FOOD SAFETY

doi: 10.14252/foodsafetyfscj.D-20-00007
Food Safety 2020; Vol. 8, No. 4, 67–89

Original Article

Attempt of Bayesian Estimation from Left-censored Data Using the Markov Chain Monte Carlo Method: Exploring Cr(VI) Concentrations in Mineral Water Products

Yoshinari Suzuki[1], Noriko Tanaka[2], and Hiroshi Akiyama[1]

[1]Division of Foods, National Institute of Health Science, Tonomachi 3-25-26, Kawasaki-ku, Kawasaki, Kanagawa 210-9501, Japan

[2]Department of Health Data Science Research, Healthy Aging Innovation Center, Tokyo Metropolitan Geriatric Medical Center, Sakae-cho 35-2, Itabashi-ku, Tokyo 173-0015, Japan

Hexavalent chromium (Cr(VI)) is toxic, carcinogenic, and mutagenic substances. Oral exposure to Cr(VI) is thought to be primarily from drinking water. However, under the certain reporting limit (~0.1 μg/L), percentage of Cr(VI) concentration in mineral water products under the reporting limit were estimated higher than 50%. Data whose values are below certain limits and thus cannot be accurately determined are known as left-censored. The high censored percentage leads to estimation of Cr(VI) exposure uncertain. It is well known that conventional substitution method often used in food analytical science cause severe bias. To estimate appropriate summary statistics on Cr(VI) concentration in mineral water products, parameter estimation using the Markov chain Monte Carlo (MCMC) method under assumption of a lognormal distribution was performed. Stan, a probabilistic programming language, was used for MCMC. We evaluated the accuracy, coverage probability, and reliability of estimates with MCMC by comparison with other estimation methods (discard nondetects, substituting half of reporting limit, Kaplan-Meier, regression on order statistics, and maximum likelihood estimation) using 1000 randomly generated data subsets (n = 150) with the obtained parameters. The evaluation shows that MCMC is the best estimation method in this context with greater accuracy, coverage probability, and reliability over a censored percentage of 10-90%. The mean concentration, which was estimated with MCMC, was 0.289×10^{-3} mg/L and this value was sufficiently lower than the regulated value of 0.05 mg/L stipulated by the Food Sanitation Act.

Key words: nondetects, left-censored data, Bayesian model, MCMC, Stan, Cr(VI)

Received: 19 March 2020; Accepted: 14 October 2020; Published online: 25 December 2020
Corresponding author: Yoshinari Suzuki, Division of Foods, National Institute of Health Science, Tonomachi 3-25-26, Kawasaki-ku, Kawasaki, Kanagawa 210-9501, Japan (szk-yoshi@nihs.go.jp)
The contents of this article reflect solely the view of the author(s).
Abbreviations: CDF: cumulative distribution function, CI: confidence interval, CP: coverage probability, CrI: credible interval, DN: discarding nondetects, EAP: expected a posteriori, GM: parameter of lognormal distribution, geometric mean, GSD: parameter of lognormal distribution, geometric standard deviation, IQR: interquartile range, KM: Kaplan-Meier, LCI: lower confidence interval, LOD: limit of detection, LOQ: limit of quantification, $\hat{m}$: estimate of geometric mean. The letters on the upper right indicate the estimation method, MCMC: Markov chain Monte Carlo, MLE: maximum likelihood estimation, MSE: mean of squared error, MW: mineral water, PDF: probability density function, RIQR: relative inter quartile range, RL: reporting limit, RL/2: substitution of half of RL, RSD: relative standard deviation, $\hat{s}$: estimate of geometric standard deviation. The letters on the upper right indicate the estimation method, SD: standard deviation, UCI: upper confidence interval, WAIC: widely applicable information criterion, $\hat{\mu}$: estimate of expected value. The letters on the upper right indicate the estimation method, $\hat{\sigma}$: estimate of standard deviation. The letters on the upper right indicate the estimation method,
Suggested citation: Yoshinari Suzuki, Noriko Tanaka, Hiroshi Akiyama. Attempt of Bayesian Estimation from Left-censored Data Using the Markov Chain Monte Carlo Method: Exploring Cr(VI) Concentrations in Mineral Water Products. *Food Safety.* 2020; 8 (4) 67–89. doi: 10.14252/foodsafetyfscj.D-20-00007

1. Introduction

Chromium (Cr) is a metal widely distributed in the environment. Chromium has various oxidation states from −2 to +6, but the most predominant forms are +3 and +6[1]. Trivalent Cr (Cr(III)) is an essential nutrient for cholesterol, fat and glucose metabolism in human bodies, while hexavalent Cr (Cr(VI)) is toxic, carcinogenic, mutagenic, and mobile in nature[2]. Continuous exposure to Cr(VI) may cause pulmonary congestions, liver damage, skin irritation, and kidney failure[3]. Once Cr(VI) enters inside the cell, it undergoes a rapid metabolic reduction and is converted ultimately to Cr(III)[4]. Thus, the Food Safety Commission of Japan estimated daily intake of Cr(VI) from consumption of mineral water and tap water[5]. According to the reports on Cr(VI) concentration in mineral water (MW) products[6–8], detected concentrations ranged from 0.1 to 3.4 μg/L, and the data contained many non-detected values with censored percentage from 50 to 75% under the certain reporting limit (~0.1 μg/L). When nondetects are present in data, they lead to difficulties in computing statistical metrics. Estimates from the data with nondetects could vary widely depends upon not only the censored proportion due to the properties of the measured substance and the performance of the analytical instrument, but also the performance and the assumption of the statistical analysis methods. In other words, estimation of Cr(VI) exposure via MW products becomes uncertain.

Since foods are major exposure source to chemicals, assessment of chemical exposure via foods is important for health protection. Non-detected values often occur for chemical concentrations in foods because the analytical methods for chemicals in foods always have the limit of detection (LOD) and the limit of quantification (LOQ)[9–13]. Data lacking certain values which are lower than LOD and LOQ, are known as left-censored. In the field of food analytical science, many studies used substitution (0[10,11], LOD/2[10,13], or LOD[9,12]) for left-censored data on the exposure estimation. The estimation of chemical exposure via foods is likely to be evaluated by the conservative method using substitution of LOD or LOQ for left-censored data on the safety side. Therefore, such conservative assessment using those methods may cause a severe overestimation that is meaningless and impractical from the viewpoint of health risk and food distribution. False concerns and excessive regulations lead to strict rules onto the society and the food community. In 2006, EPA guidance supported the use of 0, LOD/2, or LOD substitution in data with less than 15% nondetects[14]. However, in 2015 the EPA revised its guidelines stating that substitution by LOD/2 should be used only if the percentage of censored data is less than 5% and if the data are mildly skewed[15]. Helsel[16] advocates that academic journals should consider substitution as flawed and reject the papers that such methods are implemented. Helsel recommended three methods—Kaplan-Meier (KM), maximum likelihood estimation (MLE), and regression on order statistics (ROS)—as more accurate alternatives for computing statistics on data with nondetects[17,18]. Helsel[18] gave recommendations of usage of KM, MLE, and ROS methods as follows: 1: KM should be used for data with censored percentage of <50%; 2: ROS should be used for small sample size (<50) and 50-80% censored percentage; 3: MLE should be used for large sample size (≥50) and 50-80% censored percentage. It can be expected that the MLE shows good performance for large sample size, however it is also reported that MLE shows poor performance for high-skewed data[19]. On the contrary, it is well known that incorporating prior information in Bayesian framework result in a reduced sample size, and Bayesian method allows for great flexibility in dealing with missing data[20]. Thus, we expect that Bayesian framework may give a unified solution for different situations of left-censored data.

In singular models with many hidden parameters, such as multinomial mixture model, change-point model, neural networks, hidden Markov models, and so on, it was reported that Bayesian estimation make the generalization error smaller[21]. Since left-censored data have many hidden values as nondetects, Bayesian estimation is expected to provide appropriate estimates. Thus, the estimation method based on Bayesian modeling using the Markov chain Monte Carlo (MCMC) method has recently been applied in left-censored data. To compute Bayesian models, probabilistic programming languages such as BUGS and JAGS have been used. For example, Bayesian modeling using BUGS and JAGS has been applied to left-censored data for microbiological contamination data[22], hexabromocyclododecane diastereomer compositions in water[23], fluoride in drinking water[24], and so on.

While there are various reports on comparisons among non-Bayesian methods (substitution, non-parametric, semi-parametric, and parametric) in the context of left-censored data analysis[25–30], the reports comparing the performance of Bayesian estimation with other estimation methods are limited. Huynh et al[31] evaluated the estimation performance of a Bayesian method and substitution by simulation with real data in the specific area except for food safety science. Nie et al[32], Hady and Rain[33], and Feroze and Aslam[34] also evaluated by comparison of Bayesian and likelihood-based estimation for left-censored data using simulations. Although Nie el al[32] adopted Bayesian approach to the real data in the specific area except for food safety science, Hady and Rain[33], and Feroze and Aslam[34] performed a simulation

study without real data. They revealed that Bayesian method resulted in less bias. Although Bayesian estimation has been reported to have satisfactory coverage probability, as well as less bias, the comparable performance and relative reliability (i.e., the stability of results given alternative datasets) of the Bayesian method approach in food safety science have been never evaluated. A method with less reliability may fail to fit additional data or predict future observations. We consider that it is important to validate the Bayesian method approach for the standardization in food safety science.

In this study, in order to elaborate exposure assessment, we applied Bayesian modeling using MCMC, in addition to non-Bayesian methods, to obtain less biased summary statistics on Cr(VI) concentrations in MW products from Kataoka et al (2017)[8], which has over 50% nondetects and no specified criterion concerning nondetects. Since the true values of probability density distribution parameter are unknown, we have conducted a simulation study. In addition, we validated the proposed MCMC method on the left-censored data analysis by comparison of several estimation methods to explore the accuracy, coverage probability, and reliability of the proposed MCMC method in food safety science. The process was divided into two stages as follows: 1: Estimate the concentration distribution from the original data using MCMC and other methods; 2: Evaluate MCMC performance compared with other statistical methods from simulation study. Finally, this study can contribute to the reliability of food safety evaluation by indicating the possibility of evaluating more accurate concentration in a similar situation to Cr(VI).

2. Methods

2.1 Data

Data on Cr(VI) concentrations in MW products reported by Kataoka et al[8] were used. One hundred and fifty MW products were purchased in 2016. Among them, 110 products are domestic, and 40 products are imported. The number of purchases was assigned according to the number of products for each prefecture and country in the Japanese market. The lower LOQ was calculated and data below LOQ are marked as "Tr.". However, no description about the value of LOQ can be found in the text. Although summary statistics for the observed data are reported in the paper, appropriate analysis for data containing nondetects has not been performed. Herein, unless otherwise specified, the censoring criterion is described as reporting limit (RL).

2.2 Inferential Analysis

Statistical analysis was performed using R (ver. 3.4.0). R packages EnvStat (ver. 2.3.1) and rstan (ver. 2.16.2) were used to estimate parameters from left-censored data.

2.3 Parameter Estimation

To compare parameters estimated from left-censored data with MCMC, we used 5 other methods (1: Discard nondetects (DN); 2: Substitution of one half of RL (RL/2); 3: KM; 4: MLE; 5: ROS). ROS is a method corresponding to a normal or lognormal distribution. We performed statistical analysis under assumption that the original data follows lognormal distribution (see **Supporting Information S4**. for further detail). KM, MLE, and ROS methods were implemented using the EnvStats package. The minimum value of detected data was used as a censoring point for convenience. Then a location parameter (geometric mean, $\hat{m}$), a shape parameter (geometric standard deviation, $\hat{s}$), a mean ($\hat{\mu}$), a standard deviation ($\hat{\sigma}$), and lower (LCI, $\hat{L}$) and upper (UCI, $\hat{U}$) 95% confidence intervals of mean Cr(VI) concentration were estimated for data reported by Kataoka et al[8] using the 6 methods.

2.3.1 Discard nondetects

Entries with non-detected values are eliminated and only detected values are used for further calculation. This approach is attractive because of its simplicity but the data may be distorted.

2.3.2 Substitution of one half of reporting limit

Non-detected values are replaced with one half of RL and both detected and replaced values are used for further calculation. Substitution is still widely used in various fields because of its tractability, and it is perhaps acceptable under certain conditions.

2.3.3 Kaplan-Meier

Kaplan-Meier is a nonparametric estimator since it does not make distributional assumptions. More specifically, KM is a non-parametric version of maximum likelihood estimation, and this method estimates the percentiles, or cumulative distribution function (CDF), for the data, where the mean equals the area beneath the CDF[35]. It is widely used in survival analysis to estimate survival functions, which are then used to estimate different summary statistics[35] (see **Supporting Information S6.1** for the detailed algorithm). The KM method can provide useful estimates when sample sizes are small; however, it does not perform well if more

than 50% of data are nondetects or if fewer than eight detections are available for evaluation[18].

2.3.4 Regression on order statistics

Regression on order statistics is a semi-parametric, simple imputation method that transforms nondetects on the basis of a probability plot of detects[17,18,36,37]. ROS is also known as "Imputation Using Quantile-Quantile Regression" (see **Supporting Information S6.2** for the detailed algorithm). ROS performs better than MLE and some substitution methods when the sample size is small (less than 50) and where data do not fit a distribution[38].

2.3.5 Maximum likelihood estimation

Maximum likelihood estimation solves a likelihood equation to estimate the parameter(s) using both detected observations and the proportion of data falling below RL[17]. The observed data (x) enter the likelihood function through the probability density function (PDF, $f(x|m, s)$) of lognormal distribution and the censored observations can be accounted for by the CDF ($F(\mathrm{RL}|m,s) = P(x \leq \mathrm{RL}|m,s)$) of lognormal distribution as follows:

$$L\left(m,s|x_1,x_2,\cdots,x_n\right)=\prod_{x\in D} f\left(x|m,s\right)\cdot\prod_{x\in C} F\left(\mathrm{RL}|m,s\right), \quad \text{Equation 1}$$

where D is the set of all observed values and C is the set of all left-censored values. MLE finds the parameter values (m, s) that maximize the likelihood function against the observations (see **Supporting Information S6.3** for the detailed algorithm).

2.3.6 Markov chain Monte Carlo

Among probabilistic programing languages, Stan which was developed by Gelman et al[39] was used in this study. A key feature of Stan is that NUTS (No-U-Turn Sampler), which is an implementation of Hamiltonian Monte Carlo, one of the algorithms of the MCMC method, is adopted. Various Bayesian modeling sources using Stan have been published and documentation is abundantly available online[40].

In addition to two parameters of distribution (GM and GSD), we estimated RL with using MCMC (see **Supporting Information S1.** for furter details). Four parallel Markov chains are calculated using MCMC to check convergence. The numbers of iterations, warm-up, and thinning were set to 2×10^3, 1×10^3, and 2, respectively (see **Supporting Information S2.** for further details). For parameters estimated by MCMC, we used an expected a posteriori (EAP), which is mean of the posterior predictive distribution, as a point estimate. The reason is that an EAP reflect the information of entire distribution, while maximum a posteriori can be significantly affected by a small number of outliers. Then EAP was used for further calculation. After execution of the Stan model, we calculated the widely applicable information criterion (WAIC)[21] (see **Supporting Information S3.** for further details).

2.4 Monte Carlo Simulations

We generated 10^5 random numbers according to the lognormal distribution with the parameters estimated in section 2.3. Using MCMC, estimated parameter(s) can be obtained as posterior predictive distribution(s). Since we obtained a set of 2×10^3 estimates for each parameter, 50 random numbers are generated against posterior predictive distributions, and a total of 10^5 random numbers are obtained (see line 42 of **Fig. S1**).

2.5 Simulation and Comparison of Various Methods

To explore the accuracy, coverage probability, and reliability of parameter estimation from left-censored data by MCMC, simulation and comparison with other methods (DN, RL/2, KM, ROS, and MLE) was carried out. To rank the 6 methods for estimating left-censored data, simulations were performed with the following algorithm:

1. Generate a random number of size 150 according to a lognormal distribution with the estimated parameters in section 2.3.6 as true values.
2. Simulate censoring by using a particular theoretical percentile of the sample data as an RL.
3. Execute the various methods to estimate parameters.
4. Repeat steps 1-3 1000 times.
5. Using the true values, compute some criteria.

When simulating censoring, we used theoretical percentiles of simulated data. For example, to simulate data with 40% nondetects, we considered any value falling below the theoretical 40^{th} percentile of the data as nondetects. We considered the minimum value of detected data as RL. The simulations cover multiple censored percentages from 10 to 90%. Then 6 parameters ($\hat{m}$, $\hat{s}$, $\hat{\mu}$, $\hat{\sigma}$, $\hat{L}$, and $\hat{U}$) were estimated using the 6 methods.

To evaluate accuracy, bias and mean squared error (MSE) were calculated. The bias of an estimator $\hat{\theta}$, for estimating parameter θ, is defined as follows:

$$B\left(\hat{\theta}\right)=E\left(\hat{\theta}-\theta\right). \quad \text{Equation 2}$$

Table 1. Summary statistics and estimated parameters from left-censored data based on six inferential methods.

Method	Mean ($\times10^{-3}$)	SD ($\times10^{-3}$)	LCI[b] ($\times10^{-3}$)	UCI[b] ($\times10^{-3}$)	GM ($\times10^{-3}$)	GSD
DN	0.472	0.433	0.380	0.587	0.348	2.18
RL/2	0.233	0.353	0.187	0.291	0.114	2.97
KM	0.261	0.338	0.216	0.305	0.160	2.69
ROS	0.257	0.343	0.208	0.398	0.148	2.78
MLE	0.266	0.846	0.201	0.342	0.080	4.72
MCMC	0.289[a]	1.06 [a]	0.194	0.460	0.082[c]	4.90[c]

a: Expected a posteriori of generated quantities (mean_est or sd_est) calculated by MCMC.
b: Lower and upper 95% confidence interval of the mean. For MCMC, 95% credible interval of posterior predictive distribution for generated quantities (mean_est).
c: Expected a posteriori of target parameters (GM and GSD).

MSE of an estimator $\hat{\theta}$ is defined as follows:

$$\mathrm{MSE}\left(\hat{\theta}\right) = E\left[\left(\hat{\theta} - \theta\right)^2\right]. \qquad \text{Equation 3}$$

We calculated the coverage probability of 95% CIs for mean (CP), which indicates the fraction of computed CIs that include the desired but unobservable parameter value. The 95% CIs for means estimated by DN and RL/2 were calculated using Cox's method[41] (see **Supporting Information S5.** for further details). The 95% CIs for means estimated by KM, ROS, and MLE were calculated by the bootstrap with bias-corrected and accelerated method (with 1000 repetitions). For the MCMC method, 95% credible intervals (CrI) for the mean were calculated (see line 43 of **Fig. S1**). Then the CP was evaluated by calculating whether the expected value was included within the 95% CI or 95% CrI (see **Supporting Information S7.** for further information of CI and CrI).

To evaluate reliability, which indicates the stability of results given alternative data, we calculated the relative interquartile range (RIQR) and relative standard deviation (RSD) for estimated parameters of GM and GSD from 1000 different datasets.

3. Results

3.1 Method Inter-comparison for Cr(VI) concentrations with Nondetects

To validate the MCMC results, we conducted a method inter-comparison. **Table 1** shows summary statistics and estimated parameters from left-censored data by various methods. The following describes the results from each estimation method.

3.1.1 Discard nondetects

Among the estimation methods, $\hat{\mu}^{\mathrm{DN}}$ (0.472×10^{-3}) and $\hat{s}^{\mathrm{DN}}$ (2.18) were highest and lowest, respectively. Given that nondetects are ignored, it seems reasonable to have such a result.

3.1.2 Substitution of one half of reporting limit

Although the 95% CI of $\hat{\mu}^{\mathrm{RL/2}}$ overlapped with that estimated by other methods except for DN, $\hat{\mu}^{\mathrm{RL/2}}$ (0.233×10^{-3}) showed the minimum value among the estimation methods. Since the dataset used in this study does not satisfy the EPA's proposed requirements for RL/2[15], we should not use the RL/2 method.

3.1.3 Kaplan-Meier

If there is only one RL, the estimation result is that $\hat{\mu}^{\mathrm{KM}}$ is identical to a substitution of the RL for censored data[18].

Thus, it is reasonable that $\hat{\mu}^{\mathrm{KM}}$ (0.261×10^{-3}) was higher than $\hat{\mu}^{\mathrm{RL/2}}$ (0.233×10^{-3}). The value of $\hat{\mu}^{\mathrm{KM}}$ was similar to $\hat{\mu}^{\mathrm{ROS}}$ (0.257×10^{-3}) and $\hat{\mu}^{\mathrm{MLE}}$ (0.266×10^{-3}). The value of $\hat{\sigma}^{\mathrm{KM}}$ (0.338×10^{-3}) was similar to $\hat{\sigma}^{\mathrm{RL/2}}$ (0.353×10^{-3}) and $\hat{\sigma}^{\mathrm{ROS}}$ (0.343×10^{-3}). Since the data used herein have more than 50% nondetects, KM estimates may be less relevant.

3.1.4 Regression on order statistics

$\hat{\mu}^{\mathrm{ROS}}$ (0.257×10^{-3}) was relatively close to what was obtained using other methods except for DN. $\hat{\sigma}^{\mathrm{ROS}}$ (0.355×10^{-3}) was lower than $\hat{\sigma}^{\mathrm{MLE}}$ (0.846×10^{-3}) and $\hat{\sigma}^{\mathrm{MCMC}}$ (1.06×10^{-3}).

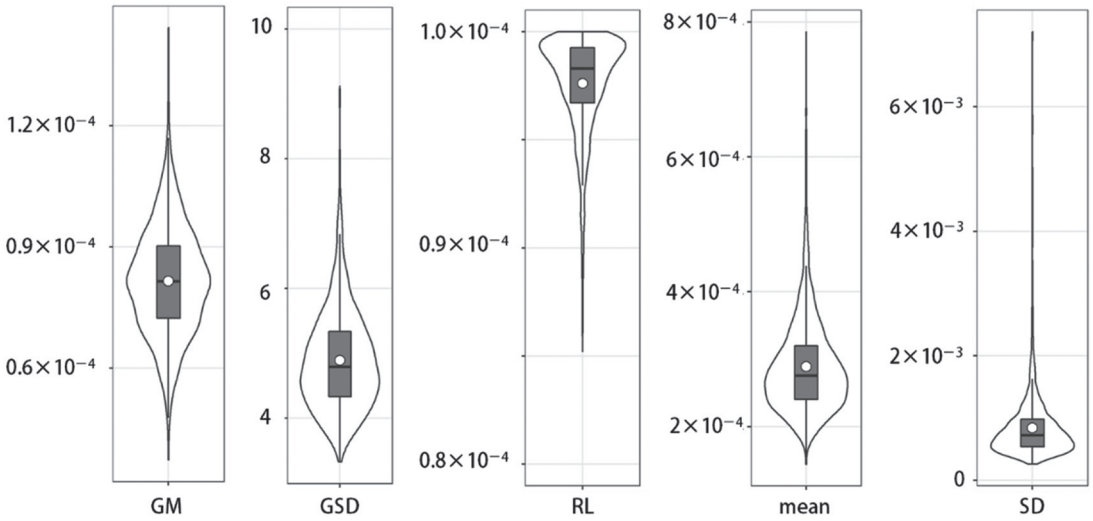


Fig. 1. Violin plots for posterior predictive distribution of parameters (geometric mean (GM), geometric standard deviation (GSD), reporting limit (RL), mean, and standard deviation (SD)) about Cr(VI) concentration in mineral water products reported by Kataoka et al (2017) under assumption that original data follow a lognormal distribution. Violins, boxes, horizontal solid lines, and open circles indicate probability density, interquartile range, median, and arithmetic mean, respectively.

3.1.5 Maximum likelihood estimation

$\hat{\mu}^{MLE}$ (0.266×10^{-3}) was comparable to what was obtained using other methods except for DN. However, $\hat{\sigma}^{MLE}$ (0.846×10^{-3}) and $\hat{s}^{MLE}$ (4.72) were higher compared to the other methods except for MCMC.

3.1.6 Markov chain Monte Carlo

The posterior predictive distributions of parameters (GM, GSD, and RL) and generated quantities (means and SD) are shown in **Fig. 1**. The value of $\hat{\mu}^{MCMC}$ (0.289×10^{-3}) was slightly higher compared to KM, ROS, and MLE. $\hat{m}^{MCMC}$ (0.082×10^{-3}) showed the second lowest value among all other methods. $\hat{\sigma}^{MCMC}$ (1.06×10^{-3}) and $\hat{s}^{MCMC}$ (4.94) were higher than corresponding values from all other methods.

The 95% CrI for the mean by MCMC was wider than other methods. There are two possible reasons for this. The first is that, according to the Cox method, the larger GSD resulted in the wider CrI. Another reason is that the estimated parameters themselves have a distribution in MCMC, and the uncertainty was combined, resulting in wider CrI than other methods.

3.1.7 Distribution of Cr(VI) concentrations in mineral water products

To estimate the distribution of Cr(VI), a total of 10^5 Cr(VI) random numbers were generated (**Fig. 2**). Regardless of estimation method, Cr(VI) concentrations in MW products were sufficiently lower than the regulated value of 0.05 mg/L stipulated by the Food Sanitation Act[42]. The mean values were higher in the order of DN (0.472×10^{-3}) > MCMC (0.289×10^{-3}) > MLE (0.268×10^{-3}) > KM ($0.260_2\times10^{-3}$) > ROS ($0.259_9\times10^{-3}$) > RL/2 (0.205×10^{-3}). The 97.5th percentile values were higher in the order of MCMC (1.82×10^{-3}) > MLE (1.69×10^{-3}) > DN (1.61×10^{-3}) > ROS ($1.12_3\times10^{-3}$) > KM ($1.11_7\times10^{-3}$) > RL/2 (0.96×10^{-3}). Since the actual 97.5th percentile value is 1.38×10^{-3}, it can be concluded that the methods explored herein express the original data reasonably well. The estimated mean values showed similar values except for DN, but the 97.5th percentile value was larger when MCMC was used.

3.2 Comparison the Different Methods

Since the true values for parameters of probability density distribution are unknown, we conducted the simulation study to determine the most appropriate methods. We compared the performances of each method with using assessment criteria (bias, MSE, CP, RIQR, and RSD). Ideally, we want the methods being studied to have a bias as close to 0 as possible and an MSE as small as possible. In addition to high accuracy, we want to know that a particular method is appropriate in terms of estimating CI for means, and reliable when faced with alternative data sets.

3.2.1 Accuracy

The results of accuracy assessment under censored percentages from 10 to 90% are shown in **Fig. 3**. The DN method overestimates GM and underestimates GSD. This tendency became more pronounced as the censored percent-

doi: 10.14252/foodsafetyfscj.D-20-00007

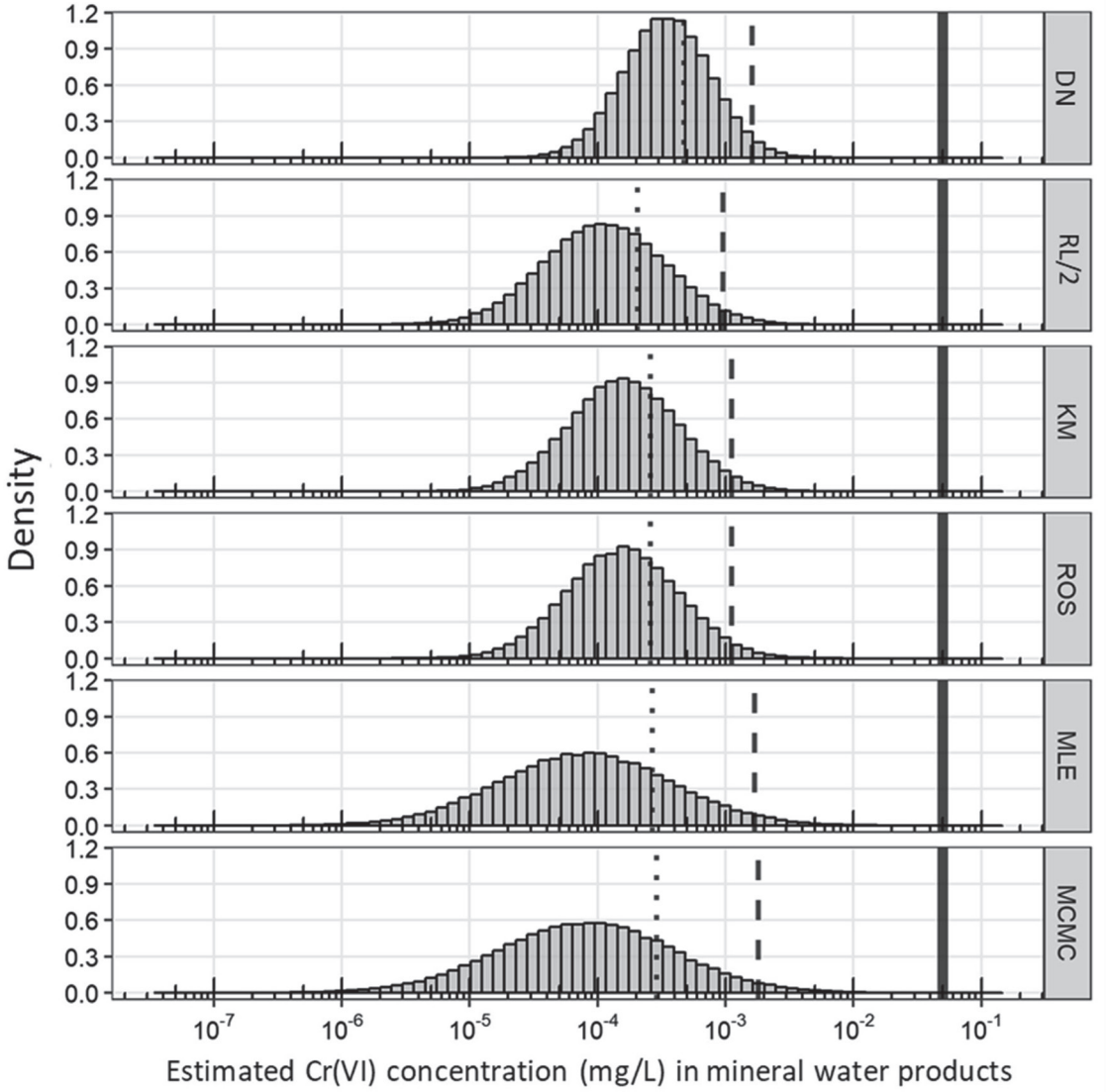


Fig. 2. Histograms of generated random numbers for Cr(VI) concentrations in mineral water products following a lognormal distribution with estimated parameters by DN, RL/2, KM, ROS, MLE, and MCMC. Vertical solid, dashed, and dotted lines indicate regulated concentration (0.05 mg/L), 97.5th percentiles, and mean, respectively.

age increases. We should refrain from calculating summary statistics after discarding nondetects. MCMC tended to overestimate the GSD slightly (<101% overestimated when the censored percentage was between 10 and 50), but other methods tended to underestimate the GSD.

Even when the censored percentage is small, $\hat{s}^{DN}$ and $\hat{s}^{KM}$ exhibited high MSE values. Where the censored percentage was between 10 and 50, the ROS, MLE, and MCMC methods exhibited similar MSE values for both GM and GSD. However, in terms of the ROS method, MSE increased from around 50% censoring upwards for both GM and GSD.

$\hat{m}^{MLE}$ and $\hat{m}^{MCMC}$ showed similar values in the censoring range of 10 to 90%, and these estimation methods exhibited better accuracy. The MSE of both $\hat{m}^{MCMC}$ and $\hat{s}^{MCMC}$ were lower compared to other estimation methods when the censored percentage was in the range of 10 to 90%. From these results, we concluded that MCMC showed the best accuracy to estimate distribution parameters from the original data.

3.2.2 Coverage probability

The results of CPs assessment under censored percentages from 10 to 90% are shown in **Fig. 4**. The CPs decreased sharply from 10% for DN and 60% for the KM method. In terms of the RL/2 method, the CP gradually decreased from 10%, and rapidly decreased around 70%. In ROS and MLE, the CP showed similar values (75 to 85%) depending on the censoring ratio. MCMC showed the most stable CPs, and the CPs were generally close to the target coverage of 0.95.

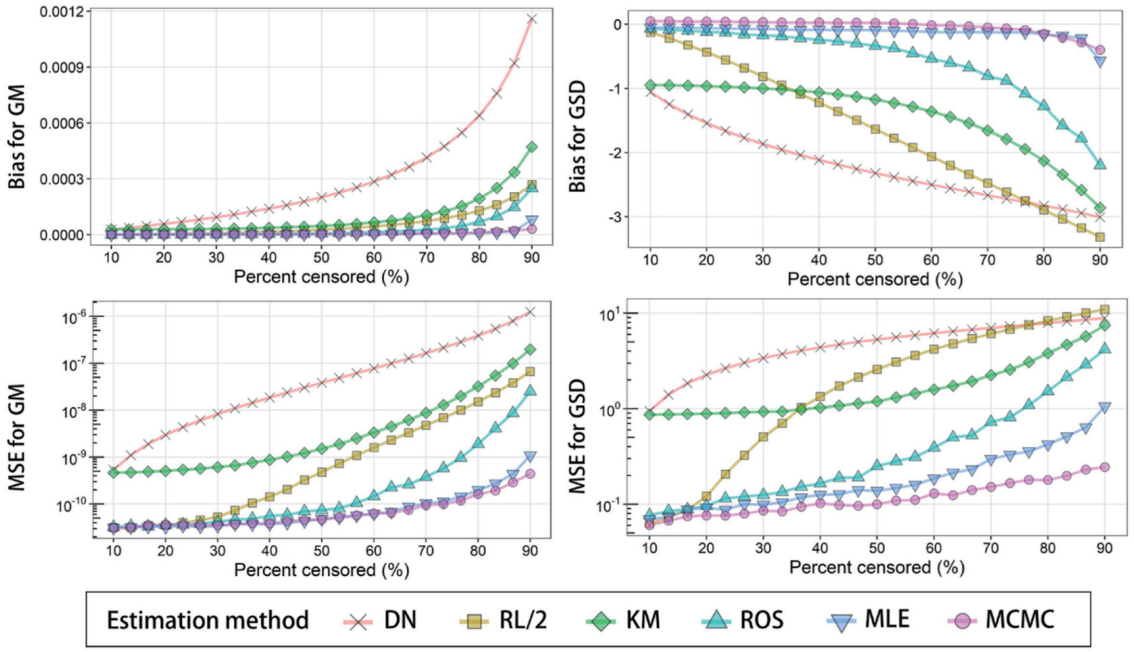


Fig. 3. Bias and mean squared error (MSE) for GM and GSD estimated by DN, RL/2, KM, ROS, MLE, and MCMC from 1000 randomly generated left-censored data subsets (n = 150) which follow a lognormal distribution with certain parameters (GM = 0.082×10^{-3}; GSD = 4.9) over a 10 to 90% censoring range. For MCMC, the expected a posteriori was used for further calculation.

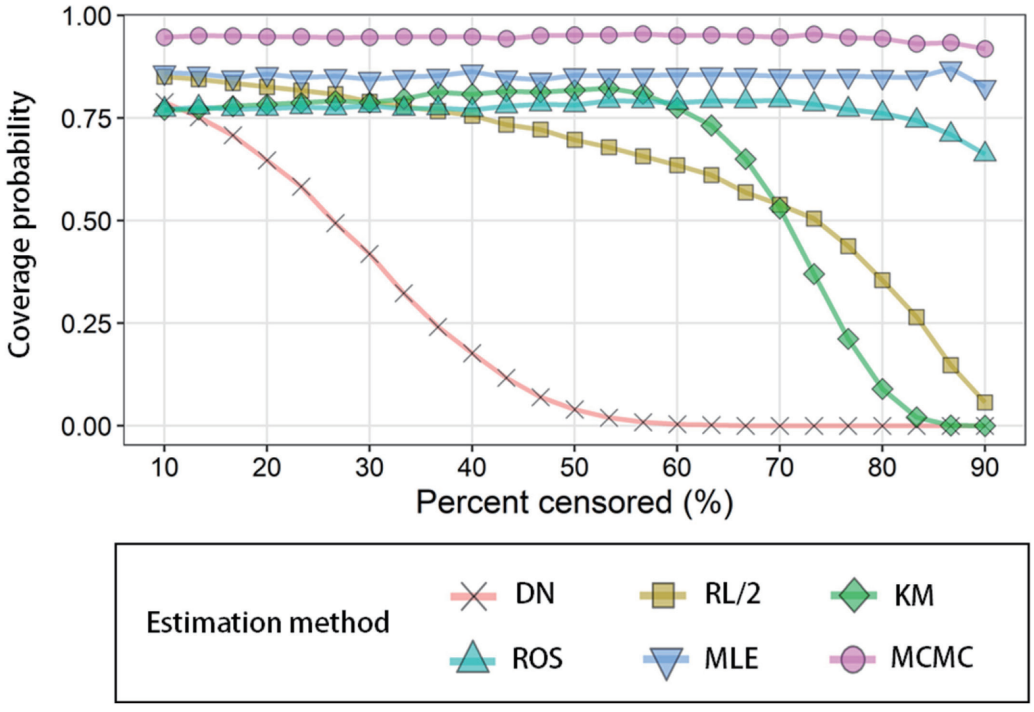


Fig. 4. Coverage probability of 95% CI or 95% CrI for mean estimated by DN, RL/2, KM, ROS, MLE, and MCMC from 1000 randomly generated left-censored data subsets (n = 150) which follow a lognormal distribution with certain parameters (GM = 0.082×10^{-3}; GSD = 4.9) over a 10 to 90% censoring range.

3.2.3 Reliability

The violin plots for $\hat{m}$ and $\hat{s}$ at the censored ratio of 85/150 are shown in **Fig. 5**. The true values were distributed within the IQR estimated by ROS, MLE, and MCMC, but not within those by DN, RL, and KM. Among these three estimation methods, MCMC showed the smallest variation

doi: 10.14252/foodsafetyfscj.D-20-00007

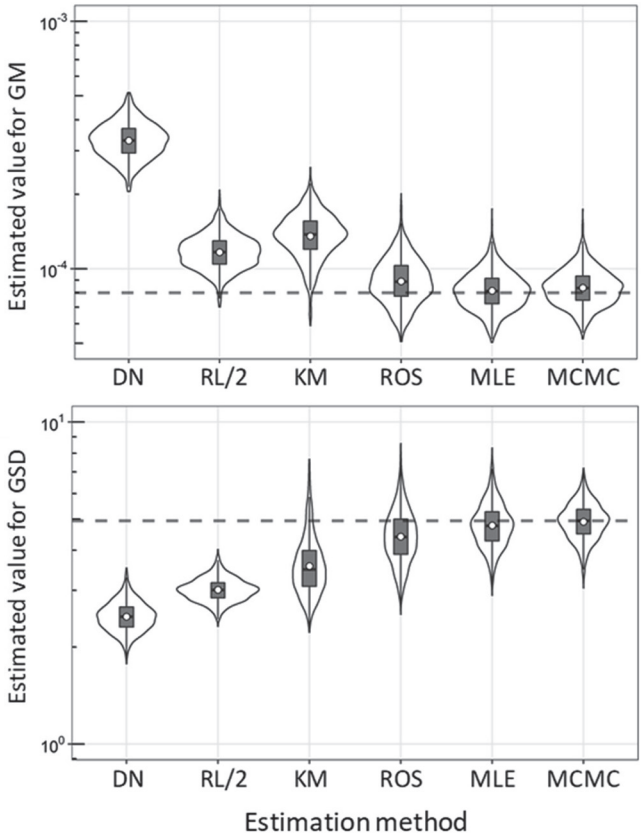


Fig. 5. Violin plots of GM and GSD estimated by DN, RL/2, KM, ROS, MLE, and MCMC from 1000 randomly generated left-censored data subsets (n = 150) which follow a lognormal distribution with certain parameters (GM = 0.082×10^{-3}; GSD = 4.9) at a censoring ratio of 85/150. For MCMC, the expected a posteriori was used for further calculation. Horizontal dashed lines indicate true values. Violins, boxes, horizontal solid lines, and open circles indicate probability density, IQR, median, and arithmetic mean, respectively.

in RIQR and RSD for both $\hat{m}$ and $\hat{s}$.

The results of reliability assessment under censored percentages from 10 to 90% are shown in **Fig. 6**. The variations for both $\hat{m}$ and $\hat{s}$ were generally in the order of RL/2 < DN < MCMC < MLE < ROS. When the censored rate was high, KM exhibited lower RSD and RIQR compared to MLE and ROS. Even when the censored percentage was low, KM exhibited higher variation in both RSD and RIQR compared to other methods.

Since DN, RL/2, and KM showed lower accuracy, when these three methods were excluded, the MCMC estimation exhibited lower variation in both RSD and RIQR compared to ROS and MLE at censored percentages from 10 to 90%. From these results, it was concluded that MCMC is the most reliable estimation method.

Although $\hat{s}^{MCMC}$ showed reasonable reliability (the maximum RSD of $\hat{s}^{MCMC}$ was 0.18 at 90% censoring), the variation in $\hat{m}^{MCMC}$ was as high as RSD=0.37 and RIQR=0.44 at a censoring rate of 90%. A reliable estimation method for left-censored data with a high censoring rate is required.

4. Discussions

We have described the MCMC method, as well as the commonly used methods, for summarizing and analyzing left-censored Cr(V) concentration in MW products. We used a simulation study to demonstrate the properties of these methods. MCMC showed the most stable CPs with the highest accuracy, and the CPs were generally close to the target coverage of 0.95. These results are consistent with previous reports[32,34]. MLE showed the second-best performance next to MCMC. Since the original data is sufficient for MLE requirements (more than 50 observations and censoring ratio of 50–80%) recommended by Helsel[18], the performance of MLE was superior to those of ROS and KM. Other studies have also reported that MLE has a smaller bias than ROS (or probability plotting method) and KM[25,27]. Although RL/2 showed the best reliability, RL/2 showed lower accuracy and coverage probability under censored percentage from 20 to 90%. The RL/2 estimation results are considered to have a serious bias. As reported previously, substitution is essentially creating data which does not have an empirical basis. Since our simulation results are consistent with previous reports, the simulation in this study is considered to be valid.

Although the reliability of Bayesian approaches on left-censored data is less understood, our study shows that MCMC performed with the smaller variation in estimates compared to KM, ROS, and MLE. Since MCMC showed better performance on accuracy, coverage probability, and reliability compared to other methods, estimates by MCMC for Cr(VI) concentration in MW products are the most appropriate estimates among the methods used in this study. The most appropriate estimates of mean and those of 97.5th percentile of Cr(VI) concentration, which were estimated with MCMC, were 0.289×10^{-3} and 1.82×10^{-3} mg/L, respectively. These estimates were sufficiently lower than the regulated value of 0.05 mg/L stipulated by the Food Sanitation Act[42].

Huynh et al[31] reported that the use of more informative priors generally improved the Bayesian method's performance, making both the bias and the root MSE lower. The adoption of the appropriate prior distribution is sometimes difficult under the conditions when they are needed most, i.e. small sample sizes or high censored percentage. Since MCMC showed better performance on accuracy, coverage probability, and reliability, we have successfully adopted

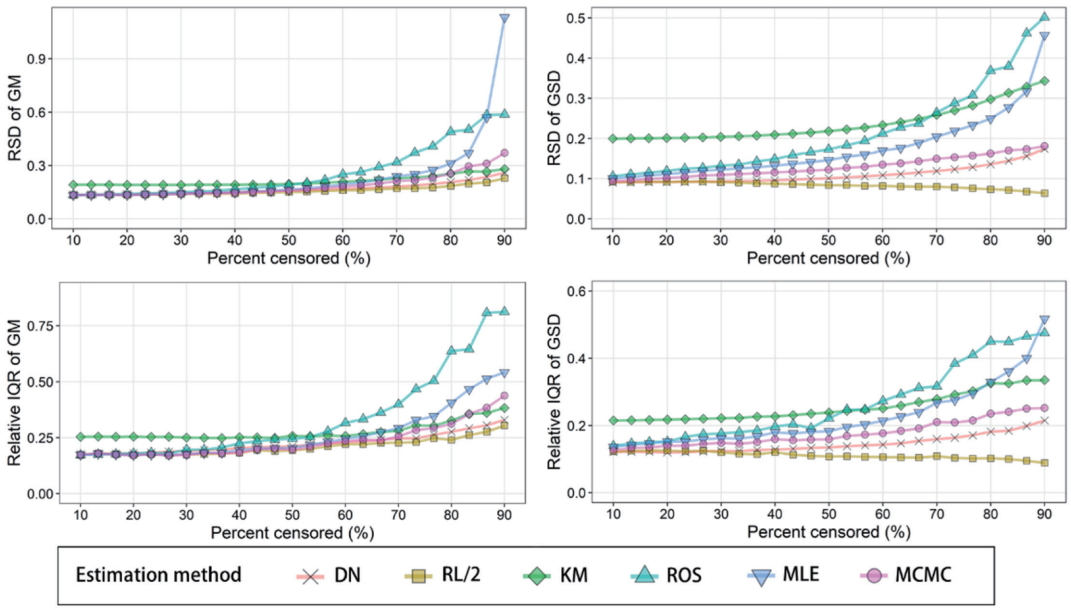


Fig. 6. Relative standard deviation (RSD) and relative interquartile range (IQR) of GM and GSD estimated by DN, RL/2, KM, ROS, MLE, and MCMC from 1000 randomly generated left-censored data subsets (n = 150) which follow a lognormal distribution with certain parameters (GM = 0.082×10^{-3}; GSD = 4.9) over a 10 to 90% censoring range. For MCMC, the expected a posteriori was used for further calculation.

informative prior distribution to the original data. We consider that this is the main reason why the MCMC showed better performance compared to other methods. It is well known that MCMC can accommodate outliers by describing data via heavy-tailed distributions to the extent implied by the data[43]. Since lognormal distributions with GSD greater than 4 are considered to have heavy tails, the performance of MCMC was superior to that of the other methods. In this study, we used the EAPs as point estimator in MCMC. Other point estimators (maximum a posteriori and median of posterior predictive distribution) for Bayesian estimation might show different performances. These are other reasons why the MCMC is better than other methods.

To perform safe assessment of exposure to chemicals, substituting RL for nondetects may be a conservative and preferable method. However, when the dissociation between reality and assumption is large, such a conservative method may generate overestimates that are meaningless for health risk, leading to false concern and excessive regulation. Rather than imposing strict rules onto society and the food community, it is important to introduce new, scientifically sound methods and resolve the lack of transparency arising from inadequate evaluation. The use of prior distributions in Bayesian estimation can be considered as bridging a prior knowledge to MLE. Although it is necessary to carefully check whether the prior distribution has been adopted appropriately, utilizing knowledge of society and food community, as well as researchers, is able to help to reduce the gap between science-technology and society. Bayesian estimation provides the whole (posterior predictive) distribution for each parameter estimated, instead of one single-value estimate in MLE, thus it can provide estimates on the safe side with considering uncertainty. We expect that Bayesian estimation with incorporating consumer and regulatory science reduce this gap in food safety fields.

Since the original data are univariate, we have not extended the regression model in this study. However, it may be possible to estimate individual values of nondetects using data covering signal intensity, production location, and other information. If sufficient information is available on the manufacturing process, storage method, factory, manufacturer, and so on, a hierarchical Bayesian model that considers these factors as random effects and incorporates them into the model could be considered (see **Supporting Information S8.** as example of comparison between Japanese and imported MW products). By virtue of such a complex analysis, it may be possible to evaluate manufacturing process risks such as contamination or changes in chemical species.

This study demonstrates that using the MCMC method provide the most appropriate estimation result for Cr(VI) concentration in MW products. On the other hands, as mentioned in many articles[18,27,44], our simulation result indicate that the use of RL/2 (and DN method) should be avoided.

Evaluation, prediction, and judgment based on scientific evidences are the basis of regulatory science and are important for risk assessment in food safety administration. MCMC, as well as other methods based on statistical validity such as MLE, is important in the field of food safety as an estimation method instead of conventional substitution method. On the other hand, it should be discussed carefully whether the MCMC estimation results are appropriate for other data. Evaluation of the validity for different probability density distributions, sample sizes, and parameters is necessary.

5. Conclusions

We demonstrate that parameter estimation from left-censored data with MCMC show better performance on accuracy, coverage probability, and reliability compared to other methods (DN, RL/2, KM, MLE, and ROS) for Cr(VI) concentration in MW products. The most appropriate estimate of mean concentration, which was estimated with MCMC, was 0.289×10^{-3} mg/L and this estimate was sufficiently lower than the regulated value of 0.05 mg/L stipulated by the Food Sanitation Act.

In the field of food safety, although chemical substances in food are rarely detected, there are many chemical substances that have to be periodically investigated for protecting public health. This study suggests that MCMC could be a powerful estimation method in exposure assessment contexts in food safety science. However, there is ample scope of future research for the standardization to explore this further for the model with various sample size, alternative parameters, and different distributions, as well as multiple RLs due to the performance difference of analytical instruments.

Acknowledgments

This work was partially supported by a Health Sciences Research Grant (19KA1002 and 19KA2001) from the Ministry of Health, Labour, and Welfare in Japan.

Conflict of interest

The authors declare no conflicts of interest.

References

1. Marqués MJ, Salvador A, Morales-Rubio A, de la Guardia M. Chromium speciation in liquid matrices: a survey of the literature. *Fresenius J Anal Chem*. 2000; **367**(7): 601–613. PMID:11221921, doi:10.1007/s002160000422
2. Zhitkovich A. Chromium in drinking water: sources, metabolism, and cancer risks. *Chem Res Toxicol*. 2011; **24**(10): 1617–1629. PMID:21766833, doi:10.1021/tx200251t
3. Sun H, Brocato J, Costa M. Oral chromium exposure and toxicity. *Curr Environ Health Rep*. 2015; **2**(3): 295–303. PMID:26231506, doi:10.1007/s40572-015-0054-z
4. Zhitkovich A. Importance of chromium-DNA adducts in mutagenicity and toxicity of chromium(VI). *Chem Res Toxicol*. 2005; **18**(1): 3–11. PMID:15651842, doi:10.1021/tx049774+
5. Food Safety Commission of Japan. Hexavalent chromium (Contaminants). *Food Safety*. 2019; **7**(2): 56–57. PMID:31998588, doi:10.14252/foodsafetyfscj.D-1900002
6. Catalani S, Fostinelli J, Gilberti ME, Apostoli P. Application of a metal free high performance liquid chromatography with inductively coupled plasma mass spectrometry (HPLC–ICP-MS) for the determination of chromium species in drinking and tap water. *Int J Mass Spectrom*. 2015; **387**: 31–37. doi:10.1016/j.ijms.2015.06.015
7. Marcinkowska M, Komorowicz I, Barałkiewicz D. New procedure for multielemental speciation analysis of five toxic species: As(III), As(V), Cr(VI), Sb(III) and Sb(V) in drinking water samples by advanced hyphenated technique HPLC/ICP-DRC-MS. *Anal Chim Acta*. 2016; **920**: 102–111. PMID:27114229, doi:10.1016/j.aca.2016.03.039
8. Kataoka Y, Watanabe T, Hayashi K, Ozawa R, Takizawa K, Akiyama H. Surveillance of chromium(VI) concentrations in mineral water products. *Food Hyg. Saf. Sci*. 2017; **58**(6): 275–280. doi:10.3358/shokueishi.58.275
9. Nabeshi H, Tsutsumi T, Uekusa Y, Hachisuka A, Matsuda R, Teshima R. Surveillance of strontium-90 in foods after the Fukushima Daiichi Nuclear Power Plant Accident. *Food Hyg. Saf. Sci*. 2015; **56**(4): 133–143. PMID:26346857, doi:10.3358/shokueishi.56.133
10. Tsutsumi T, Matsuda R, Yanagi T, et al. Dietary intake of dioxins in Japan in 2016 with time trends since 1998. *Food Addit Contam Part A Chem Anal Control Expo Risk Assess*. 2018; **35**(8): 1553–1564. PMID:29843576, doi:10.1080/19440049.2018.1469052
11. Uekusa Y, Takatsuki S, Tsutsumi T, et al. Determination of polychlorinated biphenyls in marine fish obtained from tsunami-stricken areas of Japan. *PLoS One*. 2017; **12**(4): e0174961. PMID:28376117, doi:10.1371/journal.pone.0174961
12. Cao P, Liu AD, Yang DJ, et al. Assessment of dietary exposure of young Chinese children to aluminium residues. *Food Addit Contam Part A Chem Anal Control Expo Risk Assess*. 2019; **36**(4): 582–591. PMID:30865572, doi:10.1080/19440049.2019.1579369
13. Oguri T, Yoshinaga J, Tao H, Nakazato T. Daily intake of inorganic arsenic and some organic arsenic species of Japanese subjects. *Food Chem Toxicol*. 2012; **50**(8): 2663–2667. PMID:22617717, doi:10.1016/j.fct.2012.05.010
14. US Environmental Protection Agency. Data quality assessment: Statistical methods for practitioners. EPA/240-B-06/003, Office of Environmental Information, Washington, DC, USA. 2006.
15. US Environmental Protection Agency. ProUCL Version 5.1.002 Technical Guide, EPA/600/R-07/041, Office of Research and Development, Washington, DC. October. 2015.
16. Helsel D. Much ado about next to nothing: incorporating nondetects in science. *Ann Occup Hyg*. 2010; **54**(3): 257–262. PMID:20032004

17. Helsel DR. More than obvious: better methods for interpreting nondetect data. *Environ Sci Technol*. 2005; **39**(20): 419A–423A. PMID:16295833, doi:10.1021/es053368a
18. Helsel DR. Statistics for Censored Environmental Data Using Minitab® and R, 2nd edition. Hoboken, NJ: Wiley; 2012.
19. Gilliom RJ, Helsel DR. Estimation of distributional parameters for censored trace level water quality data: 1. Estimation techniques. *Water Resour. Res*. 1986; **22**(2): 135–146. doi:10.1029/WR022i002p00135
20. Food and Drug Administration. Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. U.S. Department of Health and Human Services, ed. 2010.
21. Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J Mach Learn Res.* 2010; **11**: 3571–3594.
22. Busschaert P, Geeraerd AH, Uyttendaele M, Van Impe JF. Hierarchical Bayesian analysis of censored microbiological contamination data for use in risk assessment and mitigation. *Food Microbiol*. 2011; **28**(4): 712–719. PMID:21511131, doi:10.1016/j.fm.2010.06.006
23. Ichihara M, Yamamoto A, Kakutani N, Sudo M, Takakura K. A Bayesian approach for estimating hexabromocyclododecane (HBCD) diastereomer compositions in water using data below limit of quantification. *Environ Sci Pollut Res Int*. 2017; **24**(3): 2667–2674. PMID:27830419, doi:10.1007/s11356-016-7990-5
24. Qian SS, Schulman A, Koplos J, Kotros A, Kellar P. A hierarchical modeling approach for estimating national distributions of chemicals in public drinking water systems. *Environ Sci Technol*. 2004; **38**(4): 1176–1182. PMID:14998034, doi:10.1021/es020686q
25. She N. Analyzing censored water quality data using a nonparametric approach. *J Am Water Resour As*. 1997; **33**(3): 615–624. doi:10.1111/j.1752-1688.1997.tb03536.x
26. Sinha P, Lambert MB, Trumbull VL. Evaluation of statistical methods for left-censored environmental data with nonuniform detection limits. *Environ Toxicol Chem*. 2006; **25**(9): 2533–2540. PMID:16986810, doi:10.1897/05-548R.1
27. LaFleur B, Lee W, Billhiemer D, Lockhart C, Liu J, Merchant N. Statistical methods for assays with limits of detection: Serum bile acid as a differentiator between patients with normal colons, adenomas, and colorectal cancer. *J Carcinog.* 2011; **10**(1): 12. PMID:21712958, doi:10.4103/1477-3163.79681
28. Huynh T, Ramachandran G, Banerjee S, et al. Comparison of methods for analyzing left-censored occupational exposure data. *Ann Occup Hyg.* 2014; **58**(9): 1126–1142. PMID:25261453
29. Antweiler RC. Group comparisons evaluation of statistical treatments of left-censored environmental data using coincident uncensored data sets. II. *Environ Sci Technol*. 2015; **49**(22): 13439–13446. PMID:26490190, doi:10.1021/acs.est.5b02385
30. Shoari N, Dubé JS. An investigation of the impact of left-censored soil contamination data on the uncertainty of descriptive statistical parameters. *Environ Toxicol Chem*. 2016; **35**(10): 2623–2631. PMID:26946089, doi:10.1002/etc.3420
31. Huynh T, Quick H, Ramachandran G, et al. A comparison of the β-Substitution method and a Bayesian method for analyzing left-censored data. *Ann Occup Hyg.* 2016; **60**(1): 56–73. PMID:26209598
32. Nie L, Chu H, Feng S. Estimating variance parameters from multivariate normal variables subject to limit of detection: MLE, REML, or Bayesian approaches? *Stat Med*. 2009; **28**(20): 2605–2616. PMID:19598183, doi:10.1002/sim.3644
33. Hady DHA, Rania MS. Bayesian and non bayesian parameter estimation for bivariate pareto distribution based on censored samples. Int J Eng Res Appl. 2015; 5(10): 01-06.
34. Feroze N, Aslam M. Comparison of Bayesian and maximum likelihood estimation for analysis of burr type V distribution under left censored samples. *J Stat.* 2015; **22**(1): 116–192.
35. Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. New York: Springer; 2003.
36. Cohen AC. Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics*. 1959; **1**(3): 217–237. doi:10.1080/00401706.1959.10489859
37. Helsel DR, Cohn TA. Estimation of descriptive statistics for multiply censored water quality data. *Water Resour. Res*. 1988; **24**(12): 1997–2004. doi:10.1029/WR024i012p01997
38. Shumway RH, Azari RS, Kayhanian M. Statistical approaches to estimating mean water quality concentrations with detection limits. *Environ Sci Technol*. 2002; **36**(15): 3345–3353. PMID:12188364, doi:10.1021/es0111129
39. Gelman A, Lee D, Guo J. Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*. 2015; **40**(5): 530–543. doi:10.3102/1076998615606113
40. Carpenter B, Gelman A, Hoffman MD, et al. Stan: A probabilistic programming language. *Journal of Statistical Software*. 2017; **76**(1): 1–29. doi:10.18637/jss.v076.i01
41. Land CE. An evaluation of approximate confidence interval estimation methods for lognormal means. *Technometrics*. 1972; **14**(1): 145–158. doi:10.1080/00401706.1972.10488891
42. Ministry of Health, Labour and Welfare. Partial revision of ministerial ordinance concerning ingredient specifications of milk and dairy products (Notification No. 106) and standards of foods and additives (Notification No. 302) [In Japanese]. https://www.mhlw.go.jp/web/t_doc?dataId=00tc3580&dataType=1&pageNo=1. Published on August 2018. Accessed on January 17, 2020.
43. Kruschke JK. Bayesian estimation supersedes the t test. *J Exp Psychol Gen*. 2013; **142**(2): 573–603. PMID:22774788, doi:10.1037/a0029146
44. Hewett P, Ganser GH. A comparison of several methods for analyzing censored data. *Ann Occup Hyg.* 2007; **51**(7): 611–632. PMID:17940277

doi: 10.14252/foodsafetyfscj.D-20-00007

Supporting Information

S1. Description of the Stan Program

The Stan code assuming a lognormal distribution is shown in **Fig. S1**. This code contains 5 blocks (data, transformed data, parameter, model, and generated quantities). In the data block (**Fig. S1**, lines 1-6), we specified 3 data dimensions (sample number of observed (N_{obs}) and censored (N_{cen}) data, and value of observed data (Y_{obs})). Moreover, we specified repetition number to generate random numbers (N_{new}) for 2-dimensional Monte Carlo simulations. In the transformed data block (**Fig. S1**, lines 8–14), we calculated the minimum and maximum value of Y_{obs}. In the parameter block (**Fig. S1**, lines 16–21), we declared 3 parameters (GM, GSD, and RL) and 85 non-detected values (Y_{cen}) to estimate.

In the model block (**Fig. S1**, lines 23–31), Stan can estimate declared parameter(s). In the Bayesian framework, all parameters follow a probability distribution, so a prior distribution must be specified. In lines 24–27 of **Fig. S1**, the informative prior distributions for all parameters are specified. Since GM is equivalent to median of lognormal distribution, we used the normal distribution with mean of half maximum value of Y (Y_{max}) and standard deviation of Y_{max}/4. The natural total Cr content of dissolved waters is approximately 0.02–0.3 μg/L[1]. If the Cr concentration follows a logarithmic distribution and this range covers 80% of the concentration distribution, the GSD is estimated to be 2.9. Thus, we used the lognormal distribution with GM of 3 and GSD of 2 for prior distribution of GSD. It is expected that RL is close to minimum of observed data. We used the normal distribution with mean of minimum value of Y_{obs} ($Y_{obs,\ min}$) and standard deviation of $Y_{obs,\ min}$/5. In lines 29 and 30 of **Fig. S1**, both Y_{obs} and Y_{cen} are specified as stochastically generated from a lognormal distribution with certain parameters, and Stan seeks to find the optimum parameter values from the data.

In the generated quantities block (**Fig. S1**, lines 33–45), we calculated the posterior predictive distribution of log(arithmic) pointwise density for both observed and censored data. For the likelihoods of censored data, we used the cumulative distribution function (CDF). Moreover, in this block, we generated random numbers according to the declared distribution having estimated parameters.

In addition to lognormal distribution, we performed Bayesian estimation under assumption that the original data follow a gamma distribution with the stan code in **Fig. S2**.

S2. Convergence of MCMC

After the MCMC iteration finished, the three parameters (GM, GSD, and RL) declared in the parameter block and the sum of the log(arithmic) posterior predictive probabilities (lp__) had converged (**Figs. S3a and b**). Besides these four values, in all calculation results such as log posterior predictive probabilities, $\hat{R}$, which is the ratio of inter-chain variance to intra-chain variance, satisfied 1.1 or less, which is a general criterion of convergence. In addition, the relative effective sample number and the relative Monte Carlo standard deviation also satisfied general criteria (0.1 or more and 0.1 or less, respectively) (**Fig. S3c**).

In addition, the estimated values about these parameters were similar even assuming different prior distributions (**Table S1**). From these results, the dissociation of the estimated values among the chains and the influence of autocorrelation are small, and each chain started from different initial values and finally arrived at a similar value. We conclude that all the calculated values converged. Even when assuming a gamma distribution, parameter convergence was confirmed under identical conditions.

S3. Calculation of Widely Applicable Information Criterion (WAIC)

The widely applicable information criterion (WAIC) derived by Watanabe[2] as an information criterion for Bayesian inference, is defined as follows:

$$\mathrm{WAIC} = L_{\mathrm{WAIC}} + p_{\mathrm{WAIC}}, \qquad \text{Equation S1}$$

where L_{WAIC} denotes average of the pointwise predictive density, which shows the Bayes training loss, and p_{WAIC} denotes fluctuation of the posterior distribution, which indicates the estimated effective number of parameters.
L_{WAIC} and p_{WAIC} are defined as follows:

$$L_{\mathrm{WAIC}} = -\frac{1}{n}\sum_{i=1}^{n}\log E\left(L_i\right), \qquad \text{Equation S2}$$

$$p_{\mathrm{WAIC}} = \frac{1}{n}\sum_{i=1}^{n}V\left(\log L_i\right), \qquad \text{Equation S3}$$

Where n represents the sample size, L_i represents the point-

```
1.  data {
2.    int N_obs; // sample number of detected data
3.    int N_cen; // sample number of nondetects
4.    real<lower=0> Y_obs[N_obs]; // Values of detected data
5.    int N_new;
6.  }
7.
8.  transformed data {
9.    real<lower=0> Y_obs_min; // minimum of observed value
10.   real<lower=0> Y_obs_max; // maximum of observed value
11.
12.   Y_obs_min = min(Y_obs);
13.   Y_obs_max = max(Y_obs);
14. }
15.
16. parameters {
17.   real<lower=0> GM; // location parameter, geometric mean
18.   real<lower=1> GSD; // shape parameter, geometric standard deviation
19.   real<lower=0, upper=Y_obs_min> RL; // Reporting limit
20.   real<lower=0, upper=RL> Y_cen[N_cen]; // Values of nondetects
21. }
22.
23. model {
24.   GSD ~ lognormal( log(3), log(2) );
25.   GM ~ normal(Y_obs_max/2, Y_obs_max/4);
26.   RL ~ normal(Y_obs_min, Y_obs_min/5);
27.   Y_cen ~ uniform(0, Y_obs_min);
28.
29.   Y_obs ~ lognormal( log(GM), log(GSD) );
30.   Y_cen ~ lognormal( log(GM), log(GSD) );
31. }
32.
33. generated quantities {
34.   real log_lik_obs[N_obs];
35.   real log_lik_cen[N_cen];
36.   vector[N_new] Y_new;
37.   real mean_est;
38.   real sd_est;
39.
40.   for (i in 1:N_obs) log_lik_obs[i] = lognormal_lpdf(Y_obs[i] | log(GM), log(GSD));
41.   for (i in 1:N_cen) log_lik_cen[i] = lognormal_lcdf(RL | log(GM), log(GSD));
42.   for (i in 1:N_new) Y_new[i] = lognormal_rng( log(GM), log(GSD) );
43.   mean_est = exp( log(GM) + 0.5*( log(GSD) )^2 );
44.   sd_est = sqrt( exp( 2 *log(GM) + (log(GSD))^2 ) * (exp( (log(GSD))^2 )-1) );
45. }
46.
```

Fig. S1. Stan code for estimating parameters from left-censored data assuming a lognormal distribution

wise predictive density for each data point i obtained from the posterior predictive distribution, $E(X)$ indicates the mean of variable X, and $V(X)$ indicates the variance of variable X. To calculate complete likelihood according to the likelihood function (Equation 1), we used log(arithmic) pointwise density (**Fig. S1**, line 40) for log L_i of observed data and log(arithmic) lower-tail probability (**Fig. S1**, line 41) for log L_i of censored data.

Unlike other criteria, WAIC has a theoretical foundation as follows[2]:

$$E(G) \equiv E(\mathrm{WAIC}) + O\left(\frac{1}{n}\right), \qquad \text{Equation S4}$$

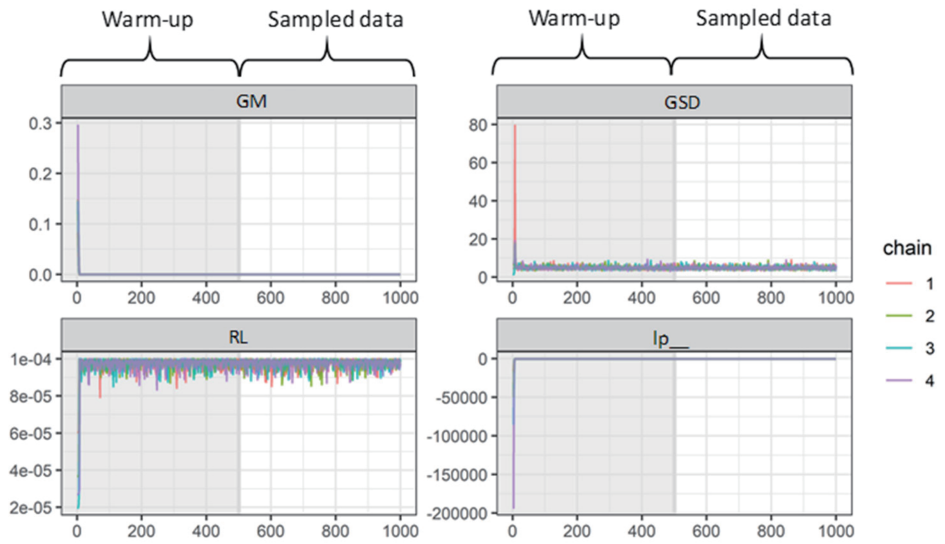
```
data {
  int N_obs; // sample number of observed data
  int N_cen; // sample number of censored data
  real<lower=0> Y_obs[N_obs]; // Values of detected data
  int N_new;
}

transformed data {
  real<lower=0> Y_obs_min; // minimum of observed value
  real<lower=0> Y_obs_max; // maximum of observed value

  Y_obs_min = min(Y_obs);
  Y_obs_max = max(Y_obs);
}

parameters {
  real<lower=0> shape; // shape parameter of gamma distribution
  real<lower=0> rate; // rate parameter of gamma distribution
  real<lower=0, upper=Y_obs_min> RL; // Reporting limit
  real<lower=0, upper=RL> Y_cen[N_cen]; // Values of nondetects
}

model {
  shape ~ double_exponential(1, 5);
  target += normal_lpdf( shape/rate | Y_obs_max/2, Y_obs_max/4 );
  RL ~ normal(Y_obs_min, Y_obs_min/5);
  Y_cen ~ uniform(0, Y_obs_max);

  Y_obs ~ gamma(shape, rate);
  Y_cen ~ gamma(shape, rate);
}

generated quantities {
  real log_lik_obs[N_obs];
  real log_lik_cen[N_cen];
  vector[N_new] Y_new;
  real mean_est;
  real sd_est;

  for (i in 1:N_obs) log_lik_obs[i] = gamma_lpdf( Y_obs[i] | shape, rate);
  for (i in 1:N_cen) log_lik_cen[i] = gamma_lcdf( RL | shape, rate);
  for (i in 1:N_new) Y_new[i] = gamma_rng(shape, rate);
  mean_est = shape/rate;
  sd_est = sqrt(shape)/rate;
}
```

Fig. S2. Stan code for estimating parameters from left-censored data assuming a gamma distribution

Where G represents Bayes generalization loss and $O(g(x))$ represents mathematical notation that describes the limiting behavior of a function ($f(x)$) when the argument tends toward a particular value or infinity. The expectation value of the Bayes generalization loss is asymptotically equal to WAIC. And Watanabe[2] also showed the cross-validation loss is asymptotically equivalent to the WAIC, even in singular statistical model. WAIC can be used for any true distribution, probability model, and prior distribution. The WAIC is a generalized version of Akaike's information criterion (AIC) and can be calculated from posterior predictive distributions by MCMC.

S4. Probability Distribution Assumption

(a)



(b) Convergence indices for MCMC parameters (iterations=2000, warm-up=1000, chain=4, thinning=2)

Parameter	Mean	SD	$\hat{R}$	n_{eff}/N	mcse/sd
GM	0.082×10^{-3}	0.014×10^{-3}	1.001	0781	0.025
GSD	4.90	0.80	0.999	0.739	0.026
RL	0.098×10^{-3}	0.002×10^{-3}	0.9996	0.874	0.024
lp__	−225	8.4	1.001	0.475	0.032

(c)

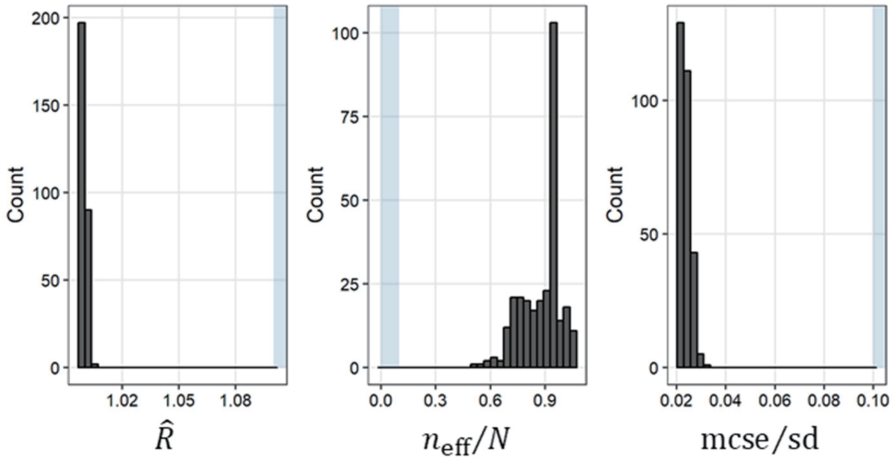


Fig. S3. Confirmation of MCMC convergence. (a) Trace plot for declared parameters (GM, GSD, and RL) in stan code (**Fig. 1**) and the sum of log posterior predictive probabilities (lp__); (b) Summary of estimation and convergence indices for parameters and lp__; (c) Histograms of three convergence indices ($\hat{R}$, n_{eff}/N, and mcse/sd) for all calculated values, where $\hat{R}$ is the ratio of intra-chain variation to inter-chain variance, n_{eff}/N is the ratio of effective sample number to MCMC sample, and mcse/sd is the ratio of Monte Carlo standard error to standard deviation. If calculated data are not distributed in the light blue area, it can be concluded that MCMC calculations have converged.

doi: 10.14252/foodsafetyfscj.D-20-00007

Table S1. MCMC results assuming a lognormal distribution using various prior distributions. Mean ± standard deviation and 95% credible interval are shown.

Prior distribution	GM (× 10^{-3})	GSD	WAIC
GM ~ normal(Y_max/2, Y_max/4)[a*] GSD ~ lognormal(log(3), log(2))[b*]	0.082±0.014 0.055–0.111	4.90±0.80 3.65–6.71	−2.28
GM ~ cauchy(0, 5)[c] GSD ~ lognormal(log(3), log(2))	0.081±0.013 0.057–0.108	4.94±0.80 3.76–6.90	−2.28
GM ~ uniform(0, 10^5)[d] GSD ~ uniform(0, 10^5)	0.078±0.014 0.052–0.106	5.24±0.98 3.88–7.61	−2.28
GM ~ normal(0, 10^5) GSD ~ normal(0, 10^5)	0.078±0.014 0.054–0.109	5.20±0.93 3.83–7.36	−2.28
GM ~ cauchy(0, 5) GSD ~ cauchy(0, 5)	0.079±0.014 0.053–0.108	5.08±0.87 3.81–7.04	−2.28

a: "X ~ normal(μ, σ)" means that variable X follows a normal distribution with location parameter μ and shape parameter σ.
b: "X ~ lognormal(μ, σ)" means that variable X follows a lognormal distribution with location parameter μ and shape parameter σ.
c: "X ~ cauchy(μ, σ)" means that variable X follows a Cauchy distribution with location parameter μ and shape parameter σ.
d: "X ~ uniform(a, b)" means that variable X follows a uniform distribution with range from a to b.
* The prior distributions used in **Fig. 1**.

Some EPA guidance documents strongly recommend against using a lognormal model for environmental data and instead suggest a gamma distribution[3]. We examined whether a lognormal or a gamma distribution is more appropriate for the original data. The WAIC assuming a lognormal and a gamma distribution resulted in similar values for WAIC (**Table S2**).

The WAIC (−2.28) of original data assuming a lognormal distribution was in good agreement with WAICs (−2.33±0.09) from 1000 randomly generated left-censored data subsets (n = 150) which follow a lognormal distribution with certain parameters (GM = 0.082×10^{-3}; GSD = 4.9) at a censoring ratio of 85/150 (**Fig. S4**). This result indicate that the data fit a lognormal distribution reasonably well.

Although WAIC assuming a lognormal distribution was slightly higher than that assuming a gamma distribution (−2.29), to compare as many estimation methods as possible, we proceeded with further analysis under the assumption that the original data follow a lognormal distribution.

S5. Cox's Method for 95%CIs for Means Which Follow a lognormal Distribution

The 95% CIs for means estimated by DN and RL/2 were calculated by Cox's method[4] as follows:

$$\left[\hat{\mu}/\exp\left\{1.96\sqrt{\frac{\hat{s}^2}{n}+\frac{\hat{s}^4}{2(n-1)}}\right\},\ \hat{\mu}\times\exp\left\{1.96\sqrt{\frac{\hat{s}^2}{n}+\frac{\hat{s}^4}{2(n-1)}}\right\}\right], \quad \text{Equation S5}$$

where $\hat{\mu}$ indicates the estimated mean, $\hat{s}$ denotes the log-transformed GSD estimate, and n is total observations.

S6. Estimation Using Left-censored Data

S6.1 The Kaplan-Meier Algorithm

The Kaplan-Meier (KM) method is a nonparametric technique for dealing with censored data. It is widely used in survival and lifetime data analysis to estimate survival functions, which are then used to estimate different summary statistics. The KM method can be summarized via the following steps:[5]

1) This method requires the use of right-censored data. Therefore, right-censored data (Flip_i) are constructed by subtracting all observations (x_i) from M.

$$\text{Flip}_i = M - x_i, \quad \text{Equation S6}$$

Where M equals a flipping constant larger than the maximum of x_i.

2) The survival function probability ($S(t)$) is the product of the j = 1 to k incremental survival probabilities to the point, proceeding from high to low concentration for the k uncensored observations.

$$S(t_j) = \prod_{j=1}^{k} \frac{b_j - d_j}{b_j}, \quad \text{Equation S7}$$

Where t_j is the flipped uncensored observations; b_j is the

Table S2. MCMC results assuming both lognormal and gamma distributions for estimating censored values. Widely applicable information criterion, mean ± standard deviation, maximum a posteriori (MAP), and 95% credible interval (CrI) of posterior predictive distribution for target parameters

	Lognormal	Gamma
WAIC	−2.28	−2.29
Parameter	RL ($\times 10^{-3}$) 0.098±0.002 MAP: 0.099_3, 95% CrI: 0.091–0.099_9	RL ($\times 10^{-3}$) 0.096±0.004 MAP: 0.099_0, 95% CrI: 0.087–0.099_9
	GM ($\times 10^{-3}$) 0.082±0.014 MAP: 0.080, 95% CrI: 0.055–0.111	Shape 0.353±0.059 MAP: 0.354, 95% CrI: 0.254–0.487
	GSD 4.90±0.80 MAP: 4.60, 95% CrI: 3.65–6.71	Rate 1625±347 MAP: 1589, 95% CrI: 1019–2378
	Mean ($\times 10^{-3}$) 0.290±0.71 MAP: 0.249, 95% CrI: 0.197–0.459	Mean ($\times 10^{-3}$) 0.221±0.31 MAP: 0.212, 95% CrI: 0.168–0.288
	SD ($\times 10^{-3}$) 1.09±0.70 MAP: 0.715, 95% CrI: 0.453–2.88	SD ($\times 10^{-3}$) 0.375±0.60 MAP: 0.353, 95% CrI: 0.275–5.16

number of observations, both detected and censored, at and higher than each t_j; and d_j is the number of uncensored observations at that value.

3) The KM estimate of the mean is calculated by integrating the area under the KM survival curve. Since this is the average value for flipped right-censored data, the average value ($\hat{\mu}^{\mathrm{KM}}$) for the original left-censored data is calculated as follows:

$$\hat{\mu}^{\mathrm{KM}} = M - \int_0^{t_{\max}} \hat{S}(t)\,dt \approx M - \sum_{j=1}^{k} \left\{ S\left(t_{j-1}\right) \times \left(t_j - t_{j-1}\right) \right\}. \qquad \text{Equation S8}$$

4) Following Lee and Wang[6], the variance of the mean is computed using the formula inside the square root sign of the following equation for data with *m* censored values under the Kaplan-Meier curve:

$$\text{Standard error}^{\mathrm{KM}} = \sqrt{\left(\frac{m}{m-1}\right) \sum_{r=1}^{m} \frac{{A_r}^2}{(n-r)(n-r+1)}}, \qquad \text{Equation S9}$$

Where *n* is total observations and A_r is the cumulative area. Finally, the standard deviation ($\hat{\sigma}^{\mathrm{KM}}$) is calculated as follows:

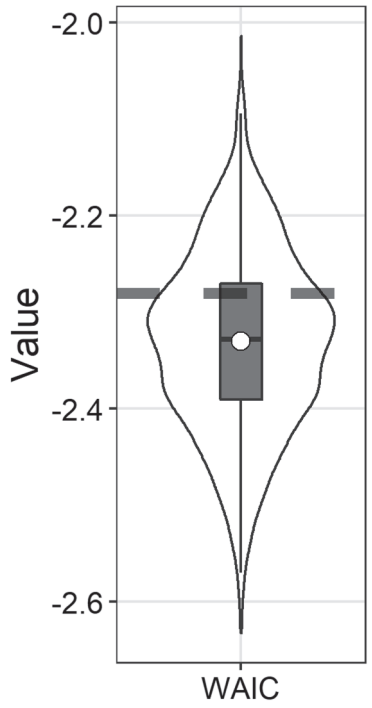


Fig. S4. Violin plot of WAICs from 1000 randomly generated left-censored data subsets (n = 150) which follow a lognormal distribution with certain parameters (GM = 0.082×10^{-3}; GSD = 4.9) at a censoring ratio of 85/150. Horizontal dashed lines indicate WAIC calculated from original data. Violins, boxes, horizontal solid lines, and open circles indicate probability density, interquartile range, median, and arithmetic mean for simulation data, respectively.

doi: 10.14252/foodsafetyfscj.D-20-00007

$$\hat{\sigma}^{\mathrm{KM}} = \sqrt{n} \times \text{Standard error}^{\mathrm{KM}}. \quad \text{Equation S10}$$

In the KM method, the location parameter of the geometric mean (GM) and the shape parameter of the geometric standard deviation (GSD) for a lognormal distribution were calculated using $\hat{\mu}^{\mathrm{KM}}$ and $\hat{\sigma}^{\mathrm{KM}}$ with the following equations.

If x is a random variable following a lognormal distribution with certain parameters (GM and GSD), and GM and GSD are expressed as e^m and e^s, respectively, and the mean $E(x)$ and variance $V(x)$ are expressed as follows:

$$E(x) = \exp\left[m + \frac{s^2}{2}\right], \quad \text{Equation S11}$$

$$V(x) = \exp\left[2m + s^2\right]\left(\exp\left[s^2\right] - 1\right). \quad \text{Equation S12}$$

The GM and GSD parameters estimated by the KM method assuming a lognormal distribution can be transformed as follows:

$$\hat{m}^{\mathrm{KM}} = \log \frac{\hat{\mu}^{\mathrm{KM}}}{\sqrt{\frac{\left(\hat{\sigma}^{\mathrm{KM}}\right)^2}{\left(\hat{\mu}^{\mathrm{KM}}\right)^2} + 1}}, \quad \text{Equation S13}$$

$$\hat{s}^{\mathrm{KM}} = \sqrt{\log\left(\frac{\left(\hat{\sigma}^{\mathrm{KM}}\right)^2}{\left(\hat{\mu}^{\mathrm{KM}}\right)^2} + 1\right)}. \quad \text{Equation S14}$$

S6.2 The Robust Regression on Order Statistics Algorithm

There are two versions of ROS, one is a fully parametric and the other is a semi-parametric robust implementation ROS. These are sometimes confused in the literature[5]. In this study, we used robust ROS (rROS) developed by Helsel and Cohn[7]. Robust ROS is also known as "Imputation Using Quantile-Quantile Regression". This method involves using quantile-quantile regression on the log-transformed observations to fit a regression line. Using this method, a more limited assumption of normal or lognormal distribution is used. The algorithm for the ROS method can be summarized according to the following steps:

1) In general, the probability of exceeding the j^{th} detection limit ($^{\mathrm{E}}P_j$) is

$$^{\mathrm{E}}P_j = {}^{\mathrm{E}}P_{j+1} + \frac{A_j}{A_j + B_j}\left[1 - {}^{\mathrm{E}}P_{j+1}\right], \quad \text{Equation S15}$$

where Aj is the number of observations in the range of j^{th} and $(j + 1)^{\mathrm{th}}$ reporting limits, and Bj is the number of observations, censored and uncensored, below the j^{th} reporting limit.

When j is the highest reporting limit, $^{\mathrm{E}}P_{j+1} = 0$ and $A_j+B_j=n$. The number of censored observations below the j^{th} reporting limit is defined as C_j:

$$C_{\mathrm{j}} = B_j - B_{\mathrm{j-1}} - A_{j-1}. \quad \text{Equation S16}$$

2) Plotting positions for uncensored observation $^{\mathrm{D}}p_j$ can be calculated by

$$^{\mathrm{D}}p_j = \left(1 - {}^{\mathrm{E}}P_j\right) + \left(\frac{i}{A_j + 1}\right)\left[{}^{\mathrm{E}}P_j - {}^{\mathrm{E}}P_{j+1}\right] \text{ for } i = 1 \text{ to } A_j, \quad \text{Equation S17}$$

and for censored observations $^{\mathrm{C}}p_{\mathrm{j}}$ are generally given by

$$^{\mathrm{C}}p_j = \left(\frac{i}{C_j + 1}\right)\left[1 - {}^{\mathrm{E}}P_j\right] \text{ for } i = 1 \text{ to } C_j. \quad \text{Equation S18}$$

3) The normal quantiles of the plotting positions are known as the order statistics of the ROS method. Assuming a lognormal distribution, a linear regression of the uncensored observations against the normal quantiles of the uncensored plotting positions is formed as

$$\log\left(x_{\mathrm{obs}}\right) = \beta + \alpha \times {}^{\mathrm{D}}p_j. \quad \text{Equation S19}$$

Then the regression equation for predicting the unobserved data can be obtained as

$$\log\left(\hat{x}_{\mathrm{cens}}\right) = \beta + \alpha \times {}^{\mathrm{C}}p_j. \quad \text{Equation S20}$$

4) Using the log-transformed detected data ($\log\left(x_{\mathrm{obs}}\right)$) and the predicted log-value ($\log\left(\hat{x}_{\mathrm{cens}}\right)$), calculate the mean and standard deviation for the log-transformed x, and finally estimates of the geometric mean ($\hat{m}^{\mathrm{ROS}}$) and geometric standard deviation ($\hat{s}^{\mathrm{ROS}}$) are obtained. After retransforming the log-scale imputed values, compute the usual method of moments estimates of $\hat{\mu}^{\mathrm{ROS}}$ and $\hat{\sigma}^{\mathrm{ROS}}$.

Gilliom and Helsel[8] reported that the robust ROS method perform better on high-skew distributions than did MLE. Helsel and Cohn[7] reported that the robust ROS generally produced better estimates for the mean and standard devia-

tion whenever data did not follow the distribution assumed by maximum likelihood.

S6.3 The Maximum Likelihood Algorithm

Maximum likelihood estimation solves a likelihood equation to estimate the parameter(s) using both detected observations and the proportion of data falling below RL[9]. The observed data (x) enter the likelihood function through the probability density function $f(x|\theta)$ and the censored observations can be accounted for by the cumulative distribution function $F(\mathrm{RL}|\theta) = P(x \leq \mathrm{RL}|\theta)$ as follows:

$$L(\theta|x_1, x_2, \cdots, x_n) = \prod_{x \in D} f(x|\theta) \cdot \prod_{x \in C} F(\mathrm{RL}|\theta), \qquad \text{Equation S21}$$

where D is the set of all observed values; and C is the set of all left-censored values. MLE finds the parameter values (θ) that maximize the likelihood function against the observations. In the case of lognormal distribution, $f(x|m_{\log}, s_{\log})$ and $F(\mathrm{RL}|m_{\log}, s_{\log})$ are defined as follows:

$$f(x|m_{\log}, s_{\log}) = \frac{1}{\sqrt{2\pi}\, s_{\log} x} \exp\left\{ -\frac{(\log x - m_{\log})^2}{2 s_{\log}^2} \right\}, \qquad \text{Equation S22}$$

$$F(\mathrm{RL}|m_{\log}, s_{\log}) = \frac{1}{2} \mathrm{erfc}\left(-\frac{\log x - m_{\log}}{\sqrt{2} s_{\log}} \right), \qquad \text{Equation S23}$$

where $m_{\log}$ is log-transformed geometric mean, $s_{\log}$ is log-transformed geometric standard deviation, and erfc(x) indicates complementary error function.

In the EnvStat package, a transformation is adopted as follows[10]. Let x be a vector of n observations from a lognormal distribution with location parameter μ_x (mean) and scale parameter σ_x (standard deviation). Set $y = \log(x)$. Then y is a vector of observations from a normal distribution with location parameters $\mu_y = m$ ($\mathrm{GM}_x = e^m$) and scale parameter $\sigma_y = s$ ($\mathrm{GSD}_x = e^s$). The maximum likelihood estimators of $\hat{\mu}_x^{\mathrm{MLE}}$ and $\hat{\sigma}_x^{\mathrm{MLE}}$ are given by

$$\hat{\mu}_x^{\mathrm{MLE}} = \exp\left[\hat{m}^{\mathrm{MLE}} + \frac{\left(\hat{s}^{\mathrm{MLE}}\right)^2}{2} \right], \qquad \text{Equation S24}$$

$$\hat{\sigma}_x^{\mathrm{MLE}} = \hat{\mu}_x^{\mathrm{MLE}} \sqrt{\left(\hat{\sigma}^{\mathrm{MLE}}\right)^2 - 1}, \qquad \text{Equation S25}$$

where $\hat{\mu}^{\mathrm{MLE}}$ and $\hat{\sigma}^{\mathrm{MLE}}$ denote the maximum likelihood estimators of μ and σ, respectively.

S7. Confidence Interval and Credible Interval

A credible interval (CrI) is an important concept in Bayesian statistics to describe and summarize the uncertainty. In this regards, CrI is quite similar to the frequentist "confidence Intervals (CI)". However, while their goal is similar, their statistical meaning is different.

- 95% CI: with a large number of repeated samples, 95% CI represents 95% frequency (i.e. 95% proportion) of possible confidence intervals that contain the true estimate of the unknown parameter.
- 95% CrI: given the observed data, there is a 95% probability that the true estimate of unknow parameter would lie within the 95% CrI.

Thus, it is inappropriate to compare CI and CrI directly. In this study, to compare uncertainty of mean estimates, we calculated the coverage probability of 95% CIs for mean (CP), which indicates the fraction of computed CIs and CrIs that include the desired but unobservable parameter value.

S8. Comparison between Japanese and Imported MW Products

To compare Cr(VI) concentration in Japanese MW products and imported ones, we used another Stan code (**Fig. S5**). **Fig. S6A** presents histograms of the Cr(VI) concentrations of Japanese and imported MW products based on empirical data. **Table S3** shows the estimated parameters. Both GM and GSD parameters of Japanese and imported MW products (**Table S3**) showed good agreement within $\pm 1\sigma$ with those in **Table 1**.

Fig. S6B shows the distribution of expected values (**Fig. S5**, lines 47 and 48). Although the distribution of mean Cr(VI) concentrations in imported products was wider, the probability that the expected value of Cr(VI) concentrations in imported MW products is higher than in Japanese products was 0.632. Kataoka et al[11] investigated imported MW products from 14 countries and, based on that, the present results may reflect differences in concentration depending on the country of origin.

Finally, 10^5 random numbers for Cr(VI) concentrations in imported MW and Japanese products were generated, and

```
1.  data {
2.    int N_obs_F; // Obseved sample size of foreign domestic product
3.    int N_obs_J; // Obseved sample size of Japanese product
4.    int N_cen_F; // Sample size of nondetects of foreign domestic product
5.    int N_cen_J; // Sample size of nondetects of Japanese product
6.    real<lower=0> Y_min; // Minimum value of observed data
7.    vector<lower=Y_min>[N_obs_F] Y_obs_F; // Observed data of foreign domestic product
8.    vector<lower=Y_min>[N_obs_J] Y_obs_J; // Observed data of Japanese product
9.    int N_new;
10. }
11.
12. parameters {
13.   real<lower=0> GM_F; // location parameter of foreign domestic product
14.   real<lower=0> GM_J; // location parameter of Japanese product
15.   real<lower=1> GSD_F; // shape parameter of foreign domestic product
16.   real<lower=1> GSD_J; // shape parameter of Japanese product
17.   real<lower=0, upper=Y_min> RL; // Value of limit of quantification
18.   vector<lower=0, upper=RL>[N_cen_F] Y_cen_F; // Censored data of foreign domestic product
19.   vector<lower=0, upper=RL>[N_cen_J] Y_cen_J; // Censored data of Japanese product
20. }
21.
22. model {
23.   RL ~ normal(0.000098, 0.0000024);
24.   GSD_F ~ normal(4.9, 0.8);
25.   GSD_J ~ normal(4.9, 0.8);
26.   GM_F ~ normal(8.2*10^-5, 1.4*10^-5);
27.   GM_J ~ normal(8.2*10^-5, 1.4*10^-5);
28.
29.   Y_obs_F ~ lognormal( log(GM_F), log(GSD_F) );
30.   Y_cen_F ~ lognormal( log(GM_F), log(GSD_F) );
31.   Y_obs_J ~ lognormal( log(GM_J), log(GSD_J) );
32.   Y_cen_J ~ lognormal( log(GM_J), log(GSD_J) );
33. }
34.
35. generated quantities {
36.   real Y_F[N_new];
37.   real Y_J[N_new];
38.   real mean_est_F;
39.   real mean_est_J;
40.
41.   for ( i in 1:N_new) {
42.     Y_F[i] = lognormal_rng( log(GM_F), log(GSD_F) );
43.     Y_J[i] = lognormal_rng( log(GM_J), log(GSD_J) );
44.   }
45.   mean_est_F = exp( log(GM_F) + 0.5*( log(GSD_F) )^2 );
46.   mean_est_J = exp( log(GM_J) + 0.5*( log(GSD_J) )^2 );
47. }
48.
```

Fig. S5. Stan code for estimating parameters from left-censored data assuming a lognormal distribution: comparing differences between Japanese and imported mineral water products

the difference between the groups was verified. Histograms of the generated data are shown in **Fig. S6C**. These results suggest that the probability of Cr(VI) concentrations being higher in imported MW products compared to Japanese products was 0.507.

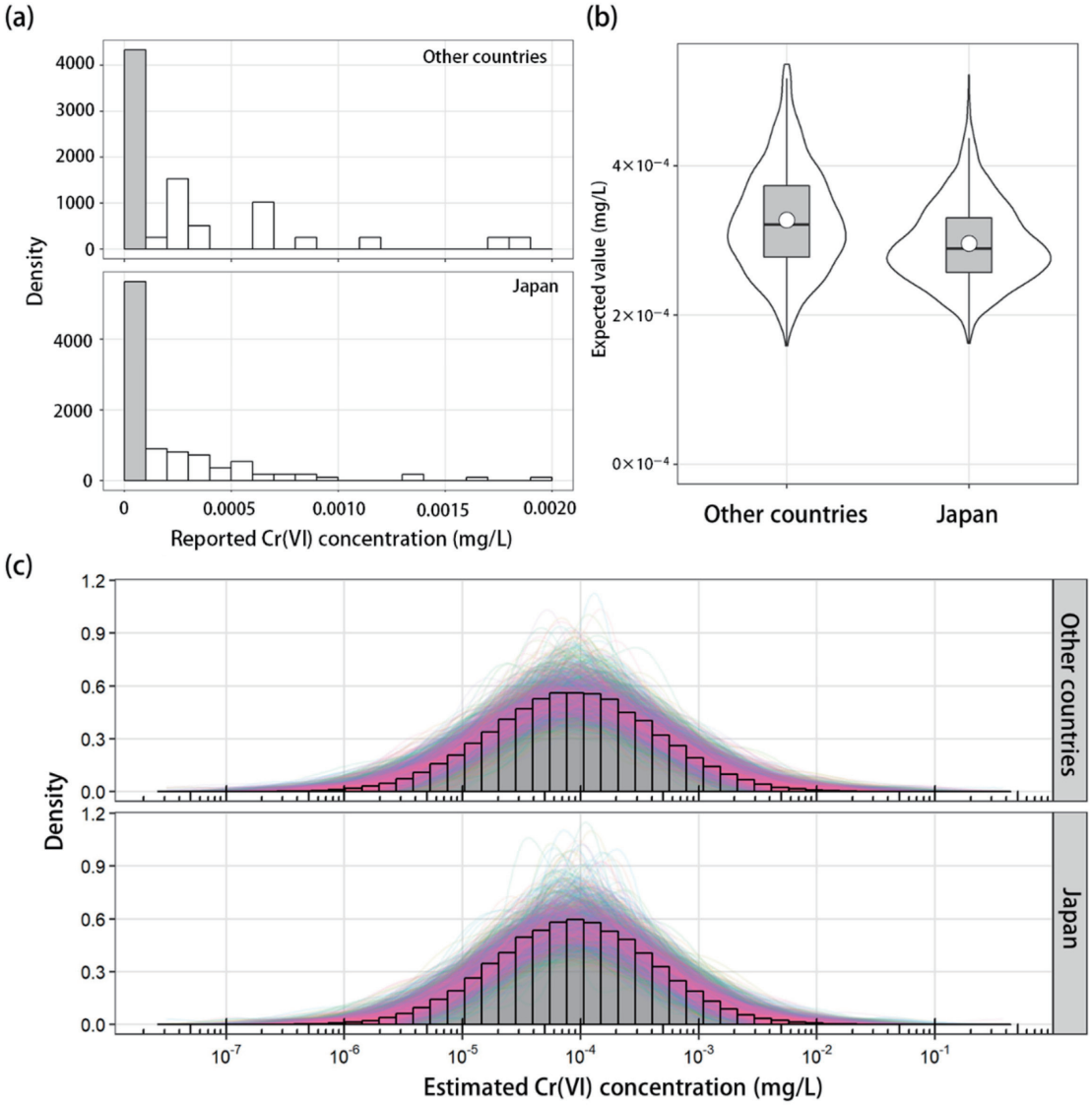


Fig. S6. Histograms of Cr(VI) concentrations in mineral water products from Japan and other countries. (a) Original data from Kataoka et al (2017). Gray bars indicate nondetects; (b) Violin plots for expected Cr(VI) concentrations in Japanese and imported mineral water products; (c) 50 random numbers generated by 2-dimensional Monte Carlo simulations from 2000 parameters estimated by MCMC. Color indicates density curves obtained from 2000 parameters.

Table S3. Group comparison of estimated parameters and summary statistics by MCMC assuming a lognormal distribution. Mean ± standard deviation and 95% credible interval are shown.

Group	GM ($\times 10^{-3}$)	GSD	Expected value ($\times 10^{-3}$ mg/L)	Predicted concentration[a] ($\times 10^{-3}$ mg/L)
Imported MW products	0.088±0.013 0.065–0.113	5.07±0.60 3.91–6.29	0.331±0.074 0.205–0.508	0.333±1.442 0.003–2.111
Japanese MW products	0.089±0.011 0.068–0.110	4.71±0.54 3.74–5.82	0.296±0.054 0.209–0.418	0.297±1.474 0.004–1.825

a: 50 random numbers generated by 2-dimensional Monte Carlo simulations from 2000 parameters estimated by MCMC.

References for supporting information

1. The World Health Organization. Chromium in drinking-water background document for development of WHO Guidelines for drinking-water quality. 2003. https://www.who.int/water_sanitation_health/dwq/chemicals/chromium.pdf.
2. Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular Learning Theory. *J Mach Learn Res.* 2010; **11**: 3571–3594.
3. US Environmental Protection Agency. ProUCL Version 5.1.002 Technical Guide, EPA/600/R-07/041. Office of Research and Development, Washington, DC. October. 2015.
4. Land CE. An evaluation of approximate confidence interval estimation methods for lognormal means. *Technometrics*. 1972; **14**(1): 145–158. doi:10.1080/00401706.1972.10488891
5. Helsel DR. Statistics for Censored Environmental Data Using Minitab® and R, 2nd edition. Hoboken, NJ: Wiley; 2012.
6. Lee ET, Wang JW. Statistical Methods for Survival Data Analysis. 3rd ed. Hoboken, NJ: John Wiley & Sons, Inc.; 2003.
7. Helsel DR, Cohn TA. Estimation of descriptive statistics for multiply censored water quality data. *Water Resources Research*. 1988; **24**(12): 1997–2004. doi:10.1029/WR024i012p01997
8. Gilliom RJ, Helsel DR. Estimation of distributional parameters for censored trace level water quality data: 1. Estimation Techniques. *Water Resources Research*. 1986; **22**(2): 135–146. doi:10.1029/WR022i002p00135
9. Helsel DR. More than obvious: better methods for interpreting nondetect data. *Environ Sci Technol*. 2005; **39**(20): 419A–423A. PMID:16295833, doi:10.1021/es053368a
10. Package 'EnvStats': Package for Environmental Statistics, Including US EPA Guidance. Version 2.3.12018. Millard, NY: Springer; 2013.
11. Kataoka Y, Watanabe T, Hayashi K, Ozawa R, Takizawa K, Akiyama H. Surveillance of chromium(VI) concentrations in mineral water products. *Shokuhin Eiseigaku Zasshi*. 2017; **58**(6): 275–280. doi:10.3358/shokueishi.58.275