한국통계학회논문집
2010, 17권, 6호, 791–801

DOI: 10.5351/CKSS.2010.17.6.791

마이크로데이터 제공에 따른 임계모집단 크기 결정

남궁 평[1,a], 소정현[b]

[a]성균관대학교 통계학과, [b]충청지방통계청 서산사무소

요 약

마이크로데이터 제공시 발생될 수 있는 노출(disclosure)과 노출위험을 나타내는데 사용되는 측도인 유일성(uniqueness) 그리고 모집단 유일성의 개수를 추정하기 위한 초모집단 모형으로 Multinomial-Dirichlet 모형, Takemura의 Poisson-Gamma 모형, Modified Multinomial-Dirichlet 모형, Bethlehem의 Poisson-Gamma 모형을 다룬다. 이 4개의 모형에 대해 마이크로데이터 제공에 따른 임계모집단 크기(critical population size)를 결정한다.

주요용어: Superpopulation 모형, Multinomial-Dirichlet 모형, Poisson-Gamma 모형, 임계모집단 크기.

1. 서론

통계청 MDSS(Micro Data Service System)의 용어 정의에 의하면, 통계자료는 원자료(raw data), 마이크로데이터(micro data), 매크로데이터(macro data) 및 메타데이터(meta data)로 구분한다. 원자료는 통계조사 자료에서 최초 입력한 전산화일 자료로서 입력오류, 조사오류 등이 걸러지기 이전 단계의 자료를 말한다. 마이크로데이터는 원자료에서 입력오류, 조사오류 등을 수정하여 통계표 작성 등 데이터 가공의 기초 자료로 사용되는 자료로서, 통계표에서 얻을 수 없는 심층적인 분석을 원하는 다양한 계층의 이용자에게 제공하고 있다. 특히, 마이크로데이터는 과거 미시자료, 원시자료 또는 통계원시자료로 혼용해 사용되어 왔으나 통계청의 통계자료제공 규정(통계청 훈령 제186호, 2009.3.11)에서 마이크로데이터로 명명하기로 규정하였다.

매크로데이터는 마이크로데이터를 임의의 기준에 따라 집계한 자료로써 집계의 정도에 따라 세분된 자료부터 통합된 자료까지 다양하게 제공된다. 메타데이터는 통계자료 이용자의 이해를 돕기 위해 제공되는 통계의 작성 개요 및 자료의 측정 단위나 집계 기준 등 통계 전반의 내용을 설명하는 자료이다. 즉, 원자료, 마이크로데이터, 매크로데이터는 모두 수치로 이루어져 있어 의미 파악이 곤란하기 때문에 이 자료들을 설명하기 위한 자료가 필요한데 이를 메타데이터라고 한다. 예를 들어, 통계청 지정통계(제10,111호)인 광업·제조업 동향조사의 경우 조사 단위가 톤(ton), 킬로리터(kl), 제곱미터(m^2) 등 조사 품목에 따라 조사 단위가 달라지는데 이와 같이 통계자료에 대한 설명 자료가 메타데이터이다. 또한, 각 통계조사에 대한 지침서 및 조사표 등도 메타데이터에 속하며, 최근 들어 통계청에서는 메타데이터의 중요성을 인식하고 활발한 연구가 이루어지고 있다.

마이크로데이터에는 개인이나 법인 또는 단체에 대한 정보들이 모두 수록되어 있기 때문에 마이크로데이터를 그대로 이용자에게 제공할 경우 응답한 개인이나 법인 또는 단체에 대한 정보들이 노출되게 되어 개인의 사생활 침해와 경쟁업체에 정보 노출로 인해 극심한 사회문제의 발생 위험이 있다. 그래서 통계법 제33조(비밀의 보호)에는 통계작성기관이 통계의 과정에서 알려진 사항으로써 개인이나

[1] 교신저자: (449-791) 서울시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 교수. E-mail: namkung@skku.edu

법인 또는 단체 등의 비밀에 속하는 사항은 보호하고, 통계의 작성을 위하여 수집된 개인이나 법인 또는 단체 등의 비밀에 속하는 자료는 통계작성 이외의 목적으로 사용해서는 안 된다고 규정하고 있다. 하지만, 통계이용자들이 경제, 사회현상에 대한 다양하고 심층적인 분석을 위해 각종 통계조사에 대한 마이크로데이터 제공 요구가 급증하고 있으며, 이에 부응하여 통계청에서는 2006년 1월 1일부터 통계법 제30조(통계자료의 제공) 제2항, 제31조(통계자료의 이용) 제2항 및 제33조(비밀의 보호)에 의거 개인, 가구, 사업체 등의 통계조사 응답자가 제3자에 의해 식별되거나 비밀정보가 노출되지 않도록 통계적 노출조절기법(Statistical Disclosure Control)을 적용한 제공용 마이크로데이터를 제공하고 있다. 통계적 노출조절기법은 자료의 종류와 형태에 따라 다양한 방법이 있는데, 자료의 종류에 따라 매크로데이터인 경우 셀 감추기(cell suppression), 반올림(rounding), 임의변조(random perturbation) 등의 방법이 있고, 마이크로데이터인 경우 익명화(anonymisation), 표본추출(sampling), 그룹화(grouping), 자료교환(data swapping) 등의 방법이 널리 활용되고 있다. 자료의 형태에 따라 이산형 자료인 경우 자료교환, 코딩접근법(coding approach), 그룹화 등의 방법이 있고, 연속형 자료인 경우 자료교환, 반올림, 구간그룹화(grouping into interval), 가법잡음(additive noise), 승법잡음(multiplicative noise) 등의 방법이 활용된다 (정동명 등, 2008).

본 연구에서는 마이크로데이터 제공시 발생될 수 있는 노출(disclosure)과 노출위험을 나타내는데 보편적으로 사용되는 측도인 유일성(uniqueness)을 이용해 임계모집단크기를 결정하는 방안을 제안하고자 한다. 그래서 2장에서는 유일성의 개념과 표본자료를 이용하여 모집단 유일성의 개수를 추정하기 위한 초모집단 모형 방법으로 Multinomial-Dirichlet 모형, Takemura의 Poisson-Gamma 모형, Modified Multinomial-Dirichlet 모형, Bethlehem의 Poisson-Gamma 모형을 정리하고, 3장에서는 임계모집단크기 결정 기준을 설명하고, 초모집단 가정에서 유일성을 추정하는 4개의 모형을 이용하여 마이크로데이터 제공에 따른 임계모집단 크기(critical population size)를 결정하는 방안을 제시한다.

2. 모집단의 유일성 개수의 추정

2.1. 유일성 개수의 개념

노출이란 통계작성기관이 자료를 수집, 정리하여 다양한 형태의 통계정보로 제공할 경우, 이를 통해서 응답자의 특성이 파악되는 경우를 말하며 어떤 경우에도 노출은 발생되지 않도록 하는 것이 바람직하다. 이에 대해 Bethlehem 등 (1990)은 노출이 바람직하지 않은 것은 개인의 사생활이 보호되어야 하는 법적이고 윤리적인 이유뿐만 아니라 조사 응답률에 영향을 줄 수 있다는 실질적인 이유가 있기 때문이며, 가능하면 마이크로데이터를 제공할 때 노출을 제한해야 한다고 하였다.

식별은 외부 이용자가 그들의 자료파일을 이용하여 레코드와 특정 개인 간의 1 : 1 대응을 통하여 특정 레코드가 어떤 사람에 대한 정보인지를 식별 가능하게 하는 주요 변수(key variables)들과 관련이 있다. 잘 알려진 주요 변수로는 이름이나 주소 뿐만 아니라 가구 구성, 연령, 인종, 성별, 거주 지역, 직업 등이 있다.

유일성이란 전체 자료파일에서 조사단위의 특성이 유일하게 존재하는 것을 말하며 어떤 조사 단위가 식별될 가능성을 나타내는 측도로 사용된다. 유일성을 정의하기 위해서는 먼저 주요 변수(식별을 위해 사용된 변수들의 집합)들을 정의해야 하며, 여기서는 범주형 변수만을 고려하며 주요 변수에 속하는 범주의 수를 곱하면 서로 다른 K개의 셀(cells)들이 존재한다. 예를 들어, 주요 변수가 나이(6개 범주)와 성별(2개 범주)로 구성되어 있다면, 주요변수에 속하는 범주의 수를 곱한 12개의 서로 다른 셀이 존재하고, K는 12가 된다. 주요변수의 범주 i에 속하는 모집단에서의 원소들의 수를 $F_i (i = 1, 2, \ldots, K)$로 나타내고, 이에 대응되는 표본에서의 원소들의 수를 $f_i (i = 1, 2, \ldots, K)$로 나타낸다면, 누군가가 모집단에서 유일한 경우는 $F_i = 1$이며, 표본에서 유일한 경우는 $f_i = 1$이 된다. 모집

단의 크기를 N이라 할 때, 모집단에서 랜덤하게 선택된 한 사람이 주요변수의 범주 i에 속하는 확률은 $\pi_i = F_i/N$이며, 유일성은 주요 변수 또는 고려할 변수가 많아질수록 점점 커진다.

N개의 개체로 이루어진 유한 모집단(finite population) $\{y_1, y_2, \ldots, y_N\}$을 고려하자. 여기서 y_j는 j번째 개체의 특성값이고, $y_j = i$는 j번째 개체가 i번째 셀에 있는 것을 의미한다. 유일성의 개념에서 정의한 바와 같이 K는 주요 변수에 속하는 범주의 수를 곱한 서로 다른 셀의 총 개수이며, F_i는 셀 i의 모집단 빈도수로써 $F_i = \sharp\{y_j = i | j\}$로 나타낸다. 모집단의 빈도 벡터를 $\boldsymbol{F} = (F_1, F_2, \ldots, F_K)$라 하자. 셀 i에 유일한 개체가 존재한다면 $F_i = 1$이 되고, 이 개체를 모집단 유일성이라 하며, 모집단 유일성 개수를 다음과 같이 나타낸다.

$$U_P = \theta = S_1 = \sharp\{i | F_i = 1\}. \tag{2.1}$$

즉, 모집단 유일성 개수는 모집단에서 주요변수의 범주에 속한 모집단의 원소가 "1"인 범주의 수를 의미한다.

이제 표본에서 유일성 개수 θ를 추정하는 방안을 생각해 보자. 모집단으로부터 크기가 n인 특정한 표본을 추출 한다면, 모집단 유일성 총 개수보다는 오히려 표본에 포함된 모집단 유일성 개수를 추정할 수 있지만 단순임의추출 하에서는 표본에 포함된 모집단 유일성 개수의 분명한 추정량은 $\widehat{U_P} = \hat{\theta}n/N$으로 주어진다 (Bethlehem 등, 1990).

마이크로데이터의 노출 문제에 있어서 셀의 총 개수 K가 매우 크다면 각 셀들의 빈도 벡터 $\boldsymbol{F} = (F_1, F_2, \ldots, F_K)$를 추정하는 것은 어려운 문제이다. 표본자료로부터 셀의 범주가 많은 K개 모집단의 빈도를 추정하는 것이 어렵기 때문에 초모집단(superpopulation) 접근 방법을 이용해 추정하는 방법을 살펴보기로 한다. $\boldsymbol{F}$는 적은 수의 초모수(hyper parameters)들에 의해 결정되는 사전분포의 특정한 하나의 값으로 고려할 수 있으며, 이 사전분포는 초모집단으로부터 뽑은 가상의 표본으로 간주한다.

τ를 초모수라 하고 $P(F|\tau)$를 F의 확률밀도함수라 하면 주어진 τ값을 가진 특별한 초모집단 모형하에서 모집단 유일성 개수 θ는 사전분포에 의한 기대값 $E_r(\theta) = E_\tau(S_1) = \sum_{i=1}^{K} P(F_i = 1|\tau)$로 추정된다 (Takemura, 1997). 만약 표본으로부터 초모수 τ를 $\hat{\tau}$로 추정한다면 $\hat{\theta}$은 다음과 같다.

$$\hat{\theta} = E_{\hat{\tau}}(S_1) = \sum_{i=1}^{K} P(F_i = 1|\hat{\tau}). \tag{2.2}$$

2.2. 모형기반에서의 유일성 개수 추정

초모집단 모형에서는 크기 N의 유한 모집단을 초모집단으로부터 뽑은 크기 N의 확률표본으로 모집단으로 간주한다. 즉, 모집단 보다 더 큰 가상의 모집단을 초모집단이라 하며 이 모집단으로부터 뽑은 크기 N의 확률표본을 모집단으로 간주하는 것이다. 따라서 본 연구에서는 초모집단 모형기반에서 모집단 유일성을 추정하는 방법을 이용하여 표본으로부터 모집단 유일성 개수를 추정하는 방법을 제안하고자 한다. 본 절에서 검토하는 모형들은 모집단의 셀 빈도들이 초모집단 분포의 현실값이라는 가정에 근거를 둔다.

2.2.1. Multinomial-Dirichlet 모형 (Takemura, 1997; Pereira와 Stern, 2008; Bruce와 Peter, 2000)

모집단에서 랜덤하게 선택된 한 사람이 주요 변수의 범주 i에 속하는 확률은 $\pi_i = F_i/N$이므로, $\sum_{i=1}^{K} \pi_i = 1$인 것을 감안하여 $F_i = (F_1, \ldots, F_K)$의 사전분포가 Multinomial$(N, \pi_1, \pi_2, \ldots, \pi_K)$를 따른다고 가정하자. 또한, 확률 벡터 $\pi_i = (\pi_1, \ldots, \pi_K)$는 Dirichlet$(\alpha_1, \alpha_2, \ldots, \alpha_K)$를 따르는 확률변수 $\Pi_i = (\Pi_1, \Pi_2, \ldots, \Pi_K)$의 특정한 값으로 고려할 수 있다. 여기서 $\alpha_1, \alpha_2, \ldots, \alpha_K$은 F_i의 사전분포의 초모수이

다. F_i와 Π_i의 결합확률밀도함수는 다음과 같다.

$$P(F_i = x_i, \Pi_i = \pi_i) = \begin{pmatrix} N \\ x_1, \ldots, x_K \end{pmatrix} \frac{\Gamma(\alpha_1 + \cdots + \alpha_K)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \pi_1^{\alpha_1 + x_1 - 1} \cdots \pi_K^{\alpha_K + x_K - 1},$$

여기서, $\pi_K = 1 - \pi_1 - \cdots - \pi_{K-1}$이고, x_i는 확률변수 F_i에 대응되는 특정한 값이다. 위의 결합확률밀도함수를 π_i에 대해서 적분하면 F_i의 주변분포를 얻을 수 있다.

$$P(F_1 = x_1, \ldots, F_K = x_K) = \frac{N!\Gamma(\alpha_1 + \cdots + \alpha_K)}{\Gamma(\alpha_1 + \cdots + \alpha_K + N)} \frac{\Gamma(\alpha_1 + x_1)}{\Gamma(\alpha_1) x_1!} \cdots \frac{\Gamma(\alpha_K + x_K)}{\Gamma(\alpha_K) x_K!}. \tag{2.3}$$

위에서 구한 F_i의 주변분포를 Dirichlet Compound Multinomial 분포라고 한다. 여기서 $\sum_{i=1}^{K} \alpha_i = A$라 하면,

$$P(F_i = 1) = N\alpha_i \frac{\Gamma(A)\Gamma(A - \alpha_i + N - 1)}{\Gamma(A + N)\Gamma(A - \alpha_i)}$$

이고, 모집단 유일성 개수의 기대값은

$$U_P = E(S_1) = \sum_{i=1}^{K} P(F_i = 1|\alpha) = N \frac{\Gamma(A)}{\Gamma(A + N)} \sum_{i=1}^{K} \alpha_i \frac{\Gamma(A - \alpha_i + N - 1)}{\Gamma(A - \alpha_i)}$$

이 된다.

이 경우 셀 교환가능성(cell exchangeable)을 가정하여 $\alpha_1 = \alpha_2 = \cdots = \alpha_K = \alpha$라 놓으면, $A = K\alpha$가 되므로 U_P는 식 (2.4)와 같이 표현할 수 있다. 마이크로데이터의 특성상 많은 그룹의 부모집단으로 이루어진 모집단에서 부모집단이 모수가 같은 어떤 확률분포를 가진다고 가정하면 계산이 편리하기 때문에 초모집단 모형에서는 셀교환가능성을 인정하고 있다. Takemura (1997)의 연구를 보면, 셀교환가능성을 가정하지 않는 것이 보다 좋은 모형으로 적합시킬 수 있지만 계산과정이 복잡하고 어렵기 때문에 셀교환가능성을 가정하여 유일성 개수의 기대값을 유도하고 있다.

$$U_P = E(S_1) = \frac{N}{N-1} K\alpha \frac{B(K\alpha, N)}{B((K-1)\alpha, N-1)}, \tag{2.4}$$

여기서 B는 beta 분포를 의미한다.

U_P를 추정하기 위해서 Bethlehem 등 (1990)이 Poisson-Gamma 모형에서 사용하였던 적률추정법을 사용하여 표본적률과 모적률이 같다고 가정하면, $(n-1)/(K\hat{\alpha} + 1) = K/n\, s_f^2 - 1$이 된다. 여기서 $s_f^2 = \sum_{i=1}^{K} (f_i - \bar{f})^2/(K-1)$이고, $\hat{\alpha}$에 대해 정리하면 $\hat{\alpha} = (n^2 - Ks_f^2)/\{K(Ks_f^2 - n)\}$된다. 즉, 모집단 유일성 개수의 기대값에 대한 추정량 $\widehat{U_P}$는 식 (2.5)와 같다.

$$\widehat{U_P} = \frac{N}{N-1} \frac{n^2 - Ks_f^2}{\left(Ks_f^2 - n\right)} \frac{B\left(\frac{\left(n^2 - Ks_f^2\right)}{\left(Ks_f^2 - n\right)}, N\right)}{B\left(\frac{(K-1)}{K} \frac{\left(n^2 - Ks_f^2\right)}{\left(Ks_f^2 - n\right)}, N-1\right)}. \tag{2.5}$$

2.2.2. Takemura의 Poisson-Gamma 모형 (Takemura, 1997; Hoshino와 Takemura, 1998)

모집단에서 랜덤하게 선택된 한 사람이 주요 변수의 범주 i에 속하는 확률 $\pi_i = F_i/N$은 확률밀도함수가 $1/\{\Gamma(\alpha)\beta^{\alpha}\}x^{\alpha-1}e^{-x/\beta}$인 Gamma$(\alpha,\beta)$분포를 따르는 확률변수 Π_i의 특정한 값으로 간주할 수 있고, F_i는 기대값이 $N\pi_i$인 Poisson분포를 따른다고 가정하자. $\sum_{i=1}^{K}\Pi_i = 1$은 필수 조건이지만 모형 전개를 위해서 $\sum_{i=1}^{K}E(\Pi_i) = 1$이라 가정한다. 또한, 제약조건으로 모수 α를 $\alpha = 1/K\beta$라고 하면, Π_i의 공통 분포(common distribution)에서 한 모수만 결정되지 않은 상태가 된다. 결정되지 않은 모수 β는 Π_i의 공통평균(common mean) $1/K$의 주위에 퍼져있는 정도를 나타낸다. Takemura의 Poisson-Gamma 모형은 다음과 같이 요약할 수 있다.

$$F_i \sim \text{Poisson}(N\pi_i)|\pi_i = \Pi_i, \quad \Pi_i \sim \text{Gamma}(\alpha,\beta). \tag{2.6}$$

위의 가정 하에서 F_i와 Π_i의 결합확률밀도함수는 다음과 같다.

$$P(F_i = x_i, \Pi_i = \pi_i) = \frac{N^{x_i}\pi^{x_i+\alpha-1}e^{-\left(\frac{\beta N+1}{\beta}\right)\pi_i}}{x_i!\Gamma(\alpha)\beta^{\alpha}}, \quad x_i = 0,1,\ldots,N. \tag{2.7}$$

위의 결합밀도함수를 π_i에 대해서 적분하여 F_i의 주변분포를 구하면 다음과 같은 음이항분포(Negative-Binomial Distribution)가 된다.

$$P(F_i = x_i) = \frac{\Gamma(x_i+\alpha)}{x_i!\Gamma(\alpha)}\frac{(N\beta)^{x_i}}{(N\beta+1)^{x_i+\alpha}}, \quad x_i = 0,1,\ldots,N. \tag{2.8}$$

Multinomial-Dirichlet 모형에서와 같이 셀교환가능성을 가정하고, 식 (2.8)을 이용하여 모집단 유일성 개수의 기대값을 구하면 다음과 같다.

$$U_P = E(S_1) = KP(F_i = 1) = \frac{K\alpha N\beta}{(N\beta+1)^{1+\alpha}} = N(N\beta+1)^{-(\alpha+1)}, \tag{2.9}$$

여기서, $\alpha = 1/K\beta$이다.

표본 적률과 모집단 적률이 같다고 놓으면 $\bar{f} = n\hat{\alpha}\hat{\beta} = n/K$, $s_f^2 = n\hat{\alpha}\hat{\beta}(n\hat{\beta}+1)$이 되고, 위의 제약조건 $\alpha = 1/K\beta$이 있기 때문에 β만 추정한다.

$$\hat{\beta} = \frac{1}{n}\left(\frac{Ks_f^2}{n} - 1\right), \qquad \hat{\alpha} = \frac{n}{K}\left(\frac{n}{Ks_f^2 - n}\right). \tag{2.10}$$

이렇게 구한 $\hat{\alpha}$과 $\hat{\beta}$를 U_P에 대입하여 모집단 유일성의 개수의 기대값 U_P를 추정할 수 있으며, 모집단 유일성 개수의 기대값에 대한 추정량 $\widehat{U_P}$는 다음과 같다.

$$\widehat{U_P} = N\left[\frac{N}{n}\left(\frac{Ks_f^2}{n} - 1\right) + 1\right]^{-\left[\frac{n}{K}\left(\frac{n}{Ks_f^2-n}\right)+1\right]}. \tag{2.11}$$

식 (2.10)을 $1/\hat{\alpha}$에 대해서 정리하면, $1/\hat{\alpha} = K/n(K/ns_f^2 - 1)$이 된다. 이 결과는 Multinomial-Dirichlet 모형에서 $\hat{\alpha}$를 추정할 때 $(n-1)/(K\hat{\alpha}+1)$ 대신에 $n/(K\hat{\alpha})$로 놓으면, Poisson-Gamma 모형의 $1/\hat{\alpha} = K/n(K/ns_f^2 - 1)$과 같아지므로 두 모형은 동일한 결과를 보이게 된다.

따라서 Takemura의 Poisson-Gamma 모형의 적률 추정량 $\hat{\alpha}$는 Multinomial- Dirichlet 모형의 변형된 적률 추정량임을 알 수 있으며, 이 모형의 적률 추정량 $1/\hat{\alpha} = K/n(K/ns_f^2 - 1)$을 이용하여 Multinomial-Dirichlet 모형의 모집단 유일성 개수의 기대값 U_P를 추정할 수 있다. 즉, 모집단 유일성 개수의 기대값에 대한 추정량 $\widehat{U_P}$는 다음과 같다.

$$\hat{U}_P = \frac{N}{N-1}\frac{n^2}{\left(Ks_f^2 - n\right)}\frac{B\left(\dfrac{n^2}{\left(Ks_f^2 - n\right)}, N\right)}{B\left(\dfrac{(K-1)}{K}\dfrac{n^2}{\left(Ks_f^2 - n\right)}, N-1\right)}. \tag{2.12}$$

이 경우 Poisson-Gamma 모형의 적률 추정량 $\hat{\alpha}$을 이용하여 Multinomial-Dirichlet 모형의 모집단 유일성 개수의 기대값 U_P를 추정한 추정량 $\widehat{U_P}$을 Modified Multinomial-Dirichlet 모형의 $\widehat{U_P}$이라 한다.

2.2.3. Bethlehem의 Poisson-Gamma 모형 (Bethlehem,Keller 와 Pannekoek,1990)

Poisson-Gamma 모형에서 Takemura (1997)는 Π_i가 Gamma(α, β)를 따른다고 가정하고 표본에서의 셀의 개수 K를 공셀(empty cells)을 포함한 모든 셀의 개수로 정의하였지만, Bethlehem 등 (1990)은 Π_i가 Gamma$(N\alpha, \beta/N)$을 따른다고 가정하였고, 표본에서의 셀의 개수 K를 0이 아닌 셀의 개수로 정의하였다.

$$F_i \sim \text{Poisson}(N\pi_i)|\pi_i = \Pi_i, \quad \Pi_i \sim \text{Gamma}\left(N\alpha, \frac{\beta}{N}\right). \tag{2.13}$$

위의 가정 하에서 F_i와 Π_i의 결합확률밀도함수는 다음과 같다.

$$P(F_i = x_i, \Pi_i = \pi_i) = \frac{N^{x_i+N\alpha}\pi_i^{x_i+N\alpha-1}e^{-\left(\frac{N\beta+N}{\beta}\right)\pi_i}}{x_i!\Gamma(N\alpha)\beta^{N\alpha}}, \quad x_i = 0, 1, \ldots, N. \tag{2.14}$$

위의 결합밀도함수를 π_i에 대해서 적분하여 F_i의 주변분포를 구하면 다음과 같은 음이항분포가 된다.

$$P(F_i = x_i) = \frac{\Gamma(x_i + N\alpha)}{\Gamma(x_i+1)\Gamma(N\alpha)}\frac{(\beta)^{x_i}}{(\beta+1)^{x_i+N\alpha}}, \quad x_i = 0, 1, \ldots, N. \tag{2.15}$$

식 (2.15)를 이용하여 모집단 유일성 개수의 기대값을 구하면 다음과 같다.

$$U_P = E(S_1) = KP(F_i = 1) = KN\alpha\frac{\beta}{(\beta+1)^{1+N\alpha}} = N(\beta+1)^{-(N\alpha+1)}, \tag{2.16}$$

이 경우 표본적률과 모적률이 같다고 놓으면 $\bar{f} = n\hat{\alpha}\hat{\beta} = n/K$, $s_f^2 = n\hat{\alpha}\hat{\beta}(\hat{\beta}+1)$이 되고, 제약조건 $\alpha = 1/K\beta$을 이용하여 β와 α를 추정하면 다음과 같다.

$$\hat{\beta} = \frac{K}{n}s_f^2 - 1, \qquad \hat{\alpha} = \frac{n}{K}\left(\frac{1}{Ks_f^2 - n}\right). \tag{2.17}$$

이렇게 구한 $\hat{\alpha}$과 $\hat{\beta}$를 U_P에 대입하여 모집단 유일성 개수의 기대값 U_P를 추정할 수 있으며, 모집단 유일성 개수의 기대값에 대한 추정량 $\widehat{U_P}$는 식 (2.18)과 같다.

$$\widehat{U_P} = N\left(\frac{K}{n}s_f^2\right)^{-\left[N\frac{n}{K}\left(\frac{1}{Ks_f^2-n}\right)+1\right]}. \tag{2.18}$$

3. 임계모집단 크기 결정

3.1. 임계모집단 크기

임계모집단 크기란 지역분류(행정구역코드)에 있어서 가장 소규모로 구별되는 지역의 인구 규모 또는 사업체 규모가 최소한 임계모집단 크기 $N_{c(r)}$보다는 커야 한다는 임계 크기를 의미한다. 마이크로데이터 이용자들은 전체 모집단 자료 대신에 어떤 특정한 지역과 관련된 자료에 더 많은 관심을 가지고 있다. 그러므로 지역분류(행정구역코드)와 관련하여 각각의 지역에 대한 안전한 마이크로데이터 파일을 제공할 기준이 필요하다.

수집 자료의 마이크로데이터 파일에서 식별되는 자료들만이 모집단에서 유일하다고 가정할 때, 식별 가능한 자료들의 기대값을 U_{PS}이라고 하자. 이 U_{PS} 값은 특정한 마이크로데이터 파일의 노출위험을 반영하며, U_{PS}는 $\widehat{U_{PS}} = n\widehat{U_P}/N$으로 추정할 수 있고, $\widehat{U_P}$은 U_P의 추정량이다.

임계모집단 크기 결정에 있어서 첫 번째 기준은 절대기준 C_a로 크기를 고려한 $\widehat{U_{PS}} < C_a$이고, 두 번째 기준은 상대적으로 덜 엄격한 기준으로 식별 가능한 자료들의 비율 ($U_{PS}/n = U_P/N$)이 어떤 임계값 C_r보다 작아야 한다는 기준이다. 이 기준을 적용할 때에는 $\widehat{U_P}/N < C_r$이 된다.

상대기준 C_r에 대응되는 임계모집단 크기 $N_{c(r)}$를 상대기준이 충족되는 모집단 크기로 정의하면

$$\frac{U_P(N_{c(r)})}{N_{c(r)}} = C_r \tag{3.1}$$

이 된다. 여기서 $N_{c(r)}$은 상대 기준(relative criterion)을 만족하는 모집단의 크기로 정의된다. $U_P(N_{c(r)})$은 U_P를 $N_{c(r)}$의 함수로 간주한다.

Multinomial Dirichlet 모형을 이용한 임계모집단 크기 결정에서 모집단 유일성 개수의 기대값이 $U_P = E(S_1) = N/(N-1)K\alpha B(K\alpha, N)/B((K-1)\alpha, N-1)$이므로 상대기준 C_r은 식 (3.2)와 같다.

$$\frac{U_P(N_{c(r)})}{N_{c(r)}} = C_r = \frac{1}{N_{c(r)}-1} K\alpha \frac{B(K\alpha, N_{c(r)})}{B((K-1)\alpha, N_{c(r)}-1)}. \tag{3.2}$$

Takemura의 Poisson-Gamma 모형을 이용한 임계모집단 크기 결정에서 모집단 유일성 개수의 기대값이 $U_P = E(S_1) = KP(F_i = 1) = N(N\beta+1)^{-(\alpha+1)}$이므로, 임계모집단 크기 $N_{c(r)}$은 다음과 같다.

$$N_{c(r)} = \frac{C_r^{-\left(\frac{K\beta}{K\beta+1}\right)} - 1}{\beta}. \tag{3.3}$$

즉, 임계모집단 크기 $N_{c(r)}$은 β의 추정값인 $\hat{\beta} = 1/n(Ks_f^2/n - 1)$을 이용하여 추정할 수 있다. Modified Multinomial-Dirichlet 모형에서는 모집단 유일성 개수의 기대값이 $U_P = E(S_1) = N/(N-1)K\alpha B(K\alpha, N)/B((K-1)\alpha, N-1)$이므로 임계모집단 크기 결정에서 상대기준 C_r은 Multinomial-Dirichlet 모형을 이용한 C_r과 같다.

Bethlehem의 Poisson-Gamma 모형을 이용한 임계모집단 크기 결정에서 모집단 유일성 개수의 기대값이 $U_P = E(S_1) = N(\beta+1)^{-(N\alpha+1)}$이므로, 임계모집단 크기 $N_{c(r)}$는 식 (3.4)와 같다.

$$N_{c(r)} = -K\beta\left[\frac{\ln(C_r)}{\ln(1+\beta)} + 1\right]. \tag{3.4}$$

즉, 임계모집단 크기 $N_{c(r)}$는 추정량 $\hat{\alpha}$ 또는 $\hat{\beta}$, 상대기준 C_r과 셀의 수 K를 이용하여 구할 수 있다.

표 1: Multinomial-Dirichlet 모형의 임계모집단 크기

주요변수	셀의 개수	유일성개수	추정된 $\hat{\alpha}$	임계모집단 크기
I	18	0	0.444264	53
$I \times S$	101	9	0.476428	138
$I \times S \times E$	181	33	0.235182	153
$I \times S \times E \times Y$	373	93	0.128447	177

표 2: Takemura의 Poisson-Gamma 모형의 임계모집단 크기

주요변수	셀의 개수	유일성개수	추정된 $\hat{\beta}$	임계모집단 크기
I	185	0	0.111096	891
$I \times S$	101	9	0.020348	5,074
$I \times S \times E$	181	33	0.022941	11,362
$I \times S \times E \times Y$	373	93	0.020435	21,914

표 3: Modified Multinomial-Dirichlet 모형의 임계모집단 크기

주요변수	셀의 개수	유일성개수	추정된 $\hat{\alpha}$	임계모집단 크기
I	18	0	0.500069	54
$I \times S$	101	9	0.486573	138
$I \times S \times E$	181	33	0.240828	154
$I \times S \times E \times Y$	373	93	0.131194	178

표 4: Bethlehem의 Poisson-Gamma 모형의 임계모집단 크기

주요변수	셀의 개수	유일성개수	추정된 $\hat{\beta}$	임계모집단 크기
I	18	0	222.1914	1,109
$I \times S$	101	9	40.69686	3,501
$I \times S \times E$	181	33	45.88229	6,605
$I \times S \times E \times Y$	373	93	40.87032	12,953

3.2. 임계모집단 크기 결정 사례

임계모집단 크기 결정을 위해 사용한 자료는 2007년 기준 사업체 모집단 데이터베이스 자료 중 행정구역코드 34050에 해당하는 서산시에 존재하는 9,691개의 사업체 중에서 20% 정도인 2,000개의 사업체이다.

주요 변수는 산업분류(I), 년 간 총매출액(S), 월평균 종사자수(E), 창설년도(Y) 순서로 4개이다. 산업분류의 범주는 한국표준산업분류(제 9차 개정)에 의하여 산업분류코드 5자리 숫자 중에서 앞자리 2자리를 이용한 대분류 기준으로 21개의 범주로 나누었다. 년 간 총매출액과 월평균 종사자수는 각각 6개와 4개의 범주로 구성하였다. 창설 년도는 1910년 이전을 하나의 범주로 1910년부터 20년 주기로 2008년 까지 6개의 범주로 구성하였다.

4개의 주요변수에 대한 모든 가능한 범주들의 조합을 고려하지만, 구조적 "0"셀은 분석에서 제외하며 표본에서의 공셀(empty cell)은 제외하지 않고 포함하였다. 즉, 주요변수에 대한 모든 범주들의 조합들이 가능한 것은 아니기 때문에 조합들의 수가 관련 변수들의 범주의 곱보다는 작다. 임계모집단 크기 결정을 위한 상대기준 C_r은 0.1%를 사용하였다. 즉, 어떤 하위모집단에서 식별할 수 있는 레코드들의 수가 0.1%보다 작아야 한다. 이와 같은 조건하에서 4개의 모형에 대한 임계모집단 크기를 구하였다. 표 1에서 부터 표 4까지 주요변수의 셀의 수에 따라 유일성을 가지는 셀의 수가 유일성의 개수이다.

상대기준 C_r으로 0.1%를 고려한 것이 다소 작은 것처럼 보이지만 2007년 기준 전국사업체 조사의 전국 사업체 수가 6,221,268개 인 것을 고려한다면 6,221개의 사업체가 유일하며, 노출의 위험이 있음을 의미한다.

위의 표 1부터 표 4까지 산업분류(I)의 셀의 개수가 18인 이유는 2007년 기준 사업체 모집단 데이터베이스 자료에서 산업분류(I) 대분류 기준 21개 중에서 구조적 "0"셀인 A(농업, 임업 및 어업), B(광업), U(국제 및 외국기관)를 분석에서 제외했기 때문이다. 전국사업체조사 자료 중에서 가장 중요하게 생각되는 변수는 각 사업체들의 산업분류(I)와 년 간 총매출액(S)이다. 산업분류(I) 하나 만을 고려할 경우 임계모집단 크기는 53, 891, 54, 1,109개이므로 지역분류(행정구역코드)에 있어서 4개의 모형 모두를 고려할 경우 1,109개 이하의 사업체를 가진 지역은 자료를 제공할 수 없다. 또한, 산업분류(I)와 년 간 총매출액(S)의 임계모집단 크기는 138, 5,074, 138, 3,501개인 것을 감안하면 서산시에 존재하는 사업체 수가 9,691개이므로 5,074개 보다 크므로, 두 변수에 대한 자료의 제공이 가능하다.

4개의 주요 변수(산업분류(I), 년 간 총매출액(S), 월평균 종사자수(E), 창설년도(Y)) 모두를 고려할 경우 임계모집단 크기가 177, 21,914, 178, 12,953개이므로, 사업체 수가 각각의 모형을 고려할 때 임계모집단 크기 이하인 지역은 지역분류(행정구역코드)를 감안하여 4개의 변수 모두를 제공할 수는 없다.

4. 결론

본 연구에서는 마이크로데이터 파일을 통계이용자들에게 제공 할 때 발생할 수 있는 노출과 식별에 대해서 살펴보았고, 식별의 측도로 사용되는 유일성의 개념을 살펴보았다. 모집단 유일성 개수는 마이크로데이터 파일의 제공 여부를 결정하는데 중요한 지표이며, 이를 추정하는 방법인 초모집단 모형으로는 Multinomial-Dirichlet 모형, Takemura의 Poisson-Gamma 모형, Modified Multinomial-Dirichlet 모형, Bethlehem의 Poisson-Gamma 모형 등이 있다. 임계모집단 크기는 상대기준 C_r, 모집단의 셀의 수 K, 추정량 $\hat{\alpha}$ 또는 $\hat{\beta}$에 의해 결정되며, 상대기준 C_r을 고정시켰을 때 모집단으로부터 20%의 표본을 추출할 경우 MSE가 안정적인 것을 감안하면 분산과 연관된 추정량 $\hat{\alpha}$과 $\hat{\beta}$의 변화는 미비하며, 결과적으로 모집단의 셀의 수 K에 의해 임계모집단 크기는 좌우됨을 알 수 있었다.

임계모집단 크기 결정 사례에서는 전국사업체조사 자료를 이용하였으며, 사업체조사에서 가장 중요하고 사업체 입장에서 노출을 꺼리는 산업분류(I), 년 간 총매출액(S), 월평균 종사자수(E)를 주요 변수로 하였으며, 식별이 용이할 수 있는 창설년도(Y)도 포함하였다. 각 주요변수에 대한 범주의 구성은 기준이 마련되어 있는 산업분류(I)는 대분류 기준에서 A, B, U를 제외한 18개로, 기준이 마련되어 있지 않은 년 간 총매출액(S)은 6개, 월평균 종사자수는 4개, 창설년도(Y)는 6개의 범주로 구성한 결과 4개의 주요 변수에 대한 임계모집단 크기는 Multinomial-Dirichlet 모형을 이용할 경우 177개, Takemura의 Poisson-Gamma 모형을 이용할 경우 21,914개, Modified Multinomial-Dirichlet 모형을 이용할 경우 178개, Bethlehem의 Poisson-Gamma 모형을 이용할 경우 12,953개였다. 결과적으로 어떤 모형을 이용하는가에 따라서 사업체 수가 적게는 177개에서 많게는 사업체 수가 21,914개 이하인 지역은 지역분류(행정구역코드)를 감안하여 4개의 주요 변수 모두를 제공할 수 없다는 것을 알 수 있다. 네덜란드는 인구주택총조사(Population and Housing Census)의 경우 주요 변수를 4개 변수(가구구성: 24개 범주, 나이: 14개 범주, 혼인상태: 2개 범주, 성별: 2개 범주)로 정하고 4개 변수 모두를 제공할 경우의 임계모집단 크기를 46,228명으로 하고 있다. 네덜란드에서 주로 사용하는 지역분류는 COROP(Coördinatie Commissic Regionaal Onderzoeks Programma)이며, 이 지역분류 COROP에서의 주민 수는 54,747명에서 1,057,095명의 범위 이다. 이 범위는 4개의 주요 변수에 대한 임계모집단 크기 46,228명의 조건을 만족하므로 마이크로데이터 파일을 제공할 수 있다 (Bethlehem 등, 1990).

미국은 마이크로데이터 파일에서 최소한 지역분류 규모로 주민수 100,000명을 선택했다 (Cox 등, 1986). 한국 통계청은 인구주택총조사의 경우 주요변수를 8개 변수(가구구분, 나이, 혼인상태, 성별, 가구주와의 관계, 점유형태, 주인가구, 거처의 종류)로 정하고 있지만, 임계모집단 크기와 최소 지역분류 규모에 대한 기준은 마련되지 않았다. 따라서 응답 노출 및 식별을 제한하기 위한 임계모집단 크기 기준을 설정할 수 있으므로 본 연구는 다양한 마이크로데이터 파일에 대해 상대기준 C_r값에 대응되는 최소 규모를 결정하는데 활용할 수 있다.

일반적으로 주요변수의 선정은 연구자에 따라 다소 주관적이며 일정하지 않을 수도 있다. 또한, 주요변수들을 범주화 할 때에도 자료의 제공 범위 등에 따라 다양하게 할 수도 있다. 따라서 선정한 4개의 변수와 그에 따른 범주화 방법이 최적이라고 할 수는 없으며, 자료의 특성이나 제공의 범위에 따라 적정한 방법을 적용하면 될 것이라 판단된다.

참고 문헌

정동명, 정남수, 한승훈 (2008). 마이크로데이터 활용연구 및 통계를 이용한 현황분석, <연구보고서 2008-03>, 제 2장 가계조사 마이크로데이터의 비밀보호.

통계청 (2008). 2008 한국표준산업분류(제 9차 개정)

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

Bruce G. S. H. and Peter S. F. (2000). Applied probability models in marketing research, *Supplementary materials for the A/R/T furum tutorial*.

Hoshino, N. and Takemura, A. (1998). On the relation between logarithmic series model and other superpopulation models useful of microdata disclosure risk assessment, *Journal of the Japan Statistical Society*, **28**, 125–134.

Pereira, C. A. B. and Stern, J. M. (2008). Special characterizations of standard discrete models, *REVSTAT-Statistical Journal*, **6**, 199–230.

Takemura, A. (1997). Some superpopulation models for estimating the number of population uniques. Discussion Paper 97-F-29, Faculty of Economics, University of Tokyo.

2010년 5월 접수; 2010년 9월 채택

The Decision of Critical Population Size for Releasing Micro Data Files

Pyong Namkung[1,a], Joung-Hyun So[b]

[a]Department of Statistics, Sungkyunkwan University
[b]Seosan branch, Chungcheong Regional Statistics Office

Abstract

This study reviews the concept of disclosure, disclosure risks, and uniqueness. The number of uniqueness in the population is of great importance in evaluating the disclosure risk of micro data files. We approach this problem by considering some basic superpopulation models including the Multinomial-Dirichlet model, the Poisson-Gamma model of Bethlehem *et al.* (1990) and Takemura (1997), and the Modified Multinomial-Dirichlet model. We decided the critical population size of each superpopulation model for four different superpopulation models.

Keywords: Superpopulation models, Multinomial-Dirichlet model, Poisson-Gamma model, critical population size.

[1] Corresponding author: Professor, Department of Statistics, Sungkyunkwan University, 53, Myeongnyun-Dong 3-Ga, Jongno-Gu, Seoul 110-745, Republic of Korea. E-mail: namkung@skku.edu