응용통계연구 (2011)
24(1), 127–136
DOI: 10.5351/KJAS.2011.24.1.127

승법잡음모형을 이용한 통계적 노출조절기법의 적용

김영원[1] · 김태연[2] · 김계남[3]

[1]숙명여자대학교 통계학과, [2]숙명여자대학교 통계학과, [3]숙명여자대학교 통계학과

(2010년 12월 접수, 2011년 1월 채택)

요 약

본 연구에서는 통계기관에서 마이크로자료를 제공할 때, 연속형 변수를 마스킹하는 기법으로 잘 알려진 승법잡음모형을 적용하는 경우 원자료의 평균과 분산을 유지할 수 있는 변수 변환 방안을 제시하고, 제시된 방법의 적절성과 다양한 잡음생성 분포에 따른 마스킹자료의 유용성을 검토하였다. 아울러 여러 변수들을 대상으로 승법잡음모형을 적용하는 경우 변수들 간의 상관관계를 유지하기 위해서는 잡음생성과정에서 어떤 측면이 고려되어야 하는지 살펴보았다. 본 연구에서는 제시된 변수 변환 방법의 적절성과 자료의 유용성 등을 평가하기 위해 우리나라 가계조사자료를 이용한 모의실험을 수행하였다.

주요용어: 노출위험, 승법잡음모형, 자료유용성, 통계적 노출조절기법.

1. 서론

최근 개인용 컴퓨터의 성능이 향상되어 연구자들이 손쉽게 대용량 자료를 분석할 수 있게 되고, 통계작성기관에서 일괄적으로 제공하는 단순 집계 형식의 통계보다 연구목적에 따른 맞춤형 통계분석에 대한 수요가 대폭 증가함에 따라 마이크로자료(micro-data)를 직접 분석하여 활용하는 이용자들이 늘어나고 있다. 이런 수요 때문에 통계기관에서는 수집된 자료를 개별 레코드 단위의 마이크로자료 형식으로 제공하는 것을 피할 수 없게 되었다. 이와 같이 통계기관이 개인이나 사업체 단위의 마이크로자료를 제공하는 경우 응답자 비밀이 노출될 위험이 있기 때문에 응답자의 비밀을 보호하는 동시에 원래 자료가 갖고 있는 특성이 그대로 유지될 수 있는 마스킹(masking) 자료를 생성할 수 있는 기법이 필요하다. 최근 마이크로자료 제공과 관련해 통계적 노출조절기법(disclosure control techniques)에 대한 연구가 국내외에서 활발히 이루어지고 있다.

연속형 변수에 대한 마이크로자료를 제공하는 경우 사용되는 통계적 노출조절기법은 잡음추가기법, 반올림방법, 자료교환(data swapping)방법, 마이크로애그리게이션(micro-aggregation), 재표집(re-sampling)방법 등이 있다. 이런 기법들에 대한 내용은 Dalenius 등 (1982), Torra (2004), 정동명 등 (2007), 정동명과 정미옥 (2008), 김규성 (2009) 등을 참고하기 바란다.

이 중 잡음추가기법은 연속형 변수에 적용 가능한 기법으로 우리나라에서도 최근 관심이 높아지고 있다. 잡음추가기법은 크게 가법잡음(additive noise)모형과 승법잡음(multiplicative noise)모형으로 구분된다. 원자료를 x_i, 잡음을 e_i라고 하고 잡음이 반영된 마스킹자료를 y_i라고 하면, 가법모형의 경우 $y_i = x_i + e_i$ 형식을 갖는다. 많은 조사에서 연속형 자료의 경우 '0'의 값을 갖는 경우가 많은데 가법모

본 연구는 숙명여자대학교 2009학년도 교내연구비 지원에 의해 수행되었음.
[1]교신저자: (140-742) 서울특별시 용산구 효창원길 52, 숙명여자대학교 통계학과, 교수.
E-mail: ywkim@sookmyung.ac.kr

형의 경우 마스킹(masking)된 자료가 음수가 될 수 있다는 점에 유의할 필요가 있다. 가법모형에 대한 자세한 내용은 Kim (1986)과 Fuller (1993) 등을 참고하기 바란다.

승법잡음모형의 경우 원자료를 $y_i = x_i \cdot e_i$ 형식으로 변환하는 것으로 이론적으로 입증이 된 것은 아니지만 가법잡음모형에 비해 비밀보호 측면에서 효과적이라고 알려져 있으며 (Kim과 Winkler, 2001), 가법모형과는 달리 음수로 변환되는 경우도 발생하지 않는다는 장점이 있다. 승법잡음모형에서는 잡음을 발생시키기 위해 '1'을 중심으로 한 다양한 분포가 사용될 수 있는데, Kim과 Winkler (2001), Kim (2007), 정동명 등 (2009)은 잡음 발생을 위해 절단된 정규분포, 삼각분포, 절단된 삼각분포, 사다리꼴분포, 이중삼각분포의 활용 방안을 제시하고 있다.

한편 잡음추가기법을 사용함에 있어서 일반적으로 두 가지 조건이 요구되는데, 첫째 조건은 비밀보장이 되어야 한다는 것이고, 둘째는 변환된 자료와 원자료에서 얻어지는 통계적인 분석결과가 가능한 일치해야 한다는 점이다. 이런 측면을 고려해 Kim (1986) 등은 가법잡음모형에서 평균과 분산 등이 유지되는 변수변환방법을 제시하고 있다. 한편 Kim과 Winkler (2001)는 승법잡음모형으로 변환된 자료에서 얻어진 결과를 토대로 추가적인 계산과정을 통해 이용자가 원자료의 평균과 분산을 구하는 방안을 제시하였고, 정동명 등 (2009)은 이 방법을 이용해 우리나라 가계조사자료에 대한 사례분석 결과를 제시하고 있다. 하지만, 이들이 제시한 방법은 변환된 자료로 원자료의 평균이나 분산을 직접 추정할 수 없고, 이용자들이 잡음발생 분포의 모수값을 제공 받아 추가적으로 계산해야 하는 불편함이 있을 뿐만 아니라 변환 자료를 갖고는 평균과 분산 이외의 다양한 통계를 이용자 스스로 계산할 수 없다는 매우 큰 결함을 갖고 있다.

따라서 본 연구에서는 승법잡음모형을 적용하는 경우 변환된 자료를 이용한 통상적인 통계분석 과정을 통해 원자료의 평균과 분산 등을 포함한 주요 통계를 얻을 수 있는 승법잡음모형 적용 방안을 제시한다. 또한 이런 기법을 적용함에 있어서 다양한 잡음 발생 분포에 따른 승법잡음모형의 효율성을 비교하기 위해 신뢰구간 및 성향점수 개념을 토대로 한 자료의 유용성(data utility) 척도를 활용해 잡음생성 분포의 차이에 따른 효과를 살펴본다. 또한 2개 이상 변수를 대상으로 승법잡음모형을 동시에 적용하는 경우 원자료에서의 변수간의 상관관계가 유지되기 위해서는 어떤 측면이 잡음생성 과정에서 고려될 필요가 있는지 추가적으로 검토한다. 본 연구에서는 제시된 방안이 실제로 작동할 수 있는 유효한 방법임을 확인하기 위해 통계청의 가계조사자료를 이용한 모의실험을 통해 제시된 방법의 효과를 확인하기로 한다.

본 연구에서는 우선 2절에서 기존 승법잡음모형이 갖고 있는 한계를 극복하기 위한 새로운 승법잡음모형 자료변환 방법을 제시한다. 3절에서는 마스킹자료의 유용성 평가를 위한 척도를 소개하고, 4절에서는 우리나라 가계조사자료를 이용한 모의실험을 통해 제시된 승법잡음모형을 토대로 한 추가적인 변환과정의 적절성을 평가하고, 자료의 유용성과 함께 상관관계 유지를 위한 승법잡음모형의 적용방법에 대해 살펴본다.

2. 승법잡음모형에 의한 자료 변환

2.1. 기존 승법잡음모형의 한계

일반적인 승법잡음모형을 통해 생성되는 자료는 다음과 같이 표현될 수 있다.

$$y_i = x_i \cdot e_i, \quad i = 1, 2, \ldots, n$$

여기서 x_i는 통계조사에서 i-번째 단위(사람, 가구 또는 사업체 등)에 대한 원래 관측 자료이고, y_i는 원자료에 잡음이 곱해진 마스킹된 자료, e_i는 잡음을 나타낸다. 여기서 잡음 e_i는 원자료 x_i와 무관하게

발생되기 때문에 x_i와 e_i는 서로 독립적이다. 따라서 다음 관계식이 성립한다.

$$E(Y) = E(X) * E(e), \tag{2.1}$$

$$\mathrm{Var}(Y) = \mathrm{Var}(X)\mathrm{Var}(e) + [E(e)]^2\mathrm{Var}(X) + [E(X)]^2\mathrm{Var}(e) \tag{2.2}$$

여기서 잡음 e_i를 발생시키는 분포의 평균이 1, 즉 $E(e) = 1$이면, $E(Y) = E(X)$이지만, 원자료 x_i의 분산 $\mathrm{Var}(X)$와 승법잡음이 반영된 y_i의 분산 $\mathrm{Var}(Y)$는 일치하지 않는다. 따라서 원자료 대신 단순히 승법잡음이 반영된 y_i를 이용자들에게 제공하는 경우 이용자는 통계작성기관이 제공한 마이크로자료의 분석을 통해 원자료의 분산을 구할 수 없다는 심각한 문제를 갖게 된다. 이런 문제를 해결하기 위해 Kim (2007)과 정동명 등 (2009)에서는 잡음발생분포의 평균($E(e)$)과 분산($\mathrm{Var}(e)$)을 이용자들에게 제공하고, 이용자들은 y_i의 평균과 분산을 구한 후 식 (2.2)에서 유도된 식 (2.3)을 이용해 원자료의 분산을 구하는 방안을 제시하고 있다.

$$\mathrm{Var}(X) = \frac{\mathrm{Var}(Y) - [E(X)]^2\mathrm{Var}(e_i)}{\mathrm{Var}(e_i) + [E(e_i)]^2} \tag{2.3}$$

하지만 이런 마스킹 마이크로자료를 제공하면, 이용자는 식 (2.3)을 통해 원자료의 분산을 추정할 수는 있지만, 결과적으로 평균과 분산을 제외한 다른 자료의 특성을 나타내는 다양한 통계를 구할 수 없다는 중요한 결함을 갖게 된다. 따라서 마이크로자료 제공을 위해 승법잡음모형을 적용하는 경우 기존 연구에서 제시한 대로 단순히 원자료에 잡음이 곱해진 형식의 자료인 y_i를 제공하는 것은 현실적으로 좋은 방안이 될 수 없다. 이런 문제점을 해결하기 위해서는 이용자들이 제공된 마이크로자료의 분석을 통해 손쉽게 원자료와 같은 평균과 분산을 얻을 수 있도록 승법잡음모형을 변환해서 마스킹 하는 새로운 방안을 구현할 필요가 있다.

2.2. 승법잡음모형 변수의 변환

앞에서 설명한 것처럼 잡음발생 과정에서 평균이 1인 분포를 사용하게 되면 원자료 x_i와 단순 승법잡음 자료 y_i의 평균은 일치하게 되지만, 분산은 차이가 발생한다. 따라서 단순 승법잡음 자료 $y_i = x_i \cdot e_i$를 생성한 후, 추가적으로 식 (2.4)와 같은 선형 변환을 통해 얻은 z_i를 이용자들에게 제공하는 새로운 마스킹 방법을 고려할 수 있을 것이다.

$$z_i = \frac{y_i}{\alpha} + E(X)\left(1 - \frac{1}{\alpha}\right) \tag{2.4}$$

여기서 $\alpha = \sqrt{\mathrm{Var}(Y)/\mathrm{Var}(X)}$로 마스킹을 수행하는 기관은 원자료를 이용해 쉽게 $E(X)$와 $\mathrm{Var}(X)$를 구할 수 있고, 마스킹 과정에서 사용한 잡음발생 분포의 평균과 분산도 알고 있기 때문에 식 (2.2)을 이용해 $\mathrm{Var}(Y)$를 계산할 수 있으며, 이를 이용하면 손쉽게 α를 구할 수 있다.

잡음 발생과정에서 $E(e) = 1$인 분포를 이용하면, $E(Y) = E(X)$이기 때문에 $E(Z) = E(X)$가 되고, 또한 $\mathrm{Var}(Z) = \mathrm{Var}(X)$가 성립한다. 따라서 식 (2.4)와 같은 방식으로 일반적인 승법잡음모형을 변환한 마스킹자료 z_i를 생성해 이용자들에게 제공하면, 마스킹자료와 원자료의 평균과 분산은 같아지기 때문에 식 (2.3)과 같은 추가적인 계산과정 없이도 이용자들이 원자료의 평균과 분산을 쉽게 구할 수 있다. 또한 마스킹된 자료 z_i를 이용한 일반적인 통계분석 과정을 통해 원자료 x_i와 유사한 통계분석 결과를 얻는 것이 가능해질 수 있다.

참고로 Kim (1986)은 가법잡음모형을 이용하는 경우, 가법잡음모형을 그대로 사용하게 되면 원자료 보다 마스킹자료의 분산이 커지게 된다는 점을 고려해, 마스킹자료와 원자료의 평균과 분산이 같아지도록

가법잡음 변수의 선형변환방법을 제안하였다. 하지만 승법잡음모형에 대해서는 이런 변환방법이 아직 개발되어 있지 않다는 점을 고려해, 본 연구에서는 식 (2.4)의 변수변환방법을 제시하고, 새로운 변환방법의 적절성을 검토해 보고자 한다.

2.3. 잡음생성을 위한 분포

마이크로자료 제공을 위해 본 연구에서 새로 제시한 식 (2.4)의 변환을 기초로 마스킹 자료를 생성하기 위해서는 잡음 e_i를 적절한 분포에서 생성해야 한다. 잡음을 추가하는 마스킹기법에 있어서 어떤 잡음을 생성해 사용하느냐에 따라 마스킹에 의한 비밀보호 효과와 함께 마스킹 자료와 원자료의 통계적 특성상의 유사성 정도에 영향을 주기 때문에 잡음생성을 위해 어떤 분포를 사용할 것이지를 결정하는 것이 중요하다.

승법잡음모형에서는 우선 원자료와 마스킹 자료의 평균이 같아지도록 하기 위해 평균이 1인 잡음분포를 사용하는 것이 바람직하다. 물론 e_i가 1이면 y_i가 원자료 x_i와 같게 되어 마스킹의 의미가 없어진다는 점이 고려될 필요가 있다(참고로, 식 (2.4)를 이용하면 이런 문제도 해결될 수 있음). 또한 e_i가 1보다 너무 크거나 작은 경우에는, 다시 말해 잡음분포의 변동이 너무 커지게 되면 원자료와 마스킹 자료 사이의 차이가 심해져 마스킹 자료의 유용성(data utility)이 떨어지게 된다.

이런 점을 고려해 본 연구에서는 Kim (2007), 정동명 등 (2009)에서 사용한 삼각분포, 절단된 삼각분포, 사다리꼴분포 및 이중삼각분포를 고려하기로 한다. 이들 분포는 모두 좌우대칭이면서 공통점을 갖고 있고, 절단된 삼각분포와 이중 삼각분포의 경우 삼각분포나 사다리꼴 분포와는 달리 e_i가 1이라는 값을 취할 수 없다는 특성이 있다.

한편 바람직한 노출조절기법은 응답자의 노출위험을 줄이는 동시에 마스킹자료가 원자료와 동일한 특성(모수 추정결과나 변수들 간의 상관관계 등)을 갖도록 해야 한다. 마스킹자료가 원자료와 특성이 같아진다는 것은 마스킹 과정에서 정보 손실이나 왜곡이 적다는 것을 의미하면 이런 경우 마스킹 자료의 유용성이 높다고 말한다. 따라서 승법잡음모형의 적절성을 평가함에 있어서도 노출위험을 얼마나 줄일 수 있는지와 함께 자료의 유용성이 함께 고려되어야 한다. 실제 노출조절기법의 자료의 유용성을 평가하기 위해서 다양한 척도들이 제시되고 있으며, 본 연구에서는 최근 개발된 자료 유용성 척도를 기준으로 가계조사자료를 이용한 모의실험을 통해 잡음분포에 따른 효과를 비교해 보기로 한다.

3. 마스킹 자료 유용성 척도

마스킹 자료의 유용성이 높아지려면 공개된 자료와 원자료에서 산출되는 분석결과는 가능한 동일해야 한다. 결국 마스킹 자료의 유용성을 평가하기 위해서는 마스킹 자료의 분석을 통해 얻어지는 결과와 원래 자료를 분석해서 얻어지는 결과의 차이를 파악해야 한다. 본 연구에서는 우리나라 가계조사자료를 대상으로 승법잡음모형을 적용하는 경우 잡음분포에 따른 효과를 자료 유용성 측면에서 평가하기 위해 Karr 등 (2006)이 제안한 신뢰구간중복 척도와 정규분포 가정에 크게 영향을 받지 않는 자료 유용성 평가 방법인 Woo 등 (2009)이 제안한 성향점수 척도를 사용했다.

3.1. 승법잡음모형 변수의 변환

Karr 등 (2006)이 제안한 신뢰구간중복(confidence interval overlap) 척도는 마스킹자료 D_{mask}와 원자료 D_{orig}에서 얻은 신뢰구간들 사이의 중복의 정도를 살펴보는 방법이다. 기본적으로 마스킹 자료와 원자료를 이용해 구한 신뢰구간의 중복의 정도가 클수록 자료 유용성이 높다고 볼 수 있다. 회귀분석에서 회귀계수 β_k에 대한 중복 확률을 고려함으로써 자료의 유용성을 파악하는 용도로 주로 사용되지만, 본

연구에서는 특정 모수(평균)에 대한 중복된 신뢰구간의 길이를 이용하는 방법을 사용하도록 한다. 모수 β_k에 대한 중복구간 $(L_{over,k}, U_{over,k})$는 두 자료에서 구한 신뢰구간이 중복되는 다음 구간을 의미하며,

$$\{b \ : \ b \geq L_{orig,k},\ b \geq L_{mask,k},\ b \leq U_{orig,k},\ b \leq U_{mask,k}\}\,.$$

모수 β_k에 대한 두 자료의 신뢰구간 중복 척도는 다음 같이 정의한다.

$$J_k = \frac{1}{2}\left(\frac{U_{over,k} - L_{over,k}}{U_{orig,k} - L_{orig,k}} + \frac{U_{over,k} - L_{over,k}}{U_{mask,k} - L_{mask,k}}\right).$$

따라서 p개의 모수에 대한 신뢰구간중복 기준 자료 유용성은 다음 척도로 설명될 수 있다.

$$J = \frac{1}{p}\sum_{i=1}^{p} J_k.$$

제시된 척도는 각 모수에 대한 신뢰구간을 독립적으로 간주한 것인데, 모수 추정량들의 결합분포를 기초로 한 동시 신뢰구간을 고려하게 되면 타원중복(ellipsoid overlap) 개념의 척도로 자료 유용성을 설명할 수 있다. 이와 관련된 구체적인 내용은 Karr 등 (2006)을 참고하기 바란다.

3.2. 성향점수 척도

성향점수(propensity score 척도는 원자료와 마스킹자료의 관계를 로지스틱 회귀모형을 이용해 비교하는 자료 유용성 척도이다 (Woo 등, 2009). 성향점수 계산을 위해 원자료는 $r = 1$로 마스킹자료는 $r = 0$으로 놓으면, 조건이 x로 주어졌을 때, 성향점수는 $e(x) = P(r = 1 \,|\, x)$와 같이 표현된다. 일반적으로 성향점수는 아래와 같은 로지스틱모형에 의해 추정될 수 있다.

$$e(x) = P(r = 1 \,|\, x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

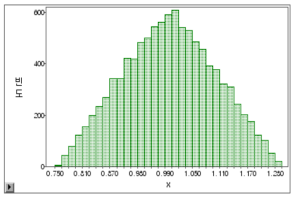
위 식에서 계산된 성향점수를 기준으로 자료 유용성 평가는 $\sum_i (\hat{e}(x_i) - c)^2$로 하게 된다. 여기서 c는 상수로 합쳐진 자료 중 원자료 레코드의 비율을 사용하게 된다. 만약 원자료와 마스킹자료가 완전히 동일한 경우, 성향 점수의 값은 0이 되고, 이 값이 증가할수록 자료 유용성이 떨어진다고 볼 수 있다.

4. 통계청 가계조사 적용사례

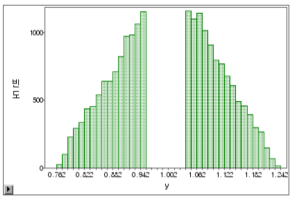
4.1. 모의실험 개요

본 연구가 정동명 등 (2009)의 연구가 갖고 있는 한계를 해결하기 위한 후속연구라는 점을 고려해 정동명 등의 연구와 마찬가지로 우리나라 가계조사자료에 대한 모의실험을 통해 식 (2.4)의 승법잡음모형 변수의 변환을 통한 노출조절기법의 적절성을 검토하는 동시에 자료 유용성 측면에서 잡음분포 가정에 따른 효율성을 비교하기로 한다.

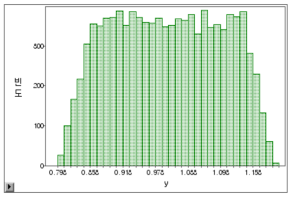
여기서는 통계청에서 제공하는 2008년 가계조사 원자료를 사용한다. 가계조사는 매월 우리나라 소득과 소비 수준 및 변화 추이를 분석하기 위한 조사로 매월 전국 8,000가구를 대상으로 조사가 수행된다. 가계조사의 조사 항목이나 방법 및 표본설계 등에 대한 내용은 통계청 (2008)을 참고하기 바란다. 본 연구에서는 가계조사 항목 중 가구 총소득과 총지출 그리고 조세 변수를 이용하였다. 2008년 연간 가계조사 원자료를 보면 총 84,908개의 레코드에 매우 많은 변수가 포함되어 있지만, 여기서는 이 중에서 서울특별시에서 관측된 10,285개 레코드에서 월별 소득, 지출과 조세 변수만을 모의실험 대상으로 사용하였다.



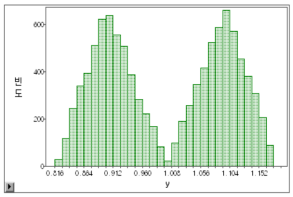
(a) 삼각분포 잡음



(b) 절단된 삼각분포 잡음



(c) 사다리꼴분포 잡음



(d) 이중삼각분포 잡음

그림 4.1. 잡음분포별로 생성된 난수의 분포형태

모의실험을 위해 우선 4개의 잡음분포에서 난수(잡음)를 생성하고, 식 (2.4)의 변환 과정을 통해 잡음분포별로 마스킹자료를 만들었다. 우선 본 연구에서 제시한 변환과정을 적용하게 되면 정동명 등 (2009)의 연구와는 달리 사후적으로 평균과 분산을 추정하는 과정 없이도 마스킹자료만을 이용해 원자료의 평균과 분산을 이용자가 직접 계산할 수 있다는 점을 확인하기로 한다. 또한 제시된 변환 방법에 따른 승법잡음모형을 적용하는 경우 잡음생성 분포에 따른 효율성 비교를 위해 자료 유용성 측면에서 잡음분포 가정에 따른 차이를 분석한다. 마지막으로 2개 변수에 대해 승법잡음모형 기법을 적용하는 경우 두 변수들 간의 상관관계를 유지하기 위해서는 승법잡음모형 적용 과정에서 어떤 점이 추가적으로 고려되어야 하는지 살펴보기로 한다.

4.2. 잡음생성 및 승법잡음 변수의 변환

승법잡음모형을 기반으로 원자료의 평균과 분산이 그대로 유지되는 마스킹자료를 만들기 위해 본 연구에서는 3절에 세시된 식 (2.4)를 이용하였으며, 잡음 생성을 위한 분포로 정동명 등 (2009)에서 사용된 삼각분포, 절단된 삼각분포, 사다리꼴분포 및 이중삼각분포를 사용하였다.

정동명 등 (2009)에서는 우선 원자료와 마스킹자료의 평균을 일치시키기 위해 $E(e) = 1$이 되도록 각각의 잡음분포에서 최빈값(m)이 1이 되게 했다. 또한 잡음이 너무 커지게 되면 원자료와 마스킹자료의 차이가 커져서 자료의 유용성이 떨어지기 때문에 잡음의 범위가 $0.01 \leq |e-1| \leq 0.4$를 만족하도록 각 분포의 최소값(a)과 최대값(d), 사다리꼴분포와 이중삼각분포의 변곡점(b와 c)을 설정해 잡음을 생성하였는데, 이들 분포의 확률밀도함수와 기댓값 및 분산은 정동명 등 (2009)을 참고하기 바란다.

하지만 이런 방식으로 잡음을 발생시키게 되면 4개 잡음분포에서 발생하는 난수의 범위는 갖지만 분포형태에 차이가 있기 때문에 분산이 달라진다. 따라서 본 연구에서는 좀더 공정한 잡음분포들 간의 비교를 위해 잡음의 범위를 설정하는 대신 각 분포에서 $E(e) = 1$이고, $\mathrm{Var}(e) = 0.01$이 되도록 각 분포별로 모수를 구해서 잡음을 생성하였다. 여기서 $\mathrm{Var}(e) = 0.01$를 가정한 이유는 정동명 등의 선행연구에서 생성된 난수와 표준편차가 비슷한 수준이 되도록 설정한 것이다.

모의실험을 위해 4개 잡음분포에서 $E(e_i) = 1$, 분산은 $V(e_i) = 0.01$이 되도록 각각의 분포에서 모수 a, b, c, d 값을 구한 후 잡음을 생성한다. 각 분포에서 2008년 가계조사에서 서울시 월별 레코드 수에 해당하는 10,285개의 난수를 생성하여 소득, 지출, 조세 변수에 승법잡음이 반영된 변수를 만든 후 식

표 **4.1.** 원자료와 마스킹자료의 변수별 평균과 표준편차 비교

통계량	변수	원자료	변환된 자료			
			삼각분포	절단된 삼각분포	사다리꼴분포	이중삼각분포
평균	연간소득	36290.05	36290.18	36290.14	36289.22	36290.17
	총지출	6861416.89	6861357.01	6860872.37	6861005.70	6861616.73
	조세	102804.14	102817.99	102794.08	102810.62	102812.82
표준편차	연간소득	25483.97	25481.41	25608.92	25483.75	25478.92
	총지출	6627447.83	6597169.20	6616683.60	6591669.78	6596263.33
	조세	337220.30	334516.97	334408.39	334286.54	334176.85

표 **4.2.** 마스킹 자료의 유용성 평가

척도	변수	변환된 자료			
		삼각분포	절단된 삼각분포	사다리꼴 분포	이중삼각 분포
신뢰구간중복	연간소득	0.9664003	0.9451994	0.9642888	0.9668063
	총지출	0.9712755	0.9575541	0.9697287	0.9718144
	조세	0.9770899	0.9447155	0.9773396	0.9767074
성향점수	연간소득	0.0000001826	0.0000004178	0.0000001862	0.0000001709
	총지출	0.0000001215	0.0000002663	0.0000001357	0.0000001169
	조세	0.0000001429	0.0000001076	0.0000000613	0.0000001343

(2.4)의 변환과정을 통해 최종 마스킹 자료를 생성하는 과정을 1,000번 반복수행한 결과를 토대로 변환방법의 적절성, 자료의 유용성 및 상관관계 관련 모의실험을 수행하였다.

제시된 방법에 따라 SAS 프로그램을 이용해 4개 잡음분포별로 생성된 난수의 분포형태는 그림 4.1의 (a)∼(d)와 같다. 제시된 그림은 각 분포별로 1,000번의 모의실험 중 1회의 시행에서 발생된 10,285개의 난수를 히스토그램 형식으로 정리한 것이다.

4.3. 자료 유용성 평가

우선 본 연구에서 제시한 승법잡음모형의 변환 과정을 통해 생산된 자료와 원자료의 평균과 분산이 일치하는지를 살펴보기 위해 각 잡음분포별로 1,000번의 모의실험을 통해 얻은 마스킹 자료의 평균과 표준편차를 구해 원자료와 비교해 보면 표 4.1과 같다. 본 연구에서 제시한 변환방법을 적용하면 정동명 등 (2009)에서와 같이 잡음분포의 평균과 분산을 별도로 제공해 사후적으로 원자료의 평균과 분산을 추정하는 과정을 거치지 않더라도 이용자들이 별 어려움 없이 직접 마스킹 자료만을 이용해 원자료의 평균과 분산을 파악할 수 있다는 것을 알 수 있다.

한편 간단하게 평균에 대한 추정문제만을 관심대상으로 하는 경우, 3절에서 설명한 신뢰구간중복 및 성향점수 척도를 기준으로 잡음분포에 따른 자료 유용성 측면에서의 차이를 비교해 보면 표 4.2와 같다. 잡음에 따른 마스킹자료들의 신뢰구간중복 비율은 대부분 0.94∼0.97로 매우 높고, 성향점수는 대부분 0에 가깝기 때문에 결국 두 가지 척도 어떤 경우에나 마스킹자료의 유용성이 높은 것으로 나타났다. 잡음분포에 따른 차이는 거의 없는 것으로 판단되며, 신뢰구간중복 척도를 기준으로 보면 아주 미세한 차이지만 삼각분포 또는 이중삼각분포의 경우 자료 유용성이 상대적으로 높은 것으로 보인다.

한편, 가계조사와 같이 대부분의 마이크로자료에는 여러 개의 변수가 포함되어 있다. 가계조사의 경우에도 소득, 지출, 조세 등과 같은 변수들이 있고, 이들 변수들을 대상으로 승법잡음모형을 개별적으로 적용해 마스킹자료를 만들면, 원자료에서 변수들의 상관관계가 유지되지 않게 되기 때문에 이용자들이 흔히 관심을 갖는 상관분석이나 회귀분석을 통해 원자료가 갖고 있는 변수들 간의 관계를 파악할 수 없다.

표 **4.3.** 잡음생성 방법에 따른 마스킹 변수들 간의 상관관계

잡음생성 방법	변수	원자료	변환된 자료			
			삼각분포	절단된 삼각분포	사다리꼴 분포	이중삼각 분포
[방법 1]	소득 & 지출	0.56490	0.82385	0.87631	0.82701	0.80905
	소득 & 조세	0.34273	0.48851	0.49185	0.41477	0.41496
[방법 2]	소득 & 지출	0.56490	0.00100	0.00835	0.00033	−0.08348
	소득 & 조세	0.34273	0.00082	0.03450	0.02771	−0.01638
[방법 3]	소득 & 지출	0.56490	0.39437	0.45406	0.44042	0.35157
	소득 & 조세	0.34273	0.29060	0.27403	0.27636	0.20184
[방법 4]	소득 & 지출	0.56490	0.57713	0.62605	0.62282	0.55389
	소득 & 조세	0.34273	0.36313	0.35282	0.33658	0.30964

따라서 두 개 이상의 변수들을 대상으로 승법잡음 변수 변환 방법을 적용하는 경우 두 변수들 간에 상관관계가 유지될 수 있도록 하기 위해서는 어떤 측면이 고려되어야 하는지 검토해 보기로 한다. 여기서는 2개 이상의 변수를 대상으로 다음 4가지 승법잡음 적용 방법을 비교해 본다.

[방법 1] 하나의 잡음을 생성해 두 개 변수에 동시에 적용해 변환.

[방법 2] 변수별로 독립적인 잡음을 생성해 변환.

[방법 3] 변수별로 독립적인 잡음과 동일한 잡음을 생성해 두 잡음의 평균을 적용해 변환.

[방법 4] 두 번째 변수에 첫 번째 변수에 적용한 잡음과 독립적인 잡음의 평균을 적용해 변환.

가계소득 원자료에서 소득과 지출 및 조세의 상관관계와 위의 4가지 방법으로 변환된 마스킹 자료에서 얻어진 상관관계를 잡음분포별로 산출해 정리하면 표 4.3과 같다. 우선 [방법 1]의 경우 동일한 잡음을 각 변수에 적용하기 때문에 상관관계가 원자료보다 훨씬 커지는 현상이 발생하고, [방법 2]의 경우에는 상관관계가 거의 없어지는 현상이 발생함을 볼 수 있다. 반면에 [방법 3] 또는 [방법 4]의 경우에는 원자료의 상관관계가 상당 부분 유지될 수 있으며, 특히 [방법 4]를 적용하는 경우 원자료와 거의 동일한 상관관계가 마스킹자료에서도 유지될 수 있다는 것을 알 수 있다. 물론 제시된 결과는 이론적인 측면보다는 경험적인 방식으로 변수들 간의 상관관계를 유지할 수 있는 승법잡음모형의 적용 방안을 직관적으로 제시한 것이다. 따라서 이론적인 측면에서 제시된 방법의 적절성을 평가하고, 보다 효율적인 방법을 개발하기 위해서는 향후 보다 심층적인 연구가 필요할 것이다.

5. 결론

본 연구에서는 승법잡음모형을 토대로 한 마스킹자료 생성을 위해 추가적인 변수변환 과정을 통해 기존의 승법잡음모형 관련 연구와는 달리 원자료의 평균과 분산이 마스킹 자료에서도 그대로 유지될 수 있는 노출조절기법을 제시했다.

우리나라 2008년도 가계조사자료를 이용한 모의실험을 통해 제시된 방법이 평균과 분산이 원자료 그대로 유지될 수 있는 마스킹자료 작성방법이라는 점을 확인할 수 있었다. 아울러 잡음분포에 따라 마스킹 자료의 유용성에 어떤 차이가 있는지 검토해 본 결과 검토된 4가지 잡음분포에 따른 차이는 거의 없는 것으로 나타났다. 아울러 여러 개의 연속형 변수를 대상으로 승법잡음모형을 기반으로 마스킹을 하는 경우 원자료가 지니고 있는 변수들 간의 상관관계를 유지하기 위해서는 잡음 발생 과정에 상당한 주의가 필요하다는 점을 보여주었고, 동시에 경험적인 접근을 통해 상관관계를 유지할 수 있는 아이디어를 제시했다.

본 연구에서 제시한 승법잡음모형을 토대로 한 마스킹자료 변환 방법이나 상관관계 유지를 위한 방안은 관련 선행연구가 갖고 있는 한계를 극복했다는 측면에서 의의가 크지만, 아직 이론적으로 다듬어져야 할 부분이 많다. 예를 들어 일반적으로 마이크로자료 이용자들은 단순히 변수별 평균이나 분산에 대한 분석보다 사분위수 또는 변수들 간의 관계를 규명하기 위한 회귀분석이나 다변량분석 등에 관심을 가질 수 있기 때문에, 이런 측면에서 원자료가 갖고 있는 특성이 마스킹자료에 유지될 수 있을 뿐만 아니라 응답자 노출위험이 일정 수준을 넘지 않도록 해야 한다. 결국 이런 효과적인 승법잡음모형을 개발해 실제 통계작성기관에서 활용하기 위해서는 향후 보다 심층적인 연구가 수행되어야 할 것이다.

참고문헌

김규성 (2009). 마이크로데이터 제공과 통계적 노출조절기법, <한국통계학회논문집>, **16**, 1–11.

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

정동명, 김종익, 김경미 (2009). 잡음을 이용한 가계조사자료의 정보노출제한방법, <응용통계연구>, **22**, 141–151.

정동명, 정미옥 (2008). 인구주택총조사 마이크로자료의 개인정보 노출제한방법, <응용통계연구>, **21**, 313–325.

통계청 (2008). 가계조사 조사지침서.

Dalenius, T. and Reiss, S. P. (1982). Data swapping: A technique for disclosure control, *Journal of Statistical Planning and Inference*, **6**, 73–85.

Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation, *Journal of Official Statisitcs*, **9**, 383–406.

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

Kim, J. (1986). A method for limiting disclosure in microdata based on random noise and transformation, *American Statistical Association Proceedings of the Section on Survey Research Methods*, 303–308.

Kim, J. (2007). Application of the truncated triangular and trapezoidal distributions for developing multiplicative noise, *American Statistical Association Proceedings of the Section on Survey Research Methods*, 2723–2729.

Kim, J. and Winkler, W. E. (2001). Multiplicative noise for masking continuous data, *American Statistical Association Proceedings of the Section on Survey Research Methods*, CD-ROM.

Torra, V. (2004). Microaggregation for categorical variavbles: A median based approach, In Domingo-Ferrer, J. and Torra, V. Editors, Privacy in Statistical Databases, *Lecture Notes in Computer Science*, **3050**, 162–174.

Woo, M. J., Reiter, P., Anna, O. and Karr, A. F. (2009). Global measures of data utility for microdata masked for dislosure limitation, *The Journal of Privacy and Confidentiality*, **1**, 111–124.

Application of a Statistical Disclosure Control Techniques Based on Multiplicative Noise

Young-Won Kim[1] · Tae-Yeon Kim[2] · Kye-Nam Kim[3]

[1]Department of Statistics, Sookmyung Women's University
[2]Department of Statistics, Sookmyung Women's University
[3]Department of Statistics, Sookmyung Women's University

(Received December 2010; accepted January 2011)

Abstract

Multiplicative noise model is the one of popular method for masking continuous variables. In this paper, we propose the transformation on the variable to which random noise was multiplied. An advantage of the masking method using proposed transformation is that the masking data users can obtain the unbiased values of mean and variance of original (unmasked) data. We also consider the data utility and correlation structure of variables when we apply the proposed multiplicative noise scheme. To investigate the properties of the method of masking based on multiplicative noise, a simulation study has been conducted using the 2008 Householder Income and Expenditure Survey data.

Keywords: Data utility, disclosure risk, multiplicative noise model, statistical disclosure control.

This Research was supported by the Sookmyung Women's University Research Grants 2009.
[1]Corresponding author: Professor, Department of Statistics, Sookmyung Women's University, Hyochangwon-gil, 52, Yongsan-gu, Seoul 140-742, Korea. E-mail: ywkim@sookmyung.ac.kr