한국산업보건학회지, 제32권 제1호(2022) ISSN 2384-132X(Print) ISSN 2289-0564(Online)
Journal of Korean Society of Occupational and Environmental Hygiene, 2022: 32(1): 21-30
https://doi.org/10.15269/JKSOEH.2022.32.1.21

Original Article

# 불검출 자료를 포함한 작업환경측정 자료의 분석 방법 비교

박주현 · 최상준[1*] · 고동희[2] · 박동욱[3] · 성예지[1]

동국대학교 통계학과, [1]가톨릭대학교 보건의료경영대학원, [2]가톨릭관동대학교, [3]한국방송통신대학교

# A Comparison of Analysis Methods for Work Environment Measurement Databases Including Left-censored Data

Ju-Hyun Park · Sangjun Choi[1*] · Dong-Hee Koh[2] · Donguk Park[3] · Yeji Sung[1]

*Department of Statistics, Dongguk University*
*[1]Graduate School of Public Health and Healthcare Management, The Catholic University of Korea*
*[2]Department of Occupational and Environmental Medicine, International St.Mary's Hospital, Catholic Kwandong University*
*[3]Department of Environmental Health, Korea National Open University*

## ABSTRACT

**Objectives:** The purpose of this study is to suggest an optimal method by comparing the analysis methods of work environment measurement datasets including left-censored data where one or more measurements are below the limit of detection (LOD).

**Methods:** A computer program was used to generate left-censored datasets for various combinations of censoring rate (1% to 90%) and sample size (30 to 300). For the analysis of the censored data, the simple substitution method (LOD/2), $\beta$-substitution method, maximum likelihood estimation (MLE) method, Bayesian method, and regression on order statistics (ROS)were all compared. Each method was used to estimate four parameters of the log-normal distribution: (1) geometric mean (GM), (2) geometric standard deviation (GSD), (3) 95th percentile (X95), and (4) arithmetic mean (AM) for the censored dataset. The performance of each method was evaluated using relative bias and relative root mean squared error (rMSE).

**Results:** In the case of the largest sample size (n=300), when the censoring rate was less than 40%, the relative bias and rMSE were small for all five methods. When the censoring rate was large (70%, 90%), the simple substitution method was inappropriate because the relative bias was the largest, regardless of the sample size. When the sample size was small and the censoring rate was large, the Bayesian method, the $\beta$-substitution method, and the MLE method showed the smallest relative bias.

**Conclusions:** The accuracy and precision of all methods tended to increase as the sample size was larger and the censoring rate was smaller. The simple substitution method was inappropriate when the censoring rate was high, and the $\beta$-substitution method, MLE method, and Bayesian method can be widely applied.

**Key words:** Left-censored data, limit of detection, maximum likelihood estimation, $\beta$-substitution

*Corresponding author: Sangjun Choi, Tel:02-2258-7379, E-mail: junilane@gmail.com
Graduate School of Public Health and Healthcare Management, The Catholic University of Korea, 222 Banpo-daero, Seocho-gu, Seoul, 06591,Korea
Received: February 2, 2022, Revised: March 3, 2022, Accepted: March 17, 2022

Ju-Hyun Park https://orcid.org/0000-0001-9675-6475
Sangjun Choi https://orcid.org/0000-0001-8787-7216
Donghee Koh https://orcid.org/0000-0002-2868-4411
Donguk Park https://orcid.org/0000-0003-3847-7392
Yeji Sung https://orcid.org/0000-0003-2267-5490

## I. 서 론

직업적 노출 평가를 위해 공기 중 유해인자 농도를 측정할 때 실제 농도가 측정 및 분석 기기가 검출 할 수 있는 검출한계(Limit Of Detection, LOD) 미만 수준인 경우 불검출(non-detected) 자료가 생성된다. 불검출 자료는 전체 자료 분포에서 왼쪽 영역(작은 값) 자료가 측정이 안 된 경우이기 때문에 왼쪽 검열(left-censored) 자료라고 하는데, 이런 왼쪽 검열 자료를 '0' 또는 검출한계 값으로 처리하여 분포의 특성값(예, 평균)을 추정할 경우 편향성(bias)이 발생할 수 있다(Mulhausen and Damiano, 1998). 이러한 편향성을 줄이기 위해 LOD/2나 LOD/$\sqrt{2}$로 검열자료들을 대체하여 분석하는 단순 대체법(simple substitution), 최대우도추정법(Maximum Likelihood Estimation method, MLE), 순위통계량 회귀분석(Regression on Order Statistics, ROS), 베이지안 분석법(Bayesian method) 및 베타 대체법($\beta$-substitution) 등을 이용한 검열 자료 처리 방법이 제안되고 평가되었다(Finkelstein & Verma, 2002; Ganser & Hewett, 2010; Glass & Gray, 2001; Hewett & Ganser, 2007; Hornung & Reed, 1990; Huynh et al., 2014; Huynh et al., 2016; Perkins et al., 1990).

국내 작업환경측정 자료는 사업주가 주기적으로 작업환경 중 유해인자의 노출수준이 노출기준 미만으로 유지하는지 확인하고 관리하도록 하는 규제 적합성(compliance based)에 기초한 노출평가에 의해 생성된다(Byeon et al., 2009; Choi, 2008). 이렇게 전국 단위로 측정된 자료는 2002년부터 안전보건공단에 전산 시스템(K2B, https://k2b.kosha. or.kr/)을 통해 작업환경측정 데이터베이스(Work Environment Measurement Data, WEMD)로 축적되어 오고 있어 국가 노출감시체계로 활용 가치가 크며(Choi et al., 2019), 석면(Choi et al., 2017), 벤젠(Koh et al., 2015), 납(Koh et al., 2017; Koh et al., 2018; Koh et al., 2021)에 대한 직무노출 매트릭스(Job-Exposure Matrix, JEM) 구축에 활용되었다. 그러나 노출기준의 강화, 사업주의 규제 순응도 향상 등 복합적 요인에 의해 측정자료 중 불검출 자료(왼쪽 검열 자료)의 비율이 높아 측정 자료를 활용한 JEM 구축 시 불검출 자료의 처리는 해결해야 할 중요한 문제이다.

본 연구에서는 작업환경 측정 자료의 분포 특성, 검열 정도(censoring rate) 및 표본 크기를 고려하여 모의실험(simulation)을 통해 검열된 자료를 생성 후 단순 대체법(LOD/2), MLE, ROS, 베타 대체법, 베이지안 방법에 의해 추정한 산술평균(Arithmetic Mean, AM), 기하평균(Geometric Mean, GM), 기하표준편차(Geometric Standard Deviation, GSD), 그리고 95백분위수(95$^{th}$ percentile, X95)를 참값과 비교함으로써 최적의 검열 자료 처리 방법을 제안하고자 한다.

## II. 대상 및 방법

### 1. 연구 설계

검열자료에 대한 분석 결과에 대한 정확도는 자료의 분포, 검열 정도, 표본 크기에 따라 결정된다. 우리는 선행연구(Koh et al., 2021)를 통해 납의 2015년, 2016년 WEMD를 이용하여 자료 분포를 분석한 결과 대수정규분포(log-normal distribution)에 가깝다는 것을 확인하였고, 납 노출 자료가 많은 '일차전지 및 축전지 제조업'의 노출 프로 파일에 기초하여 Fig. 1과 같이 모의실험의 모집단에서 납의 노출은 모수들의 참값이 각각 AM=6.98, GM=3.76, GSD=3.04, X95=34.41인 대수 정규분포를 따르고, 단일 LOD만 존재하는 상황을 가정하였다.

표본의 크기와 검열 정도가 추정 결과에 어떤 영향을 미치는지 확인하기 위해 모의실험에서는 다양한 표본 크기(30, 60, 100, 300)와 4개(낮음, 10%; 보통, 40%; 높음, 70%; 매우 높음, 90%)의 대표적 검열율(이하 불검출율)을 조합하여 총 16개의 시나리오를 고려하였다.

### 2. 검열 자료 분석

검열자료 분석방법은 다음과 같이 5가지 방법을 이용하였다.

- 단순 대체법(LOD/2 substitution)
- 최대우도 추정방법(maximum likelihood estimation methods)
- 순위 통계량 회귀분석(regression on order statistics, ROS)
- 베타 대체법($\beta$-substitution)
- 베이지안 분석법(Bayesian method)

#### (1) 단순 대체법

대체방법은 LOD 값이 왼쪽 검열되었다는 사실을 기반으로 실제 값은 LOD 값보다 더 작을 것이라 예상해서 특정 값으로 대체한다. 가장 많이 사용되는 대체 값

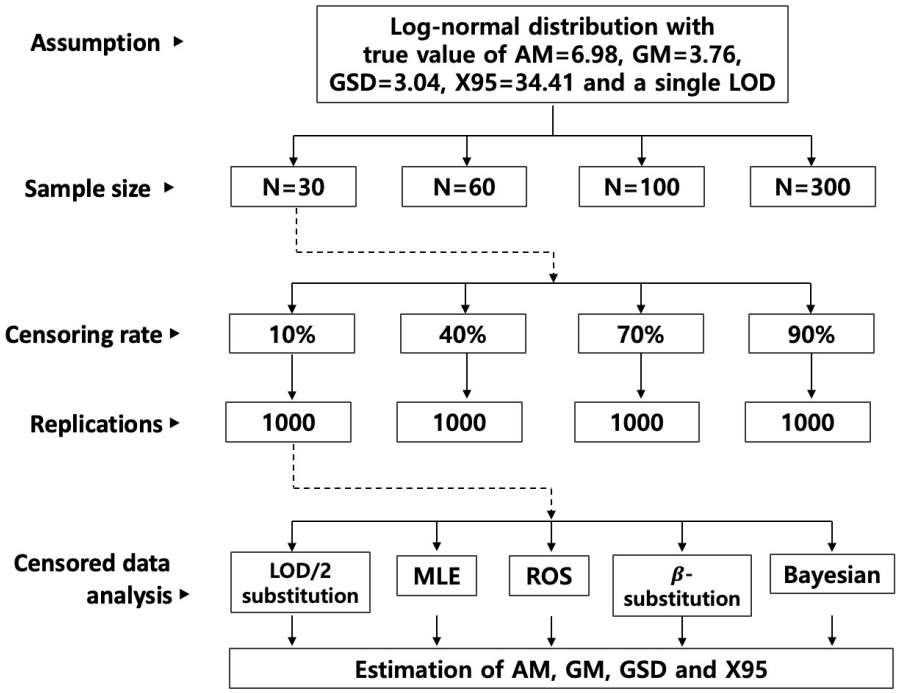


Fig. 1. A graphical diagram on the simulation study design. The dashed line means that the rest of the items proceed the same (AM: arithmetic mean, GM: geometric mean, GSD: geometric standard deviation, X95: 95th percentile, LOD: limit of detection, MLE: maximum likelihood estimation, ROS: regression on order statistics).

으로는 LOD, LOD/2, LOD/$\sqrt{2}$가 있다. 일반적으로 보수적(conservative)인 방법으로 알려져 있으며, 평균과 분산에 각각 양과 음의 편향성을 보인다. Hornung & Reed (1990)은 GSD가 3보다 작은 경우는 LOD/$\sqrt{2}$를, GSD≥3 또는 LOD이하의 비율이 전체의 50% 미만인 경우에는 LOD/2를 사용할 것을 추천하였다. 본 연구에서는 모집단의 GSD가 3 이상으로 가정해서 LOD/2를 이용하여 대체하는 방법을 사용하였다. AM, GM, GSD 및 X95는 기존 값과 대체된 값들을 모두 이용해서 표본 추정방법으로 구한다.

(2) MLE

MLE는 자료가 대수 정규분포를 따를 때 가장 좋은 방법으로 알려져 있다. n개의 자료($x_1,...,x_n$) 중 처음 k개가 LOD인 경우, 우도함수(Likelihood Function, LF)는 다음 식(1)과 같이 표현된다.

$$LF(\mu,\sigma)=\prod_{i=k+1}^{n}\frac{1}{\sigma}\phi\left(\frac{\ln(x_i)-\mu}{\sigma}\right)\times\prod_{j=1}^{k}\Phi\left(\frac{\ln(x_i)-\mu}{\sigma}\right) \quad (1)$$

여기서 $\phi(\cdot)$와 $\Phi(\cdot)$는 각각 표준정규분포의 확률밀도함수(Probability density function)와 누적분포함수(Cumulative distribution function)가 정보적 사전분포이다. 이 우도함수를 최대로 하는 $\hat{\mu}$와 $\hat{\sigma}$를 최대우도추정법(Maximum Likelihood Estimation)으로 구한 후, GM과 GSD는 다음 식(2)와 식(3)과 같이 표현할 수 있다.

$$GM=\exp(\hat{\mu}) \quad (2), \qquad GSD=\exp(\hat{\sigma}) \quad (3)$$

(3) ROS

ROS 방법은 우선 LOD를 포함한 모든 자료를 작은 것부터 큰 순서로 재배열한 후, 대수정규분포의 분위수에 대한 회귀분석을 실시해서 평균과 표준편차를 각각 y절편과 기울기로 추정하는 방법이다. $x_i$를 순서대로 정렬한 자료라고 할 때, 이를 모형으로 표현하면 다음 식(4)와 같다.

$$y_i=\hat{\mu}+\hat{\sigma}\cdot\Phi^{-1}(p_i) \quad (4),$$

여기서 $y_i=\ln(x_i)$이고 $\Phi^{-1}(\cdot)$은 표준정규분포의 누적분포함수의 역함수이다. 일반적으로 Blom의 공식

(Blom's formula)를 이용하여 $p_i = \frac{(i-3/8)}{(n+1/4)}$로 지정하고 위 식에서 구한 $\hat{\mu}$와 $\hat{\sigma}$를 이용하여 MLE 방법과 동일하게 GM과 GSD를 구할 수 있다.

(4) 베타 대체법($\beta$-substitution)

LOD에 일정 상수값(예, $1/2$ 또는 $1/\sqrt{2}$)을 곱해서 보정을 하는 단순 대체법을 발전시켜 자료가 대수정규분포를 따른다는 가정 하에서 적률법(Method of moments)을 적용시켜 추정하고자 하는 모수에 따라 다른 상수값($\beta$)을 곱해 보정을 방법이다(Ganser and Hewett, 2010). k개의 LOD들과 n-k개의 관측된 값들이 있다고 할 때, AM을 구할 때 사용하는 상수값 $\beta_{AM}$은 다음의 식(5)로 구할 수 있다;

$$\beta_{AM} = \frac{n}{k} \cdot \Phi(z-s) \cdot \exp(-s \cdot z + s^2/2) \quad (5),$$

여기서 $z = \Phi(k/n)$, $s = \frac{\tilde{y} - \log(LOD)}{f(z) - z}$, $f(z) = \phi(z)/(1-\Phi(z))$이고, 평균 $\tilde{y} = \sum_{i=1}^{n-k} y_i/(n-k)$는 관측된 값들로만 계산한다. k개의 LOD 값들을 $\beta_{AM} \cdot LOD$으로 대체한 후 표본 AM을 계산한다. GM을 계산할 때는 또 다른 상수값 $\beta_{GM}$을 다음 식(6)과 같이 계산한다;

$$\beta_{GM} = \exp\left[\frac{-(n-k) \cdot n}{k} \cdot \log(g(z,s)) - s \cdot z - \frac{n-k}{2kn} \cdot s^2\right] \quad (6),$$

여기서 $g(z,s) = \frac{1-\Phi(z-s/n)}{1-\Phi(z)}$이고, AM을 구할 때와 동일하게 k개의 LOD 값들을 모두 $\beta_{GM} \cdot LOD$으로 대체한 후 표본 GM을 계산한다. GSD와 X95는 다음의 식(7)과 (8)을 이용해서 계산한다;

$$GSD = \exp(s_y) \text{ with } s_y = \sqrt{\frac{2n}{n-1} \cdot \log(AM/GM)} \quad (7),$$

$$X_{95} = \exp\left[\log(GM) - s_y^2/(2n) + 1.645 \cdot s_y\right] \quad (8)$$

(5) 베이지안 분석법

베이지안 방법은 앞서 소개된 MLE 방법과 비교할 때 접근 방법과 결과의 해석이 가장 극명하게 다르다. MLE 방법은 빈도론적 관점으로 분석하고자 하는 자료에서만 추정하고자 하는 모수의 정보를 이끌어 내는 데 반해, 베이지안 방법은 분석하고자 하는 자료 뿐 아니라 전문가의 의견 및 판단을 사전 분포(prior distribution)를 통해 분석에 반영할 수 있다. MLE 방법과 같이 자료가 대수정규분포를 따른다는 가정하에서 베이지안 추론은 사후 분포(posterior distribution)를 기반으로 이뤄지는데 사후 분포는 다음 식(9)와 같이 표현할 수 있다;

$$\pi(\mu,\tau \mid x_1,\cdots,x_n) = \frac{LF(\mu,\tau)\pi(\mu,\tau)}{\int\int LF(\mu,\tau)\pi(\mu,\tau)d\mu d\tau} \quad (9),$$

여기서 정밀도 $\tau = 1/\sigma^2$이며, $\pi(\mu,\tau)$는 $\mu$와 $\tau$의 사전 분포이고 본 연구에서는 공액사전분포(conjugate prior distribution)를 지정하여 $\mu$와 $\tau$에 정규-감마분포(Normal-Gamma distribution)을 가정하였다. 이때 초모수들(hyperparameters)로서 $\tau$의 사전분포 평균과 분산은 모두 1, $\mu$의 사전 분포 평균과 분산은 각각 0과 100이 되도록 설정하였다. 모수가 여러 개이거나 LOD와 같이 왼쪽 검열이 있는 경우, 위 사후 분포의 정확한 형태가 알려져 있지 않고, 베이지안 추론은 사후 분포에서 마코브 연쇄 몬테칼로 방법(Markov chain Monte Carlo method)을 통해 얻은 $\mu$와 $\tau$의 사후 표본(posterior sample)을 기반으로 이뤄진다.

지금까지 소개한 5개의 방법 중에서 베타 대체법을 제외하고 자료가 대수정규분포를 따른다고 가정하는 MLE, ROS, 베이지안 방법은 GM과 GSD를 먼저 추정하고 그 값들을 이용해서 AM 및 X95를 추정하는데 일반적으로 알려져 있는 변환식 $AM = \exp\left(\hat{\mu} + \hat{\sigma^2}/2\right)$을 이용하면 AM에서 편향된 결과를 얻는 것이 알려져 있다(Cohn et al., 1989). 따라서, 이러한 편향성을 보정하기 위해 다음의 최소분산불편추정량을 구하는 식(10)을 이용하여 AM을 계산하였다.

$$AM = \exp(\hat{\mu}) \times g_{n-1}\left(\frac{\hat{\sigma^2}}{2}\right) \quad (10),$$

여기서 $g_{n-1}\left(\frac{\hat{\sigma^2}}{2}\right) \approx \sum_{j=0}^{4} \frac{(n-1)^i(n-1+2j)}{(n-1)(n-1+2)\cdots(n-1+2j)} \left(\frac{(n-1)\times\hat{\sigma^2}/2}{n}\right)^j \frac{1}{j!}$이다. 모의실험과 연산은 통계 프로그램 R (R Core Team, 2021)을 이용하였고, MLE와 ROS는 R package인 EnvStats(Millard, 2013)의 함

수를, 베타대체법과 베이지안 방법은 R 함수들을 만들어서 적용하였다.

### 3. 평가 지표

검열자료 분석 결과의 정확도와 정밀도 수준은 상대편향(relative bias)과 상대 제곱근평균제곱오차(root mean square error, rMSE)를 이용하여 평가하였다.

상대편향은 식(11)과 같이 각 모의실험 시나리오에서 1,000개 반복 자료(replicated data)에 대해 5가지 분석방법 각각에 의해 추정한 통계 추정치(AM, GM, X95)의 평균($\overline{\theta_E}=\sum_{i=1}^{1000}\hat{\theta_i}/1000$)과 참값($\theta$)의 차이가 참값($\theta$)에 대해 차지하는 비율(%)로 계산되며 추정값이 참값에 대한 상대적인 양(+) 혹은 음(-)의 편향(bias) 크기를 나타낸다(Huynh et al., 2014).

$$\text{Relative bias, \%} = \left(\frac{\overline{\theta_E}-\theta}{\theta}\right)\times 100 \qquad (11)$$

상대 제곱근 평균 제곱 오차 (이하 rMSE)는 식(12)와 같이 편향(bias)과 정밀도(표준편차)의 조합이 참값에 대해 차지하는 상대 비율(%)로 계산되며, 정확도와 정밀도가 조합된 개념으로 양(+)의 값만 갖게 된다((Huynh et al., 2014).

$$\text{Relative rMSE, \%} = \frac{1}{\theta}\sqrt{(\overline{\theta_E}-\theta)^2+\frac{\sum_{i=1}^{1000}(\hat{\theta_i}-\overline{\theta_E})^2}{1000-1}}\times 100 \qquad (12)$$

검열자료에 대한 각 분석방법에 따라 추정한 통계량의 상대편향과 rMSE가 작을수록 추정값이 정확하다고 평가 할 수 있다.

## Ⅲ. 연구결과

검열 자료 분석방법에 따른 AM, GM, GSD, X95 추정치의 상대편향과 rMSE의 비교 결과는 각각 Fig. 2와 Fig. 3과 같으며, 참값($\theta$) 대비가 아닌 실제 편향과 rMSE 결과는 Supplementary Table 1에서 확인 할

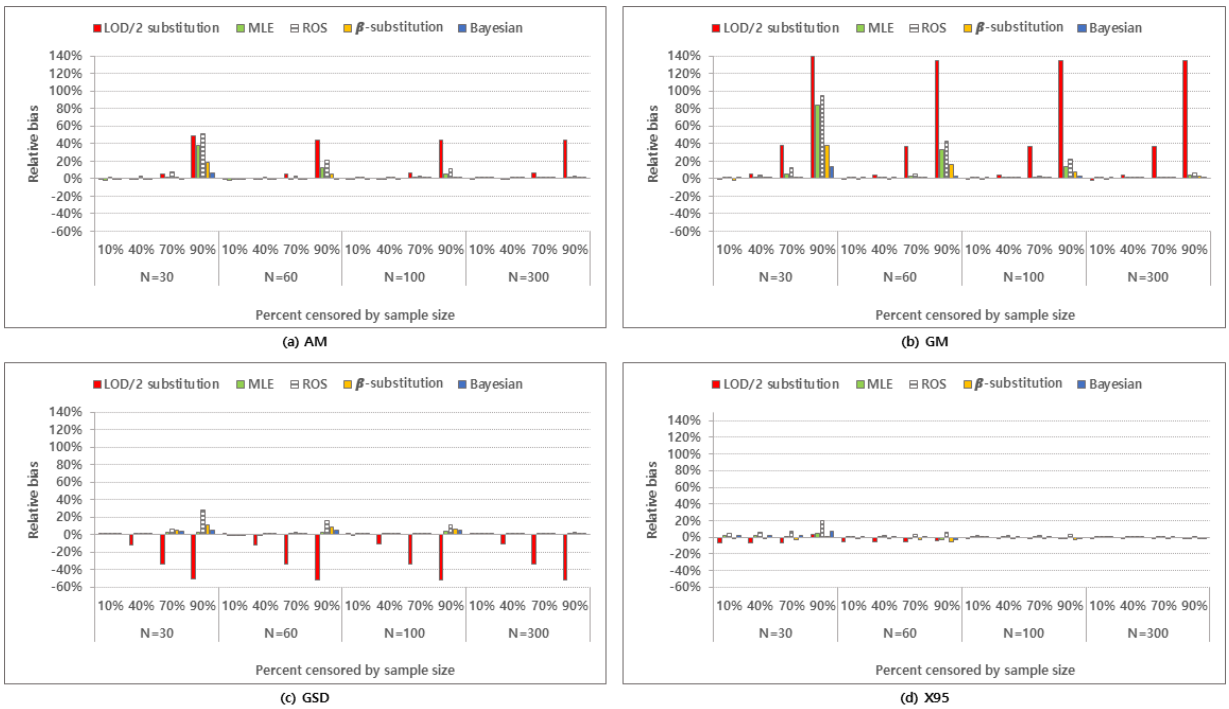


Fig. 2. Relative bias of AM (a), GM (b), GSD (c), and X95 (d) estimated by five censoring data analysis methods for datasets generated by a combination of censoring rate and sample size (AM: arithmetic mean, GM: geometric mean, GSD: geometric standard deviation, X95: 95$^{th}$ percentile, LOD: limit of detection, MLE: maximum likelihood estimation, ROS: regression on order statistics).

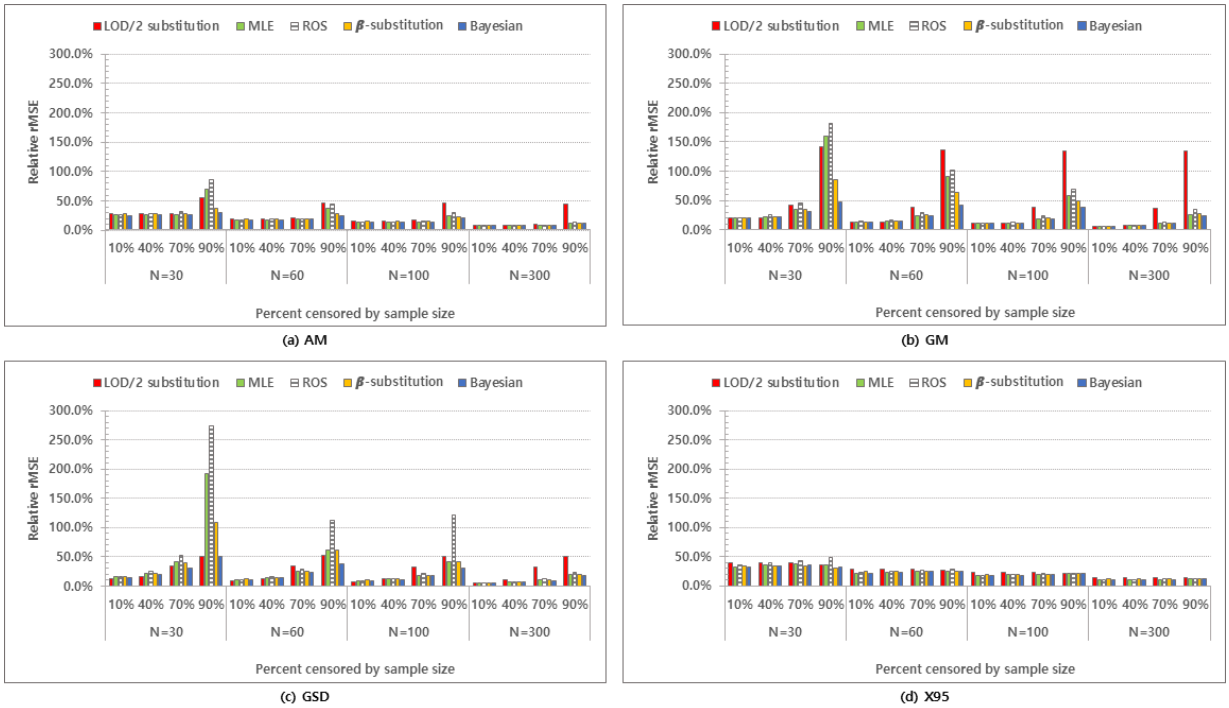


Fig. 3. Relative rMSE of AM (a), GM (b), GSD (c), and X95 (d) estimated by five censoring data analysis methods for datasets generated by a combination of censoring rate and sample size (AM: arithmetic mean, GM: geometric mean, GSD: geometric standard deviation, X95: 95$^{th}$ percentile, LOD: limit of detection, MLE: maximum likelihood estimation, ROS: regression on order statistics, rMSE: root mean square error).

수 있다.

상대편향은 5가지 분석방법 모두 표본 크기가 작고, 불검출률이 클수록 증가하는 경향을 보였다(Fig. 2). 그러나 단순 대체법(LOD/2)으로 추정한 AM, GM, GSD의 상대편향은 불검출률이 큰(70%, 90%) 경우 표본 크기에 상관없이 다른 분석법과 비교하여 높은 수준으로 유지되었다(Fig. 2(a)~2(c)). AM, GM, X95 평균 추정값의 상대편향은 분석방법과 관계없이 GM이 가장 크고, X95가 가장 작았다. rMSE의 경우도 상대편향과 비슷한 결과를 보였으나, 표본의 크기가 작으면서(n=30, 60) 불검출율이 큰(70%, 90%) 경우 ROS가 AM과 GSD의 추정에 있어서 단순 대체법과 비슷하거나 더 높은 rMSE를 보였다(Fig. 3).

Table 1은 표본크기와 불검출율의 조합에 따라 고려한 16개의 모의 실험 시나리오에서 5가지 방법에 의해 분석한 결과 각 모수추정치에 대한 상대편향은 〈±5%, rMSE는 〈20%으로 추정값의 정확도가 높은 분석법들을 요약하였다. Table 1에서 회색 음영 배경에 굵은 글씨로 표현된 분석법은 상대편향은 〈±1%, rMSE는 〈15%의 가장 정확한 추정방법을 나타낸다. 표본크기가 가장 큰(n=300) 경우 불검출율이 40% 이하에서는 5가지 방법 모두 상대편향과 rMSE가 작게 나타났다. 그러나, 불검출율이 큰 경우(70%, 90%)는 표본 크기가 커도 단순 대체법으로 AM, GM, GSD를 추정하는 경우 상대편향이 가장 크게 나타나서 부적절했다. 표본 크기가 작고, 불검출율이 큰 경우엔 비교적 베이지안 방법과 베타 대체법, MLE 방법이 가장 작은 상대편향을 보였다. rMSE의 경우 15% 미만으로 가장 정확도와 정밀도가 크게 나타난 경우는 표본 크기가 크고(N=100, 300) 불검출율이 70% 이하인 시나리오에서의 AM, GM 추정치에 대한 베이지안 분석법과 MLE 방법이었다.

## Ⅳ. 고 찰

과거 노출과 질병과의 관계를 평가하는데 있어 작업환경측정과 같이 정량적인 노출 평가 자료는 가장 정확한 노출 특성을 확인하는데 활용할 수 있다(Chung et al., 2015). 그러나 불검출(LOD 미만) 자료가 있는 경우 불검출 자료를 무시하고 분석할 경우 실제 분포 특성을 과대평가할 우려가 있기 때문에 다양한 검열자료

**Table 1.** Censoring data analysis methods with good performance by sample size and the degree of censoring

| Parameter | Percent censored | Sample size | | | |
|---|---|---|---|---|---|
| | | 30 | 60 | 100 | 300 |
| *Relative bias ⟨ ±5%* | | | | | |
| AM | 10 | **S**/M/**R/BS**/Ba | **S**/M/**R/BS**/Ba | **S/M/R/BS/Ba** | **S/M/R/BS/Ba** |
| | 40 | **S**/M/R/**BS/Ba** | S/M/**R/BS/Ba** | **S/M**/R/**BS/Ba** | **S/M/R/BS/Ba** |
| | 70 | **M/BS/Ba** | **M**/R/**BS/Ba** | **M**/R/**BS/Ba** | **M/R/BS/Ba** |
| | 90 | | BS/**Ba** | BS/**Ba** | M/R/**BS/Ba** |
| GM | 10 | **S/M**/R/BS/**Ba** | S/**M**/R/**BS/Ba** | S/**M/R/BS/Ba** | S/**M/R/BS/Ba** |
| | 40 | M/R/**BS**/Ba | S/**M**/R/**BS/Ba** | S/**M/R/BS/Ba** | S/**M/R/BS/Ba** |
| | 70 | BS/Ba | M/BS/Ba | M/R/**BS/Ba** | **M**/R/**BS/Ba** |
| | 90 | | Ba | Ba | M/BS/Ba |
| GSD | 10 | S/**M/R**/BS/**Ba** | S/**M/R/BS/Ba** | S/**M/R/BS/Ba** | S/**M/R/BS/Ba** |
| | 40 | **M**/R/BS/Ba | **M/R/BS/Ba** | **M/R/BS/Ba** | **M/R/BS/Ba** |
| | 70 | M/BS/Ba | **M**/R/BS/Ba | **M**/R/BS/Ba | **M/R/BS/Ba** |
| | 90 | M | M | M/Ba | M/R/BS/Ba |
| X95 | 10 | M/R/**BS**/Ba | **M**/R/**BS/Ba** | S/**M**/R/**BS/Ba** | **S/M/R/BS/Ba** |
| | 40 | M/**BS**/Ba | S/**M**/R/BS/**Ba** | S/**M**/R/**BS/Ba** | **S/M/R/BS/Ba** |
| | 70 | **M**/BS/Ba | S/**M**/R/BS/**Ba** | S/**M**/R/**BS/Ba** | **S/M**/R/**BS/Ba** |
| | 90 | S/M/BS | S/M/Ba | S/M/R/BS/Ba | **S/M**/R/**BS/Ba** |
| *Relative rMSE ⟨ 20%* | | | | | |
| AM | 10 | | M/R/Ba | S/**M/R**/BS/**Ba** | **S/M/R/BS/Ba** |
| | 40 | | M/R/BS/Ba | S/**M/R**/BS/**Ba** | **S/M/R/BS/Ba** |
| | 70 | | M/R/BS/Ba | S/**M**/R/BS/**Ba** | **S/M/R/BS/Ba** |
| | 90 | | | | **M/R/BS/Ba** |
| GM | 10 | | **S/M/R/BS/Ba** | **S/M/R/BS/Ba** | **S/M/R/BS/Ba** |
| | 40 | | **S**/M/R/BS/Ba | **S/M/R/BS/Ba** | **S/M/R/BS/Ba** |
| | 70 | | | M/Ba | **M/R/BS/Ba** |
| | 90 | | | | |
| GSD | 10 | **S**/M/R/BS/**Ba** | **S/M/R/BS/Ba** | **S/M/R/BS/Ba** | **S/M/R/BS/Ba** |
| | 40 | S/Ba | **S/M**/R/**BS/Ba** | **S/M/R/BS/Ba** | **S/M/R/BS/Ba** |
| | 70 | | | M/BS/Ba | **M/R/BS/Ba** |
| | 90 | | | | M/BS/Ba |
| X95 | 10 | | | M/R/BS/Ba | **S/M/R/BS/Ba** |
| | 40 | | | M/Ba | **S/M/R/BS/Ba** |
| | 70 | | | Ba | **S/M/R/BS/Ba** |
| | 90 | | | | **S/M/R/BS/Ba** |

AM: arithmetic mean, GM: geometric mean, GSD: geometric standard deviation, X95: 95$^{th}$ percentile, rMSE: root mean square error, S: simple substitution, M: maximum likelihood estimator, R: regression on order statistics, BS: beta-substitution, Ba: Bayesian estimator, The bold text on a grey background means ⟨±1% for relative bias or ⟨15% for relative rMSE.

분석 방법이 제안되고 평가되어왔다. 본 연구에서는 Hornung & Reed(1990)이 제안한 이후 간단하므로 불검출 자료 처리에 많이 활용되어 왔던 LOD/2 대체법과 Ganser & Hewett(2010)이 개발한 베타 대체법 및 MLE, ROS, 베이지안 분석법 등 총 5가지 분석법을 비교하였다.

2015년과 2016년 WEMD의 납 자료를 분석한 선행연구 결과를 살펴보면 64개의 소분류 산업별 납 자료 수의 중위수는 75개로 산업별로 자료의 수는 적으면서 불검출율의 중위수는 84.6%로 불검출율이 매우 높음을 알 수 있다(Koh et al., 2021). 산업 내 공정 별 자료의 수는 더욱 작아지게 된다. 따라서, 소 표본의 높은 불검출율인 자료들을 분석하는 경우 편향성과 정확도를 모두 고려할 때 베이지안 방법을 사용하는 것이 더욱 정확한 평균(AM 또는 GM) 추정 결과를 생성할 것이다. 본 연구에서 베이지안 방법을 적용하기 위해 사용한 사전 분포는 $\mu$와 $\tau$ $(=1/\sigma^2)$에 대해서는 무정보 사전분포(noninformative prior distribution)이지만 이들의 함수인 GM $(=\exp(\mu))$와 GSD$(=\exp(\sigma))$에 대해서는 정보적 사전분포(informative prior distribution)를 사용한 것과 동일하다. 정보적 사전분포를 이용한 것과 관련해서 Huynh et al. (2016)은 베타 대체법과 베이지안 방법을 비교하였는데 정보적 사전정보를 이용한 베이지안 방법과 베타 대체법이 유사한 정도의 편향성과 정확도를 나타낸다는 결과를 보였고, 본 연구의 모의실험 결과와 일치함을 알 수 있다. 본 연구에서 고려한 그 외 방법들에 대한 분석결과 특징을 요약하면 다음과 같다.

단순 대체법은 LOD미만의 값을 LOD/2로 대체하였기 때문에 모집단의 노출 분포를 LOD/2에서 절단한 것과 유사하고 그 결과 불검출율이 높을수록 GM은 양의 편향성을, GSD는 음의 편향성을 갖는 것을 모의실험에서 확인하였다. 단순 대체법은 다른 방법들과 다르게 자료가 대수정규분포를 따른다고 가정하지 않으므로 AM과 X95를 구할 때 GM과 GSD를 이용하지 않지만, AM은 GM과 비슷하게 분포가 절단되었으므로 불검출율이 높을수록 양의 편향성을 보인다. 단순 대체법에서 X95는 자료를 가장 작은 것부터 큰 순서대로 나열한 후 상위 5% 분위 수에 해당하는 자료값을 X95로 한다. 이와 같은 방법은 표본의 크기가 크고, 불검출율이 95% 미만인 경우에는 거의 편향성이 없는 결과를 나타낸다. Supplementary Table 1에서 n=300인 모의실험에서 불검출율에 상관없이 X95의 상대적 편향성은 −0.6%, rMSE는 13.7%로 소수점 첫째 자리에서 동일하게 나왔다. 하지만, 표본의 크기가 작을 때는 모수를 정확하게 추정할 정보가 부족하고 정확하게 상위 5%의 값이 존재하지 않는 경우가 발생하기 때문에 일정 정도의 편향성이 발생했다(n=30, 60에서 평균적으로 5% 정도의 편향성을 보임).

ROS는 표본의 크기가 큰 경우를 제외하고 불검출율이 높은 모든 모의실험 시나리오에서 AM, GM, GSD의 편향성과 rMSE가 크게 나왔다. 이는 ROS가 대수정규분포의 분위수를 독립변수로, 정렬된(ordered) 관측값을 종속변수로 하는 회귀모형에서 y-절편과 기울기로 ln(GM)과 ln(GSD)를 추정하는 방법인데, LOD 미만의 정렬된 관측값들은 모두 동일하게 LOD 값을 갖기 때문에 y-절편과 기울기의 추정에서 모두 편향성이 나타나고 그 정도는 불검출율이 높아질수록 높아진다. ROS의 단점을 개선해서 대체법(imputation)과 적률법(method of moments)을 조합한 견고한(roubst) ROS가 제안되었다(Helsel, 2012).

MLE는 가장 많이 알려지고 추천되는 방법이나(Hewett & Ganser, 2007; Huynh et al., 2014) 표본의 크기가 작고, 불검출율이 높은 경우, 모든 모수에서 편향성과 rMSE가 베타 대체법이나 베이지안 방법보다 모두 높게 나오는 것이 관측되었다. 이는 MLE가 정보의 양이 적어서 발생할 수도 있지만, 추가로 EnvStats의 함수가 MLE를 계산할 때 수렴문제로 인할 가능성이 있다. 실제, n=30 또는 60이고 불검출율이 90%인 모의실험에서 MLE 수렴문제에 대한 경고가 여러 번 관측되었다. 이러한 프로그램적 문제는 추가적으로 검증이 필요하다.

베타 대체법은 단순 대체법을 발전시킨 방법으로 고려한 모든 모수에서 모의 실험에서 좋은 추정 결과를 보여주었고, 베타 대체법을 제안한 Ganser & Hewett (2010)의 연구결과와 유사했다. 하지만, 단점으로 AM을 제외한 나머지 모수들에 대한 표준 오차(standard error)가 알려져 있지 않기 때문에 구간 추정을 하려면 붓스트랩(bootstrap)과 같은 방법을 적용해야만 한다(Huynh et al., 2016).

검열자료 분석법을 비교한 선행연구들의 경우 자료의 분포, 표본 크기, 불검출율(검열정도)에 따라 분석법 선정을 다르게 할 것을 제안하고 있다(Hewett & Ganser, 2007; Hornung & Reed, 1990; Huynh et al., 2014; Huynh et al., 2016; Tekindal et al., 2017). 본 연구에서도 표본 크기와 불검출율에 따라 어느 하나의 최적화된 방법을 선택하긴 어려웠다(Table 1). 표본 크기가 100개 이상이면서 불검출율이 40% 이하인 경우엔 5가지 방법 모두 적절하게 모수 추정이 가능하다고 판단된다. 표본 크기가 작고 불검출율이 커지게 되면 단순 대체법은 부적절했고, 베타 대체법이나 베이지안 분석법, MLE 방법이 적절하다고 할 수 있다. 그러나

X95 추정의 경우 단순 대체법도 불검출율이 큰(90%) 경우에도 상대편향 〈±5% 수준에서 적용 가능하였다.

본 연구는 몇 가지 제한점이 있다. 첫째, 검열 자료 분석법의 정확도는 자료의 분포 특성에 따라 달라질 수 있는데, 본 연구에서는 단일 LOD의 대수정규분포만 가정하여 모의실험을 하였다. Huynh 등(2014)은 LOD가 여러 개 있는 대수정규분포와 혼합된 대수정규분포(mixed lognormal distribution)를 가정하여 베타 대체법과 MLE, 비모수 추정법(the Kaplan-Meier, K-M)을 비교하였는데, 혼합된 대수정규분포는 두 개의 이질적인 노출 집단이 각각 다른 대수 정규분포를 가지면서 혼합되어 있다고 가정한 경우였다. 본 연구에서는 WEMD의 납 노출 특성의 선행연구 결과에 기초하여 단일 LOD의 대수정규분포만을 가정하였다. WEMD는 국내 약 180여개 측정기관들에 의해 측정 분석된 결과가 모인 자료이기 때문에 각 분석기관의 분석기기 종류에 따라 LOD가 다를 수 있다. 그러나 정확한 각 기관의 LOD를 파악하긴 어렵고, 고용노동부 안전보건공단의 공정시험법에 따라 표준화된 측정 분석방법을 활용한다는 공통된 특징이 있어서 단일 LOD로 가정하였다. 각 측정기관이 납 분석에 있어 보다 감도가 좋은 장비(예, ICP-OES)가 있을 수 있으나, 법정 장비보다 고가이기 때문에 대부분의 기관들은 원자흡광분석기(AAS)를 가지고 있기 때문에 단일 LOD로 가정하여도 큰 무리가 없다고 판단된다. 또한 Huynh 등(2014)도 다수의 LOD가 있는 혼합 대수정규분포를 가정한 결과나 단일 LOD 대수정규분포를 가정한 결과 모두 유사한 결과를 보고하고 있다. 둘째, 표본크기와 불검출율을 각각 4가지 조건으로 고정시켜 놓고 조합된 시나리오별로 모의실험을 하였다. 따라서 최소 표본 크기인 30개보다 더욱 적은 경우에 대해서는 표본크기 30개였을 때의 본 연구결과를 그대로 적용할 수는 없으며, 표본 수가 적어질수록 정확도가 더욱 낮아짐을 유의해야 한다.

## V. 결 론

단일 LOD가 있는 대수정규분포를 가정하고 다양한 표본 크기(30-300)와 불검출율(10%-90%)을 조합한 시나리오에 따라 다섯 가지 불검출 자료 분석법(LOD/2 대체법, 베타 대체법, MLE, ROS, 베이지안)의 모수 추정치(AM, GM, GSD, X95)에 대한 정확도를 비교 평가한 결과 표본수가 작고, 불검출율이 커질수록 각 방법들의 정확도가 낮아지는 특성이 확인되었다. 표본크기가 100개 이상이면서 불검출율이 40% 이하인 경우엔 5가지 방법 모두 적절하게 모수 추정이 가능하였으나, 표본크기가 30개로 작고 불검출율이 70% 이상 커지게 되면 단순 대체법은 부적절했고, 베타 대체법이나 베이지안 분석법, MLE 방법이 적절하다고 할 수 있다. 그러나 X95 추정의 경우 단순 대체법도 불검출율이 큰(90%) 경우에도 상대편향 〈±5% 수준에서는 적용 가능하였다.

## References

Byeon S, Yi K, Yu G, Phee Y. Regulatory compliance for the working environment measurement system in Korea. J Korean Soc Occup Environ Hyg 2009;19(3):233-239

Choi S. Assessment on work environment monitoring program in Korea. J Korean Soc Occup Environ Hyg 2008;18(4):282-292

Choi S, Kang D, Park D, Lee H, Choi B. Developing asbestos job exposure matrix using occupation and industry specific exposure data (1984-2008) in Republic of Korea. Saf Health Work 2017;8(1):105-115 (https://doi:10.1016/j.shaw.2016.09.002)

Chung DA, Yang RR, Verma DK, Luo J. Retrospective exposure assessment for occupational disease of an individual worker using an exposure database and trend analysis. J Occup Environ Hyg 2015;12(12):855-65 (https://doi.org/10.1080/15459624.2015.1072630)

Cohn TA, DeLong LL, Gilroy EJ, Hirsch RM, Wells DK. Estimating constituent loads. Water Resources Research 1989;25(5):937-942 (https://doi.org/10.1029/WR025i005p00937)

Finkelstein MM, Verma DK. Exposure estimation in the presence of nondetectable values: Another look. AIHAJ 2001;62(2):195-198 (https://doi.org/10.1080/15298660108984622)

Ganser GH, Hewett P. An accurate substitution method for analyzing censored data. J Occup Environ Hyg 2010;7(4):233-244 (https://doi.org/10.1080/15459621003609713)

Glass DC, Gray CN. Estimating mean exposures from censored data: Exposure to benzene in the Australian petroleum industry. Ann Occup Hyg 2001;45(4):275-282 (https://doi.org/10.1093/annhyg/45.4.275)

Helsel DR. Statistics for censored environmental data using minitab and R, Second Edition. Hoboken, New Jersey: John Wiley & Sons; 2012

Hewett P, Ganser GH. A comparison of several methods for analyzing censored data. Ann Occup Hyg. 2007;51(7):611–632 (https://doi.org/10.1093/annhyg/mem045)

Hornung RW, Reed LD. Estimation of average concentration in the presence of nondetectable values. Appl Occup Environ Hyg. 1990;5(1):46–51 (https://doi.org/10.1080/1047322X.1990.10389587)

Huynh T, Ramachandran G, Banerjee S, Monteiro J, Stenzel M, Sandler DP, Engel LS, Kwok RK, Blair A, Stewart PA. Comparison of methods for analyzing left-censored occupational exposure data. Ann Occup Hyg 2014;58(9):1126–1142 (https://doi.org/10.1093/annhyg/meu067)

Huynh T, Quick H, Ramachandran G, Banerjee S, Stenzel M, Sandler DP, Engel LS, Kwok RK, Blair A, Stewart PA. A comparison of the $\beta$-substitution Method and a Bayesian method for analyzing left-censored data. Ann Occup Hyg 2016;60(1):56–73 (https://doi.org/10.1093/annhyg/mev049)

Koh DH, Jeon HK, Lee SG, Ryu HW. The relationship between low-level benzene exposure and blood cell counts in Korean workers. Occup Environ Med 2015;72(6):421–427 (https://doi:10.1136/oemed-2014-102227)

Koh DH, Park JH, Lee SG, Kim HC, Choi S, Jung H, Park DU. Combining lead exposure measurements and experts' judgment through a Bayesian framework. Ann Work ExpoHealth 2017;61(9):1054–1075 (https://doi: 10.1093/annweh/wxx072)

Koh DH, Park JH, Lee SG, Kim HC, Choi S, Jung H, Park JO, Park DU. Estimation of lead exposure prevalence in Korean population through combining multiple experts' judgment based on objective data sources. Ann Work Expo Health 2018;62(2):210–220 (https://doi: 10.1093/annweh/wxx106)

Koh DH, Park JH, Lee SG, Kim HC, Jung H, Kim I, Choi S, Park D. Estimation of lead exposure intensity by industry using nationwide exposure databases in Korea. Saf Health Work 2021;12(4):439–444 (https://doi: 10.1016/j.shaw.2021.07.008)

Millard SP. EnvStats: an R package for environmental statistics. New York: Springer; 2013. ISBN 978-1-4614-8455-4, URL:https://www.springer.com.

Mulhausen J, Damiano J. A Strategy for assessing and managing occupational exposures, Second Edition. Fairfax Va: AIHA; 1998.

Perkins JL, Cutter GN, Cleveland MS. Estimating the mean, variance, and confidence limits from censored (〈Limit of Detection), lognormally distributed exposure data. Am Ind Hyg Assoc J 1990;51(8):416–419 (https://doi.org/10.1080/15298666091369871)

R Core Team. R: A language and environment for statistical computing. R Foundation forStatistical Computing, Vienna, Austria. 2021. URL https://www.R-project.org/.

Tekindal MA, Erdoğan BD, Yavuz Y. Evaluating left-censored data through substitution, parametric, semi-parametric, and nonparametric methods: a simulation study. Interdiscip Sci 2017;9(2):153–172 (https://DOI 10.1007/s12539-015-0132-9)

## <저자정보>

박주현(교수), 최상준(교수), 고동희(교수), 박동욱(교수), 성예지(박사과정)