국, 경제통계국, 사회통계국 등

Ⅲ 교육대상자 및 교육기간

교육대상자: 통계학 및 관련분야 석사학위자 또는 그에 준하는 자

o 통계적 정보보호 방법 기초 과정 이수자 또는 그에 준하는 자

Ⅳ 교재 및 참고문헌

❑ 교재 : 『통계적 정보보호 방법』 [정책연구 결과물]

Ⅴ 평가 및 결과

가.

❑ 평가 : 이론 및 리포트 평가

구분	평가	배점	평가범위 및 내용
이론 (70)	중간	30	정보노출 위험의 개념적 모형, 노출 위험의 평가, 노출위험의 통제, 매크로자료 정보보호 기법 - 이해도 평가 시험 (필기시험)
	기말	50	마이크로 자료 정보보호 기법, 위험노출과 자료의 유용성, 매스킹 소프트웨어 - 이해도 평가 시험 (필기시험)
	숙제	20	필요한 주차에 제시
리포트 (30)			o 수업을 마치면서 제공, 1주 뒤 답안 제출 o 학습 내용을 응용한 문제 해결 과제 o 분석 결과 발표 (구두시험 및 리포트)

❑ 결과 : 준거지향평가(절대평가) A, B, C 등급 부여

VI 주별 학습 계획 및 내용

주	학습주제	학습내용
1	통계적 정보보호의 개념 (I)	노출위험의 개념적 모형, 인식된 위험과 실제 위험, 정보 노출의 시나리오
2	통계적 정보보호의 개념 (II)	노출위험의 평가 및 통제 개념, 자료의 유용성 정의
3	노출위험의 평가	임계값과 대리측정값, 마이크로 및 매크로데이터에서의 노출위험 측정, 민감도
4	통계적 정보보호를 위한 기본통계	통계적 정보보호 기법의 이해를 돕기 위한 확률, 확률분포, 의사결정론 등의 통계개념
5	매크로자료 정보보호 기법 (I)	매크로데이터의 정보보호 개념, 단계, 및 기법
6	매크로자료 정보보호 기법 (II)	매크로데이터의 정보보호 기법
7	매크로자료 정보보호 기법 (III)	매크로데이터의 정보보호 기법 요약 및 소프트웨어 소개
8	중간고사	정보노출 위험의 개념적 모형, 노출 위험의 평가, 노출위험의 통제, 매크로자료 정보보호 기법의 이해도 평가

주	학습주제	학습내용
9	마이크로 자료 정보보호 기법 (I)	마이크로데이터의 정보보호 개념, 단계
10	마이크로 자료 정보보호 기법 (II)	접근의 제한, 자료감추기, 범주화
11	마이크로 자료 정보보호 기법 (III)	잡음첨가방법, 자료교환, 표본추출, 인위자료
12	마이크로 자료 정보보호 기법 (IV)	마이크로데이터의 정보보호 기법 요약 및 소프트웨어 소개
13	위험의 노출과 자료의 유용성	위험노출과 자료 유용성의 개념, 통계적 노출제한 기법에서 모수의 선택, 자료의 유용성 측정 방법, R-U 정보보호지도
14	예제 분석 (I)	매크로 자료 예제 분석
15	예제 분석 (II)	마이크로 자료 예제 분석
16	기말고사	마이크로데이터 정보보호 방법, 자료의 유용성과 정보노출의 측정방법 - 이해도 평가 시험