# 베이지안 차분 정보보호 적용 재현자료 생성 방법론 소개

**박성률**
통계개발원 주무관 / 경제학 박사
srpark08@korea.kr

**오영주**
통계청 통계데이터기획과 주무관
oyj1928@korea.kr

본 연구는 데이터 비식별화 기법 중에서 베이지안 모형 기반 차분 정보보호 기술이 적용된 재현자료를 생성하고, 생성된 재현자료와 원자료 간의 데이터 유사성(데이터 유용성)을 검토한 연구이다. 차분 정보보호의 고전적인 정의와 베이지안 모형, 베이지안 차분 정보보호 적용 재현자료 방법론에 대한 이론을 소개하였다. 또한 실제 데이터를 이용하여 연구에서 제안한 방법으로 재현자료를 생성하고, 검증 과정을 거쳐 $\epsilon$ 값이 커질수록 데이터 유용성이 높아짐을 확인하였다. 하지만 안전성과의 trade-off 관계, 시계열 문제, 다변량 변수 방법론 등의 한계가 드러났다. 본 연구 결과가 차분 정보보호 기술을 이해하는 데 도움이 되고, 아울러 향후 후속연구에서 안정성과 유용성 평가가 함께 실시되어 최적의 $\epsilon$ 값 선정에 기여하는 기초 자료가 되길 바란다.

· 본 원고는 통계개발원 연구보고서『최신 차분 정보보호 방법론 탐색과 통계 활용 방안 연구』(2023년 5월 발간)의 일부 내용을 발췌 · 정리한 것임.



## Ⅰ. 들어가며

우리가 일상생활에서 접하는 수많은 데이터는 분야별로 정의에 차이가 있지만, 속성을 나타내는 의미 있는 정보이고 민감 정보를 다수 포함하고 있다는 공통점이 있다. 민감 정보는 자료를 분석하거나 통계를 작성할 때 중요한 변수로 활용할 수 있으나 프라이버시 문제가 있어서 데이터 활용에 제약이 있다. 데이터 활용성을 높이기 위해 프라이버시를 보장하면서 유용성을 높이는 비식별화 방법이 연구되고 있다.

민감 정보를 삭제하거나 대체하는 전통적인 비식별화 방법은 자료 분석 시 데이터 손실이 발생하여 원자료 활용에 큰 제약이 있다. 특히 민감 정보에는 데이터 분석에서 중요하게 활용되는 정보가 포함된 경우가 많다. 재현자료(synthetic data)는 원자료의 특성을 고려하여 만든 가상자료로, 이를 활용하면 민감 정보가 진짜 같은 가짜 정보로 변환되어 제공되므로 정보의 누락을 해결할 수 있다. 다만 재현자료라도 연결이나 추론 공격 등으로 정보가 노출될 수도 있다. 재현자료의 문제점을 안정적으로 제어하기 위해 차분 정보보호(Differential privacy, DP) 기술이 제안되었다. 우리는 이 연구를 통해 차분 정보보호 개념과 모형(베이지안) 기반 차분 정보보호가 적용된 재현자료 기법에 대해 자세히 살펴보았다.

## Ⅱ. 차분 정보보호(Differential Privacy, DP) 정의

2006년에 D. Work이 발표한 차분 정보보호(DP)는 새로운 정보보호 방법론으로 프라이버시를 정량적으로 모델화하여 프라이버시 보호 정도를 측정할 수 있는 방법이다. 즉 DP는 데이터 분포 특성을 고려하여 데이터의 유용성은 유지하면서 개인정보를 보호하기 위해 잡음을 추가하는 방법이다.

DP는 프라이버시의 일부 노출을 감수하면서 원본 데이터와 유사한 특징을 갖도록 데이터를 익명화하는 것이 중요하다. 그 이유는 프라이버시와 데이터 유용성이 트레이드-오프(trade-off) 관계에 있기 때문이다.

다시 말하면 DP는 특정 개인의 존재 유무가 다른 두 가지 데이터베이스와 관련된 어떠한 질의의 출력값에 무작위성을 추가하여 두 가지 버전에 대한 질의응답 값이 확률적으로 일정값 이하의 차이가 나도록 함으로써 차분 공격을 어렵게 하는 프라이버시 모델이라고 할 수 있다.

$$P[K(D_1 \in S)] \leq e^{\epsilon} P[K(D_2 \in S)] \qquad [1]$$

$$\frac{P[K(D_1 \in S)]}{P[K(D_2 \in S)]} \leq e^{\epsilon} \qquad [2]$$

- 식[1]은 $\epsilon$ DP의 정의로 1개의 데이터 포함 여부(값 존재 유무, 값 차이 등)가 다른 두 가지 데이터베이스 $D_1$과 $D_2$에 대하여 확률함수 $K$에 임의의 잡음을 추가한다.
- 응답값이 $K(D_1)$과 $K(D_2)$가 될 때 $K$에 의해 생성되는 값들의 확률분포에서 특정값이 나올 확률의 차이가 $e^{\epsilon}$보다 작거나 같으면 차분 정보보호가 되었다고 본다.
- 여기서 $S$는 모든 결과값으로 평균, 총계 등의 추정값을 의미한다.

또한 DP는 프라이버시 예산(privacy budget)을 도입하여 정보 공개에 따른 프라이버시 손실을 계량화하였고, $\epsilon$으로 표기하고 있다. 여기서 $\epsilon$은 프라이버시와 데이터 유용성을 결정하는 기준이 됨을 알 수 있다.

[표 1] 차분 정보보호에서 $\epsilon$의 특징

| 구 분 | 특징 |
|---|---|
| $\epsilon$가 小 | 프라이버시 강화, 손실 감소, 잡음이 많은 데이터, 유용성 하락, 안전성 상승 |
| $\epsilon$가 大 | 프라이버시 약화, 손실 증가, 잡음이 적은 데이터, 유용성 상승, 안전성 하락 |

다음으로 DP를 이해하려면 전역 민감도(global sensitivity)를 알아야 한다. 전역 민감도는 특정 개인을 추가 또는 제거했을 때 생기는 변화량의 최대값을 말한다. 데이터 경계가 모호하거나 이상치가 큰 경우 민감도를 설정하기 어렵거나 지나치게 큰 값이 설정되는 문제가 함께 발생한다. 작은 전역 민감도는 잡음을 적게 만들고 큰 전역 민감도는 잡음을 많이 만든다.

마지막으로 식[1]에서 확률변수 $K$메커니즘의 대표적인 방법은 라플라스 메커니즘이다.





라플라스 메커니즘은 임의의 잡음을 추가하는 방법으로 $f(x)$에 $Laplace(0,b)$ 분포를 이용하여 잡음을 추가한다.

$$X = \mu - b\ sgn(U)\ \ln(1-2[U]) \qquad [3]$$

- 여기서, $X$는 잡음, $\mu$는 대체로 0으로 설정, $U$는 $-1/2 \leq U \leq 1/2$ 사이 랜덤 값을 잡음으로 생성, $sgn$은 입력값이 양수=1, 음수=−1 반환, $b$는 $\Delta f/\epsilon$로 정의한다.
- 예컨대 원본 데이터의 속성값이 0.7인 값에 잡음을 추가하여 비식별 속성값을 계산한다면 여기서 $\mu = 0$, $U = 0.3$, $\Delta f = 1$, $\epsilon = 0.2$ 값이 주어졌을 때, 라플라스 메커니즘을 적용하여 계산한 잡음($X$)를 더하여 만든 비식별화된 속성값은 2.6897이 된다.

## Ⅲ. 베이지안 적용 재현자료 생성 방법론

베이지안 통계학은 고전 통계학과 달리 선험적 정보를 반영하여 분석 및 예측하는 통계학의 한 분야이다. 고전 통계학과 베이지안 통계학의 큰 차이점은 모수를 바라보는 관점이 다르다는 것이다. 고전 통계학은 모수를 고정된 값으로 고려하는 반면, 베이지안 통계학은 모수를 분포를 가지는 확률변수로써 고려한다.

$$f(\theta|\boldsymbol{x}) = \frac{f(\theta,\boldsymbol{x})}{f(\boldsymbol{x})} = \frac{f(\boldsymbol{x}|\theta)f(\theta)}{f(\boldsymbol{x})} \qquad [4]$$

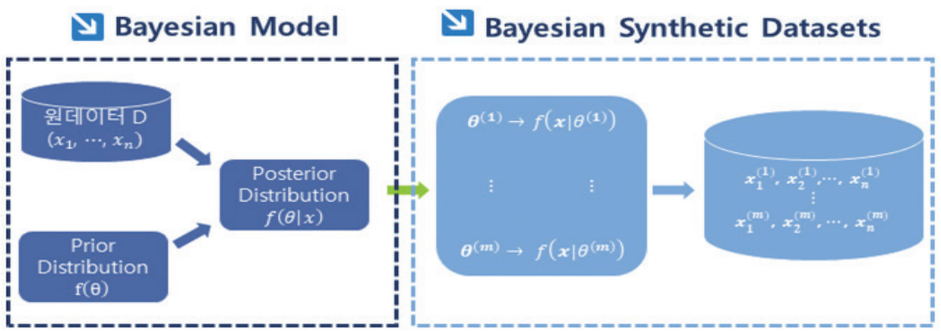
- 여기서, $f(\theta|\boldsymbol{x})$는 사후분포, $f(\theta)$는 사전분포, $f(\boldsymbol{x}|\theta)$는 우도함수를 의미한다.
- 식[4]의 사후분포를 활용하여, 관측되지 않은 새로운 자료 $x^*$에 대한 확률분포는 식[5]와 같이 추정될 수 있다. 이는 사후예측분포라고 불리며, 본 연구에서 재현자료 생성을 위해 이용된다.

$$f(x^*|\boldsymbol{x}) = \int_{\theta} f(x^*|\theta) f(\theta|\boldsymbol{x}) d\theta \qquad [5]$$

앞서 소개한 베이지안 정리는 주어진 데이터에 대해 베이지안 모델을 생성할 수 있고, 모수를 샘플링하는 과정을 통해 재현자료 생성에 이용될 수 있다. 이를 기반으로 한 다중 재현자료($m$개) 생성 절차는 다음과 같다.

- 1단계는 베이지안 모델 생성 단계로 원본 데이터(D)와 $m$개의 사전분포($f(\theta)$)를 이용하여 $m$개의 사후분포($f(\theta|x)$)를 생성한다.
- 2단계는 재현자료 생성 단계로 $m$개의 사후분포로부터 추정한 $m$개의 $\theta$들을 이용하여 사후예측분포를 통해 다중 재현자료 세트를 생성한다.

[그림 1] 베이지안 적용 재현자료 생성 알고리즘



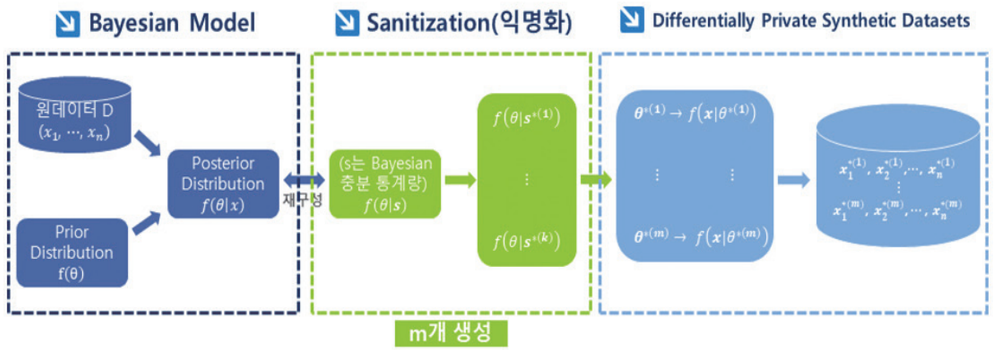

## Ⅳ. 베이지안 차분 정보보호 적용 재현자료 생성 방법론

이 방법은 기존 베이지안 적용 재현자료 생성 방법론에 차분 정보보호 적용을 위해 익명화(sanitization) 기법을 추가한 것으로 재현자료 내 원자료에 대응되는 개별 데이터가 없도록 차분 정보보호 개념을 적용하여 사용한다.

차분 정보보호를 적용한 재현자료 생성 절차는 베이지안 모델 생성, 익명화, 재현자료 생성 등 세 단계로 이루어진다.

- 1단계는 기존 방법과 동일하게 베이지안 모델을 생성하는 단계로 후보 세트 $m$개를 생성한다.
- 2단계는 익명화(sanitization) 단계로 Direct sanitization, SBS(Sanitization through Bayesian Sufficiency), sanitization of Approximate Distribution 등 세 가지 익명화 방법이 있다. 이 연구에서는 베이지안 충분통계량 $\boldsymbol{s}$를 이용하여 사후분포($f(\theta|x)$)를 $f(\theta|\boldsymbol{s})$로 재구성하고, $\boldsymbol{s}$를 익명화하는 베이지안 충분성을 통해 SBS 기법으로 분석한다.
- 3단계는 차분 정보보호 메커니즘을 통해 선정된 베이지안 모델에 대응하여 사후 표본 $\theta^{*(i)}$를 만드는 단계로 이 값을 이용하여 $f(\boldsymbol{x}|\theta^{*(i)})$에서 $\tilde{\boldsymbol{x}}^{*(i)}$를 생성한다. 최종적으로 $m$개의 재현자료 데이터 세트가 생성된다.

[그림 2] 베이지안 차분 정보보호 적용 재현자료 생성 알고리즘




## Ⅴ. 재현자료 생성 및 결과 비교

본 연구에서는 재현자료 생성을 위한 자료로 금융감독원 전자공시시스템의 2021년 사업보고서 중 재무상태표, 손익계산서, 현금흐름표 등의 데이터를 활용하였다. 재현자료 생성 대상 변수로는 'null 최소인 변수', '극단값 제거', '정규성 검정', '산업대분류 100개 이상'이라는 조건을 충족하는 변수를 선정하였다.

베이지안 모델을 활용하여 차분 정보보호를 충족하는 재현자료 생성 과정에서 가장 먼저 베이지안 모델을 생성해야 한다.

[그림 3] 최종 변수 현황

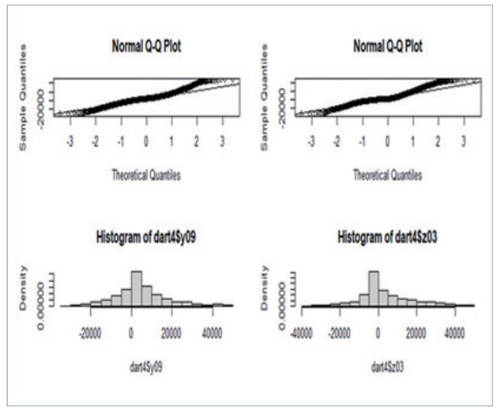
〈최종 변수〉

| 자료 구분 | 항목 코드 | 항목명 | 변수명 |
|---|---|---|---|
| 기업일반 | 업종 | 산업대분류 | KSIC1 |
| 손익 계산서 | ifrs-full_ProfitLoss | 당기순이익(손실) | y09 |
| | ifrs-full_CashFlowsFromUsedinFinancingActivities | 재무활동으로 인한 현금흐름 | z03 |

〈산업대분류별 데이터 현황〉

| 산업대분류 | 원자료 | null & 극단값 |
|---|---|---|
| C 제조업 | 1,454 | 1,135 |
| G 도매 및 소매업 | 161 | 123 |
| J 정보통신업 | 276 | 218 |
| 합계 | 1,891 | 1,476 |

〈QQplot, Histogram〉




– 앞서 최종으로 선정된 변수는 정규분포를 따르기 때문에 평균 $\mu$와 정밀도 $\lambda$를 가진 정규분포의 우도함수(likelihood function)는 식[6]과 같이 정의된다.

$$p(x_1, \cdots, x_n|\mu, \lambda) \sim \prod_{i=1}^{n} N(x_i|\mu, \lambda^{-1}) = \prod_{i=1}^{n} \sqrt{\frac{\lambda}{2\pi}} \exp\left\{-\frac{\lambda(x_i-\mu)^2}{2}\right\} \quad [6]$$

– 알려지지 않은 모수($\mu$,$\lambda$)에 대해서는 식[7]과 같이 정규-감마 사전분포(Normal-Gamma Prior distribution)를 이용한다.

$$\mu \sim N(\mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi}\,\sigma_0} \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} \qquad \lambda \sim Gam(\alpha_0, \beta_0) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \lambda^{\alpha_0-1} \exp\{-\beta_0\lambda\} \quad [7]$$

– 본 연구에서는 모수($\mu$,$\lambda$)의 분포에 대해 아무런 지식을 가지고 있지 않으므로, 사전분포의 분산($\mu$의 분산: $\sigma_0^2$, $\lambda$의 분산: $\alpha/\beta^2$)을 $1e+10$으로 매우 크게 설정함으로써 무정보 사전분포(non-informative prior distribution)로 두고 분석을 수행하였다.





– 결합 사후분포(joint posterior distribution)는 우도함수에 사전분포를 곱하여 얻을 수 있으며 식[8] 및 식[9]와 같이 구해진다.

$$p(\mu, \lambda|\boldsymbol{x}) \propto \left(\frac{\lambda}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{\sum_{i=1}^{n}\lambda(x_i-\mu)^2}{2}\right\} * \frac{1}{\sqrt{2\pi}\,\sigma_0} \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} * \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \lambda^{\alpha_0-1} \exp\{-\beta_0\lambda\} \quad [8]$$

$$\propto \lambda^{\frac{n}{2}+\alpha_0-1} \exp\left\{-\frac{1}{2}\lambda\left(\sum_{i=1}^{n}x_i^2 - \mu\sum_{i=1}^{n}2x_i + n\mu^2 + 2\beta_0\right) - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$

$$= \lambda^{\frac{n}{2}+\alpha_0-1} \exp\left\{-\frac{1}{2}\lambda\left(\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2 + 2\beta_0\right) - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} \quad [9]$$

$$= \lambda^{\frac{n}{2}+\alpha_0-1} \exp\left\{-\frac{1}{2}\lambda\left(ns^2 + n(\bar{x}-\mu)^2 + 2\beta_0\right) - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$

– 이때 식[8] 및 식[9]의 함수는 $(\bar{x},\ s^2)=\left(\frac{1}{n}\sum_{i=1}^{n}x_i,\ \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)$에 대한 함수로 나타낼 수 있으며, 베이지안 충분통계량은 $\boldsymbol{s}=(\bar{x}, s^2)$이 된다.

다음으로 $\epsilon$ DP를 충족하는 재현자료를 생성하기 위해서 라플라스 메커니즘을 활용하여 익명화(sanitization)를 진행한다.

– 라플라스 메커니즘($\mathcal{L}(0, \Delta_1\epsilon^{-1})$)으로부터 독립적인 확률표본을 생성함으로써 프라이버시 예산($(\epsilon-\epsilon_0)/m$ )을 가지고 충분통계량 $s^*$를 익명화한다.

$$\boldsymbol{s}^* = \boldsymbol{s} + \boldsymbol{e} \quad [10]$$

– 여기서, $\boldsymbol{s}=(s_1, \cdots, s_r)$는 $r$차 통계량, $\boldsymbol{e}$는 $\mathcal{L}(0, \Delta_1\epsilon^{-1})$로부터 $r$개의 독립적인 확률표본들,

$\Delta_1 = \max_{\boldsymbol{x}, \boldsymbol{x}', d(\boldsymbol{x}, \boldsymbol{x}')=1} \| \boldsymbol{s}(\boldsymbol{x}) - \boldsymbol{s}(\boldsymbol{x}') \|$는 $\boldsymbol{s}$의 $l_1$ 전역 민감성($l_1$ global sensitivity, $l_1$ GS)을 말한다.

[표 2] $\bar{x}$에 적용되는 산업대분류별 $l_1$ 전역 민감성

(단위: 백만 원)

| 항목명 \ 산업대분류 | C 제조업 | G 도매 및 소매업 | J 정보통신업 | 전체(C+G+J) |
|---|---|---|---|---|
| 당기순이익(손실) | 53.39 | 51.63 | 47.02 | 53.39 |
| 재무활동으로 인한 현금흐름 | 56.11 | 49.29 | 55.02 | 56.27 |

[표 3] $s^2$에 적용되는 산업대분류별 $l_1$ 전역 민감성

(단위: 백만 원)

| 항목명 \ 산업대분류 | C 제조업 | G 도매 및 소매업 | J 정보통신업 | 전체(C+G+J) |
|---|---|---|---|---|
| 당기순이익(손실) | 4,207,979.50 | 3,934,741.15 | 3,263,038.32 | 4,207,979.50 |
| 재무활동으로 인한 현금흐름 | 4,646,129.32 | 3,585,772.36 | 4,468,007.02 | 4,673,708.93 |

마지막으로 깁스샘플링(Gibbs sampling)에 사용되는 $\mu$와 $\lambda$의 완전 조건부 분포(Full Conditional Distribution, FCD)를 이용하여 식[11] 및 식[12]와 같이 재현자료를 생성한다.

$$
\begin{aligned}
p(\mu|\lambda,\boldsymbol{x}) &\propto \exp\left\{-\frac{1}{2}\mu^2\left(\lambda n+\sigma_0^{-2}\right)+\mu\left(\lambda\sum_{i=1}^{n}x_i+\mu_0\sigma_0^{-2}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left\{\mu^2\left(\lambda n+\sigma_0^{-2}\right)-2\mu\left(\lambda\sum_{i=1}^{n}x_i+\mu_0\sigma_0^{-2}\right)\right\}\right\} \\
&\sim N\left(\frac{\lambda n\bar{x}+\mu_0\sigma_0^{-2}}{\lambda n+\sigma_0^{-2}}, \left(\lambda n+\sigma_0^{-2}\right)^{-1}\right)
\end{aligned}
\qquad [11]
$$

$$
\begin{aligned}
p(\lambda|\mu,\boldsymbol{x}) &\propto \lambda^{\frac{n}{2}+\alpha_0-1}\exp\left\{-\frac{1}{2}\lambda\left(ns^2+n(\bar{x}-\mu)^2+2\beta_0\right)\right\} \\
&\sim Gam\left(\frac{n}{2}+\alpha_0,\ \frac{ns^2+n(\bar{x}-\mu)^2}{2}+\beta_0\right)
\end{aligned}
\qquad [12]
$$





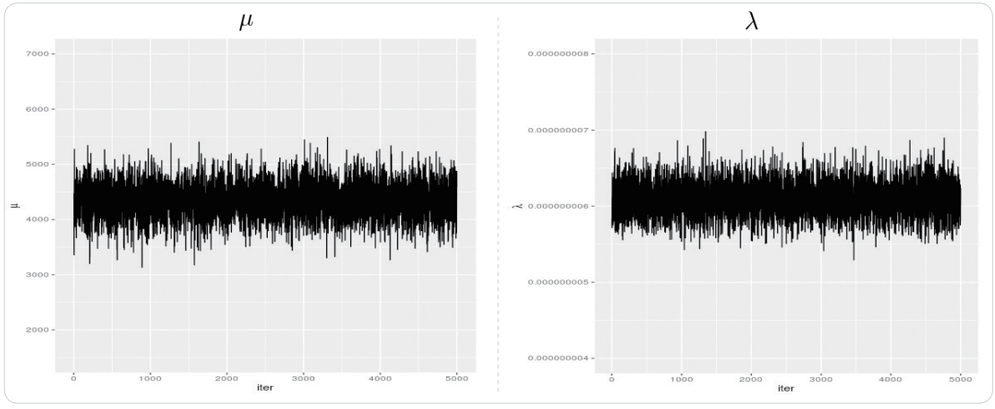
– 깁스샘플링 절차에 따라 $\mu, \lambda$의 사후 기댓값은 식[13]과 같이 얻을 수 있다.

$$
\begin{aligned}
\hat{\mu} &= \hat{E}(\mu|\underline{x}) = \frac{1}{M-b}\sum_{m=b+1}^{M}\mu_m, \\
\hat{\lambda} &= \hat{E}(\lambda|\underline{x}) = \frac{1}{M-b}\sum_{m=b+1}^{M}\lambda_m
\end{aligned}
\qquad [13]
$$

– 여기서, $b$은 burn-in 구간이고, $M=5000$, $b=4000$으로 설정하였다.

– [그림 4]의 trace plot 결과를 보면 깁스샘플링 수행 결과 $\mu$와 $\lambda$가 정상적으로 수렴함을 확인할 수 있다.

[그림 4] $\mu$와 $\lambda$의 trace plot




– [표 4]에서 베이지안 추정량은 베이지안 충분통계량에 차분 정보보호($\epsilon$=0.1, 1, 10)를 적용한 후 깁스 샘플링을 통해 추정한 사후 기댓값을 의미한다.

[표 4] 차분 정보보호 수준별 베이지안 추정량(사후 기댓값)

(단위: 백만 원)

| 항목명 \ 구분 | 모수 | 원자료 (표본평균/분산) | 베이지안 추정량(사후 기댓값) | | |
|---|---|---|---|---|---|
| | | | $\epsilon=0.1$ | $\epsilon=1$ | $\epsilon=10$ |
| 당기순이익(손실) | $\hat{\mu}$ | 4,356.87 | 4,838.92 | 4,468.82 | 4,348.23 |
| | $1/\hat{\lambda}$ | 164,224,210.03 | 152,340,333.14 | 161,545,480.38 | 164,158,464.03 |
| 재무활동으로 인한 현금흐름 | $\hat{\mu}$ | 4,051.75 | 4,558.05 | 4,170.51 | 4,042.24 |
| | $1/\hat{\lambda}$ | 202,924,845.66 | 189,690,723.97 | 199,904,953.59 | 202,824,765.91 |

– $\epsilon$ 값이 커질수록 베이지안 추정량이 표본평균 및 분산과 가까워지는 것을 확인할 수 있으며, 이에 따라 $\epsilon$ 값이 커질수록 원자료의 분포와 유사해지는 것을 알 수 있다.

## Ⅵ. 재현자료 유용성 평가

앞 절에서 생성한 재현자료의 유용성을 평가하고 그 결과를 비교하기 위해 평균, 사분위수 등의 통계량과 95% 신뢰구간 중첩성(interval overlap) 측도값을 비교하였다.

– 신뢰구간 중첩성 측도값은 원자료와 재현자료의 개별 추정량 각각에 대해 신뢰구간 공식을 이용하여 95% 신뢰구간을 추정하고, 두 신뢰구간의 교집합을 계산함으로써 얻어진다.

$$I = \frac{U_i - L_i}{2(U_o - L_o)} + \frac{U_i - L_i}{2(U_s - L_s)} \qquad [14]$$

– 여기서, $(L_o, U_o)$는 원자료의 95% 신뢰구간, $(L_s, U_s)$는 재현자료의 95% 신뢰구간을 의미한다.

[표 5] 차분 정보보호 수준별 통계량 비교

(단위: 백만 원)

| 구분 / 항목명 | 통계량 | 원자료 | 재현자료 | | |
|---|---|---|---|---|---|
| | | | $\epsilon = 0.1$ | $\epsilon = 1$ | $\epsilon = 10$ |
| 당기순이익(손실) | 평균 | 4,356.87 | 5,170.54 | 4,114.67 | 4,326.07 |
| | 분산 | 164,224,210.03 | 157,881,065.27 | 160,043,386.76 | 162,677,248.71 |
| | 최솟값 | −31,806.73 | −37,530.18 | −30,685.08 | −44,206.40 |
| | 제1사분위수 | −2,348.38 | −2,591.03 | −4,548.15 | −4,587.47 |
| | 중앙값 | 3,314.23 | 5,128.45 | 4,013.85 | 4,484.55 |
| | 제3사분위수 | 9,986.03 | 13,581.50 | 12,254.87 | 12,854.90 |
| | 최댓값 | 47,003.02 | 44,778.63 | 38,944.49 | 49,409.25 |
| 재무활동으로 인한 현금흐름 | 평균 | 4,051.75 | 4,928.10 | 3,776.56 | 4,017.61 |
| | 분산 | 202,924,845.66 | 196,589,917.79 | 198,046,183.30 | 200,994,661.01 |
| | 최솟값 | −36,258.99 | −42,720.57 | −34,934.98 | −49,928.64 |
| | 제1사분위수 | −3,202.39 | −3,732.84 | −5,860.04 | −5,890.23 |
| | 중앙값 | −16.67 | 4,881.13 | 3,664.41 | 4,193.77 |
| | 제3사분위수 | 11,195.36 | 14,313.68 | 12,831.79 | 13,497.82 |
| | 최댓값 | 46,797.58 | 49,125.78 | 42,521.55 | 54,129.79 |

– $\epsilon$ 값이 커질수록 재현자료의 평균 및 분산이 원자료와 유사해지고, 이에 따라 재현자료의 유용성이 높아지는 것을 확인할 수 있다. 이는 $\epsilon$의 값이 커질수록 $\boldsymbol{s}$에 대한 교란이 줄어들어서 나온 결과로 보인다.

– 또한 $\epsilon$의 값이 커질수록 신뢰구간의 중첩성이 1에 가까워지고, 이에 따라 재현자료의 유용성이 높아지는 것을 확인할 수 있다.

[표 6] 차분 정보보호 수준별 신뢰구간 중첩성 비교

(단위: 백만 원)

| 구분 / 항목명 | 통계량 | 원자료 | 재현자료 | | |
|---|---|---|---|---|---|
| | | | $\epsilon = 0.1$ | $\epsilon = 1$ | $\epsilon = 10$ |
| 당기순이익(손실) | 신뢰구간(하한) | 3,702.6 | 4,390.8 | 3,329.6 | 3,534.6 |
| | 신뢰구간(상한) | 5,011.2 | 5,950.3 | 4,899.7 | 5,117.6 |
| | 중첩성($I$) | 1 | 0.44 | 0.84 | 0.91 |
| 재무활동으로 인한 현금흐름 | 신뢰구간(하한) | 3,324.4 | 4,058.0 | 2,903.3 | 3,137.8 |
| | 신뢰구간(상한) | 4,779.1 | 5,798.2 | 4,649.8 | 4,897.4 |
| | 중첩성($I$) | 1 | 0.46 | 0.84 | 0.91 |

## Ⅶ. 결론

원자료의 특성이 반영된 차분 정보보호를 적용한 재현자료의 생성은 데이터 프라이버시를 보호하는 동시에 자료 분석 및 통계 작성 등 데이터 유용성을 확보하는 방법이다. 본 연구에서 베이지안 모형을 기반으로 차분 정보보호를 적용한 재현자료는 $\epsilon$값이 커질수록 데이터 유용성이 커짐을 알 수 있었다.

이 연구의 결과로 증명된 사실은 데이터 유용성이 높아지면 $\epsilon$값이 커져야 한다는 것이다. 그러나 $\epsilon$값이 커지면 데이터 유용성은 높아지지만 반대로 원자료와 비슷해져 안전성(노출위험)에 대한 문제가 발생할 수 있다. 또한 다변량 재현자료 생성 방법, 시계열 재현자료 생성 방법, 프라이버시와 유용성 관계(trade-off)에서 적정 $\epsilon$값 기준, hyper-parameter 지정 등이 정해져 있지 않다는 한계가 있다.

본 연구의 결과를 토대로 한 후속연구에서는 차분 정보보호를 적용한 재현자료를 생성할 때 데이터 유용성과 안전성을 동시에 측정함으로써 최적의 $\epsilon$값을 선정하는 데 기여할 수 있기를 바란다.

## 참고문헌


정강수, 박석. (2018). 「차분 프라이버시 기반 비식별화 기술에 대한 연구」. 「정보보호학회지」, 28(2), 61–77.

통계교육원(2022). 「2022 통계데이터 비밀보호의 이해 과정 교육」. 교육자료집.

Bowen, C. M., & Liu, F. (2016). "Comparative Study of Differentially Private Data Synthesis Methods." *Statistical Science 35(2)*, 280–307.

Brubaker, M., & Prince, S. (2021). "Tutorial #12: Differential Privacy I: Introduction." *Borealis AI.*

Dwork, C. (2006). "Differential Privacy." *International Colloquium on Automata, Languages, and Programming* (pp. 1–12). Springer.

Dwork, C. (2008). "Differential Privacy: A Survey of Results." *International Conference on Theory and Applications of Models of Computation* (pp. 1–19). Springer.

Liu, F. (2016). "Model–based Differentially Private Data Synthesis and Statistical Inference in Multiply Synthetic Differentially Private Data." *Transactions on Data Privacy, 15(3),* 141–175.


베 이 지 안 차 분 정 보 보 호 적 용

재 현 자 료 생 성 방 법 론 소 개