The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.5.743
25(5), 743–755

Study on a Measurement of Disclosure Risk of Microdata by Similarity

Hyeon-Kwan Cho[1] · Dae-Hong Kwon[2] · Suk-Hoon Lee[3]

[1]Department of Statistics, Chungnam National University; [2]Agency for Defense Development
[3]Department of Statistics, Chungnam National University

(Received July 17, 2012; Revised September 6, 2012; Accepted September 25, 2012)

Abstract

Researchers using various of statistical data want to obtain microdata for a detailed analysis. Institutes need to provide microdata after masking processes for sensitive data. Many researchers have used the proportion of unique identity for the measurement of disclosure risk. We proposed a new measurement of disclosure risk that considers the case that all identities are the same or similar. As an application example, we compare the newly proposed measurement and the existing measurement using 10667 data in 'Korea Household Income and Expenditure Survey data for 2010'

Keywords: Measurement of disclosure risk, microdata, similarity.

1. 서론

대부분 국가의 통계작성 기관에서는 많은 인원과 예산을 투입하여 생산한 자료를 정부기관, 학계, 연구기관, 일반 이용자들에게 다양한 형태로 제공하고 있다. 통계작성 기관에서 제공하는 통계자료 형태는 분할표나 집계표 등의 매크로데이터와 개인, 가구, 사업체 등(이하 '개체'라 한다) 개체별 자료인 마이크로데이터가 있다. 그런데 좀 더 세부적이고 다양한 분석을 위하여 자료 이용자들은 더 많은 마이크로데이터의 제공을 요청하고 있다. 반면에 자료제공 기관에서는 개체에 대한 민감한 정보가 공개되는 것을 막기 위해 이름, 주민등록번호, 주소 등 공식적 식별자를 제외하고 성별, 나이, 거주지역 등 개체가 식별될 가능성이 있는 식별변수와 수입, 학력 등 노출이 꺼려지는 민감변수(또는 관심변수)로 구성된 자료를 제공하고 있다. 그러나 이렇게 제공된 자료도 식별변수들의 조합에 의해서 규정되는 개체가 유일하거나 소수인 경우에는 그 개체(들)의 식별변수 값을 아는 사람들에게는 그 개체(들)의 민감한 정보가 공개될 것이다. 따라서 자료제공기관에서 마이크로데이터를 제공할 때에는 공개위험 정도에 따라 적절한 비밀보호 방법을 강구하게 된다.

통계자료의 비밀보호 방법에 관한 연구는 크게 두 가지로 나누어지는데 하나는 공개위험을 낮추기 위한 비밀보호 기법들을 연구하는 것이고 다른 하나는 제공대상인 자료가 어느 정도의 공개위험에 노출되어 있는가를 측정하는 것이다. 비밀보호 기법과 관련하여 유럽과 미국에서는 1970년대부터 다양한 기법들이 연구되었으며 유럽연합에서는 연구내용을 종합하여 비밀보호 편람을 제공하고 있고 (Eurostat,

[3]Corresponding author: Professor, Department of Statistics Chungnam National University, 99 Daehak-ro, Yuseong-gu, Daejeon 305-764, Korea. E-mail: sukhoon@cnu.ac.kr

Table 1.1. Grades of Principles of Economics course in university A

학과	성별	학년	학점	학과	성별	학년	학점
통계학과	남	2	B	행정학과	여	3	C
통계학과	남	2	B	행정학과	여	3	B
통계학과	남	2	B	행정학과	여	3	A
행정학과	남	3	C	사회학과	남	2	A
행정학과	남	3	B	사회학과	남	2	C
행정학과	남	3	A	사회학과	남	2	B+

1996), 미국에서는 1994년 제정한 비밀보호 지침서에 대해 2005년 개정판을 제공하고 있다 (FCSM, 2005). 또한 Duncan 등 (2001), Gomatam 등 (2003), Shlomo (2010) 등이 원자료 대신 비밀보호 자료를 제공함으로써 발생하는 정보손실과 공개위험을 함께 고려하는 연구를 수행하였다. 최근에는 Xiao 등 (2010)은 '공격자'(의도적으로 민감한 정보를 알려고 하는 자)가 민감정보를 제외한 모든 정보를 알고 있을 경우의 공개위험에 대해 연구하고 있다. 국내에서는 Kim (2006)이 개인정보 공개위험의 통계적 비밀보호 기법에 대해 종합적인 정리를 하였고, 마이크로데이터를 공개위험과 유용성을 함께 고려하여 공개하는 방안을 Kwon (2009)이 처음으로 연구하였다. 또한 Jeong 등 (2009)은 2006년 가계조사자료(마이크로데이터)에 대해 다양한 확률분포로 생성한 승법잡음을 추가하여 비밀보호처리를 실시한 사례를 발표하였으며, 동일한 사례에 대해 Kim 등 (2011b)은 각 분포의 분산이 일정한 값을 가지도록 모수 값을 설정하여 로지스틱 회귀모형을 통해 비밀보호를 처리하고 유용성을 측정하였다.

마이크로데이터의 공개위험 측도와 관련하여 Bethlehem 등 (1990)은 모집단에서 임의로 추출된 개체가 모집단에서 유일할 확률을 사용하였으며, 이 값을 추정하기 위해 P-G(Poisson-Gamma) 분포를 이용하였다. Zayatz (1991)는 표본의 유일한 개체가 모집단에서 유일한 개체일 확률을 제안하였으며 이를 추정하기 위해 부표본 방법과 EQC(Equivalence Class) 방법을 제안하였다. 이후 P-G 분포나 EQC 방법 대신 Chen과 Keller-McNulty (1998)은 N-B(Negative-Binomial) 분포를 사용하였고 Takemura (1997)는 D-M(Dirichlet-Multinomial) 분포를 사용하였으며 Fienberg와 Makov (1998)은 log-linear 모형을 이용하였다. 또한 Skinner와 Elliot (2002)은 모집단에서 임의로 추출한 개체가 표본의 유일한 개체와 식별변수 값이 일치할 때 표본에 속하는 유일한 개체가 모집단에서 추출된 개체와 일치할 확률을 공개위험 측도로 제안하고 있다. 최근에 Huda 등 (2010)은 평균치보다 높은 공개위험 값을 갖는 개체들의 공개위험에 대해서도 연구하고 있다. 공개위험 측도와 관련한 연구들은 Huda 등 (2010)이 살펴본 바와 같이 대부분 유일개체의 비율을 기준으로 삼고 있다.

한편 Samarati (2001)은 유일개체 수의 비율을 기준으로 자료를 공개할 때 k-익명성(k-anonymity)이라는 기준을 제안하였다. k-익명성이란 공개되는 자료들 중 식별변수들의 조합이 같은 개체수가 최소 k개는 되어야 한다는 것이다. 그러나 k-익명성의 기준도 비밀보호가 되지 않는 상황이 발생한다. Table 1.1은 경제학원론 수강생의 성적자료를 공개하고자 성명을 미기재하여 작성한 것이다. 이 경우 $k = 3$인 k-익명성을 만족하므로 자료를 공개할 수 있으나 경제학원론을 수강한 {통계학과, 남, 2학년}은 학점이 모두 같으므로 그 학생들의 학점은 공개되는 셈이다. 즉, 식별변수 값들이 동일한 개체가 k개 이상 있더라도 관심변수 값이 일치하면(또는 유사하면) 그 개체들의 관심변수 값은 공개되는 것이다. 이런 문제를 보완하기 위하여 Machanavajjhala 등 (2006)은 l-다양성(l-diversity)이라는 기준을 제시하였다. l-다양성이란 식별변수가 같은 조합에서는 최소한 l개의 관심변수 값들이 있어야 한다는 기준이다. 즉, 식별변수 뿐만 아니라 관심변수들도 고려한 공개기준을 제시하고 있다.

본 논문에서는 이러한 l-다양성 개념을 적용한 새로운 공개위험 측도를 제안하였다. 공개위험 측도와 관련한 대부분의 연구들은 유일개체의 비율을 기준으로 삼고 있으나, 본 논문에서는 유일개체가 아니더라

도 관심변수 값이 일치하거나 유사한 경우까지 고려한 개념을 적용하였다. 본 논문은 다음과 같이 구성된다. 2장에서는 본 논문에서 사용하는 기호와 함께 마이크로데이터의 공개위험 측도에 대하여 소개하고 3장에서는 유사성 개념을 적용한 마이크로데이터의 공개위험 측도를 관심변수가 이산형 변수인 경우와 연속형 변수인 경우에 대하여 각각 제안하며 4장에서는 실제 자료에 대하여 기존의 측도와 본 논문에서 제안한 측도의 적용결과를 비교하였다. 그리고 5장에서는 결론과 추후 연구방향을 언급하였다.

2. 기호의 정의 및 기존의 공개위험 측도

2.1. 기호의 정의

이번 장에서는 마이크로데이터의 공개와 관련하여 기호의 정의와 기존의 공개위험 측도에 대해 소개한다. 본 논문에서는 Kim 등 (2011a)이 정의한 기호를 사용하였으며 다음과 같다.

(1) 모집단 개체(개인, 가구, 사업체 등)의 수: N.

(2) 식별 변수: X_i, $i = 1, \ldots, m$. X_i는 K_i개의 값을 가질 수 있음 (X_i: 범주형 변수 또는 연속형 변수를 범주화한 변수).

(3) 모집단을 식별변수들로 교차 분류할 때, 크기가 1이상인 j번째 범주에 속하는 개체 수(모집단 빈도수): F_j, $j = 1, \ldots, J$ ($F_j \geq 1$).
개체가 1개 이상 있는 식별변수들의 조합의 수: J(최대값 $= \prod_{i=1}^{m} K_i$).

(4) 모집단 빈도수가 r인 범주의 수: $N_r = \sum_{j=1}^{J} I(F_j = r)$, I: 지시함수 그러면, $\sum_{r=1}^{\infty} N_r = J$, $\sum_{r=1}^{\infty} rN_r = N$이고, N_1은 유일 개체의 수이다.

(5) 식별변수로 교차 분류할 때 j번째 범주의 k번째 관심 변수: Y_{jk}, $j = 1, \ldots, J$, $k = 1, \ldots, F_j$.

(6) 모집단에서 j번째 범주의 최대, 최소 관심변수 값: Y_j^M, Y_j^m, $j = 1, \ldots, J$.

(7) 모집단에서 j번째 범주의 최대, 최소 관심변수 값의 차이: $R_j = Y_j^M - Y_j^m$, $j = 1, \ldots, J$.

(8) 표본의 개체 수: n.

(9) 표본을 식별변수들로 교차 분류할 때, j번째 범주에 속하는 개체 수(표본 빈도수): f_j, $j = 1, \ldots, J$ ($f_j \geq 1$).

(10) 표본 빈도수가 r인 셀의 수: $n_r = \sum_{j=1}^{J} I(f_j = r)$ 그러면, $\sum_{r=1}^{\infty} rn_r = n$이다.

(11) 표본을 식별변수로 교차 분류할 때 j번째 범주의 k번째 관심변수: y_{jk}, $j = 1, \ldots, J$, $k = 1, \ldots, f_j$.

(12) 표본에서 j번째 범주의 최대, 최소 관심변수 값: y_j^M, y_j^m, $j = 1, \ldots, J$.

(13) 표본에서 j번째 범주의 최대, 최소 관심변수 값의 차이: $r_j = y_j^M - y_j^m$ $j = 1, \ldots, J$.

2.2. 기존의 마이크로데이터 공개위험 측도

2.2.1. 모집단을 공개하는 경우 전체집단을 공개할 경우 발생할 수 있는 정확한 공개위험 측도는 모집단 중 유일개체 수의 비율로 나타낼 수 있다. 즉, 유일개체의 비율은 외부인이 모집단에서 임의로 개체를 추출하는 방법에서 추출된 개체가 모집단에서 유일할 확률이다. 이 측도는 집단 전체를 공개할 때 사용할 수 있으며 대부분 통계자료들이 성별, 나이, 학력 등 범주형 변수(또는 범주형으로 수정할 수 있

는 변수)를 많이 포함하고 있어 유일개체 비율을 이용한 측도는 널리 이용되고 있다. 유일개체의 비율을 P_1이라 하고 이를 수식으로 표현하면 다음과 같다.

$$P_1 = \frac{N_1}{N} = \sum_{j=1}^{J} \frac{I(F_j = 1)}{N}.$$

2.2.2. 표본을 공개하는 경우 표본을 공개하는 경우의 공개위험 측도는 Zayatz (1991), Chen과 Keller-McNulty (1998), Takemura (1997), Fienberg와 Makov (1998)이 연구한 바 있으며 이들의 단점을 보완하여 Skinner와 Elliot (2002)이 제안한 다음의 측도가 있다.

$$\theta_1 = \sum_{j=1}^{J} I(f_j = 1) \Big/ \sum_{j=1}^{J} F_j I(f_j = 1).$$

이 측도는 관심변수 값이 포함된 표본이 공개되었을 때, 모집단에 속한 어느 한 개체가 누구인지 외부인이 알고 있을 경우의 공개위험 측도로서 본 논문에서 기본적으로 활용한 방법이며 간략히 소개하면 다음과 같다. 아래의 탐색 방법을 생각해 보자.

(1) 모집단에서 임의로 개체를 추출하고 그 개체를 선택 개체라 함.

(2) 선택 개체의 식별변수 값을 표본의 개체 값과 비교하여 같은 값을 가진 개체가 하나만 있으면 유일 매치라고 하고, 표본의 개체를 매칭 개체라 함.

(3) 매칭 개체와 선택 개체가 같으면 유일 매치를 옳은 매치라 함.

이 측도는 위의 탐색방법에서 유일 매치가 옳은 매치일 조건부 확률이다. 즉, 외부인이 모집단에서 임의로 추출한 개체가 표본의 임의 개체와 식별 변수 값이 일치할 때(가능한 경우의 수 = θ_1의 분모), 표본의 유일 개체가 모집단에서 추출된 개체와 일치할(가능한 경우의 수 = θ_1의 분자) 확률이다.

θ_1을 추정하기 위하여 Skinner와 Elliot (2002)은 다음의 방법을 제시하였다.

(1) 표본에서 임의로 하나의 개체를 제거한다 (선택 개체).

(2) 주어진 표본이 모집단에서 추출된 비율(표본 추출률 p)로 그 개체를 다시 표본에 포함시킨다.

(3) 제거된 개체가 식별 변수에 대하여 표본 개체와 유일 매치가 되는지를 확인하고, 그럴 경우 유일 매치가 옳은 매치인지 확인한다.

여기서 유일 매치가 되는 것은 다음의 두 가지 경우이다.

(1) $f_j = 1$인 집단에서 제거되었던 개체가 확률 p로 포함되는 경우. 유일 매치이며 옳은 매치가 되는 경우로 기댓값은 $n_1 p$이다.

(2) $f_j = 2$인 집단에서 제거되었던 개체가 확률 $(1-p)$로 다시 포함되지 않는 경우. 두 개체 중 하나가 제거되고 다시 포함되지 않는 기댓값은 $2n_2(1-p)$이다.

이 중에서 (1)의 경우가 옳은 매치가 되는 경우이다. 즉 유일한 개체가 제거된 후 그 개체(선택 개체)가 다시 포함되므로 선택 개체와 식별변수 값이 같은 개체는 선택 개체 자신뿐이므로 유일 매치가 되고 옳은 매치가 된다. 따라서 θ_1에 대한 추정치는 다음과 같이 나타낼 수 있으며, Skinner와 Elliot (2002)은 실제 사례에 위 추정치를 적용한 결과 실제 값 θ_1과 큰 차이가 나지 않음을 보였다.

$$\hat{\theta}_1 = \frac{n_1 p}{n_1 p + 2n_2(1-p)}.$$

Table 3.1. Grades of a sample size 10 from freshmen of college A

모집단(100명)	표본(10명) 공개
- 남자, A고교: 9명 - 남자, C고교: 19명 - 남자, D고교: 20명 - 남자, F고교: 10명 - 남자, G고교: 3명 - 여자, B고교: 21명 - 여자, G고교: 3명 - 여자, H고교: 15명	

no	성별	출신고	학점	no	성별	출신고	학점
1	남자	A고교	C	6	남자	D고교	C
2	남자	A고교	C	7	남자	G고교	C
3	남자	C고교	F	8	여자	B고교	B
4	남자	D고교	A	9	여자	B고교	A
5	남자	D고교	A	10	여자	H고교	C

3. 유사성 개념을 적용한 공개위험 측도 제안

3.1. 기본 개념

어느 대학교 A단과대학 신입생 100명의 성적표 중 표본 10명에 대한 성적표를 익명으로 공개할 경우 공개위험을 측정해보자. 공개되는 정보는 Table 3.1과 같이 {성별, 출신고교, 학점}이며 성별과 출신고교는 식별변수이고 학점은 관심변수이다. Skinner와 Elliot (2002)이 제안한 측도에 따르면, 표본에서 식별변수 기준으로 유일개체는 3개(♯3, ♯7, ♯10)이며 출신고교는 각각 C고교, G고교, H고교이므로 해당 표본의 공개위험 측도는

$$\theta_1 = \frac{1+1+1}{19+3+15} = 0.081$$

이다. 하지만 표본 중 ♯1과 ♯2 개체는 유일개체가 아니어서 θ_1에서는 고려대상이 되지 않지만 이 개체들의 관심변수 값이 C학점으로 동일하므로 이 개체들도 유일개체와 같은 부류로 취급하여 공개위험을 계산하는 것이 필요하다. 따라서 ♯1과 ♯2 개체를 포함한 새로운 공개위험을 계산하면 다음과 같다

$$\frac{2+1+1+1}{9+19+3+15} = 0.109.$$

이와 같은 개념에 따라 본 논문에서는 유일개체 뿐만 아니라 유일개체가 아니더라도 같은 범주에서 관심변수 값이 모두 일치하거나 근사한 값을 가지는 경우까지 고려한 새로운 공개위험 측도를 제안하고자 한다. 새롭게 제안하는 내용은 Skinner와 Elliot (2002)이 제안한 공개위험 측도를 기반으로 표본을 공개하는 경우에 적용하기 위한 공개위험 측도이며, 모집단을 공개하는 경우의 공개위험 측도는 Hwang (2012)이 소개한 바가 있으며 다음과 같다.

$$\begin{aligned} P_2 &= \sum_{j=1}^{J} \frac{F_j I(R_j = 0)}{N} \\ &= \sum_{j=1}^{J} \frac{I(F_j = 1)}{N} + \sum_{j=1}^{J} \frac{F_j I(F_j \geq 2) I(R_j = 0)}{N}. \\ P_3 &= \sum_{j=1}^{J} \frac{F_j I(R_j \leq c)}{N} \\ &= \sum_{j=1}^{J} \frac{I(F_j = 1)}{N} + \sum_{j=1}^{J} \frac{F_j I(F_j \geq 2) I(R_j \leq c)}{N}. \end{aligned}$$

P_2는 관심변수 값이 이산형이며 모두 일치하는 경우를 고려한 측도이고 P_3는 관심변수 값이 연속형이며 허용치(c) 이하로 근사한 경우를 고려한 측도이다.

3.2. 관심변수 값이 이산형일 때

2.2.2에서 소개한 개념을 확장하여 다음의 탐색방법을 생각해 보자.

(1) 모집단에서 임의로 개체를 추출. 그 개체를 선택 개체라 함.

(2) 선택 개체의 식별변수 값을 표본의 개체 값과 비교하여 같은 값을 가진 개체가 하나이거나, 둘 이상이더라도 그 개체들의 관심변수 값이 모두 동일하면 동일 매치라고 하고 표본의 집단을 매칭 집단이라고 함.

(3) 매칭 집단의 개체가 선택 개체와 같으면 동일 매치를 옳은 매치라고 함.

이 경우 동일 매치가 옳은 매치일 조건부 확률은 다음과 같으며 θ_1과 구별하기 위해 θ_2라 하자. 즉, 외부인이 모집단에서 임의로 추출한 개체와 식별변수 값과 일치하는 표본 개체들이 있고 그 표본 개체들의 관심변수 값이 동일할 때(가능한 경우의 수 = θ_2의 분모), 표본 개체들이 모집단에서 추출된 개체와 일치할(가능한 경우의 수 = θ_2의 분자) 확률이다.

$$\begin{aligned}\theta_2 &= \sum_{j=1}^{J} f_j I(r_j = 0) \Big/ \sum_{j=1}^{J} F_j I(r_j = 0) \\ &= \left\{ \sum_{j=1}^{J} I(f_j = 1) + \sum_{j=1}^{J} f_j I(f_j \geq 2) I(r_j = 0) \right\} \Big/ \sum_{j=1}^{J} F_j I(r_j = 0).\end{aligned}$$

또한 θ_2을 추정하기 위하여 다음의 방법을 생각해 보자.

(1) 표본에서 임의로 하나의 개체를 제거한다.

(2) 주어진 표본이 모집단에서 추출된 비율(표본 추출률 p)로 그 개체를 다시 표본에 포함시킨다.

(3) 제거된 개체가 식별 변수에 대하여 표본 개체와 동일 매치가 되는지를 확인하고, 그럴 경우 동일 매치가 옳은 매치인지 확인한다.

여기서 동일 매치가 되는 것은 다음의 몇 가지 경우이다.

(1) $r_j = 0$인 집단에서 제거되었던 개체가 확률 p로 포함되는 경우. 동일 매치이며 옳은 매치가 되는 경우로 기댓값은 $\sum_{j=1}^{J} f_j I(r_j = 0)p$이다.

(2) $r_j = 0$, $f_j \geq 2$인 집단에서 제거되었던 개체가 확률 $(1-p)$로 다시 포함되지 않는 경우. 동일 매치이나, 옳은 매치가 아닌 경우로 기댓값은 $\sum_{j=1}^{J} f_j I(r_j = 0) I(f_j \geq 2)(1-p)$이다.

(3) $r_j \neq 0$이면서 하나만 제거하면 $r_j = 0$이 되는 집단(Gr^0집단이라 하자)에서 제거 되었던 개체가 확률 $(1-p)$로 다시 포함되지 않는 경우. 해당 집단에서 관심변수가 다른 개체(유일한 개체)가 제거되고 다시 포함되지 않는 경우로 기댓값은 $n_{(2)} \times (1-p)$이다. 여기서 $n_{(2)}$는 Gr^0집단의 수를 말한다.

예를 들어 관심변수가 급여일 때 {남자, 20대, 고졸, 200만원}, {남자, 20대, 고졸, 200만원}, {남자, 20대, 고졸, 250만원} 집단에서 첫 번째 혹은 두 번째 개체는 제거된 후 다시 포함되지 않더라도 동일 매치가 아니며, 세 번째 개체는 제거되고 다시 포함되지 않으면 동일 매치이나 옳은 매치는 아니다.

이 중에서 (1)의 경우가 옳은 매치가 되는 경우이며 θ_2에 대한 추정치는 다음과 같이 정리할 수 있다.

$$\hat{\theta_2} = \frac{\sum_{j=1}^{J} f_j I(r_j = 0)p}{\sum_{j=1}^{J} f_j I(r_j = 0)p + \sum_{j=1}^{J} f_j I(r_j = 0)I(f_j \geq 2)(1-p) + n_{(2)}(1-p)}.$$

위 식의 $n_{(2)}$를 구하기 위해 추가로 정의한 기호들은 다음과 같다.

(1) $y_{j(1)}, \ldots, y_{j(f_j)}$: $y_{j1}, \ldots, y_{jf_j}$의 순서통계량.

(2) $dn_j = (y_{j1}, y_{j2}, \ldots, y_{jf_j})$ 중 값이 다른 y_{ji}의 수. 즉 $dn_j = 1 \Leftrightarrow r_j = 0$.

(3) $r_j^{(-1)} = y_{j(f_j)} - y_{j(2)}$: j번째 집단에서 가장 작은 관심변수 값이 제거된 후의 관심변수 값의 범위.

(4) $r_j^{(-f_j)} = y_{j(f_{j-1})} - y_{j(1)}$: j번째 집단에서 가장 큰 관심변수 값이 제거된 후의 관심변수 값의 범위.

이 기호들을 이용하여 $n_{(2)}$ 정리하면 다음과 같다.

$$n_{(2)} = \sum_{j=1}^{J} 2I(f_j = 2)I(dn_j = 2) + \sum_{j=1}^{J} I(f_j \geq 3)I(dn_j = 2)\left[I\left(r_j^{(-1)} = 0\right) + I\left(r_j^{(-f_j)} = 0\right)\right].$$

$n_{(2)}$의 첫 번째 합은 j번째 집단에 두 개의 개체가 있고 관심변수 값이 서로 다른 경우 두 개체 중 하나가 제거되는 상황이므로 2배를 한다. 두 번째 합은 j번째 집단에 세 개 이상의 개체가 있고 관심변수 값이 다른 개체가 유일한 경우로서 이는 세 개 이상의 개체들 중에서 가장 작은 값 또는 가장 큰 값만 다르고 나머지 값들은 모두 같은 경우를 나타낸다. 예를 들어 두 번째 합의 경우 관심변수 값이 $(100, 100, 150)$, $(100, 150, 150)$, $(100, 100, 100, 150)$, $(100, 150, 150, 150)$, $(100, 100, 150, 150)$인 집단 중 처음 4개의 집단은 해당되나 5번째 집단은 해당되지 않는다.

3.3. 관심변수 값이 연속형일 때

3.2에서 정리한 방식을 확장하여 다음의 탐색방법을 생각해 보자.

(1) 모집단에서 임의로 개체를 추출. 그 개체를 선택 개체라 함.

(2) 선택 개체의 식별변수 값을 표본의 개체 값과 비교하여 같은 값을 가진 개체가 하나이거나, 둘 이상이더라도 그 개체들의 관심변수 값이 유사하면 유사 매치라고 하고 표본의 집단을 매칭 집단이라고 함.

(3) 매칭 집단의 개체가 선택 개체와 같으면 유사 매치를 옳은 매치라고 함.

이 경우 유사 매치가 옳은 매치일 조건부 확률은 다음과 같으며 θ_1, θ_2와 구별하기 위해 θ_3라 하자. 즉, 외부인이 모집단에서 임의로 추출한 개체와 식별변수 값과 일치하는 표본 개체들이 있고 그 표본 개체들의 관심변수 값이 유사할 때(가능한 경우의 수 $= \theta_3$의 분모), 표본 개체들이 모집단에서 추출된 개체와 일치할(가능한 경우의 수 $= \theta_3$의 분자) 확률이다. 관심변수 값들의 차이가 허용치(c) 이하이면 유사하다고 본다.

$$\begin{aligned}\theta_3 &= \sum_{j=1}^{J} f_j I(r_j \leq c) \Big/ \sum_{j=1}^{J} F_j I(r_j \leq c) \\ &= \left\{\sum_{j=1}^{J} I(f_j = 1) + \sum_{j=1}^{J} f_j I(f_j \geq 2)I(r_j \leq c)\right\} \Big/ \sum_{j=1}^{J} F_j I(r_j \leq c)\end{aligned}$$

또한 θ_3을 추정하기 위하여 다음의 방법을 생각해 보자.

(1) 표본에서 임의로 하나의 개체를 제거한다.

(2) 주어진 표본이 모집단에서 추출된 비율(표본 추출률 p)로 그 개체를 다시 표본에 포함시킨다.

(3) 제거된 개체가 식별 변수에 대하여 표본 개체와 유사 매치가 되는지를 확인하고, 그럴 경우 유사 매치가 옳은 매치인지 확인한다.

여기서 유사 매치가 되는 것은 다음의 몇 가지 경우이다.

(1) $r_j \le c$인 집단에서 제거되었던 개체가 확률 p로 포함되는 경우. 유사 매치이며 옳은 매치가 되는 경우로 기댓값은 $\sum_{j=1}^{J} f_j I(r_j \le c)p$이다.

(2) $r_j \le c, f_j \ge 2$인 집단에서 제거되었던 개체가 확률 $(1-p)$로 다시 포함되지 않는 경우. 유사 매치이나, 옳은 매치가 아닌 경우로 기댓값은 $\sum_{j=1}^{J} f_j I(r_j \le c)I(f_j \ge 2)(1-p)$이다.

(3) $r_j > 0$이면서 하나만 제거하면 $r_j \le c$이 되는 집단(Gr^c집단이라 하자)에서 제거 되었던 개체가 확률 $(1-p)$로 다시 포함되지 않는 경우. 해당 집단에서 관심변수가 다른 개체(유일한 개체)가 제거되고 다시 포함되지 않는 경우로 기댓값은 $n_{(3)} \times (1-p)$이다. 여기서 $n_{(3)}$는 Gr^c집단의 수를 말한다.

예를 들어 관심변수가 급여이고 $c = 30$만원일 때 {남자, 20대, 고졸, 200만원}, {남자, 20대, 고졸, 210만원}, {남자, 20대, 고졸, 250만원} 집단에서 첫 번째 혹은 두 번째 개체는 제거된 후 다시 포함되지 않더라도 유사 매치가 아니며, 세 번째 개체는 제거되고 다시 포함되지 않으면 유사 매치이나 옳은 매치는 아니다.

이 중에서 (1)의 경우가 옳은 매치가 되는 경우이며 θ_3에 대한 추정치는 다음과 같이 정리할 수 있다.

$$\hat{\theta}_3 = \frac{\sum_{j=1}^{J} f_j I(r_j \le c)p}{\sum_{j=1}^{J} f_j I(r_j \le c)p + \sum_{j=1}^{J} f_j I(r_j \le c)I(f_j \ge 2)(1-p) + n_{(3)}(1-p)}.$$

이 식에서 $n_{(3)}$는 다음과 같다.

$$n_{(3)} = \sum_{j=1}^{J} 2I(r_j > c)I(f_j = 2)I(dn_j = 2) + \sum_{j=1}^{J} I(r_j > c)I(f_j \ge 3)\left[I\left(r_j^{(-1)} \le c\right) + I\left(r_j^{(-f_j)} \le c\right)\right].$$

$n_{(3)}$의 첫 번째 합은 j번째 집단에 두 개의 개체가 있고 관심변수 값의 차이가 c를 초과하며, 두 개체 중 하나가 제거되는 상황이므로 2배를 한다. 두 번째 합은 j번째 집단에 세 개 이상의 개체가 있고 관심변수 값이 가장 작거나 가장 큰 경우를 제거하고 나머지 개체들의 관심변수 값의 범위이가 c이하인 경우를 나타낸다. 예를 들어 두 번째 합의 경우 관심변수 값이 $(100, 120, 150)$, $(100, 120, 140)$, $(100, 125, 150)$인 집단에 대해 $c = 20$을 적용하면 처음 2개의 집단은 해당되나 3번째 집단은 해당되지 않는다.

4. 실제 자료에 의한 적용 사례 비교

본 논문에서 제안한 마이크로데이터의 공개위험 측도를 기존 측도와 비교하기 위해 '2010년 가계 동향 조사'에서 획득한 총 10,667가구의 자료를 모집단으로 사용하여 각각 20%, 10%, 5%, 2% 표본에 대해

Table 4.1. Variables of interest

변수 종류	변수	변수 값
식별변수	가구주의 성별	1: 남자, 2: 여자
	가구주의 나이	16세∼90세(17세 없음)
	가구원 수	1인∼9인
관심변수 (민감변수)	가구주의 학력	0: 무학, 1: 초, 2: 중, 3: 고, 4: 전문대, 5: 대학, 6: 석사, 7: 박사
	가구주의 월평균 소득	단위: 원

Table 4.2. Recoding strategy for the variable of Age and Members of Household

식별변수	원자료	리코딩 방법1	리코딩 방법2
나이	16세∼90세 개별 값	5세 단위로 (20세 미만 통합, 80세 이상 통합)	10세 단위로 (20세 미만 통합, 80세 이상 통합)
변수명	a0	a1	a2
가구원 수	1인∼9인 개별 값	4개 값으로 (1인, 2∼3인, 3∼5인, 6인 이상)	3개 값으로 (1∼2인, 3∼4인, 5인 이상)
변수명	m0	m1	m2

공개위험을 계산하였다. 식별변수로는 가구주의 성별, 가구주의 나이, 가구원 수를 사용하였고 관심변수로는 가구주의 학력(이산형)과 가구주의 월평균 소득(연속형)으로 처리하였다. 원자료는 Table 4.1과 같이 구성되어 있으며, Table 4.2와 같이 가구주의 나이와 가구원 수를 리코딩한 자료에 대해서도 공개위험을 계산하였다. 공개위험 측도에 대한 추정치의 정확도를 비교하기 위해 리코딩 방법에 따른 9가지 식별변수 조합에 대해 4가지 표본 추출률별로 각각 1,000개의 표본을 임의로 추출하여 MAPE 값을 산출하였으며 그 결과는 Table 4.3과 Table 4.4에 정리하였다. MAPE(mean absolute percentage error)는 다음과 같이 구하며, $\theta_{ij} = 0$인 경우 θ_{ij}값은 제외하였다.

$$\text{MAPE}(\theta_i) = \frac{\sum_{j=1}^{1000} |\hat{\theta}_{ij} - \theta_{ij}|}{\theta_{ij} \times 1000} \times 100\%, \quad i = 1, 2, 3.$$

Table 4.3과 Table 4.4에서 보는 바와 같이 θ_2와 θ_3는 36종류의 모든 표본에서 기존의 측도인 θ_1와 평균값이 유사하게 나타나 새로운 공개위험 측도로 활용할 수 있음을 알 수 있으며, MAPE(θ_2)와 MAPE(θ_3)는 36종류 표본 모두에 대해 MAPE(θ_1) 보다 작게 나타나 $\hat{\theta}_2$와 $\hat{\theta}_3$는 $\hat{\theta}_1$에 비해 양호한 추정치임을 알 수 있다. Table 4.5와 Figure 4.1은 20% 표본 1,000개에 대해 공개위험 측도와 그 추정치간의 차이값의 빈도를 나타낸 것으로 새로 제안한 공개위험 측도가 기존 공개위험 측도에 비해 참값을 잘 추정함을 알 수 있다.

5. 결론

마이크로데이터의 공개위험과 관련한 기존의 공개위험 측도는 식별변수의 조합으로 이루어진 범주내에서 유일개체의 비율을 활용하지만, 본 연구에서는 범주내에 2개 이상의 개체가 있는 경우에도 관심변수의 값들이 일치하거나 어느 기준치 이하로 유사한 경우를 추가로 고려한 새로운 공개위험 측도를 제안하였다. 새로운 공개위험 측도는 기존의 측도가 측정하지 못하는 공개위험까지 추가로 고려하므로 더 발전된 개념의 측도로 볼 수 있다.

본 논문에서 제안한 공개위험 측도를 비교하기 위해 통계청의 '2010 가계동향 조사'의 결과를 사례로 활용하였다. 식별변수로는 가구주의 성별, 나이, 가구원수로 정하여 기존의 공개위험 측도와 새로이 제

Table 4.3. Comparison of Disclosure Risk Measurements for educational attainment

리코딩 방법	표본 추출률	$\bar{\theta}_1$	$\bar{\hat{\theta}}_1$	$\bar{\theta}_2$	$\bar{\hat{\theta}}_2$	MAPE(θ_1)	MAPE(θ_2)
a0m0 (원자료)	20%	0.1790	0.1805	0.1864	0.1871	11.41%	7.11%
	10%	0.0970	0.0982	0.1014	0.1019	12.02%	8.36%
	5%	0.0580	0.0585	0.0603	0.0605	12.79%	9.42%
	2%	0.0355	0.0362	0.0361	0.0366	17.73%	14.10%
a0m1	20%	0.1662	0.1666	0.1787	0.1785	13.59%	8.96%
	10%	0.0970	0.0986	0.1015	0.1019	14.22%	10.00%
	5%	0.0565	0.0576	0.0572	0.0576	16.80%	11.84%
	2%	0.0277	0.0281	0.0280	0.0283	20.16%	14.53%
a0m2	20%	0.1788	0.1803	0.1784	0.1783	18.53%	11.30%
	10%	0.0780	0.0788	0.0837	0.0836	18.10%	11.18%
	5%	0.0423	0.0427	0.0470	0.0472	15.73%	11.39%
	2%	0.0261	0.0270	0.0271	0.0275	19.28%	14.32%
a1m0	20%	0.2034	0.1995	0.1973	0.1932	30.03%	19.83%
	10%	0.0917	0.0925	0.0911	0.0905	31.01%	18.99%
	5%	0.0437	0.0450	0.0449	0.0449	30.81%	19.40%
	2%	0.0193	0.0202	0.0202	0.0204	30.60%	19.95%
a1m1	20%	0.1787	0.1816	0.1781	0.1716	43.54%	27.44%
	10%	0.0823	0.0847	0.0833	0.0821	42.02%	25.12%
	5%	0.0398	0.0423	0.0425	0.0424	40.80%	24.14%
	2%	0.0194	0.0212	0.0202	0.0209	40.13%	25.63%
a1m2	20%	0.1747	0.1804	0.1837	0.1804	53.76%	32.75%
	10%	0.0948	0.1123	0.0980	0.0994	66.87%	36.79%
	5%	0.0491	0.0569	0.0459	0.0451	69.85%	33.09%
	2%	0.0151	0.0159	0.0162	0.0160	45.51%	26.52%
a2m0	20%	0.2187	0.2239	0.2153	0.2099	45.33%	30.14%
	10%	0.1072	0.1172	0.1018	0.1007	58.86%	32.89%
	5%	0.0446	0.0461	0.0436	0.0425	50.57%	27.90%
	2%	0.0180	0.0194	0.0189	0.0192	42.39%	27.26%
a2m1	20%	0.2103	0.2533	0.2071	0.2119	83.89%	50.24%
	10%	0.1131	0.1515	0.1005	0.1015	115.81%	54.51%
	5%	0.0431	0.0499	0.0400	0.0381	89.29%	36.94%
	2%	0.0164	0.0204	0.0178	0.0182	69.57%	33.47%
a2m2	20%	0.2133	0.2636	0.2001	0.1863	130.83%	55.66%
	10%	0.0995	0.1685	0.0983	0.1021	167.38%	62.80%
	5%	0.0545	0.1031	0.0513	0.0499	192.56%	57.79%
	2%	0.0186	0.0411	0.0171	0.0179	218.54%	46.41%

안한 공개위험 측도를 원자료와 비밀보호 처리된 자료로 비교하였다. 사례연구 결과 θ_2와 θ_3는 기존의 θ_1과 유사한 값을 나타내어 새로운 공개위험 측도로 활용할 수 있음을 알 수 있으며, 추정치 또한 기존의 측도에 비해 상당히 양호한 것으로 나타났다.

본 연구는 마이크로데이터의 공개위험 측도를 새롭게 제안한 것에 국한하였지만 향후에는 새로 제안한 측도와 연계하여 해당 마이크로데이터의 유용성을 추가로 연구할 예정이다.

Table 4.4. Comparison of Disclosure Risk Measurements for monthly incomes

리코딩 방법	표본 추출률	$\bar{\theta}_1$	$\bar{\hat{\theta}}_1$	$\bar{\theta}_3$	$\bar{\hat{\theta}}_3$	MAPE(θ_1)	MAPE(θ_3)
a0m0 (원자료)	20%	0.1790	0.1805	0.1925	0.1931	11.41%	5.49%
	10%	0.0970	0.0982	0.1052	0.1057	12.02%	6.93%
	5%	0.0580	0.0585	0.0624	0.0625	12.79%	8.64%
	2%	0.0355	0.0362	0.0369	0.0374	17.73%	13.44%
a0m1	20%	0.1662	0.1666	0.1886	0.1888	13.59%	6.38%
	10%	0.0970	0.0986	0.1069	0.1078	14.22%	8.12%
	5%	0.0565	0.0576	0.0598	0.0603	16.80%	10.34%
	2%	0.0277	0.0281	0.0289	0.0292	20.16%	13.83%
a0m2	20%	0.1788	0.1803	0.1819	0.1825	18.53%	8.22%
	10%	0.0780	0.0788	0.0875	0.0875	18.10%	8.59%
	5%	0.0423	0.0427	0.0496	0.0498	15.73%	9.64%
	2%	0.0261	0.0270	0.0282	0.0287	19.28%	13.44%
a1m0	20%	0.2034	0.1995	0.1928	0.1891	30.03%	15.02%
	10%	0.0917	0.0925	0.0912	0.0908	31.01%	14.26%
	5%	0.0437	0.0450	0.0461	0.0460	30.81%	14.28%
	2%	0.0193	0.0202	0.0211	0.0211	30.60%	16.44%
a1m1	20%	0.1787	0.1816	0.1818	0.1779	43.54%	17.80%
	10%	0.0823	0.0847	0.0864	0.0862	42.02%	16.37%
	5%	0.0398	0.0423	0.0448	0.0445	40.80%	16.33%
	2%	0.0194	0.0212	0.0215	0.0219	40.13%	20.21%
a1m2	20%	0.1747	0.1804	0.1789	0.1727	53.76%	29.36%
	10%	0.0948	0.1123	0.0938	0.0919	66.87%	26.99%
	5%	0.0491	0.0569	0.0451	0.0440	69.85%	22.83%
	2%	0.0151	0.0159	0.0171	0.0168	45.51%	20.29%
a2m0	20%	0.2187	0.2239	0.2042	0.1985	45.33%	24.09%
	10%	0.1072	0.1172	0.0952	0.0931	58.86%	22.48%
	5%	0.0446	0.0461	0.0442	0.0433	50.57%	19.65%
	2%	0.0180	0.0194	0.0195	0.0196	42.39%	21.24%
a2m1	20%	0.2103	0.2533	0.2035	0.2001	83.89%	34.82%
	10%	0.1131	0.1515	0.0925	0.0885	115.81%	29.32%
	5%	0.0431	0.0499	0.0416	0.0403	89.29%	24.35%
	2%	0.0164	0.0204	0.0189	0.0189	69.57%	22.89%
a2m2	20%	0.2133	0.2636	0.1947	0.1745	130.83%	51.34%
	10%	0.0995	0.1685	0.0924	0.0927	167.38%	47.36%
	5%	0.0545	0.1031	0.0472	0.0444	192.56%	35.73%
	2%	0.0186	0.0411	0.0175	0.0175	218.54%	33.46%

Table 4.5. Frequencies and means of $(\hat{\theta}_i - \theta_i)$ under provision of a 20% sample

구간		$(\hat{\theta}_1 - \theta_1)$빈도	$(\hat{\theta}_2 - \theta_2)$빈도	$(\hat{\theta}_3 - \theta_3)$빈도
R1 (−0.080, −0.075]		1		
R2 (−0.075, −0.065]		2		
R3 (−0.065, −0.055]		5		
R4 (−0.055, −0.045]		23	2	1
R5 (−0.045, −0.035]		45	9	2
R6 (−0.035, −0.025]		78	48	21
R7 (−0.025, −0.015]		102	120	87
R8 (−0.015, −0.005]		151	188	225
R9 (−0.005, 0.005]		163	235	312
R10 (0.005, 0.015]		132	204	201
R11 (0.015, 0.025]		124	120	121
R12 (0.025, 0.035]		84	56	26
R13 (0.035, 0.045]		42	17	2
R14 (0.045, 0.055]		23	1	2
R15 (0.055, 0.065]		18		
R16 (0.065, 0.075]		4		
R17 (0.075, 0.085]		2		
R18 (0.085, 0.095]		1		
계		1,000	1,000	1,000
$(\hat{\theta}_i - \theta_i)$의	평균	0.0015	0.0007	0.0006
	표준편차	0.0255	0.0164	0.0133

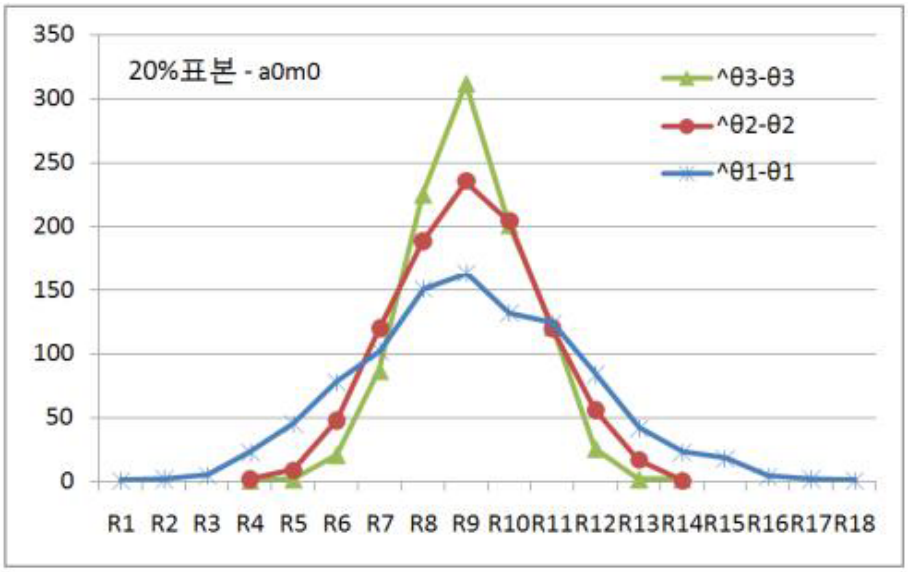

Figure 4.1. Frequencies of $(\hat{\theta}_i - \theta_i)$ under provision of a 20% sample

References

Bethlehem, J., Keller, W. and Pannekoek, J. (1990). Disclosure control of microdata, *Journal of the American Statistical Association*, **85**, 38–45.

Chen, G. and Keller-McNulty, S. (1998). Estimation of identification disclosure risk in microdata, *Journal of Official Statistics*, **14**, 79–85.

Duncan, G., Keller-McNulty, S. and Stokes, S. (2001). Disclosure risk vs data utility: The R-U confidentiality map, Technical Report Number 121 December 2001, *National Institute of Statistical Sciences*.

Eurostat (1996). Manual on Disclosure Control Methods. Luxembourg, Office for Official Publications of the European Communities, Technical Report Number 153 June 2006, *National Institute of Statistical Sciences*.

FCSM(Federal Committee on Statistical Methodology) (2005). Statistical Policy Working Paper 22(second version).

Fienberg, S. E. and Makov, U. E. (1998). Confidentiality, uniqueness, and disclosure limitation for categorical data, *Journal of Official Statistics*, **14**, 385–397.

Gomatam, S., Karr, A. and Sanil, A. (2003). A Risk_Utility Framework for Categorical Data Swapping, Technical Report Number 132 February 2003, *National Institute of Statistical Sciences*.

Huda, M. N., Yamada, S. and Sonehara, N. (2010). On Identity Disclosure Risk Measurement for Shared Microdata, *World Academy of Science, Engineering and Technology*, **70**, 310–317.

Hwang, H. S. (2012). *The Study on the Extended Measurement of Disclosure Risk of Microdata*, Master Dissertation, Chungnam National University.

Jeong, D. M., Kim, J. J. and Kim, K. M. (2009). A method of masking based on multiplicative noise, *The Korean Journal of Applied Statistics*, **22**, 141–151.

Kim, K. Y. (2006). *A Study on the Statistical Confidentiality Methodology and Variance Estimation for Census Survey Data*, Doctoral Dissertation, Chungnam National University.

Kim, K. Y., Kwon, D. H., Shin, J. E. and Lee, S. H. (2011a). *Introduction to Statistical Methods for Confidentiality*, FreeAcademy.

Kim, Y.-W., Kim, T.-Y. and Kim, K.-N. (2011b). Application of a statistical disclosure control techniques based on multiplicative noise, *The Korean Journal of Applied Statistics*, **24**, 127–136.

Kwon, D. H. (2009). *A Study on Disclosure Control Method for Disclosure Risk and Utility of Microdata*, Doctoral Dissertation, Chungnam National University.

Machanavajjhala, A., Gehrke, J. and Kifer, D. (2006). -diversity: Privacy beyond-anonymity, In *Proceedings of the International Conference on Data Engineering Atlanta Engineering*, Atlanta.

Samarati, P. (2001). Protecting respondents' identities in microdata release, *IEEE Transactions on Knowledge and Data Engineering*, **13**, 1011–1027.

Shlomo, N. (2010). Releasing Microdata: Disclosure risk estimation, data masking and assessing utility, *Journal of Privacy Confidentiality*, **2**, 73–91.

Skinner, C. J. and Elliot, M. J. (2002). A measure of disclosure risk for microdata, *Journal of Royal Statistical Society, B*, 855–867.

Takemura, A. (1997). *Some Superpopulation Models for Estimating the Number of Population Uniques*, University of Tokyo, September 1997.

Xiao, X., Tao, Y. and Koudas, N. (2010). Transparent anonymization: Thwarting adversaries who know the algorithm, *ACM Transactions on Database System*, **35**, April 2010.

Zayatz, L. (1991). Estimation of the number of unique population elements using a sample, In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Alexandria, VA, 369–373.