응용통계연구 (2009)
22(1), 141–151

잡음을 이용한 가계조사자료의 정보노출제한방법

정동명[1] · 김종익[2] · 김경미[3]

[1]통계청 통계개발원, [2]국립보건통계센터(NCHS), [3]통계청 통계개발원

(2008년 9월 접수, 2008년 10월 채택)

요 약

제공되는 마이크로자료가 어떤 형태인지에 따라 응답자의 개인정보를 보호하는 방법도 다르게 적용된다. 본 연구에서는 연속형자료의 비밀보호에 효과적인 잡음(noise)을 이용하는 방법을 소개하고, 통계청에서 실시한 2006년 가계조사 자료에 이 방법을 적용하여 응답자의 정보노출이 제한된 마이크로자료를 작성하는 과정을 설명한다. 잡음의 생성을 위해 삼각분포와 절단된 삼각분포, 사다리꼴분포 그리고 이중삼각분포를 이용하고 소지역 추정에 필요한 공식도 유도한다. 아울러 각 분포별로 얻어진 잡음을 이용하여 가계조사 자료를 변환하여 비교·분석한 결과도 보여준다.

주요용어: 노출제한, 가계조사, 절단분포.

1. 서론

최근 들어 개인정보의 비밀보호가 사회적으로 이슈화되면서 통계청을 비롯한 국가통계 작성기관에서는 마이크로자료(micro data)나 매크로자료(macro data)를 외부에 제공할 때, 사전에 미리 응답자의 개인정보가 노출되지 않도록 제한하는 방법을 적용하고 있다. 예를 들어, 통계청에서는 통계이용자들을 위해 5년마다 실시하는 인구주택총조사 결과 중 2%를 마이크로자료 파일로 작성하여 그대로 제공하였으나, 지난 2005년 인구주택총조사 결과부터는 일부 민감한 변수에 대해 노출제한방법이 적용된 2% 마이크로자료 파일을 작성하여 제공하고 있다.

통계자료의 비밀보호는 1970년대 이후부터 영국이나 네덜란드, 미국 등 여러 통계선진국에서 꾸준히 연구되고 있는 분야이며 자료교환(data swapping)이나 그룹화(grouping), 반올림(rounding) 등 다양한 방법이 개발되어 현재 널리 활용되고 있다. 우리나라의 경우에도 최근에 통계작성기관과 학계에서 이에 대한 연구가 점차 활성화되고 있으며, 특히 통계청에서는 2007년 통계법 개정으로 인한 마이크로자료 제공의무에 따라 응답자의 비밀보호를 위한 연구를 체계적으로 진행하고 있다. 정동명과 정미옥(2008)은 2005년 인구주택총조사 결과에 그룹화 등의 방법을 적용하여 응답자의 정보노출이 제한된 마이크로자료 파일을 작성하는 연구를 하였다.

본 연구에서는 연속형자료의 비밀보호에 효과적인 잡음을 이용하는 방법을 소개하고, 이 방법을 통계청에서 실시하는 가계조사의 결과에 적용하여 자료를 변환하는 과정을 설명하고자 한다. 2절에서는 잡음을 이용한 자료변환모형과 잡음의 생성에 필요한 분포를 소개한다. 3절에서는 소개된 분포별로 잡음을 생성한 후, 가계조사에서 민감변수로 선정된 항목을 대상으로 잡음을 이용하여 자료를 변환하는 과정을 설명하고 원자료와 변환된 자료를 분석하여 서로 비교해 본다. 마지막으로 본 연구의 결론은 4절에서 살펴보도록 한다.

[1]교신저자: (302-701) 대전광역시 서구 둔산동 선사로 139, 통계청 통계개발원, 연구원.
E-mail: jedomy@nso.go.kr

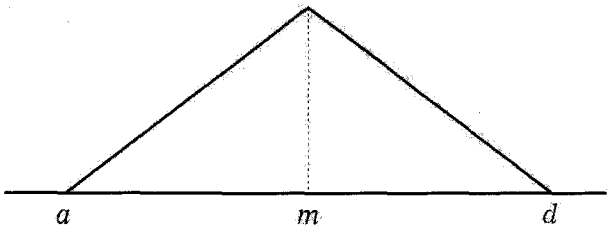

그림 2.1. 삼각분포

2. 잡음을 이용한 정보노출제한

2.1. 변환모형

연속형 자료로 이루어진 원자료를 X라 하고 이에 대응하는 잡음을 e라고 하자. 만약 원자료 X에 적절한 수준의 잡음 e를 더해주면 변환된 자료 Y가 얻어지는데, 이러한 변환모형을 가법잡음(additive noise)모형이라 하고 $Y = X + e$로 나타낸다. 이 모형은 적용하기가 편리하지만 X가 소득항목이면서 0의 값을 갖는다면 e의 값에 따라 Y가 음수가 될 수도 있다. 이에 대한 자세한 내용은 Kim (1986)과 Fuller (1993) 등을 참고하기 바란다.

한편, 원자료 X에 e를 곱해주어 자료를 변환할 수도 있는데, 이 때 변환된 자료 Y는 $Y = X \cdot e$로 나타낼 수 있으며, 이 모형을 승법잡음(multiplicative noise)모형이라 한다. 이 모형에서 e는 일반적으로 1이 아니면서 1을 중심으로 1에서 가까운 값을 가지는 것이 바람직한데, 이는 e가 1이 되면 Y는 X와 같게 되어 변환의 의미가 없고 e가 1에서 멀리 떨어진 값을 가지면 X와 Y의 차이가 너무 커서 정보의 손실이 크기 때문이다. 승법잡음모형은 연속형자료의 변환에 상당히 효과적으로 사용되는데, 본 연구에서도 이 모형을 적용하여 분석하고자 한다.

2.2. 잡음의 분포와 그 특성

잡음을 이용하여 자료를 변환시키고자 할 경우 어떤 잡음을 생성해서 이용하느냐가 매우 중요한 요인이 된다. 일반적으로 잡음은 적절한 분포를 가정하고 이를 바탕으로 난수(random number)발생 프로그램을 작성하여 생성하고 있다. Kim과 Winkler (2001)는 평균을 중심으로 좌우가 절단된 정규분포(truncated normal distribution)를 잡음의 분포로 적용하여 승법잡음모형으로 변환한 결과를 소개하였으며, Kim (2007)은 절단된 삼각분포를 이용한 결과와 그 특성에 대해 분석하였다. 본 연구에서는 승법잡음모형에서 잡음의 분포로 이용할 수 있는 4가지 분포를 소개하고 이들의 특성을 살펴보고자 한다.

2.2.1. 삼각분포 삼각분포(triangular distribution)는 그림 2.1에 나타난바와 같이 최빈값(mode) m을 중심으로 삼각형 모양으로 이루어진 분포를 말한다. 여기서 a는 자료의 최소값이고 d는 최대값이다. 잡음 e가 최빈값 m을 중심으로 삼각분포를 따른다고 하면, 확률밀도함수(probability density function: pdf) $f^t(e)$와 기대값 $E^t(e)$및 분산 $\mathrm{Var}^t(e)$은 다음과 같이 나타낼 수 있다.

$$f^t(e) = \begin{cases} \dfrac{2}{(m-a)(d-a)}(e-a), & a \le e < m, \\ \dfrac{2}{(d-m)(d-a)}(d-e), & m \le e < d, \end{cases}$$

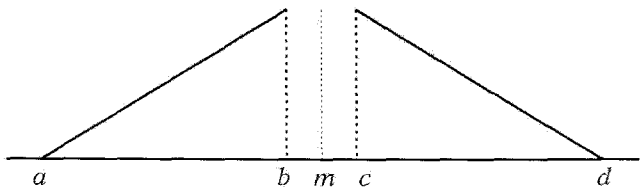

그림 2.2. 절단된 삼각분포

$$E^t(e) = \frac{m+a+d}{3},$$
$$\mathrm{Var}^t(e) = \frac{d^2+a^2+m^2-m(a+d)-ad}{18}.$$

만약 e가 m을 중심으로 좌우대칭인 삼각분포를 따른다면 $m-a=d-m$이 되므로, $E^t(e)$와 $\mathrm{Var}^t(e)$는 다음과 같이 간단히 나타낼 수 있다.

$$E^t(e) = m\ , \qquad \mathrm{Var}^t(e) = \frac{(d-m)^2}{6}.$$

2.2.2. 절단된 삼각분포 절단된 삼각분포(truncated triangular distribution)는 최빈값 m을 중심으로 좌우로 일부분이 절단된 형태로 이루어진 분포로 그림 2.2에 나타난 바와 같다. 즉, 절단된 삼각분포는 m을 중심으로 두 절사점 b와 c $(c>b)$에서 각각 절단되어 있어 m을 포함하지 않는다.

잡음 e가 m을 중심으로 절단된 삼각분포를 따른다면 확률밀도함수(pdf) $f^{tt}(e)$와 기대값 $E^{tt}(e)$ 그리고 분산 $\mathrm{Var}^{tt}(e)$는 다음과 같이 나타낼 수 있다.

$$f^{tt}(e) = \begin{cases} \dfrac{2(d-m)}{(b-a)^2(d-m)+(d-c)^2(m-a)}(e-a), & , a \le e < b, \\ \dfrac{2(m-a)}{(b-a)^2(d-m)+(d-c)^2(m-a)}(d-e), & , c \le e < d, \end{cases}$$

$$E^{tt}(e) = \frac{(d-m)(b-a)^2(2b+a)+(m-a)(d-c)^2(2c+d)}{3\{(d-m)(b-a)^2+(m-a)(d-c)^2\}},$$

$$\mathrm{Var}^{tt}(e) = \frac{1}{18\{(b-a)^2(d-m)+(d-c)^2(m-a)\}^2}\big[(b-a)^6(d-m)^2+(d-c)^6(m-a)^2 + (b-a)^2(d-m)(d-c)^2(m-a)\left\{(3b+a)^2+(3c+d)^2+2(a^2+d^2)-4(2b+a)(2c+d)\right\}\big].$$

만약 m을 중심으로 좌우대칭이면, $b-a=d-c$가 되므로 $E^{tt}(e)$와 $\mathrm{Var}^{tt}(e)$는 다음과 같이 간단히 나타낼 수 있다.

$$E^{tt}(e) = m\ , \qquad \mathrm{Var}^{tt}(e) = \frac{(d+c)^2+2\{(2m-c)^2-m(m+2d)\}}{6}.$$

절단된 삼각분포는 절단되지 않은 삼각분포와 기대값은 같지만 분산은 서로 다르다. 즉, 절단된 경우의 분산이 더 커지게 되는데, 이는 평균 주위의 자료들이 절단됨으로써 그만큼 더 나머지 자료들이 평균에서 멀리 떨어지기 때문일 것이다. 예를 들어 $a=0.5$, $b=0.75$, $m=1$, $c=1.25$, $d=1.5$라고 하면 절단된 삼각분포의 기대값은 1로서 절단되지 않은 삼각분포와 같다. 즉, $E^t(e)=1=E^{tt}(e)$가 된다. 그러나 절단된 삼각분포의 분산은 $\mathrm{Var}^t(e)=0.5/18=0.028$이고 절단되지 않은 삼각분포의 분산은 $\mathrm{Var}^{tt}(e)=8.25/72=0.115$가 된다.

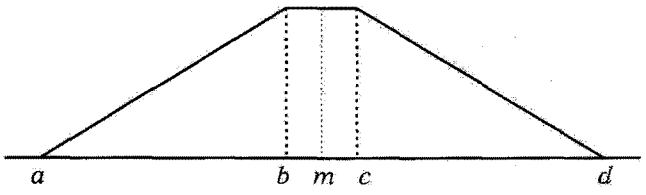

그림 **2.3.** 사다리꼴분포

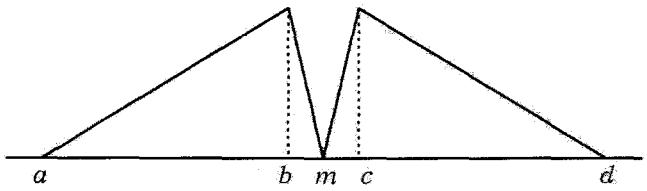

그림 **2.4.** 이중삼각분포

2.2.3. 사다리꼴분포 사다리꼴분포(trapezoidal distribution)는 자료의 최빈값 m을 중심으로 그림 2.3과 같은 모양으로 이루어진 분포를 말한다. 즉, 사다리꼴분포는 b와 c $(c > b)$사이에 일정한 값을 가지는 형태이므로 절단된 삼각분포와는 달리 최빈값을 포함하고 있다.

잡음 e가 m을 중심으로 사다리꼴분포를 따르고 두 절사점 b와 c의 높이가 동일하다고 한다면 확률밀도함수(pdf) $f^{tr}(e)$와 기대값 $E^{tr}(e)$ 및 분산 $\mathrm{Var}^{tr}(e)$는 다음과 같다.

$$f^{tr}(e) = \begin{cases} \dfrac{2}{(d+c-b-a)}\dfrac{(e-a)}{(b-a)}, & a \le e < b, \\ \dfrac{2}{d+c-b-a}, & b \le e < c, \\ \dfrac{2}{(d+c-b-a)}\dfrac{(d-e)}{(d-c)}, & c \le e < d, \end{cases}$$

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

$$\mathrm{Var}^{tr}(e) = \frac{1}{18(d+c-b-a)} \times \left[3\left(d^3+c^3-b^3-a^3-a^2b-ab^2+c^2d+cd^2\right) - \frac{2\{(d+c-b-a)(d+c+b+a)+ab-cd\}^2}{d+c-b-a} \right].$$

만약 m을 중심으로 좌우대칭이면 $b+c=d+a=2m$이 되므로 $E^{tr}(e)$와 $\mathrm{Var}^{tr}(e)$는 다음과 같이 간단하게 나타낼 수 있다.

$$E^{tr}(e) = m, \qquad \mathrm{Var}^{tr}(e) = \frac{d^3+c^3-4m^3+(6m^2+cd)(d+c)-4m(d^2+c^2+cd)}{6(d+c-2m)}.$$

2.2.4. 이중삼각분포 이중삼각분포(double triangular distribution)는 그림 2.4에 나타난바와 같이 최빈값 m을 중심으로 좌우가 각각 삼각형 모양으로 이루어진 분포를 말한다. 즉, a에서 b를 중심으로

표 **3.1.** 2006년 가계조사의 표본규모

(단위:가구)

지역	연간	1/4분기	2/4분기	3/4분기	4/4분기
전국	12,458	9,392	7,718	8,054	8,360
서울	1,638	1,259	947	965	987
기타지역	10,820	8,133	6,771	7,089	7,373

m까지 하나의 삼각형 모양이 되고, m에서 c를 중심으로 d까지 또 하나의 삼각형 모양이 되는 분포이다. 이 분포는 b와 c사이에서 적절한 값을 갖지만 사다리꼴분포와는 달리 m을 포함하지 않는다.

잡음 e가 m을 중심으로 이중삼각분포를 따른다면 확률밀도함수(pdf) $f^{dt}(e)$와 기대값 $E^{dt}(e)$ 및 분산 $\mathrm{Var}^{dt}(e)$는 다음과 같다.

$$f^{dt}(e) = \begin{cases} \dfrac{1}{(m-a)(b-a)}(e-a), & a \le e < b, \\ \dfrac{1}{(m-a)(m-b)}(m-e), & b \le e < m, \\ \dfrac{1}{(d-m)(c-m)}(e-m), & m \le e < c, \\ \dfrac{1}{(d-m)(d-c)}(d-e), & c \le e < d, \end{cases}$$

$$E^{dt}(e) = \frac{a+b+c+d+2m}{6},$$

$$\mathrm{Var}^{dt}(e) = \frac{1}{36}\left\{2a^2+2b^2+2c^2+2d^2+2m^2-m(a+b+c+d)-2(a+b)(c+d)+ab+cd\right\}.$$

만약 m을 중심으로 좌우대칭이면 $E^{dt}(e)$와 $\mathrm{Var}^{dt}(e)$는 다음과 같이 간단히 나타낼 수 있다.

$$E^{dt}(e) = m, \qquad \mathrm{Var}^{dt}(e) = \frac{(d-m)^2+(c-m)^2-m(d+c-m)+cd}{6}.$$

3. 적용사례

3.1. 2006년 가계조사

가계조사(Household Imcome and Expenditure Survey: HIES)는 우리나라 가구에 대한 가계수지의 실태를 파악하여 국민의 소득과 소비수준 변화의 측정 및 분석 등에 필요한 자료를 제공하기 위한 조사이다. 2006년 가계조사의 대상가구는 약 8,000가구이며, 2005 인구주택총조사의 10% 표본조사구를 모집단으로 하여 전국을 25개로 층화한 후 확률비례추출법을 이용하여 선정하였다. 이렇게 선정된 전국의 표본가구를 대상으로 가구의 수입과 소비지출 관련 항목들을 응답가구에서 직접 가계부에 매일 작성하도록 하고 있으며, 매월단위로 가계부의 내용을 집계, 정리한 후 그 결과를 분기별 및 연도별로 공표하고 있다. 2006년 가계조사의 대상가구는 표 3.1에 나타난바와 같이 동일한 가구의 중복을 제외하면 총 12,458가구이며, 이 중 서울이 1,638가구이고 서울이외지역이 10,820가구이다. 이 결과는 분기별 및 연별로 조사결과가 집계되어 공표되는 가구 수와 다른데, 이는 1년간 매월 조사에 응답한 동일가구는 12가구가 아닌 1가구로 처리하여 중복성을 제외하였기 때문이다.

가계조사에서 가계부를 통해 조사되는 항목은 그림 3.1에 주어진 바와 같이 가구사항을 제외하면 총수입 항목과 총지출 항목으로 구분되는데, 총수입은 다시 소득과 기타수입, 전월이월금으로 나누어지고 총지출은 가계지출과 기타지출, 월말 현금잔고로 구분된다. 조사항목들을 세분화하여 분류한 후 이를 다

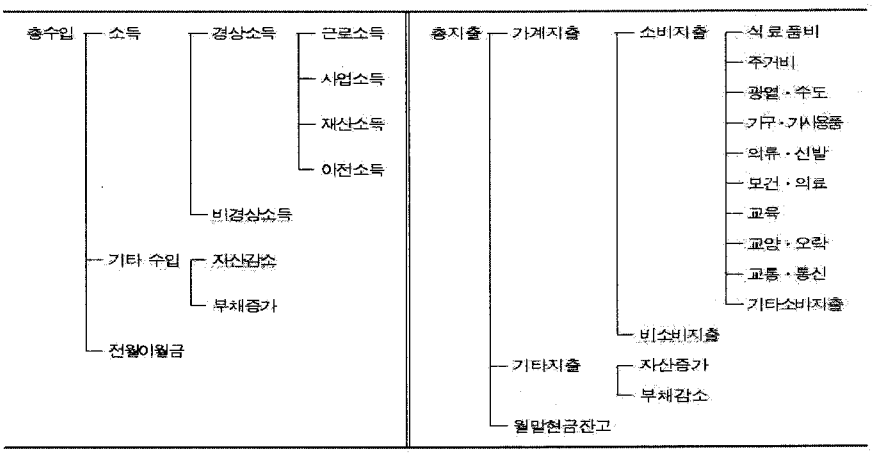

그림 **3.1.** 가계수지 항목분류

표 **3.2.** 민감변수

구분	품 목		
총수입 (9개)	· 가구주소득	· 배우자소득	· 기타 가구원소득
	· 사업소득	· 재산소득	· 이전소득
	· 비경상소득	· 자산감소로인한수입	· 부채증가로인한수입
총지출 (21개)	· 월세	· 주택설비및수선비	· 기타주거
	· 수도료	· 전기료	· 연료
	· 공동주택난방비	· 의약품	· 보건의료용품기구
	· 보건의료서비스	· 납입금	· 교재비
	· 보충교육비	· 문구류	· 교통비
	· 통신비	· 조세	· 공적연금
	· 사회보험	· 기타비소비지출	· 월말현금잔고

시 품목별로 정리하면, 수입관련 품목이 47개이고 지출관련 품목이 510개로 모두 557개가 된다. 따라서 가계조사에서 조사되는 전체항목은 가구특성 62개와 조사품목 557개를 합해 총 619개 항목이 된다. 이에 대한 자세한 내용은 가계조사 지침서 (통계청, 2006)를 참고하기 바란다.

한편, 자료변환을 위해 응답자들이 노출을 꺼려하는 항목을 외부인(intruder)이 다른 경로를 통해 정보를 얻을 수 있을 것으로 판단되는 항목들을 분석한 후, 최종적으로 수입관련 9개 품목과 지출관련 21개 품목 등 총 30개 품목을 민감변수(sensitive variable)로 선정하였으며, 이에 대한 결과는 표 3.2에 주어져 있다.

3.2. 가계조사자료의 정보노출제한

3.2.1. 잡음의 생성 잡음의 생성을 위해 잡음분포는 앞 절에서 소개한 4가지 분포로 이용하였다. 최빈값(m)은 1로 하고, 자료의 범위가 $0.01 \leq |e-1| \leq 0.4$를 만족하도록 최소값(a)과 최대값(d), 두 절사점(b, c) 등을 각각 정하였다. 이 조건이 다소 주관적일 수도 있으나, b와 c가 $m=1$에서 너무 멀어지면 분산이 커지기 때문에 가능하면 m에서 가까운 값을 갖도록 하기 위함이다. 모수 선정을 위해 원자료와 변환자료의 분산이 동일하게 되는 조건을 만족하는 값을 모수로 사용할 수도 있는데 이에 대해서는

표 **3.3.** 생성된 난수의 결과비교

분포	구분	통계량			
		평균	표준편차	최소값	최대값
삼각	이론적(A)	1.0000	0.1633	0.6000	1.4000
	난수결과(B)	1.0002	0.1624	0.6041	1.3916
	차이(B − A)	0.0002	−0.0009	0.0041	−0.0084
절단된 삼각	이론적(A)	1.0000	0.1675	0.6000	1.4000
	난수결과(B)	1.0002	0.1666	0.6040	1.3918
	차이(B − A)	0.0002	−0.0009	0.0040	−0.0082
사다리꼴	이론적(A)	1.0000	0.1634	0.6000	1.4000
	난수결과(B)	1.0002	0.1624	0.6041	1.3916
	차이(B − A)	0.0002	−0.0010	0.0041	−0.0084
이중삼각	이론적(A)	1.0000	0.1654	0.6000	1.4000
	난수결과(B)	1.0002	0.1645	0.6041	1.3917
	차이(B − A)	0.0002	−0.0009	0.0041	−0.0083

향후에 다루기로 한다. 다양한 모수값을 가정하여 비교한 결과는 정동명 등 (2008)을 참고하기 바란다.

한편, 위의 조건이 만족되면 $a = 0.6$, $d = 1.4$, $b = 0.99$, $c = 1.01$이 되는데, 이 값들을 이용하여 각 분포별로 난수발생 프로그램을 SAS로 작성하였다. 그리고 2006년도 가계조사의 월별 및 연간 표본규모를 고려하여 약 90,000개의 난수를 생성하였다. 각 분포별로 발생시킨 난수가 어느 정도 정확한지 알아보고자 기초통계량을 계산해 보면, 표 3.3에 나타난바와 같이 대부분의 경우 이론적인 값과 차이가 크지 않아 각 분포별로 난수가 비교적 잘 생성되었음을 알 수 있다.

3.2.2. 원자료의 추정식 가계조사의 원자료를 X_i, 생성된 잡음을 e_i라 하면 승법잡음모형에 따라 변환된 자료 Y_i는 $Y_i = X_i \cdot e_i$, $i = 1, \ldots, n$으로 나타낼 수 있으며, 또한 X_i와 e_i는 서로 독립이기 때문에 변환된 Y_i의 기대값 $E(Y_i)$와 분산 $\mathrm{Var}(Y_i)$은 다음과 같이 구할 수 있다.

$$E(Y_i) = E(X_i) \cdot E(e_i),$$
$$\mathrm{Var}(Y_i) = \mathrm{Var}(X_i)\mathrm{Var}(e_i) + \{E(e_i)\}^2 \mathrm{Var}(X_i) + \{E(X_i)\}^2 \mathrm{Var}(e_i).$$

통계이용자들에게 변환된 자료를 제공할 경우 이용자들은 원자료 X_i의 평균과 분산은 모르고 단지 변환된 자료 Y_i와 잡음 e_i의 평균과 분산만 알 수 있게 된다. 따라서 이용자들 중에는 추가적인 자료분석을 위해 원자료의 평균과 분산을 요구할 수도 있다. 이를 위해 X_i의 평균과 분산을 구하는 추정식이 필요한데, X_i의 기대값 $E(X_i)$와 분산 $\mathrm{Var}(X_i)$은 다음과 같다.

$$\hat{E}(X_i) = \frac{\hat{E}(Y_i)}{E(e_i)},$$
$$\hat{\mathrm{Var}}(X_i) = \frac{\hat{\mathrm{Var}}(Y_i) - \{\hat{E}(X_i)\}^2 \mathrm{Var}(e_i)}{\mathrm{Var}(e_i) + \{E(e_i)\}^2}.$$

한편, 소지역(domain)을 나타내는 첨자를 s라 하면, 승법잡음모형에 의해 변환된 소지역자료는 $Y_i^s = X_i^s \cdot e_i$, $i = 1, \ldots, n_s$로 나타낼 수 있다. 여기서 n_s는 소지역의 크기를 나타내며, 소지역의 원자료 X_i^s와 e_i는 서로 독립이기 때문에 변환된 Y_i^s의 기대값 $E(Y_i^s)$와 분산 $\mathrm{Var}(Y_i^s)$은 다음과 같이 구할 수

있다.

$$E(Y_i^s) = E(X_i^s) \cdot E(e_i),$$
$$\mathrm{Var}(Y_i^s) = \mathrm{Var}(X_i^s)\mathrm{Var}(e_i) + \{E(e_i)\}^2 \mathrm{Var}(X_i^s) + \{E(X_i^s)\}^2 \mathrm{Var}(e_i).$$

또한 소지역 s의 원자료 X_i^s에 대한 기대값 $E(X_i^s)$와 분산 $\mathrm{Var}(X_i^s)$은 다음과 같이 얻을 수 있다.

$$\hat{E}(X_i^s) = \frac{\hat{E}(Y_i^s)}{E(e_i)},$$
$$\hat{\mathrm{Var}}(X_i^s) = \frac{\hat{\mathrm{Var}}(Y_i^s) - \{\hat{E}(X_i^s\}^2 \mathrm{Var}(e_i)}{\mathrm{Var}(e_i) + \{E(e_i)\}^2}.$$

3.2.3. 변환결과 표 3.2에 주어진 민감변수를 대상으로 4가지 잡음분포를 이용하여 잡음을 생성한 후, 가계조사의 결과자료를 승법잡음모형으로 변환하였다. 즉, 가계조사의 연간자료를 원자료로 한 후 전국 및 지역별로 원자료와 변환된 자료의 평균과 표준편차 및 이들의 상대차이(relative difference)를 각각 계산하였으며, 그 결과는 표 3.4와 3.5에 각각 수록하였다. 이 표에 의하면 두 항목의 분석결과가 서로 유사하기 때문에 여기서는 가구주소득 항목에 대해서만 설명하기로 한다.

가구주소득 항목에 대한 결과를 살펴보면, 먼저 전국단위의 경우 원자료와 변환된 자료의 평균과 표준편차의 차이가 크지 않고 이들의 상대차이도 약 0.1% 정도로 나타나 각 분포별로 별 차이가 없는 것으로 나타났다. 분포별로는 삼각분포를 이용하여 변환한 결과가 원자료와 가장 차이가 적고 이어서 사다리꼴분포와 이중삼각분포 그리고 절단된 삼각분포의 순으로 나타났다. 이러한 결과는 잡음의 분포가 삼각분포나 사다리꼴분포인 경우 1이 포함되므로 원자료의 변환이 이루어지지 않는 경우도 있기 때문인 것 같다. 실제로 자료를 변환할 경우에는 잡음이 1이 포함되지 않는 절단된 삼각분포와 이중삼각분포를 이용하는 것이 더 적절한데, 표 3.4의 결과에 의하면 절단된 삼각분포를 이용하는 것보다 이중삼각분포를 이용하는 것이 원자료와의 차이가 약간 더 적음을 알 수 있다. 이러한 결과는 절단된 삼각분포의 경우 두 절사점 사이에서 난수가 전혀 생성되지 않지만, 이중삼각분포의 경우 최빈값을 중심으로 어느 정도 난수가 생성되어 그만큼 더 많은 정보가 이용되기 때문인 것으로 판단된다.

한편, 서울지역의 경우 전국단위나 기타지역의 결과와 다르게 나타났는데, 이는 원자료에 어떤 잡음의 값이 곱해지는가에 따라 변환결과가 달라질 수도 있음을 의미한다. 즉, 난수발생을 위한 잡음분포의 특성도 자료의 변환에 중요한 요인이 될 수 있지만, 무엇보다도 원자료에 곱해지는 잡음의 값 자체가 가장 절대적인 요인이 된다는 것을 보여주고 있다.

4. 결론

본 연구에서는 연속형 자료의 노출제한방법으로 널리 활용되고 있는 승법잡음모형에 의한 변환방법을 소개하고, 2006년 가계조사 자료에 직접 적용하여 자료를 변환하는 과정을 설명하고 분석하였다. 이를 위해 잡음생성에 이용할 4가지 분포를 소개하고, 이들 분포의 평균과 분산의 계산공식을 유도하였다. 그리고 가계조사 자료의 비밀보호를 위해 민감변수를 선정한 후, 변수별로 제시된 잡음분포에 따라 승법잡음모형을 적용하여 자료를 변환하였다.

그 결과 대부분의 항목에서 삼각분포를 가정한 경우가 원자료와 가장 차이가 적게 나타났다. 그러나 실제 자료의 적용에서는 잡음이 1을 포함하지 않는 것이 바람직한데, 이 경우에는 이중삼각분포를 이용하는 것이 평균차이의 측면에서 더 효과적일 수 있다. 따라서 응답가구에서 노출을 꺼려하는 민감한 정

표 3.4. 품목별변환결과

(단위:원)

품목	통계량	지역	원자료	변환된 자료			
				삼각분포	절단된 삼각분포	사다리꼴분포	이중삼각분포
가구주	평균	전국	2,258,159	2,255,944	2,255,851	2,255,942	2,255,894
		서울	2,295,665	2,299,081	2,298,817	2,299,078	2,298,942
		기타	2,252,439	2,249,366	2,249,299	2,249,364	2,249,330
	표준편차	전국	1,636,192	1,635,247	1,635,132	1,635,209	1,635,181
		서울	1,559,763	1,574,642	1,574,811	1,574,607	1,574,723
		기타	1,647,478	1,644,212	1,644,055	1,644,173	1,644,124
품목	통계량	지역	원자료	변환된 자료			
				삼각분포	절단된 삼각분포	사다리꼴분포	이중삼각분포
월세	평균	전국	181,064	180,943	180,975	180,943	180,958
		서울	246,382	247,216	247,280	247,216	247,247
		기타	171,099	170,833	170,860	170,833	170,846
	표준편차	전국	133,014	132,038	132,168	132,037	132,102
		서울	164,236	165,359	165,639	165,358	165,499
		기타	124,615	123,095	123,195	123,094	123,144

표 3.5. 품목별 변환자료의 상대차이

(단위:%)

품목	통계량	지역	상대차이			
			삼각분포	절단된 삼각분포	사다리꼴분포	이중삼각분포
가구주 소득	평균	전국	0.098	0.102	0.098	0.100
		서울	0.149	0.137	0.149	0.143
		기타	0.136	0.139	0.137	0.138
	표준편차	전국	0.058	0.065	0.060	0.062
		서울	0.954	0.965	0.952	0.959
		기타	0.198	0.208	0.201	0.204
품목	통계량	지역	상대차이			
			삼각분포	절단된 삼각분포	사다리꼴분포	이중삼각분포
월세	평균 평균	전국	0.067	0.049	0.067	0.058
		서울	0.338	0.365	0.339	0.351
		기타	0.156	0.140	0.156	0.148
	표준편차	전국	0.733	0.636	0.734	0.686
		서울	0.684	0.855	0.684	0.770
		기타	1.220	1.140	1.221	1.181

* 상대차이 = |원자료 − 변환자료|/원자료

보들이 대부분 연속형 자료로 구성된 가계조사의 경우, 이중삼각분포를 잡음분포로 가정하고 승법잡음모형을 적용하는 것이 응답자의 정보노출을 제한하는데 어느 정도 효과적이라고 할 수 있다. 끝으로 본 연구를 계기로 응답자의 정보를 보호할 수 있는 다양한 연구가 활발히 진행되고, 그 결과를 바탕으로 자료제공의 신뢰성 확보와 함께 궁극적으로 조사의 응답률도 제고될 수 있기를 기대해 본다.

참고문헌

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

정동명, 정남수, 한승훈 (2008). 가계조사 마이크로데이터의 비밀보호, 연구보고서, 통계청.

통계청 (2006). 가계조사 지침서.

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

Kim, J. (1986). A method for limiting disclosure in microdata based on random noise and transformation, In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 370–374.

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

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

A Method of Masking Based on Multiplicative Noise

Dong Myeong Jeong[1] · Jay J. Kim[2] · Kyung Mi Kim[3]

[1]Statistics Research Institute, KNSO; [2]National Center for Health Statistics;
[3]Statistics Research Institute, KNSO

(Received September 2008; accepted October 2008)

Abstract

According to the type of microdata, the various methods have been in use for masking microdata. Multiplicative noise is the one of popular schemes for masking continuous variables. In this paper, we introduce the method of masking based on multiplicative noise and show some results of the application on the 2006 Householder Income and Expenditure Survey(HIES) data. To create the multiplicative noise factor, we used the triangular distribution, truncated triangular distribution, trapezoidal distribution, and double triangular distribution. Also, formulas for the domain estimation for the data masked by the multiplicative noise are developed.

Keywords: Disclosure limitation, HIES, truncation distribution.

[1]Corresponding author: Statistician, Statistics Research Institute, KNSO, Government Complex Daejeon, 139 Seonsaro Seo-gu, Daejeon 302-701, Korea. E-mail: jedomy@nso.go.kr