한국통계학회논문집
2009, 16권, 1호, 1-11

마이크로데이터 제공과 통계적 노출조절기법

김규성[1,a]

[a]서울시립대학교 통계학과

요약

마이크로데이터를 이용자에게 제공하면 레코드 단위의 데이터가 노출되고 응답자의 정보 노출위험이 불가피하다. 통계적 노출조절기법은 통계데이터 제공시 노출위험을 줄이면서 데이터 유용성을 높이기 위한 통계적 기법이다. 본 논문에서는 노출과 노출위험, 그리고 통계적 노출조절기법을 고찰하였고 데이터 유용성과 연관하여 노출조절기법 선택 전략을 살펴보았으며, '위험-유용성 경계 지도' 방법의 예를 알아보았다. 마지막으로 마이크로데이터를 이용자에게 제공할 때 단계별로 검토할 사항을 알아보았다.

주요용어: 노출위험, 데이터 유용성, 위험-유용성 경계지도.

1. 서론

통계는 복잡한 사회 현상을 설명하고 합리적인 의사결정을 하는데 중요한 역할을 하여왔다. 통계가 이러한 역할을 통하여 사회 발전에 더욱 기여하기 위해서는 통계는 과학적으로 만들어져야 하고, 또한 많은 사람들이 손쉽게 이용할 수 있도록 널리 보급되어야 한다. 공공자원으로서의 통계는 과학적으로 작성되어야 하는 생산 측면과 널리 보급되어 이용되어야 하는 이용 측면의 두 측면이 있는데, 이제까지 통계에 대한 관심의 초점은 주로 전자에 맞추어져 왔다.

1980년대 이후 우리 사회에서 통계에 대한 활용이 급속하게 증가하는 추세를 보이고 있는데 그 배경에는 정보기술의 발달이 자리하고 있다. 개인용 컴퓨터의 성능이 개선되고 용량이 확대되면서 이전에는 거의 불가능했던 복잡한 통계 계산이 가능해지고 대용량 데이터 처리가 쉬워지면서 통계 이용의 범위가 넓어지게 된 것이다. 이러한 사회 여건 변화에 따라 통계작성기관이 생산한 통계에 대한 이용자들의 요구가 증가하고 있다. 통계개발원의 보고서에 의하면 통계청 자료에 대한 이용자들의 제공 요구는 2007년에 2,260건으로 2003년의 536건에 비하여 4.2배나 증가하였다 (김경미 등, 2008). 자료제공 요구의 빈도가 급속도로 증가한 현상과 더불어 주목할 점은 마이크로데이터 제공요구가 대부분이며 그 비율은 점점 더 높아지고 있다는 사실이다. 동 보고서에 따르면 2003년에 자료 제공 536건 중 마이크로데이터 제공 요구는 71.1%였으나 2007년에는 88.9%로 증가하였다.

통계데이터를 이용자에게 제공하여 통계이용자 혹은 연구자가 각종 통계를 생산할 수 있도록 하면서 통계이용자의 통계에 대한 만족도를 높이는 것은 통계의 보급과 이용측면에서 바람직한 일이다. 그러나 통계데이터를 제공하면 개인이나 사업체 등 레코드 단위의 정보가 노출되어 응답자의 신원이 식별될 가능성이 생긴다. 응답자의 신원이 이용자에게 노출되는 것은 법적인 측면에서는 통계법(제33조 응답자 비밀보호)을 위반하는 것이고, 윤리적인 측면에서는 응답을 통계적인 목적으로만 사용하겠다고 응답자와 한 약속을 어기는 일이며, 통계조사의 측면에서는 기관의 신뢰를 떨어뜨려 향후 응답자의 협조를 받기 어려운 상황을 스스로 초래할 수 있다.

본 논문은 2008년 통계의 날 기념 워크숍에서 발표한 내용을 수정·보완한 것임.

[1] (130-743) 서울특별시 동대문구 전농동 90, 서울시립대학교 통계학과, 교수. E-mail: kskim@uos.ac.kr

응답자의 비밀을 완벽하게 보장하고 통계데이터 제공에 따른 위험을 사전에 방지하는 방법은 통계데이터를 제공하지 않는 것이다. 그러나 이 방법은 통계데이터의 생산과 이용의 두 기능 중 전자에 충실한 방법으로 통계를 공공자원으로 간주할 때 선택할 바람직한 방법은 아니다. 그 보다는 응답자의 노출위험을 줄이고 데이터의 유용성을 통계데이터 제공방법을 선택해야 할 것이다.

본 논문은 통계데이터 제공시 응답자의 노출위험을 줄이고 데이터의 유용성을 높이는 통계적 방법론을 다룰 것이다. 그 방법은 통계데이터를 제공할 때 응답자의 노출을 조절 혹은 제한하도록 하는 방법으로, 통계적 노출조절기법 혹은 노출제한기법 (Statistical Disclosure Control/Limitation, SDC/SDL)으로 불린다. 이미 미국, 캐나다 그리고 유럽의 여러 나라에서는 최근 활발하게 연구되고 있는 분야이나 우리나라에서는 아직 그렇지 못하다. 그러나 우리나라에서도 통계작성기관이 통계데이터를 제공할 때 직면할 문제이므로 머지않은 장래에 통계적 노출조절기법의 필요성이 크게 대두될 것으로 예상된다.

통계데이터 형태는 통계표데이터(tabular data)와 마이크로데이터(microdata)로 분류할 수 있다. 매크로데이터(macrodata)라고도 불리는 통계표 데이터는 분할표 형태의 데이터로 전통적인 통계 간행물에서 주로 사용한 데이터 형태이다. 마이크로데이터는 레코드 단위의 데이터로서 일반에게 제공되는 공용화일(public use file)과 특정연구자에게 제공되는 인가된 파일(licensed file)로 구분할 수 있다. 통계적 노출조절기법은 통계표데이터와 마이크로데이터에 모두 적용 가능하지만 본 연구에서는 마이크로데이터에 적용하는 기법만을 다루기로 한다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 마이크로데이터에 적용 가능한 통계적 노출조절기법을 고찰하고, 3절에서는 통계적 노출조절기법을 선택하는 전략을 다룬다. 마지막으로 4절에서는 마이크로데이터를 이용자에게 제공할 때 검토할 사항을 단계별로 알아보았다.

2. 통계적 노출조절기법

2.1. 노출과 노출위험

마이크로데이터가 이용자에게 제공되면 이용자에게 레코드 단위의 노출은 불가피하다. 문제는 노출된 데이터로부터 개별 응답자의 신원이나 속성을 알아낼 수 있는지 여부이다. 그럴 경우 응답자의 응답 비밀보호는 지켜지지 않는다.

통계작성기관은 마이크로데이터를 이용자에게 제공할 때 이름, 주민번호, 주소 등과 같은 명시적인 식별자를 제거하고 제공하므로 제공된 마이크로데이터에서 응답자의 신원이 곧바로 노출되지는 않는다. 그러나 '직업=서울시장'과 같은 내재적인 식별자가 마이크로데이터에 포함되어 있거나 '소득=100억원'과 같은 극단값이 포함되어 있다면 개별 응답자의 신원이 노출될 수 있다. 또한 여러 문항을 조합하여 외부데이터와의 레코드 연결을 통하여 응답자의 신원이 노출될 가능성도 있다. 신원노출(identity disclosure) 이외에 노출은 응답자의 민감한 정보가 노출되어 노출된 정보로부터 개별 응답자의 속성이 추측될 때 발생하는 속성노출(attribute disclosure)과 응답자의 정보가 추정에 의하여 식별되는 추론노출을 포함한다 (Willenborg와 de Waal, 2001).

명시적인 식별자에 의한 신원노출 이외의 다른 노출에 대해서 응답자 식별 위험성을 측정하기는 쉽지 않다. 동일한 마이크로데이터라 하더라도 이용자의 사전 정보에 따라 응답자는 식별될 수도 있고 그렇지 않을 수도 있기 때문이다. 따라서 노출위험을 줄여야하는 데이터 제공자 입장에서는 노출위험에 대한 최악의 시나리오를 염두에 두어야 한다. 즉, 응답자를 식별하려고 하는 데이터 침입자(intruder)는 단일 레코드 혹은 일부 레코드에 대하여 많은 정보를 가지고 있다고 가정하고, 주요 변수를 이용하여 외부 데이터 파일과 병합할 수 있다고 가정하는 것이다. 이러한 가정에서 노출위험을 나타내는데 보편적으로 사용하는 노출위험 개념은 유일성(uniqueness) 개념이다. 즉, 제공된 통계데이터

에서 여러 변수에 대하여 중복이 없는 유일한 레코드는 노출위험이 있다고 간주한다.

모집단 유일성을 계산한 최근 사례로 2005년 인구센서스 예가 있다(정동명 외 2인, 2007). 2005년 인구센서스 결과 충남의 인구수는 1,798,397명인데 주요 분류변수를 성별(2수준), 나이(105), 가구주와의 관계(14), 혼인상태(4), 가구구분(5), 점유형태(12), 주인가구여부(6), 거처종류(12)로 하여 모집단 유일성을 계산한 결과 모집단의 0.54%인 9,664명이 주요 변수에 대하여 유일한 레코드인 것으로 나타났다. 충남 모집단을 이용자에게 제공했을 때 충남 모집단의 9,664명이 노출위험이 있다는 뜻이다. 실제로는 인구센서스 모집단을 일반에게 제공하지 않으므로 인구센서스 모집단 레코드가 노출될 위험은 없다. 대신 2% 표본 파일에 대하여 표본 유일성을 계산한 결과 38,027명 중 표본 유일성이 발견된 경우는 4.76%인 1,810명이었다. 2% 표본을 제공할 경우에 표본 레코드가 식별될 위험성이 있음을 시사하는 결과다.

통상적으로 통계표데이터보다는 레코드 단위의 익명화된 마이크로데이터가 노출위험 수준이 더 높다. 따라서 통계표데이터보다는 마이크로데이터를 제공할 때에 응답자의 정보 보호를 위한 정밀한 점검이 더 요구된다. 통계데이터 제공시 응답자의 정보를 보호하는 방법으로는 데이터에 대한 접근을 조절(access control)하는 방법과 제공하는 데이터의 노출을 조절하는 방법이 있다. 전자는 물리적으로 데이터에 대한 접근을 제한하는 방법으로 리서치 연구센터를 운영하여 그곳에서만 개인적인 허가를 받은 후 데이터를 접할 수 있게 하거나, 기관의 승인을 받은 후 원격으로 데이터를 제공하는 방법 등이 있다. 후자는 데이터를 제공하되, 제공하는 데이터의 정보를 제한하는 방법이다. 데이터의 정보를 축소하여 제공할 수도 있고, 데이터를 변형하여 제공할 수도 있는데, 두 방법 모두 응답자의 노출 위험을 줄이기 위한 방법이다. 후자의 방법들이 본 논문에서 다루고자 하는 통계적 노출조절기법들이다.

2.2. 통계적 노출조절기법

통계적 노출조절기법은 데이터 제공자가 익명화된 마이크로데이터를 대상으로 응답자 개별 정보를 알아내려고 하는 데이터 침입자의 시도에 대응하여 개인이나 사업체, 기관 등 응답자의 개별정보의 노출위험을 줄이면서 데이터를 이용자에게 제공하려고 할 때 적용하는 기법이다. 제공하는 데이터의 노출위험을 줄이면 이용자의 데이터 유용성은 떨어지고, 반대로 노출위험을 늘리면 이용자의 데이터 유용성은 증가하게 된다. 따라서 통계적 노출조절기법은 노출위험을 최소로 하면서 데이터 유용성은 극대화하도록 하여야 한다.

통계적 노출조절기법은 데이터 정보를 축소하는 기법과 데이터를 왜곡하거나 혹은 섭동(perturbation)하는 기법 그리고 합성데이터(synthetic data)를 생성하는 기법으로 분류할 수 있다. 데이터 정보를 축소하는 방법은 데이터에 포함된 정보의 일부를 제거하고 이용자에게 제공하는 방법으로 제거하는 정보의 양이 많을수록 노출위험은 줄어들지만 그만큼 데이터 유용성은 감소된다. 이 범주에 속하는 기법으로는 전체 리코딩(global recoding), 상한/하한 코딩(top/bottom coding), 부분 삭제(local suppression), 부분 리코딩(local recoding), 레코드 제거(record deletion), 변수 제거(variable deletion) 등이 있다. 데이터를 왜곡하거나 섭동하는 기법은 데이터를 일부 변형하여 레코드 연결이나 매칭 알고리즘에 의한 재식별을 어렵게 하거나 불확실하게 하는 기법이다. 이 범주에는 잡음 추가(noise addition), 마이크로애그리게이션(microaggregation), 데이터/순위 교환(data/rank swapping), 반올림(rounding), 사후 확률화 기법(post-randomization, PRAM) 등이 포함된다. 마지막으로 합성데이터 생성 기법은 데이터 시뮤레이션을 통하여 합성데이터를 생성한 후 합성데이터를 이용자에게 제공하는 것이다. 합성데이터는 실제 데이터와 동일한 데이터의 통계적 성질은 유지하되 실제 데이터가 아니므로 노출위험 문제가 발생하지 않는다는 장점이 있는 반면, 실제 데이터가 아니므로 데이터 활용과정에서 문제가 발생할 가능성이 있다.

기법을 고찰하기에 앞서 표현의 편의를 기호를 도입한다. 마이크로데이터 X는 n개의 레코드로

구성되어 있고, 각 레코드는 p개의 변수 값으로 이루어진다고 하자. 즉, $X = \{x_{ij}\}, i = 1, \ldots, n, j = 1, \ldots, p$. 그리고 원래 데이터 x_{ij}가 노출조절기법을 통하여 변형된 값을 m_{ij}라고 하고, m_{ij}들의 모임을 $M = \{m_{ij}\}$으로 표현하자. 편의상 첨자는 생략하고 '$[x] \Rightarrow [m]$'는 원래 데이터 x가 노출조절기법을 통하여 m이 됨을 나타낸다고 하자.

전체 리코딩은 모든 레코드에 대하여 연속형 변수인 경우는 연속형 값을 범주형으로 변환하는 것이고, 범주형 변수에 대해서는 여러 개 범주를 통합하여 더 넓은 범위의 범주를 만드는 기법이다. 예를 들어 소득변수는 연속형이므로 $x = 258$이라고 하면 변환된 값은 구간 $(250, 299)$에 속한다고 하는 것이다. 즉,

$$[x = 258] \Rightarrow [m \in (250, 299)].$$

상한 코딩과 하한 코딩은 전체 리코딩의 특수한 경우로서 최고값과 최소값은 희소하여 종종 식별위험이 크므로 최고 범주와 최소 범주로 변경하여 주는 기법이다. 예를 들어 월 소득이 521만원이라 하면 500만원 이상으로 바꾸는 것이다. 즉,

$$[x = 521] \Rightarrow [m \in (\geq 500)]$$

전체 리코딩과 상한/하한 코딩 기법은 노출위험이 큰 변수에 흔히 적용되는 기법이다. 반면, 이 기법은 응답값을 축소하거나 확대하기 때문에 체계적인 편향이 모수 추정과정에 개입될 위험이 있다. 이러한 체계적인 편향을 지적한 연구 결과가 있다. Burkhauser 등 (2004)는 미국 현재인구조사(Current population survey: CPS)의 공용파일 데이터를 이용하여 미국 근로자의 소득 불평등 지수로서 지니 계수(Gini coefficient)를 추정하였다. 추정 결과 0.34이던 1994년의 지니 계수가 일 년 후인 1995년에는 0.39로 급격히 높아졌는데 주된 이유는 1995년 소득조사에서 상한 코딩 기준값을 기존의 \$99,999에서 \$150,000으로 높였기 때문이었다. 상한 코딩의 기준을 높인 결과 더 \$99,999 이상의 소득이 조사되고, 이로 인해 지니 계수가 더 높게 나온 것이다. 상한 코딩을 적용하지 않은 내부 회일의 경우 1995년에 지니 계수는 도리어 1994년보다 작게 나타난 것이 이를 반증한다 (Lane, 2007).

부분 리코딩은 노출위험이 있는 일부 레코드에만 리코딩을 적용한다. 예를 들어 변수 x는 k개의 범주를 갖고 이 중 범주 $1, \ldots, a(< k)$에 속하는 레코드가 노출위험이 있다고 하자. 그러면 노출위험이 있는 범주를 하나의 범주로 통합하고 나머지 범주는 범주 숫자만 바꾸어 그대로 사용한다. 즉,

$$[x = 1, \ldots, a] \Rightarrow [m = 1]; \quad [x = j(> a)] \Rightarrow [m = j - a + 1].$$

레코드 제거는 데이터 보호 방법의 극단적인 경우로서 다른 방법을 사용하기 어려울 때에만 적용한다. 사업체 등 경제 데이터에서 예를 들어 어떤 산업 분류에 오직 한 회사만 해당한다고 하자. 이 경우 산업 분류를 제거하는 것보다는 해당 레코드를 제거하는 것이 바람직하다. 변수 제거는 데이터 보호 방법의 또 다른 극단적인 경우로서 레코드 제거 등 다른 방법 적용이 불가능할 때에만 사용한다.

잡음 추가는 연속형 변수에 적용 가능한 기법으로 변수 x에 확률오차 ϵ_0(평균 0, 분산 σ^2)을 더하여 데이터를 변형하는 잡음더하기 기법과 확률오차 ϵ_1(평균 1, 분산 σ^2) 곱하여 얻는 잡음곱하기 기법이 있다. 즉,

$$[x] \Rightarrow [m = x + \epsilon_0, \text{잡음더하기}]; \quad [m = x \times \epsilon_1, \text{잡음곱하기}]$$

잡음 추가는 쉽게 적용하여 데이터를 변형 할 수 있는 장점이 있다. 그러나 잡음이 전체 레코드에 동일하게 적용되기 때문에 작은 x값은 잡음의 영향을 크게 받고, 큰 x값은 잡음의 영향을 상대적으로 작게 받아 의 크기에 따라 잡음의 영향을 다르게 받는 성향이 있다.

마이크로애그리게이션 기법은 연속형 변수에 적용 가능한 기법이다. 일변수 마이크로애그리게이션이 경우, 연속형 변수 x를 크기순으로 정렬 한 후 순차적으로 g개의 그룹으로 나눈다. 그리고 각 그룹에 속하는 레코드의 x값을 그 그룹의 평균값으로 바꾼다. 즉,

$$\left[x \in \left\{x_{(n_{g-1}+1)}, \ldots, x_{(n_g)}\right\}\right] \Rightarrow \left[m = \frac{1}{n_g - n_{g-1}} \sum_{k=n_{g-1}+1}^{n_g} x_{(k)}\right],$$

여기에서 $x_{(k)}$는 크기순으로 정렬된 데이터이다. 변수가 k인 경우 그룹 구성은 $(x_1, \ldots, x_k)$값들 간의 유사성을 계산하여 유사성이 최대가 되도록 하고 그룹 내의 $(n_g - n_{g-1})$개 $(x_1, \ldots, x_k)$ 값은 그 그룹의 평균값으로 대체한다. 이 기법을 적용하면 적용 전 데이터의 총계와 적용 후 데이터의 총계는 동일하게 유지되는 장점이 있다.

데이터 교환은 마이크로데이터 파일에서 일정 비율의 레코드를 선정한 후 쌍으로 데이터를 상호 교환하는 기법이고, 순위 교환은 레코드를 오름차순으로 정렬한 후 각 순위 값을 다른 값과 교환하는 기법이다. 기본 교환으로 두 개의 레코드를 상호 교환한다고 하자. 그러면 먼저 n개의 레코드 중 2개를 비복원 단순확률추출하여 뽑고, (i, j)가 뽑혔다고 하자, 대응하는 x_i와 x_j를 상호 교환한다. 여러 개의 레코드를 교환할 때에는 기본 교환을 여러 번 반복하면 된다. 만일 k개의 레코드를 교환한다고 하자. 그러면 $2k$개의 레코드를 단순확률추출하고 추출된 레코드의 x값을 쌍으로 상호 교환한다. 즉, 크기 $2k$인 단순확률표본을 $s = \{i_1, \ldots, i_{2k}\}$라고 하면,

$$[s = \{i_1, \ldots, i_{2k}\}] \Rightarrow \left[m_{i_{2j}} = x_{i_{2j-1}}, m_{i_{2j-1}} = x_{i_{2j}}, j = 1, \ldots, k\right]$$

레코드 교환은 원래 변수의 값이 수정되지 않고 교환만 되므로 이용자들에게 거부감 없이 받아들여지는 장점이 있다. 그리고 교환 변수의 일변량 통계적 성질이 그대로 유지된다. 그러나 교환 변수와 다른 변수들 간의 관계는 유지되지 않고, 신원 노출 및 속성 노출의 위험이 그대로 남아 있는 단점이 있다.

전통적인 반올림에서는 나머지에 의해 올림이나 내림이 유일하게 결정된다. 만일 양의 정수 x를 밑 B, 몫 q, 나머지 r로 표현하면 $x = qB + r$가 되는데, $r \geq B/2$이면 올림을 하고 그렇지 않으면 내림을 한다. 즉,

$$[x = qB + r] \Rightarrow [r \geq B/2 \Rightarrow m_x = (q+1)B] \text{ 혹은 } [r < B/2 \Rightarrow m_x = qB]$$

이 방법은 간단하여 널리 사용되어 왔으나 노출조절기법 관점에서 보면 반올림하는 방법이 결정적이어서 노출위험이 남아 있고, 변환 전 데이터 x의 합과 변환 후 데이터 m_x의 합이 일치하지 않는 단점이 있다. 이러한 단점의 보완책으로 올림이나 내림을 확률적으로 하는 랜덤 반올림(random rounding)과 이에 더해서 두 변수이 합을 일치하도록 하는 조절된 반올림(controlled rounding) 방법이 제안되었다 (예, Cox와 Kim, 2006).

사후확률화 기법은 원래 데이터의 일부 레코드의 일부 변수의 값을 미리 지정한 전이확률 행렬(마코프 행렬)에 따라 다른 값으로 변환하는 기법이다. T_x를 x값 응답자 수, T_m을 m값 응답자 수라고 하고 전이 행렬을 $P = (p_{ij}), p_{ij} = \Pr[m_x = j | x = i]$라고 하자. 그러면 PRAM 변환 후 응답자 수 T_m의 기댓값은 다음과 같다.

$$\mathrm{E}(T_m) = P^T T_x$$

예를 들어 응답자 100명 중 '$x = 1$'인 응답자가 99명, '$x = 2$'인 응답자 1명이라고 한다면 $T_x = (99, 1)^T$이다. 이때 '$x = 2$'인 응답이 유일하여 노출위험이 크므로 노출위험을 줄일 필요가 있다. 이제 전이행렬을 지정하자.

$$P = (p_{ij}) = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}$$

그리고 전이행렬을 이용하여 PRAM 후 응답자 수의 기댓값을 구하면 다음과 같다.

$$\mathrm{E}(T_m) = \begin{pmatrix} 0.9 & 0.2 \\ 0.1 & 0.8 \end{pmatrix} T_x = \begin{pmatrix} 0.9 & 0.2 \\ 0.1 & 0.8 \end{pmatrix} \begin{pmatrix} 99 \\ 1 \end{pmatrix} = \begin{pmatrix} 89.3 \\ 10.7 \end{pmatrix}$$

원래 데이터에서 '$x = 2$'인 레코드 수는 1명이었으나 PRAM을 통하여 '$m = 2$'인 레코드 수는 11로 증가하여 노출위험이 줄어들었다.

2.3. 사례

통계적 노출조절기법을 활용한 예제로 앞의 2절에서 언급한 우리나라 2005년 인구센서스 예제로 돌아가자 (정동명 등, 2007). 위 논문에서 사용한 노출조절기법은 전체 리코딩 기법으로 변수의 수준을 병합하여 전체 수준수를 줄였다. 성별은 그대로 유지하되 나이는 105수준에서 21수준으로, 가구주와의 관계는 4수준으로, 혼인상태는 3수준으로, 가구 구분은 4수준으로, 점유 형태는 8수준으로, 주인가구는 2수준으로, 거처의 종류는 4수준으로 줄였다. 전체 리코딩을 통하여 각 변수 수준의 조합수는 $2 \times 105 \times \cdots \times 12 = 50,803,200$ 에서 $2 \times 21 \times 4 \times \cdots \times 4 = 129,024$로 줄었다. 변환전의 0.25%에 해당한다. 이렇게 하여 표본유일성의 수는 1,810명에서 553명으로 감소하였다. 전체 리코딩이라는 통계적 노출조절기법을 활용하여 노출위험을 줄인 예이다.

통계적 노출조절기법을 실제 활용한 예로는 미국 센서스 데이터가 있다. 미국의 2000년 센서스 공용 마이크로데이터 표본파일은 지리적인 경계값, 반올림, 잡음 추가, 범주형 임계기준(categorical threshold), 상한 코딩, 데이터 교환 기법을 활용하였다 (Zayatz, 2007). 지리적인 경계값의 범위는 최소 10만 명 이상이다. 금액을 나타내는 변수는 전통적인 반올림 사용하고 크기에 따라 적용 기준이 다르다. \$1–\$7은 \$4 단위로 반올림하고, \$8–\$999는 \$10 단위로, \$1,000–\$49,000은 \$100 단위로, \$50,000이상은 \$1,000 단위로 반올림하였다. 나이 변수에는 잡음 추가 기법을 사용하였고, 모든 범주형 변수는 전국 기준으로 10,000명이 통계표의 셀 경계값인데 미달이면 리코딩하였다. 소득, 출장여행시간과 같은 모든 연속형 변수에는 상한 코딩을 적용하였다. 그리고 데이터 교환 기법도 적용하였다. 또한 2000년 센서스의 모든 통계표 데이터는 노출위험기법이 적용된 마이크로데이터에서 생성되었다.

3. 통계적 노출조절기법 선택

마이크로데이터 제공시 레코드 단위의 응답자 정보의 노출위험을 최소화 하는 것은 개인정보 보호를 위한 법적인 제약 사항이기도 하고 응답자 협조를 지속적으로 유도하여 통계 품질을 유지하기 위한 필요 사항이기도 하다. 그러나 통계작성기관으로서 기관 정책사항으로 통계 결과물의 부가가치를 제고해야 하는 측면이 있다. 통계 이용자들에게 가능하면 많은 유용한 데이터를 제공하여 통계 재생산의 기회를 늘릴 필요가 있는 것이다. 마이크로데이터의 개인 정보 노출위험 최소화와 데이터 유용성 유지는 서로 상반되는 개념이다. 따라서 마이크로데이터를 제공할 때에는 이같은 상반된 두 측면을 고려하여 통계적 노출조절기법을 선택한 후 선택된 기법이 적용된 마이크로데이터를 일반 이용자에게 제공하여야 한다.

3.1. 데이터 유용성 측도

데이터 유용성(data utility)은 제공되는 데이터의 가치를 평가하는 개념으로 데이터의 해석적인 완비성과 타당성을 의미한다 (Hundepool 등, 2005). 통상적으로 노출조절기법은 데이터 유용성에 반대 효과를 준다. 마이크로데이터에 노출조절기법을 적용하면 데이터의 정보가 줄어들거나 왜곡되고 혹은 변형되기 때문에 데이터의 완비성과 타당성이 줄어들어서 결과적으로 유용성이 떨어진다. 이상적

으로는 노출조절기법이 적용된 데이터는 원래 데이터와 동일한 특성을 가져야 한다. 일변량 특성이나 민감한 변수 간의 관계, 그리고 민감한 변수와 다른 변수간의 관계가 두 데이터에서 동일하게 유지되어야 한다. 그러나 현실적으로는 노출조절기법을 적용하면 데이터 유용성을 줄어들므로 노출조절기법 선택은 노출위험과 데이터 유용성의 타협(trade-off)이라고 할 수 있다.

데이터 유용성은 보통 데이터 편향이나 혹은 정보손실(information loss)로 측정된다. 범주형 데이터의 정보손실은 분할표에서 두 데이터의 거리를 측정함으로써 계산할 수 있는데 헬링거(Hellinger) 거리, 총변동(total variation), 에트로피 변화 등이 정보손실 측도로 사용된다 (Gomatam 등, 2005). 마이크로데이터를 분할표로 만들었을 때 셀 c에서 원래 데이터 x의 도수나 혹은 크기를 $f_x(c)$라고 하고, 기법 처리된 데이터 m이 도수나 크기를 $f_m(c)$라고 하자. 그러면 위의 세 측도는 다음과 같이 정의된다.

- 헬링거 거리: $HD(X,M) = \frac{1}{\sqrt{2}}\sqrt{\sum_{\text{cell } c}\left(\sqrt{f_x(c)} - \sqrt{f_m(c)}\right)}$
- 총변동: $TV(X,M) = \frac{1}{2}\sum_{\text{cell } c} | f_x(c) - f_m(c) |$
- 엔트로피 변화: $\Delta h(X,M) = -\sum_{\text{cell } c} f_m(c)\log[f_m(c)] + \sum_{\text{cell } c} f_x(c)\log[f_x(c)]$.

연속형 데이터의 정보손실은 두 데이터를 비교한 후 그 차이를 통하여 계산할 수 있다. 원래 마이크로데이터 X의 원소가 x_{ij}일 때 변수 $\bar{x}_j$를 j번째 변수의 평균, i번째 변수와 j번째 변수에 대하여 $v_{ij}(x)$를 공분산, $r_{ij}(x)$를 상관계수라고 하자. 그리고 노출조절기법이 적용된 데이터 M에서 대응되는 값을 각각 $m_{ij}, \bar{m}_j, v_{ij}(m), r_{ij}(m)$으로 나타내자. 그러면 다음과 같은 정보손실 측도를 고려할 수 있다 (Domingo-Ferrer 등, 2001):

$$IL(M) = \frac{1}{5}\left[MV(m_{ij}) + MV(\bar{m}) + MV(v_{ij}) + MV(v_{jj}) + MAE(r_{ij})\right],$$

여기에서 MV는 평균변동(mean variation)을 의미하고, MAE는 평균절대편차(mean absolute error)를 의미하며, 5가지 측도는 다음과 같다.

$$MV(m_{ij}) = \frac{1}{np}\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{|x_{ij} - m_{ij}|}{|x_{ij}|}$$

$$MV(\bar{m}_j) = \frac{1}{p}\sum_{j=1}^{p}\frac{|\bar{x}_j - \bar{m}_j|}{|\bar{x}_j|}$$

$$MV(v_{ij}) = \frac{2}{p(p+1)}\sum_{j=1}^{p}\sum_{i\le j}\frac{|v_{ij}(x) - v_{ij}(m)|}{|v_{ij}(x)|}$$

$$MV(v_{jj}) = \frac{1}{p}\sum_{j=1}^{p}\frac{|v_{jj}(x) - v_{jj}(m)|}{|v_{jj}(x)|}$$

$$MAE(r_{ij}) = \frac{2}{p(p-1)}\sum_{j=1}^{p}\sum_{i\le j}|r_{ij}(x) - r_{ij}(m)|.$$

위의 측도 $IL(M)$은 주로 공분산과 상관계수를 이용한 측도이기 때문에 다차원 적률에 대해서는 잘 맞지 않고, 이상치에 둔감한 단점이 있다.

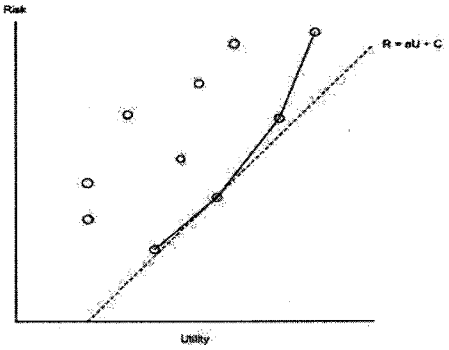

그림 1: *위험-유용성 경계 지도*

이러한 단점을 보완해 줄 수 있는 측도로 Kullback-Liebler 측도가 있다 (Karr 등, 2006). 두 데이터로부터 추정된 확률밀도함수 와 를 구한 후 두 함수의 비를 이용하여 정보손실을 측정한다.

$$d_{KL}(X, M) = \int \log\left(\frac{\hat{f}_M}{\hat{f}_X}\right)\hat{f}_M$$

이 측도는 정규분포 데이터에 잘 적용되는 특징이 있다. 또 다른 측도로는 누적확률밀도함수를 이용한 측도가 있다. 콜모고로프 통계량을 이용하여 두 데이터의 분포함수 비교하는 측도이다 (Karr 등, 2006).

3.2. 노출조절기법 선택 전략

원래 마이크로데이터에 통계적 노출조절기법을 적용한 마이크로데이터의 집합을 $\mathfrak{J}$라고 하자. 그리고 $DR(M)$을 데이터 M의 노출위험 함수라고 하고, $DU(M)$을 데이터 M의 데이터 유용성 함수라고 하자. 데이터 제공자의 목표는 제공될 마이크로데이터 후보 중 최적의 데이터 M을 선택하는 것이다. 그런데 데이터 유용성을 증가시키면 노출위험이 증가하는 상호 관계가 있기 때문에 최적의 마이크로데이터를 선택하기 위해서는 마이크로데이터 선택전략이 필요하다.

통상적으로 언급되는 데이터 선택 전략은 조건부 최적화 방법으로, 사전에 정한 노출위험의 범위 안에서 데이터 유용성을 최대로 하는 노출조절기법을 찾는 것이다. 즉, 주어진 α에 대하여 $DR(M) \leq \alpha$을 만족하는 M 중에서

$$M^* = \arg\max_{M \in \mathfrak{J}} DU(M)$$

인 M^*를 선택한다 (Willenborg와 de Waal, 2001). 이 방법은 논리적으로는 분명하지만 α에 따라 최적의 M^*이 달라지기 때문에 이 방법을 실제 데이터에 적용하기 위해서는 α를 결정해야 하는 현실적인 문제를 해결해야 한다.

이 방법보다 더 실용적인 방법은 '위험-유용성 경계(Risk-utility frontier) 지도 방법'을 이용하는 것이다 (Gomatam 등, 2003). 이 방법은 노출조절기법과 유용성 함수를 쌍으로 비교하여 노출위험은 적고 유용성은 큰 노출위험기법을 선택하는 방법이다. 즉, 노출조절기법을 적용한 마이크로데이터를 M_1과 M_2라고 할 때,

$$DR(M_2) \leq DR(M_1) \quad \& \quad DU(M_2) \geq DU(M_1)$$

인 관계가 있다고 하면 M_1보다는 M_2를 선택하는 것이다. 노출위험과 데이터 유용성의 관계식을 구하고,

$$DR(M) = \alpha \times DU(M) + b$$

표 1: 위험-유용성 경계에 접하는 변수명

교환비율	(위험 낮음, 유용성 낮음)						(위험 높음, 유용성 높음)	
0.5%	WM	MH	S*	EH	E	H		
1.0%	AR	WR*	RH	ER	R	E		
5.0%	AI	MH	WR*	WS	RH	R	H	E

* 표시한 변수는 위험과 유용성 관계 직선에 접하는 변수임.

각 쌍을 그래프로 표시하면 그림 1과 같게 된다. 그림에서 보면 10개의 쌍 중에서 실선으로 연결된 4개의 쌍이 상대적으로 유용성은 크고 노출위험은 적은 쌍이고, 다시 이중에서 노출위험과 데이터 유용성의 관계식(점선)과 만나는 쌍이 최종적으로 선택된다.

3.3. 예제

위험-유용성 경계 방법을 사용한 예제가 있다 (Gomatam 등, 2003). 데이터는 미국 CPS 데이터로서 성별(S, 2수준), 나이(A, 3), 인종(R, 2), 결혼여부(M, 2), 고용형태(W, 4), 교육정도(E, 4), 주당 평균 근로시간(H, 3), 연봉(I, 2)의 8개 변수가 사용되었고, 교환 후 데이터에서 교환 안한 레코드 총 수 중 교환 전 도수가 1이거나 2인 셀 중에 속하는 레코드 총수를 노출위험으로 정의되었다. 즉,

$$\text{Risk} = \frac{\sum_{C_1 \& C_2} \text{교환 안한 레코드 수}}{\text{교환 안한 레코드 총수}},$$

여기에서 C_1과 C_2는 교환 전에 도수가 1 혹은 2인 셀이다. 데이터 유용성 측도로는 헬링거 거리가 사용되었다. 교환 비율은 0.5%, 1%, 5%이고, 교환 변수는 변수가 1개(8가지)인 경우와 2개인 경우(28가지)가 고려되었다. 따라서 교환조건은 총 $36 \times 3 = 108$가지이다. 교환비율에 따라 위험-유용성 경계에 접하는 쌍을 노출위험이 적고 유용성이 적은 쌍부터 노출위험이 높고 유용성이 큰 순서로 나열하면 표 1과 같다.

'위험-유용성 경계 지도 방법'이 제시하는 기준에 따르면 경계에 접하는 변수 중 위험과 유용성 관계 직선에 접하는 변수는 교환비율이 0.5%일 때 성별(S), 1%와 5%일 때 인종 × 고용형태(WR)로 나타났다. 따라서 이 방법에 의하면 인종 × 고용형태 변수를 교환 변수로 하는 것이 노출위험을 줄이면서 데이터 유용성을 높이는 방법이 된다.

마이크로데이터를 제공할 때에는 노출위험과 데이터 유용성 관계에서 의사결정이 필요하다. 앞에서 소개한 위험-유용성 경계지도 방법은 좋은 선택 전략의 하나가 될 것이다.

4. 결론

통계작성기관이 마이크로데이터를 일반 이용자에게 제공할 때에는 여러 단계에 걸쳐 단계별로 세심한 검토를 하여야 한다. 먼저 법적인 문제를 검토한 후 정책적인 문제를 검토하여야 한다. 기관의 정책에 따라 이용자에게 제공하는 데이터의 범위가 달라질 수 있기 때문이다. 다음으로 제공하고자 하는 마이크로데이터를 구축한 후 데이터 노출위험을 평가한다. 노출위험 시나리오를 만들고, 변수를 검토하여 지역 변수의 수준, 직업 세분화 정도, 조사설계 가중치 제공 등을 살핀다. 그리고 계량화된 위험 측도를 만들어 위험 측도를 비교하고 한계값을 정한다. 다음으로 마이크로데이터에 통계적 노출조절기법을 적용한다. 섭동 기법으로는 레코드 교환, 잡음 추가 등을 고려하고 비섭동 방법으로는 리코딩, 삭제 방법 등을 고려하여 여러 기법을 혼합한 방법도 검토한다. 또한 온라인 접근도 고려 대상에 포함하여 검토한다. 마지막으로 통계데이터를 제공하는 단계에서는 데이터 접근 권한에 관한 모든 사항,

연구 주제에 관한 모든 사항, 그리고 이용자에 관한 사항 등을 점검한다. 통계데이터는 동일한 제공기준에 의하여 보호되어야 한다. 최량의 마이크로데이터가 신뢰할만한 이용자에게 동일한 제공기준을 적용하여 제공될 때 최선의 연구가 수행될 수 있을 것이다.

아직까지 우리나라에서는 마이크로데이터 제공 문제가 큰 사회적 혹은 통계적 이슈가 되지는 않았지만 마이크로데이터 제공 요구가 급증하는 요즘의 추세를 볼 때 머지않아 이 문제가 표면으로 부상할 것이다. 이 문제에 제기되면 제공하는 거의 모든 마이크로데이터에 통계적 노출조절기법을 필수적으로 적용될 가능성이 매우 높다.

통계적 노출조절기법은 매우 복잡하고 그 선택은 다분히 주관적이다. 게다가 응답자 정보를 완벽하게 보호하는 노출조절기법은 구현 불가능하기 때문에 현실적으로는 '응답자 정보 보호'와 '데이터 유용성 확대' 사이에서 적절하게 타협을 이루어야 한다. 그리고 그러한 타협은 다양한 경험을 필요로 한다. 향후 마이크로데이터 제공을 대비하여 마이크로데이터에 대한 통계적 노출조절기법 연구가 활발히 이루어지기를 기대한다.

참고 문헌

김경미, 이의규, 정미옥 (2008). 마이크로데이터 제공 현황에 관한 해외사례 연구. <마이크로데이터 활용연구 및 통계를 이용한 현황분석연구>, 3장, 41–87. 통계개발원.

정동명, 김종익, 강동환 (2007). 인구센서스의 비밀보호방법. <통계연구>, **12**, 95–120.

Burkhauser, R., Butler, J., Feng, S., and Houtenville, A. (2004). Long term trends in earnings inequality: what the CPS can tell us. *Economics Letters*, **82**, 295–299.

Cox, L.H. and Kim, J. J. (2006). Effects of rounding on the quality and confidentiality of statistical data. *Privacy in Statistical Database, edited by Domingo-Ferrer, J. and Torra, V.* 48–56.

Domingo-Ferrer, J. and Mateo-Sanz, J.M. and Torres, A. (2001). Comparing SDC methods for miocrodata on the basis of information loss and disclosure risk of disclosure control methods. *In Proceeding of ETK-NTTS 2001*, 807–825. Lumxembourg.

Gomatam, S., Karr, A.F. and Sanil, A.P. (2003). A risk-utility framework for categorical data swapping. *NISS technical report no. 132.*

Gomatam, S., Karr, A.F. and Sanil, A.P. (2005). Data swapping as a decision problem. *Journal of Official Statistics*, **21**, 635-655.

Hundepool, A., Nordholt, S., Tambay, J-L., Wende, T., and Elliot, M. (2005). Glossary on statistical disclosure control. *Joint UNECE/Eurostat work session on statistical data confidentiality*, WP. 45. Geneva.

Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P. and Sanil, A.P. (2006). A framework for evaluation the utility of data altered to protect confidentiality. The Americal Statistician, **60**, 1–9.

Lane, J. (2007). Optimizing the use of microdata: an overview of the issues. *Journal of Official Statistics*, **23**, 299–317.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*, Springer.

Zayatz, L. (2007). Disclosure avoidance practices and research at the U.S. Census Bureau: an update. *Journal of Official Statistics*, **23**, 253–265.

2008년 10월 접수; 2008년 11월 채택

Release of Microdata and Statistical Disclosure Control Techniques

Kyu-Seong Kim[1,a]

[a]Dept. of Statistics, Univ. of Seoul

Abstract

When microdata are released to users, record by record data are disclosed and the disclosure risk of respondent's information is inevitable. Statistical disclosure control techniques are statistical tools to reduce the risk of disclosure as well as to increase data utility in case of data release.

In this paper, we reviewed the concept of disclosure and disclosure risk as well as statistical disclosure control techniques and then investigated selection strategies of a statistical disclosure control technique related with data utility. The risk-utility frontier map method was illustrated as an example. Finally, we listed some check points at each step when microdata are released.

Keywords: Data utility, disclosure risk, risk-utility frontier map.

[1] Professor, Department of Statistics, University of Seoul, Seoul 130-743, Korea. E-mail: kskim@uos.ac.kr