



Statistical assessment of data sets for indoor air and house dust from the environmental survey GerES V using non-linear regression analysis, non-parametric methods and Monte-Carlo simulations

Tunga Salthammer ^a, Anja Daniels ^b, Wolfram Birmili ^b

^a Fraunhofer WKI, Department of Material Analysis and Indoor Chemistry, 38108 Braunschweig, Germany

^b German Environment Agency (Umweltbundesamt), Corrensplatz 1, 14195 Berlin, Germany

ARTICLE INFO

Dataset link: <https://doi.org/10.7797/17-2014-17-1-1-1-umwelt>

Keywords:

Environmental survey
Descriptive statistics
Lognormal distribution
Bootstrap
Confidence interval
Monte-Carlo simulation

ABSTRACT

Environmental surveys are essential tools to investigate the impact of pollutants on the living and non-living environment. Their data often form a basis to derive recommendations for preventive measures protecting health and ecosystems and identify the need for political action. Therefore, representative environmental data sets need to be free of systematic artifacts; their statistical structure should be explored and understood as good as possible. The German Environmental Survey (GerES) is a nationwide study conducted at irregular intervals. The data collected within the GerES V (2014–2017) campaign is important for recording and assessing pollutants in households with children and adolescents. Due to its sampling characteristics, GerES claims to be representative of the population in Germany and, with its standardized measurements and sampling protocols, as well as the selection of sampling points, provides a prime example of a study on the statistical nature of pollutants concentrations. In this work data sets from 19 pollutants in indoor air and house dust were selected from the GerES V pool. The parameters obtained from descriptive statistics were compared with the modeled data of a lognormal probability function and a lognormal cumulative density function. Confidence intervals were calculated using a bootstrap method. Monte-Carlo simulations were used to quantify uncertainties in estimators of theoretical distribution assumptions and to investigate the influence of classifying data into equidistant intervals (bins) on nonlinear regression analysis with respect to data count and bin width. The results of our study provide better insight into the general statistical nature of environmental observations, enabling a more reliable assessment of the parameters derived from the data.

1. Introduction

In the environment, the distribution of pollutant concentrations can generally be described by statistical laws. The shapes of the statistical distributions of environmental observations result from the underlying physical and chemical processes in the system but may also be influenced by the study design and the chosen sampling strategy or by random errors. Several shapes of distributions have been described and motivated in the literature, such as the normal, lognormal, gamma, Weibull and exponential distribution (see, e.g., Ott [1] for an overview). Identifying shapes of the distributions of environmental pollutants is necessary for data analysis, exposure assessment, risk management, and policy interpretation.

In many experimental applications, normal (Gaussian) distributions are observed. In examples of environmental sampling related to pollution dispersal, in contrast, lognormal distributions have been found [2]. This is because environmental data, especially pollutant concentrations,

do not assume negative values, and outliers are generally right-skewed, which can be explained by the activity and applicability of multiplicative factors. For pollutants in air and water, dilution is such a factor.

To simulate stochastic processes, the Galton board can be used as a simplistic mechanical model: Vertically falling bullets pass a grid of pegs and are collected at the bottom in n equidistantly arranged bins. The distribution of the bullets in the bins is based on the Pascal's triangle and the underlying probability distribution is binomial. As the number of repetitions of the experiment increases, the binomial distribution approximates the continuous normal (Gaussian) distribution for $n \rightarrow \infty$ according to the central limit theorem.

In the 19th century, this often led to the erroneous opinion that all natural processes are symmetrically distributed. Galton himself [3], referring to McAllister [4], argued that this is incorrect, since in many cases experimental data tend to be better described by the geometric

* Corresponding author.

E-mail address: tunga.salthammer@wki.fraunhofer.de (T. Salthammer).

mean (μ_g), which leads towards the assumption of a lognormal distribution, rather than the arithmetic mean (μ), which best represents a normal distribution. Limpert et al. [2] have summarized lognormally distributed processes for various scientific disciplines and presented a variation of the Galton board that can be used to simulate a lognormal probability distribution. In general, probability models play an important role in the natural sciences because they allow statistically based processes to be easily simulated, explained and understood [5,6].

In 1976, Ott and Mage [7] made the fundamental statement: “Univariate probability models are of considerable importance in the analysis and interpretation of environmental monitoring data for decision making purposes”. Henceforth, statistically based reference values must be calculated using statistically correct models and parameters. Ott and Mage [7,8] were also among the first to investigate the underlying physical laws which generate pollutant concentrations and to derive a probability model for environmental data analysis. A critical discussion of the methods and practical limitations for describing air pollutant concentrations by statistical distributions was published by Georgopoulos and Seinfeld in 1982 [9].

It was again Ott [10] who provided a physical explanation of the lognormality of pollutant concentrations. He used the model of the multi-stage dilution of a water-soluble pollutant in beakers, whereby statistical errors occur when a given volume is added to the next beaker. Ott showed that, due to these errors (or uncertainties), the concentration in the last beaker becomes lognormally distributed if the experiment is repeated often enough. He applied this “Theory of Successive Random Dilutions” to various scenarios, such as the dilution of a pollutant in room air through ventilation. In a later monograph, Ott [1] discussed various forms of the lognormal distribution and their application to environmental data. Nevertheless, there are relatively few studies that address the true statistical nature of indoor pollutant concentration data [11,12]. Only radon has been comparatively well examined, as large data sets are usually available for this chemical element [13,14].

Although it is now clear that data from comprehensive environmental observations are usually not normally distributed, the arithmetic mean (μ) is often given as the sole parameter in statistical analyses and for comparing data sets [15]. Others also base their results and discussions on the arithmetic mean, but additionally provide the geometric mean and/or percentiles, sometimes in graphical form as box-whisker plots [16], sometimes in tabular form [17–21]. In principle, it is not a problem to calculate the arithmetic mean as long as the median is also specified, because by comparing both parameters one can conclude the skewness of a distribution and thus the deviation from the normal distribution. Results of dust measurements often include the median, the 95th percentile and the range between minimum and maximum [22,23]. Reasons for the frequent application of the arithmetic mean could be its use as an indicator of cumulative (additive) exposure to pollutants and, partly, a desire to keep consistency with previously published data reporting the arithmetic mean. Another reason could be that the tools for inductive statistics are more popular when a normal distribution of the data is assumed. However, the arithmetic mean alone is meaningless for skewed distributions and might contribute to a misinterpretation of the data. A review of the aforementioned and other publications reveals that the underlying statistical distribution of a data set is, in most cases, only incompletely analyzed or even misinterpreted. At the same time, confidence intervals for the determined parameters are rarely provided. We therefore saw the need to apply and discuss advanced statistical tools, which is equally important for researchers and practitioners. Consequently, the current work addresses the nature of data in the fields of indoor air and house dust pollution.

We searched for an up-to-date data set that is, as far as possible, free of systematic artifacts and bias. Measured values should be subject only to random fluctuations in environmental factors. The German Environmental Survey (GerES) [24] is a population-based study to investigate human exposure to environmental pollutants. Importantly,

participants are randomly selected from German cities and municipalities in a strive to be population-representative. Indoor pollution data were collected and evaluated by the German Environment Agency (UBA), involving indoor air and house dust measurements in more than 600 private dwellings. For this work, data on indoor air and house dust from the GerES V survey conducted between 2014 and 2017 were used. We analyzed these data sets with regard to the structure of their statistical distributions using numerical methods and compared the results for their validity. To this end, we classified the data in bins, visualized them in histograms and compared the results of curve fitting with analyses of the cumulative density function. Additionally, Monte Carlo simulations were performed to investigate the influence of the histogram's bin width and the data size on statistical parameters and the distribution. For determining confidence intervals, we applied bootstrap sampling, taking into account various computational methods. Our approach aims at a better assessment of the statistical nature of indoor pollution data from surveys and the reliability of parameters extracted from them. Meanwhile, our work serves to further validate the GerES V data set by dissecting and describing its statistical structure, which contributes to further establish the data as a reference point for indoor air pollution measurements.

2. Methods

2.1. Statistical distributions

The binomial distribution (see Eq. (1)) is a discrete statistical distribution, and describes the probability P_B of k successes in a series of n Bernoulli trials with the success probability p and the failure probability $1 - p$.

$$P_B(n, p, k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad k \in (0, 1, 2, \dots, n) \quad (1)$$

For $n \rightarrow \infty$ and $p \rightarrow 0$ with $n \cdot p \rightarrow \lambda$, the binomial distribution can be approximated by the discrete Poisson distribution P_P (2).

$$P_P(\lambda, k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad k \in (0, 1, 2, \dots, n) \quad (2)$$

In Eq. (2), λ describes both the expected value and the variance of the Poisson distribution.

Both the binomial distribution, when $n \rightarrow \infty$, and the Poisson distribution, when $\lambda \rightarrow \infty$, converge to a normal (Gaussian) distribution (Eq. (3)) with the expected value μ and the standard deviation σ .

$$P_N(x, \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad x \in \mathbb{R} \quad (3)$$

The two parameter lognormal probability model P_{2LN} (Eq. (4)) describes the distribution of a random variable x if the logarithmically transformed random variable $Y = \ln x$ is normally distributed.

$$P_{2LN}(x, \mu, \sigma) = \frac{1}{x \sigma \sqrt{2\pi}} e^{-\frac{(\ln(x) - \mu)^2}{2\sigma^2}} \quad x > 0 \quad (4)$$

From Eq. (4) we obtain Eq. (5) for the median ($\tilde{x}$) and geometric mean (μ_g) and Eq. (6) for the geometric standard deviation (σ_g).

$$\mu_g = \tilde{x} = e^\mu \quad (5)$$

$$\sigma_g = e^\sigma \quad (6)$$

By substituting Eqs. (5) and (6) into Eq. (4), we obtain Eq. (7).

$$P_{2LN}(x, \mu_g, \sigma_g) = \frac{1}{x \ln(\sigma_g) \sqrt{2\pi}} e^{-\frac{[\ln(x) - \ln(\mu_g)]^2}{2[\ln(\sigma_g)]^2}} \quad x > 0 \quad (7)$$

The cumulative density function F_{2LN} (Eq. (8)) is obtained by integrating Eq. (7). F_{2LN} gives the probability that a random variable takes a value less than x .

$$F_{2LN}(x, \mu_g, \sigma_g) = \int_0^x \frac{1}{t \ln(\sigma_g) \sqrt{2\pi}} e^{-\frac{[\ln(t) - \ln(\mu_g)]^2}{2[\ln(\sigma_g)]^2}} dt \quad x > 0 \quad (8)$$

In the statistical analysis of environmental data, it has been repeatedly found that the two parameter lognormal distribution (7) cannot satisfactorily approximate the experimental results [1]. This circumstance is addressed by introducing a location parameter a , resulting in the three parameter Eq. (9) with $x \geq a$.

$$P_{3LN}(x, a, \mu_g, \sigma_g) = \frac{1}{(x-a) \ln(\sigma_g) \sqrt{2\pi}} e^{-\frac{(\ln(x-a) - \ln(\mu_g))^2}{2(\ln(\sigma_g))^2}} \quad x > 0, a \in \mathbb{R} \quad (9)$$

For a data set with n positive real numbers, the geometric mean μ_g is defined by Eq. (10).

$$\mu_g = \left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}} \quad x > 0, a \in \mathbb{R} \quad (10)$$

The equations presented in this section were essentially taken from the publications of Ott [1] and Georgopoulos and Seinfeld [9] and can be studied in detail there.

2.2. Experimental data

2.2.1. The German Environmental Survey

The German Environmental Survey (GerES) is the largest study cycle on the exposure of the general population in Germany to environmental pollutants [24]. The target group for the 5th cycle (GerES V) were children and adolescents aged 3 to 17. GerES participants are randomly selected prior to the respective study to be representative for the population in Germany in this age group. Participant-based measurements cover, amongst others, indoor air, house dust, and drinking water, as well as human biomonitoring in blood and urine samples. Questionnaire data are obtained through interviews with participants; these characterize study participants with respect to socio-economic status, consumer habits, product choices, and conditions of living. The information obtained is used to identify sources and exposure-relevant factors.

2.2.2. Indoor air and house dust measurements in GerES V

In the GerES V study cycle, indoor environmental measurements from more than 600 households with children and adolescents aged 3 to 17 were conducted between 2014 and 2017. Participants from various age groups were recruited at 167 survey locations. Indoor samples were collected during household visits by the contractor Kantar Health Munich (now Oracle Life Sciences) while chemical analysis was performed in the laboratories of the German Environment Agency (UBA) and the Fraunhofer Institute for Process Engineering and Packaging (IVV).

Air sampling took place in the room where the participant spends most of his/her time in the household. The substance concentrations determined from the 7 day passive sampling reflect normal household use conditions. Comprehensive information on the GerES V methodology is provided in several publications [21,24,25].

2.2.3. Sampling and chemical analysis

The data investigated in this work was derived using three different analytical procedures: The first method involved passive air sampling of very volatile and volatile organic compounds (VVOCs and VOCs) over a period of 7 days using Tenax TA tubes. This allowed to capture individual substances across a molecular range of approximately C_4 to C_{16} . Chemical analysis was conducted by thermal desorption and gas chromatography coupled with mass spectrometry (TD-GC/MS). See Fernandez-Lahore et al. [21] for corresponding details. The validity of the VVOC/VOC measurements using Tenax tubes was confirmed by Richter and Schüle [26].

The second method involved air sampling using DNPH-coated UME^x-type passive samplers to determine concentrations of aliphatic aldehydes. In the laboratory, elution of the filters by acetonitrile yielded aldehyde hydrazones that are quantified by high-performance liquid

chromatography (HPLC) using a triple quadrupole MS/MS. For details of the analytical procedures including validation, see Birmili et al. [25].

The third method involved the analysis of house dust from vacuum cleaner bags collected in about 600 households. Prior to analysis, the content of each vacuum cleaner bag was sieved to obtain the house dust's fine particulate fraction $\leq 63 \mu\text{m}$. This provided a relatively homogeneous sample of the house dust particles involving a large portion of the dust's surface area. Semi-volatile organic compounds (SVOCs) were subsequently analyzed, notably plasticizers and flame retardants, using gas chromatography-mass spectrometry (GC/MS) and liquid chromatography-mass spectrometry (LC/MS). A basic statistics of the corresponding data can be found in Nagorka et al. [27].

2.2.4. Selected data sets

The complete GerES V data sets for air and dust measurements were provided by the Robert Koch Institute (RKI) and the Federal Environment Agency (UBA). For the scientific purposes of this work, data sets with the highest possible number of measurements (n) and the lowest possible number of measurements below the limit of quantification (LOQ) are required. After manual review of all data sets, the following substances were selected: Air: formaldehyde, acetaldehyde, hexanal (aldehydes); α -pinene, limonene, toluene, m,p-xylene, n-butylacetate, 2-ethylhexylethanol, D5 (VOCs). Dust: DEHP, DINP, DnBP, BnBP, DiDP (phthalates), DINCH (cyclohexane dicarboxylic acid ester), DEHT (terephthalate), DEHA (adipate) and TCPP (phosphate). The full substance names and their identifiers are summarized in Table S1 of the Supporting Information. With the exception of n-butyl acetate ($n = 309$), the number of respective measurements was $n \geq 566$. Except D5 (91%) the number of measurements $\geq \text{LOQ}$ was higher than 95%. The LOQ for each substance is listed in the Supporting Information.

2.3. Statistical analysis

2.3.1. Software

OriginPro 2025 software from OriginLab Corporation (Northampton, MA) was used for statistical calculations, least-squares fitting, numerical simulations, and graphical presentations. The statistical characterization of the data was primarily based on non-parametric and robust percentiles ($P_{10}, P_{25}, P_{50}(\bar{x}), P_{75}, P_{90}, P_{95}$), with the P_{95} being of particular importance for the derivation of reference values [28]. The confidence intervals of the geometric mean and the 95th percentile were estimated using non-parametric bootstrap sampling. The upper and lower percentiles of the bootstrap distribution were calculated for the desired probability α using the percentile method from 1000 runs and the bias-corrected and accelerated (BCa) method [29]. The Marquardt-Levenberg algorithm [30] was used for non-linear regression analysis with selected model functions. Where necessary, the fitting procedure was supported by a manual grid search. The coefficient of determination R^2 , the reduced χ^2 , and analysis of the residues served as goodness-of-fit criteria for the least squares method. Further statistical methods were applied where necessary and are discussed accordingly in the text.

2.3.2. Random number generator

The random number generator used by OriginPro 2025 is deterministic. Therefore, the starting conditions must be changed for each run to obtain different sequences of random numbers. Two types of generators are implemented: a) for uniformly distributed random numbers between 0 and 1; b) for normally distributed random numbers with a mean of 0 and a standard deviation of 1. Five sets of 5000 normally distributed random numbers were generated and analyzed using the Shapiro-Wilk test and the Kolmogorov-Smirnov test at a significance level of 0.05. The null hypothesis of a normal distribution could not be rejected in any case. Analogously, five sets of 5000 uniformly distributed random numbers were generated. In all cases, the null hypothesis of a normal distribution was rejected at the 0.05

Table 1

Descriptive statistics (percentiles and geometric mean μ_g) of data sets for organic compounds selected from indoor air and house dust measurements within GerES V.

Substance	<i>n</i>	<i>P</i> ₁₀	<i>P</i> ₂₅	<i>P</i> ₅₀	μ_g	<i>P</i> ₇₅	<i>P</i> ₉₀	<i>P</i> ₉₅
$\mu\text{g}/\text{m}^3$ (indoor air)								
Formaldehyde	636	12.6	17.4	25.2	23.2	36.0	49.1	56.9
Acetaldehyde	636	2.1	3.5	5.7	5.3	8.7	12.8	16.0
Hexanal	636	3.3	5.9	10.3	9.3	16.5	28.2	33.9
α -Pinene	591	1.7	3.4	6.9	7.1	15.1	28.2	42.3
Limonene	591	2.3	4.7	11.5	11.1	28.0	56.0	91.9
Toulene	591	1.9	2.8	4.5	5.2	8.6	19.1	30.1
m,p-Xylene	591	0.8	1.1	1.9	2.3	3.7	9.8	17.2
n-Butylacetate	309	1.0	1.9	4.4	4.6	10.4	25.0	34.4
2-Ethylhexanol	566	1.7	2.8	4.8	4.6	7.8	13.2	18.0
D5	566	1.1	3.9	14.0	12.2	36.0	100.3	152.3
$\mu\text{g}/\text{g}$ (house dust)								
DEHP	646	60.6	91.9	147.4	160.2	264.6	508.0	671.7
DINP	645	57.2	105.5	201.5	220.6	561.6	1270.7	1879.7
DnBP	646	2.6	4.2	7.3	7.8	13.5	25.4	41.9
BnBP	646	1.7	1.7	2.9	4.2	7.1	26.2	56.4
DIDP	645	8.2	14.9	28.7	26.8	50.5	88.2	127.0
DINCH	633	5.1	8.8	17.4	18.0	32.5	77.1	137.4
DEHT	646	24.1	41.2	77.0	92.0	196.5	505.5	830.5
DEHA	646	0.6	0.6	0.7	1.0	1.6	3.3	5.5
TCPP	645	0.4	1.0	2.5	2.3	4.8	9.7	15.7

significance level. The runs test [31] showed that the sequences of all data sets (uniformly and normally distributed) are random. The random number generators implemented in OriginPro 2025 can therefore be considered suitable for the calculations to be carried out here.

For the Monte-Carlo simulations, *n* normally distributed random numbers with the selected expected value μ and the standard deviation σ were first generated. These were then transformed into a lognormal distribution using the exponential function.

3. Results

3.1. Descriptive statistics

Table 1 shows the statistical analysis (percentiles and geometric mean) of the selected data sets. All distributions are skewed, so the arithmetic mean is not a useful parameter and is therefore not reported. In some cases, slight deviations from the results published by Fernandez Lahore et al. [21] and Birmili et al. [25] are observed. This is due to the fact that the published GerES V results were subjected to a weighted adjustment based on population data. Since this is a purely statistical analysis, the weighting was not applied.

According to Eq. (5), the geometric mean should match the median for a lognormal distribution. Table 1 shows that this is the case for most substances in air and house dust. Notable deviations occur for DEHP, DINP, BnBP and DEHT. This indicates deviations from the lognormal distribution. In all four cases is $\mu_g > P_{50}$. The lognormal distribution significantly underestimates the measured data in the range of higher values (see Figures S11, S12, S14 and S17 in the Supporting Information), which leads to an increase of the geometric mean in the descriptive statistics. In the case of BnBP, it is also striking that *P*₁₀ and *P*₂₅ are identical. The reason for this is that a total of 207 BnBP measurement data were assigned the same value of 1.67 $\mu\text{g}/\text{g}$ (see Supporting Information).

3.2. Non-linear regression analysis

To fit a model to the respective data set, the air and house dust data collected in GerES V must first be classified into equidistant bins. This is shown in Fig. 1A for the example of formaldehyde. The concentration range is between 0.47 $\mu\text{g}/\text{m}^3$ and 113.33 $\mu\text{g}/\text{m}^3$. Thus, a bin width of $\Delta x = 5 \mu\text{g}/\text{m}^3$ between 0 $\mu\text{g}/\text{m}^3$ and 120 $\mu\text{g}/\text{m}^3$ is favorable, resulting in $DF = 24 - 3 = 21$ degrees of freedom. Note that an additional amplitude

parameter *A* is required for the non-linear regression analysis with Eq. (7) because the data are not normalized. The choice of bin width is always a compromise. A small bin width increases the degrees of freedom and decreases the probability that a data point is classified in a particular bin, which favors the use of a Poisson distribution for error estimation. When using a wide bin width, the differences between the counts become larger, but the degrees of freedom decrease, which increases the uncertainty of the fit [32]. Fig. 1A also shows that a log-normal distribution cannot represent the first bin with the class center at 2.5. This becomes even clearer in the logarithmic representation of Fig. 1B. The reason is the limit of quantification (LOQ). For the formaldehyde data set, it is 0.7 $\mu\text{g}/\text{m}^3$, so all concentrations below the LOQ are counted in this bin [25]. In principle, the bin width must be chosen small enough to result in a skewed distribution. This can lead to a high number of degrees of freedom, as shown in Table 2. For some data sets with many values near the limit of quantification, a lognormal fit was not possible even with a small bin width (see Table 2).

This also raises the question of how to assess the goodness of fit. A simple indicator is the adjusted coefficient of determination R_a^2 [33]. It can be used to quantify how well a model fits the data and considers the number of predictors in the fit function. A frequently used criterion is the reduced χ_r^2 according to Eq. (11).

$$\chi_r^2 = \frac{1}{DF} \cdot \sum_{j=1}^m \frac{(N_j - y_{0,j})^2}{(\sigma(y_{0,j}))^2} \quad (11)$$

N_j is the observed number of counts in bin *j*, $y_{0,j}$ is the expected number of counts in bin *j* and $\sigma(y_{0,j})$ is the standard deviation of $y_{0,j}$. The standard deviation values for the counts in the individual bins are unknown in most experiments. If the probability of classifying a measurement into bin *m* is small compared to the total number of measurements *n*, a Poisson distribution with $\sigma(y_{0,j})^2 = y_{0,j}$ can be assumed [30]. In this case, the value expected for a good fit to the model function is $\chi_r^2 \approx 1$. Fig. 1C shows the residuals $R_m = N_j - y_{0,j}$ for the 24 bins of the formaldehyde data. The outlier in the first bin is clearly visible, driving χ_r^2 to a value of 447. On the other hand, $\chi^2 \ll 1$ indicates that the number of counts in the bins is not large enough [34]. Therefore, one essentially relies on the analysis of the residuals, which, with the exception of the first bin, are approximately normally distributed. Finally, the fit of the cumulative density function F_{2LN} to the 636 formaldehydes is shown in Fig. 1D. The lower part shows the expected deviations of the experimental data from the fit curve, but $P_{50} = 25.2 \mu\text{g}/\text{m}^3$ and $P_{95} = 58.3 \mu\text{g}/\text{m}^3$ are in good agreement

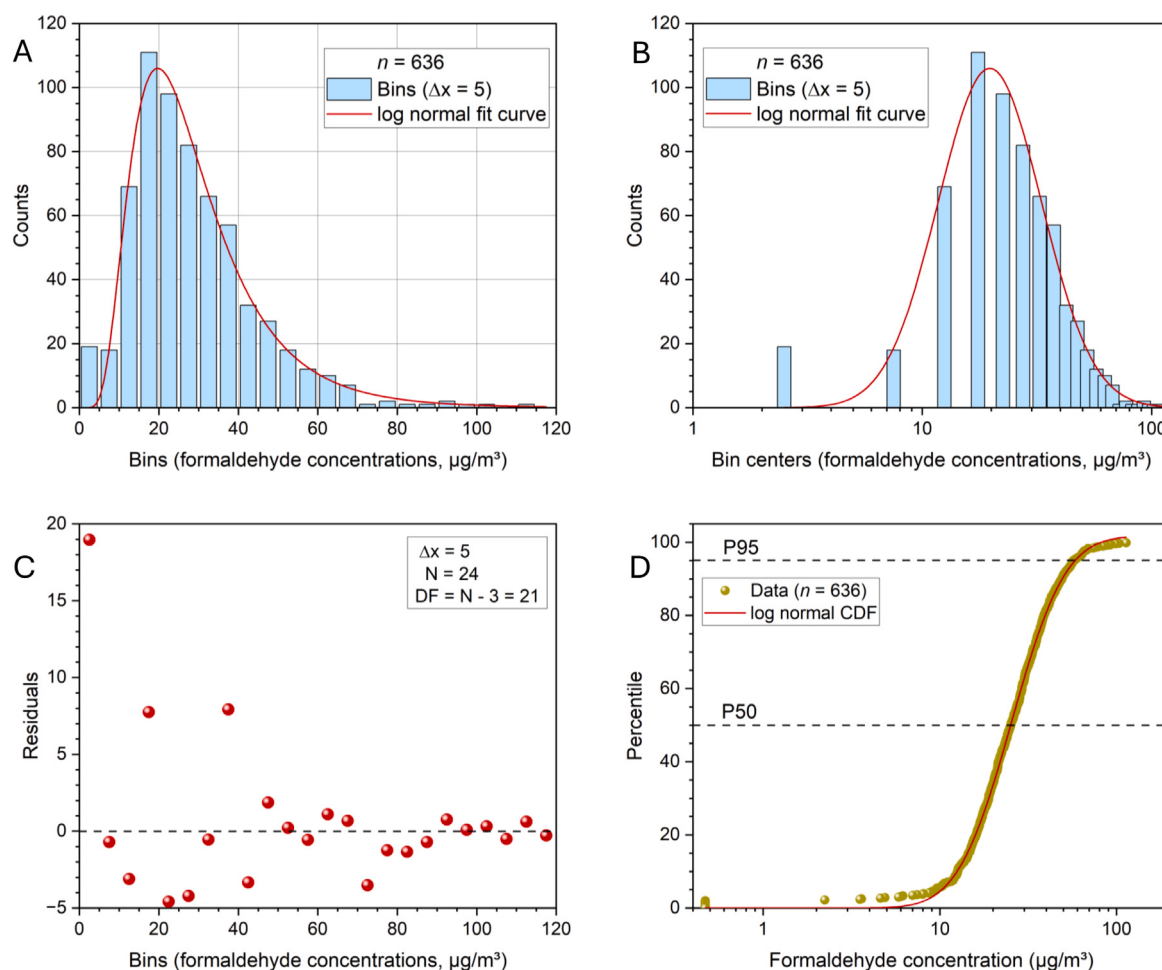


Fig. 1. A: Histogram of the 636 formaldehyde data points with a bin width of $\Delta x = 5 \mu\text{g}/\text{m}^3$ and the two parameter lognormal fit curve (Eq. (7)); B: Logarithmic representation of the histogram bin centers and two parameter lognormal fit curve (Eq. (7)); C: Residuals of the non-linear least squares fit (Eq. (7)) to the histogram data; D: Cumulative plot of the 636 formaldehyde data points with cumulative lognormal fit curve (Eq. (8)).

with the results of the descriptive statistics. Considering the experimental limitations, it can be concluded that the formaldehyde data are very well represented by a two parameter lognormal distribution. The fit parameters obtained using Eqs. (7) and (8) for the data sets of all selected substances in air and house dust are listed in Table 2. For other substances, a further problem can arise from individual very high values. This requires a large number of bins, resulting in the number of counts in many of these bins being 0. Consequently, χ_r^2 decreases with increasing DF, which also leads to bias.

A different picture emerges for the distribution of DEHP in house dust (see Fig. 2). Here, no measured concentration is below the detection limit, but there are two significantly elevated values at 6052 $\mu\text{g}/\text{g}$ and 7382 $\mu\text{g}/\text{g}$. All other concentrations are below 2000 $\mu\text{g}/\text{g}$. For the non-linear regression analysis, a bin width of $\Delta x = 50 \mu\text{g}/\text{g}$ is appropriate. However, this leads to a misleadingly high number of degrees of freedom, since many bins contain no data points at all. On the other hand, for the fitting with Eq. (7), it makes no difference whether the two elevated values are considered or not.

The two parameter lognormal distribution provides a very good fit to the histogram data (see Figs. 2A and 2B). Up to 1850 $\mu\text{g}/\text{g}$, the residuals scatter approximately around 0 (see Fig. 2C). However, the cumulative lognormal function (8) can only inadequately approximate the experimental data in the upper part (see Fig. 2D). Thus, significant deviations from the expected lognormal distribution occur in this range, which particularly influences the 95th percentile.

A comparison of Tables 1 and 2 shows that there are differences between the parameters obtained with the descriptive statistics, the lognormal fitting with Eq. (7) after classifying the data into classes and the cumulative lognormal fitting with Eq. (8). There are several reasons for this. The effects of the LOQ have already been discussed, and the data often only approximately follow a lognormal distribution. Examples are shown in Fig. 3. For toluene in air, the high values are underestimated by the cumulative fit function (8), while for D5 in air, they are overestimated. For TCP and DEHA in house dust, significant deviations from the sigmoidal curve are evident. However, this has little impact on the fit curve for TCP, whereas for DEHA, the data are hardly represented by Eq. (8). Furthermore, steps in the concentration data are noticeable for TCP and DEHA. These result from the chemical analysis of sample subsets with different LOQs. It is also observed that a data distribution is dominated by high values that do not conform to the expected lognormal distribution. This is the case, for example, with DEHT in house dust (see Supporting Information). The different calculation methods lead to significant differences between P_{50} and μ_g . The cumulative function (8) underestimates the DEHT data, making it impossible to determine the P_{95} value from the fitted curve.

Fitting the classified data with the three parameter lognormal distribution (9) did not result in any significant improvement. In most cases, the location parameter a tended toward 0; in some cases was $a \leq x$, which led to termination of the fitting routine. Finally, the attempt to fit the cumulative air and house dust data using a cumulative Weibull

Table 2

Two parameter lognormal (Eq. (7)) and cumulative lognormal (Eq. (8)) non-linear regression analysis of data sets for organic compounds selected from indoor air and house dust measurements.

Substance	DF	μ_g	σ_g	R_a^2	P_{50}	μ_g	σ_g	P_{95}	R_a^2
	lognormal				lognormal CDF				
	$\mu\text{g}/\text{m}^3$ (indoor air)								
Formaldehyde	21	25.8	1.8	0.977	25.2	25.3	1.7	58.3	0.999
Acetaldehyde	22	6.0	2.2	0.986	5.6	5.8	2.1	15.7	0.998
Hexanal	21	10.4	2.2	0.994	10.1	10.2	2.2	33.9	0.997
α -Pinene	107	7.1	2.8	0.990	6.3	6.9	2.9	43.0	0.999
Limonene	477	10.6	3.2	0.970	11.0	11.6	3.8	88.4	0.999
Toulene	87	4.0	1.9	0.992	4.6	4.3	2.1	–	0.995
m,p-Xylene	97	1.7	1.9	0.991	1.9	1.8	2.1	–	0.993
n-Butylacetate	77	–	–	–	4.4	4.2	3.3	38.8	0.999
2-Ethylhexanol	30	5.0	2.3	0.992	4.7	4.7	2.2	16.7	0.999
D5	82	–	–	–	12.9	14.4	6.1	154.5	0.997
$\mu\text{g}/\text{g}$ (house dust)									
DEHP	145	144.0	2.1	0.994	151.3	144.0	2.1	686.8	0.998
DINP	197	184.3	2.6	0.979	224.5	214.7	3.2	2510.8	0.996
DnBP	222	7.0	2.2	0.993	7.3	7.1	2.3	41.2	0.999
BnBP	897	1.5	1.5	0.913	3.1	2.8	2.4	–	0.938
DIDP	596	32.6	2.9	0.957	28.0	28.4	2.5	117.8	0.998
DINCH	357	16.1	2.4	0.994	17.0	16.3	2.6	136.2	0.999
DEHT	321	63.0	2.3	0.978	80.6	74.1	2.8	–	0.995
DEHA	121	0.7	1.5	0.974	0.9	0.8	1.7	–	0.921
TCPP	197	–	–	–	2.4	2.6	2.9	14.8	0.994

Table 3

Descriptive statistics and lognormal CDF fit results (percentiles and geometric mean μ_g) of Monte-Carlo generated data sets with the default parameters $\mu_g = 25.0 \mu\text{g}/\text{m}^3$ and $\sigma_g = 2.0 \mu\text{g}/\text{m}^3$.

Substance	n	P_{10}	P_{25}	P_{50}	μ_g	P_{75}	P_{90}	P_{95}
Monte-Carlo generated data ($\mu\text{g}/\text{m}^3$)								
Descriptive statistics	125	9.5	13.5	21.5	22.0	35.7	56.9	76.7
lognormal CDF	125	8.7	13.4	21.7	21.6	35.4	55.6	74.2
Descriptive statistics	250	10.4	14.2	23.8	24.4	37.9	63.8	86.3
lognormal CDF	250	9.6	14.6	23.4	22.8	38.0	61.6	90.6
Descriptive statistics	500	10.6	16.2	25.5	25.7	41.5	63.7	76.7
lognormal CDF	500	10.5	16.0	25.5	25.5	40.8	62.7	81.3
Descriptive statistics	1000	9.7	15.6	25.1	25.1	40.9	64.3	84.1
lognormal CDF	1000	9.8	15.4	25.1	25.2	41.0	63.5	82.0

function [1,9], was unsuccessful in all cases, so the approach was not pursued. However, this does not mean that indoor related data are always lognormally distributed. For example, Dodson et al. [12] found that data sets from field studies can also follow the Weibull and Gamma distribution.

3.3. Monte-Carlo simulations

In non-linear regression analysis, the question arises to what extent purely statistical fluctuations, the number of data n and their division into bins have influence on the results when there are no systematic errors. For the investigation, lognormally distributed data sets with $n = 125, 250, 500$ and 1000 were generated using the Monte-Carlo method. Based on the GerES V formaldehyde data the default parameters were $\mu_g = 25.0 \mu\text{g}/\text{m}^3$ and $\sigma_g = 2.0 \mu\text{g}/\text{m}^3$. These data sets were analyzed using descriptive statistics, fitted with Eq. (7) after classifying them into bins of $\Delta x = 2.5, 5$ and $10 \mu\text{g}/\text{m}^3$, and fitted with the cumulative lognormal function (8).

The results of descriptive statistics and non-linear regression with the cumulative lognormal function (8) are shown in Table 3. For $n = 125$ and $n = 250$, geometric mean μ_g and P_{50} are underestimated by both methods. For all n and methods, the P_{95} value varies within a relatively large range of $74.2\text{--}90.6 \mu\text{g}/\text{m}^3$. The reasons for this are evident in Fig. 4. With $n = 250$, the Monte-Carlo data can only be roughly fitted to the cumulative lognormal distribution (8). The data are underestimated in the higher range and overestimated in the lower range, which consequently leads to a higher P_{95} and a lower P_{10} value. With $n = 1000$, the fit using Eq. (8) is significantly better. One obtains the expected geometric mean of $\mu_g = 25.1 \mu\text{g}/\text{m}^3$ and normally distributed residuals (not shown).

The effect of n and the classification into different bin widths on the fitting result is shown in Fig. 5. Similar to descriptive statistics and cumulative analysis, μ_g is underestimated for $n = 125$ and $n = 250$. For $n = 500$ and $n = 1000$, μ_g is significantly closer to the standard value. For the same n , μ_g increases with a larger Δx , and the standard error of μ_g decreases with increasing n . The standard error is a function of the degrees of freedom, and it is statistically determined that for data sets

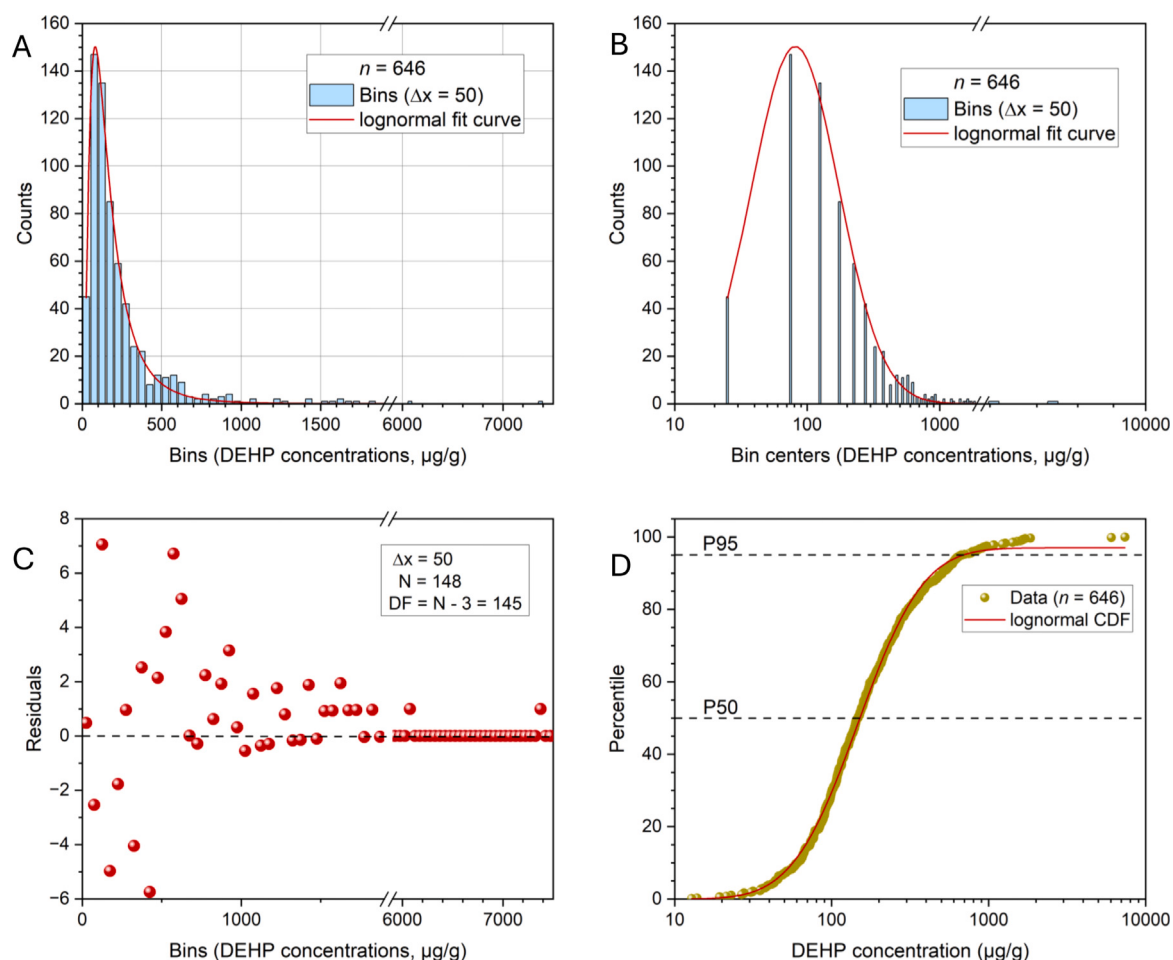


Fig. 2. A: Histogram of the 646 DEHP house dust data points with a bin width of $\Delta x = 50 \mu\text{g/g}$ and the two parameter lognormal fit curve (Eq. (7)); B: Logarithmic representation of the histogram bin centers and two parameter lognormal fit curve (Eq. (7)); C: Residuals of the non-linear least squares fit (Eq. (7)) to the histogram data; D: Cumulative plot of the 646 DEHP data points with cumulative lognormal fit curve (Eq. (8)).

representing the same function, the standard error always decreases as DF increases. Several trends are responsible for the other observations. As the bin width decreases, the differences between the counts within the individual bins become smaller. This flattens the function curve. In the limiting cases $\Delta x \rightarrow 0$, each bin contains only a maximum of one count. For $\Delta x \rightarrow \infty$, all counts are finally accumulated in a single bin. Therefore, when classifying, one must always choose a Δx that best represents the data. This effect is also indicated in Fig. 5. For $n = 125$, the standard error is largest with $\Delta x = 2.5 \mu\text{g/m}^3$, while the geometric mean is furthest from the default value. For $n = 1000$, μ_g is closest to the default value with $\Delta x = 2.5 \mu\text{g/m}^3$, while the differences in the standard error are marginal. In principle, the larger the number of data sets, the smaller the bin width can be chosen. However, this also depends on the individual quality of the data, and the optimum must be determined for each specific case.

4. Discussion

4.1. Calculating reference values and confidence intervals

Environmental surveys are conducted not only to determine the general population's exposure to pollutants, but also to derive reference values. These are usually statistically based and therefore dependent on the quality of the collected data set. Comparison of an individual observation with a reference enables to assess whether a particular value can be considered "typical", "usual", or "conspicuous". Exceeding or falling below a reference value does not imply any health assessment.

Various definitions for reference values have been proposed, which are similar but not identical [35]. A general definition by Heinzow and Sagunski [36] states that "...a reference value for a chemical substance in an environmental medium is a value which has been derived from a series of corresponding measured values of a random sample from a population on the basis of a specified procedure". Reference values are closely linked to the use and abundance of chemicals and products. If one chemical substance is replaced by another, the concentration of the substituted substance in the medium under consideration will decrease, while the concentration of the substitute will increase [27,37]. Reference values are therefore not constant, but must be redetermined from time to time.

In a series of publications, Solberg [38,39] explains methods for collecting data and deriving reference values in the medical sector. The advantages and disadvantages of parametric and non-parametric derivation are critically discussed. In the case of parametric derivation, Solberg [39] recommends estimating confidence intervals from the underlying distribution. In practice, however, it is often the case that a given distribution function only approximately represents the data set (see discussion below). For the harmonized determination of reference values, the International Federation of Clinical Chemistry (IFCC) recommends the 0.95 central interval, defined as the non-parametrically derived interval between the 2.5th and the 97.5th percentile of the respective data set [40]. Horn and Pesce [41] argue that with a two-sided interval, the smaller value is usually unimportant and recommend instead the one-sided 95th percentile (P_{95}). This reasoning has long been recognized in human biomonitoring [42] and indoor sciences [28].

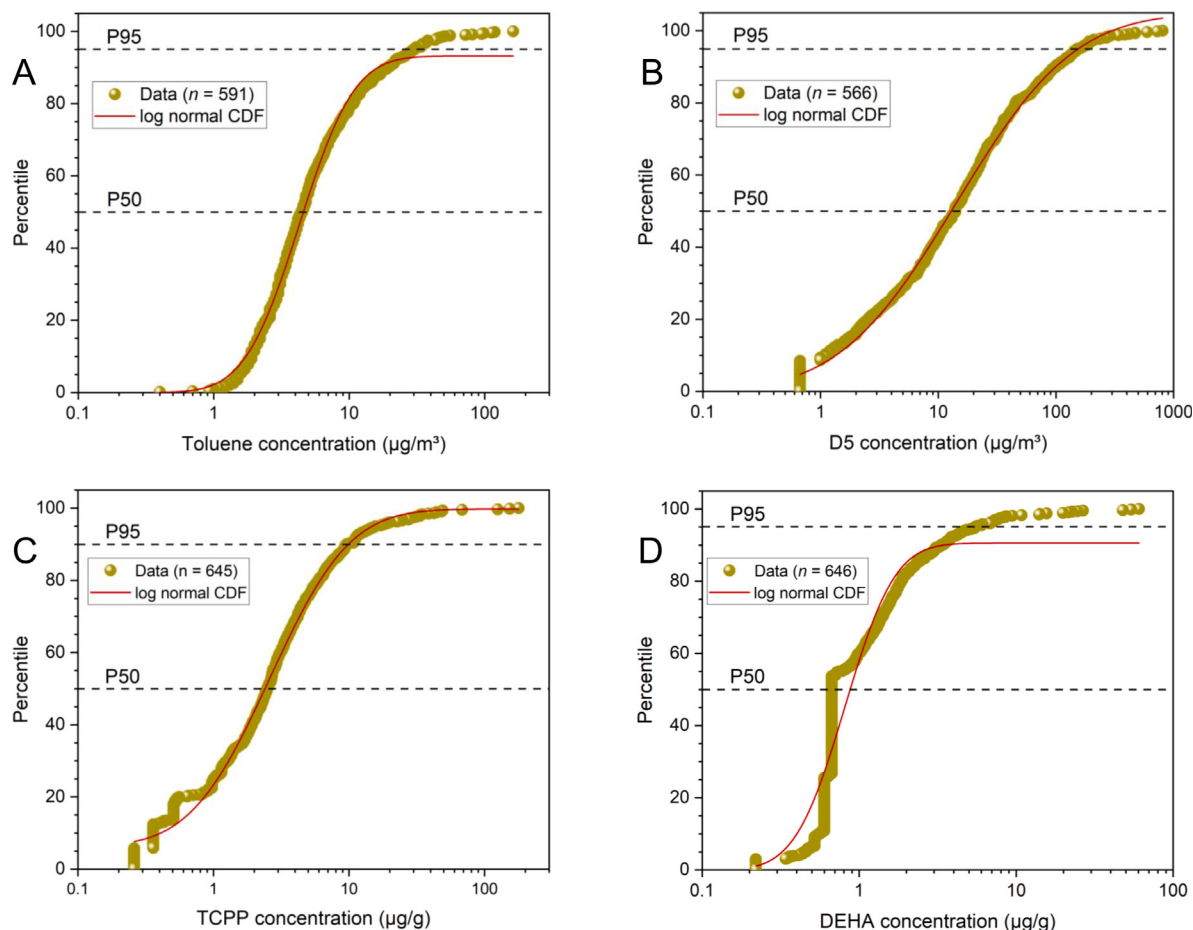


Fig. 3. Cumulative plot of air and dust data from GerES V with cumulative lognormal fit curve (Eq. (8)). A: Toluene in air; B: D5 in air; C: TCPP in house dust; D: DEHA in house dust. See Table 2 for fit parameters.

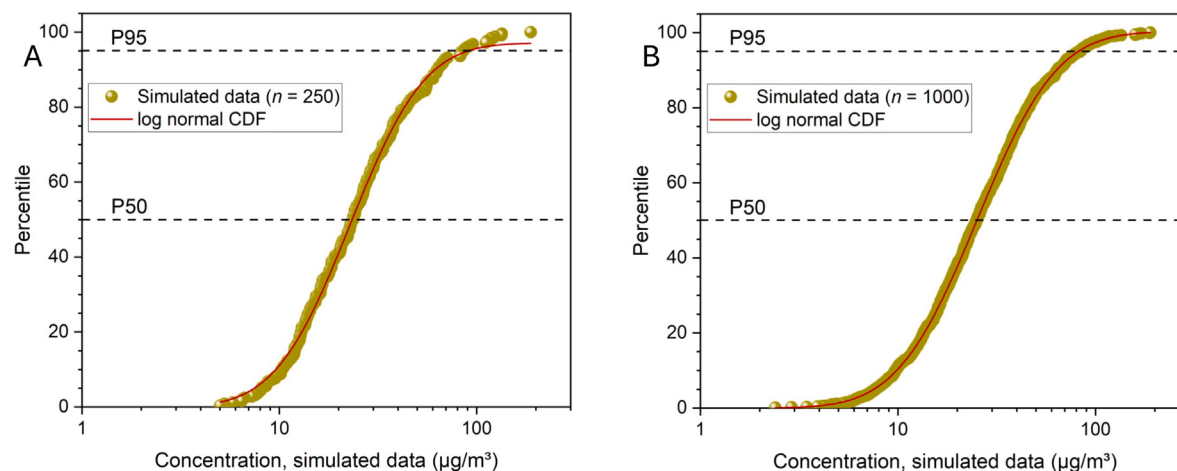


Fig. 4. Cumulative plot of Monte-Carlo simulated lognormally distributed data with cumulative lognormal fit curves (Eq. (8)). The default parameters are $\mu_g = 25.0 \mu\text{g}/\text{m}^3$ and $\sigma_g = 2.0 \mu\text{g}/\text{m}^3$. A: $n = 250$; B: $n = 1000$.

P_{95} is a purely statistically determined value that indicates whether exposure to a particular pollutant is higher or below a background level under the given conditions. However, it must be ensured that the determined value truly represents the P_{95} , i.e., that it is not influenced by systematic errors. Reed et al. [43] recommend that at least a data set of $n \geq 120$ is necessary to reliably determine percentiles. To derive P_{95} reference values from human biomonitoring (HBM) measurements

in the GerES V program, Hoopmann et al. [42] recommend a minimum sample size of 80, which should not be misinterpreted as a correction of Reed's value. In addition to HBM, reference values are suitable for house dust measurements. For indoor air, toxicologically derived guide values are often available [44], which provide information on potential health risks from exposure.

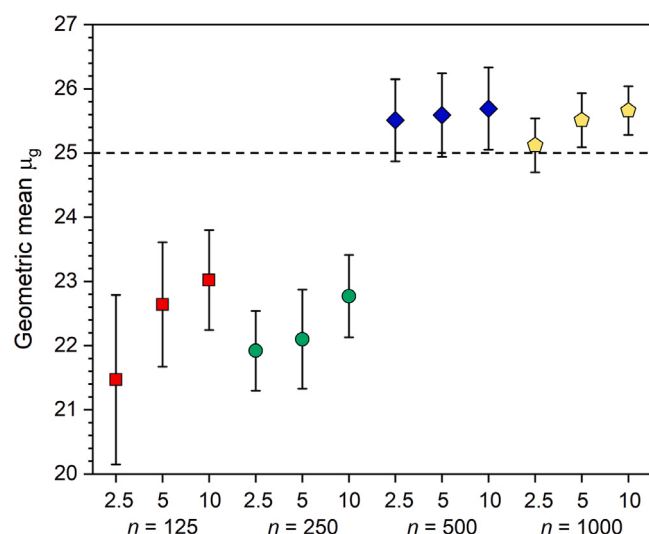


Fig. 5. Non-linear regression analysis (Eq. (7)) of Monte-Carlo generated lognormally distributed data with default parameters of $\mu_g = 25.0 \mu\text{g}/\text{m}^3$ and $\sigma_g = 2.0 \mu\text{g}/\text{m}^3$. The four data sets with $n = 125, 250, 500$, and 1000 were each classified into three bin widths of $\Delta x = 2.5, 5$, and $10 \mu\text{g}/\text{m}^3$. The dots represent the fitted geometric mean μ_g for each combination of n and bin width Δx , and the whiskers represent the standard error of μ_g .

A reference value is naturally subject to statistical uncertainties, which can be estimated using the confidence interval. This involves determining an interval that contains the desired value with a given probability (usually 95%). To estimate the confidence interval, Hoopmann et al. [42] suggest the non-parametric bootstrap procedure [45], which requires no assumptions about the distribution function. Bootstrap uses the method of resampling a data set of size n by randomly drawing a value from the given data set n times with replacement. By repeating this process frequently (1000 times or more), one obtains a distribution for the desired parameter. The confidence interval can then be calculated from the respective percentiles of this distribution for the probability α . Alternatively, methods have been developed that take a bias correction into account [29].

4.2. Assessment of GerES V data

The indoor-related data sets collected within the framework of GerES V are of high quality. This applies both to the number of measurements and the randomized selection of locations and individuals. Analytics, a key part of the study, is appropriate, well-validated, and documented. It is easy to see that the data sets selected here are right-skewed and therefore cannot be described by a normal (Gaussian) distribution. The lognormal distribution expected according to Ott [10] provides a good approximation for all data sets, but the goodness of fit varies (see Figs. 1–3 and Figures S1–S19 in the Supporting Information). This depends, among other things, on the analytical conditions and the number of measured values above the limit of quantification. In their statistical analysis of the cumulative frequency distribution from measurements of different VOCs, Jia et al. [11] also observed deviations from the lognormal distribution, which they likewise attributed to the respective detection limit. Moreover, despite all precautions, potential systematic errors are inevitable in measurements. It is also important to carefully examine outliers and check for errors before calculating statistical parameters and their associated confidence intervals. If high values can be proven to be due to measurement errors or similar factors, these outliers should be removed from the data set. This procedure requires a transparent justification.

From Fig. 4, it is evident that even the Monte-Carlo generated log-normally distributed data can only be approximately described by the cumulative function (8) for small n in the upper range. The calculated geometric mean μ_g approaches the default value as n increases (see Fig. 5). The different P_{95} values are also striking (see Table 3). For a log-normal distribution with $\mu_g = 25.0 \mu\text{g}/\text{m}^3$ and $\sigma_g = 2.0 \mu\text{g}/\text{m}^3$ without random noise, the expected value is $P_{95} = 78.2 \mu\text{g}/\text{m}^3$. The deviations in the simulated Monte-Carlo curves are statistically determined and can be explained by the standard errors of μ_g and σ_g . The analyses of the measured and simulated data demonstrate the uncertainties that can be associated with determining percentiles. It is therefore advisable to check whether the data set generated as part of an environmental survey is subject to a physically meaningful statistical distribution. The information content of the cumulative function (8) is higher because it allows trends of the individual data to be observed.

Fig. 6 shows the histograms of bootstrap analyses (1000 runs each) for the formaldehyde data in air and the DEHP data in house dust. The geometric mean and the 95th percentile with the corresponding upper and lower bounds of the 95% confidence intervals were calculated using the percentile method. It is striking that the geometric means (see 6A and 6C) exhibit largely symmetrical distributions, with confidence intervals within the expected range. A different picture emerges for the P_{95} values (see 6B and 6D). Here, both data distributions are clearly asymmetrical; in the case of DEHP, there are even bins without counts. The reason is that the 95th percentile becomes significantly more influenced by high values. In the case of DEHP, with $P_{95} = 671.7 \mu\text{g}/\text{g}$, a 95% confidence interval of 594.2–856.5 $\mu\text{g}/\text{g}$ results. Similarly broad confidence intervals were also obtained for other target compounds with high concentrations in air and dust, especially for D5, DINP, DIDP, DINCH and DEHT. This is a general disadvantage of the bootstrap method, since outliers (see Fig. 2D) distort the distribution and broaden the confidence interval. Therefore, the more advanced calculation method BCa (bias-corrected accelerated) was also used. The bias correction parameter accounts for a systematic deviation of the median of the bootstrap distribution from the observed value of the original sample. The acceleration parameter estimates the skewness of the bootstrap distribution using the jackknife resampling method [29]. The results of the bootstrap analyses percentile method and BCa of all 19 data sets at the 0.95 level for the geometric mean (μ_g) and P_{95} are listed in the Supporting Information. However, it is obvious that there are only small differences between the percentile method and BCa, so it can be concluded that high values in the data sets have little influence on the respective confidence intervals.

5. Conclusion

Data sets on pollutant concentrations in indoor air and house dust collected as part of the GerES V program were subjected to advanced statistical analyses. This required that the majority of the values be greater than LOQ. Importantly, the data of all selected target compounds follow the shape of a two parameter lognormal distribution, although the quality of the individual fits vary. Neither a three parameter lognormal distribution (Eq. (9)) nor the Weibull distribution yielded better results. It can therefore be assumed that the data sets examined were collected randomly. The agreement with the results of descriptive statistical analysis was satisfactory in most cases. The bootstrap method, which can be applied without knowledge of the respective distribution, is suitable for calculating confidence intervals. No significant differences were observed between the percentile method and the BCa method in our data sets.

In general, choosing an appropriate bin width is necessary for non-linear regression analysis using the two parameter lognormal density function (Eq. (7)). Analysis using the cumulative density function (Eq. (8)) is also recommended because it takes the ranked original data into account. By comparing the values, it is possible to determine whether

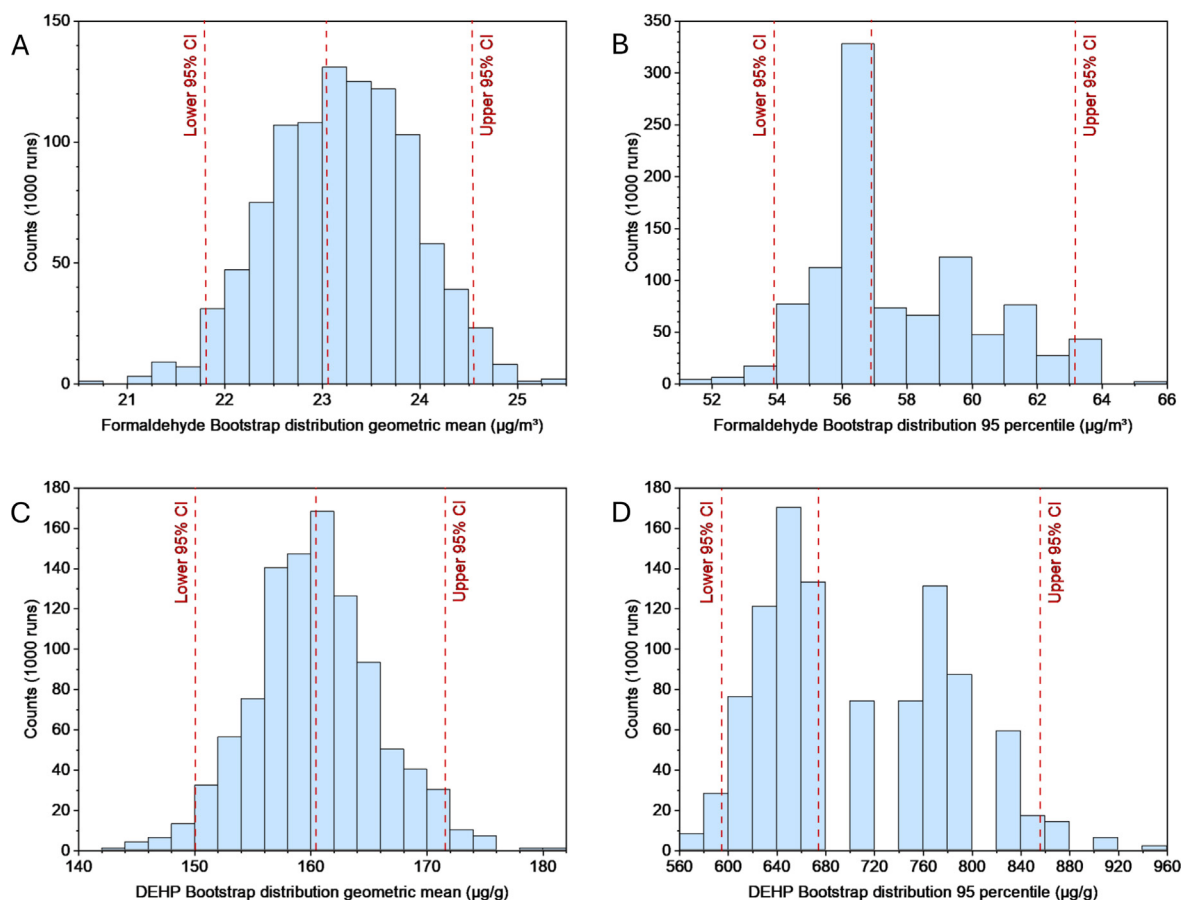


Fig. 6. Bootstrap histograms for the geometric mean (μ_g) and the 95th percentile (P_{95}) of the data for formaldehyde in air and DEHP in house dust. Each analysis includes 1000 runs. The 95% confidence intervals ($\alpha = 0.95$) were calculated using the percentile method. Upper limit: $1 - (1 - \alpha)/2$; lower limit: $(1 - \alpha)/2$.

the non-parametrically and parametrically obtained percentiles correspond to the expected range. It would then be necessary to check whether the deviations can be explained statistically and/or by analytical limitations or whether the data set is biased. This aspect should definitely be taken into account when deriving reference values. Specifying a confidence interval using the bootstrap procedure is recommended.

Finding that the investigated indoor pollutants are lognormally distributed on the GerES V study-level can imply that:

- The data behave as expected for complex environmental systems governed by multiplicative factors, even if the system involves, in principle, a myriad of dissociated dwellings across Germany.
- The collection of data underwent a certain degree of representativity.
- The data's statistical description can be meaningfully simplified, such as through the median ($\bar{x}$), the geometric mean (μ_g), the geometric standard deviation (σ_g) and the 95th percentile (P_{95}). This may help risk-assessment methods, for example through a simplified and reliable estimation of upper percentiles.
- The occurrence of skewed lognormal distributions may provide a rationale for focusing health interventions on the small subset of dwellings with very high pollutant concentrations.

In addition to the detailed assessment of data sets from the GerES V program, our work demonstrates which statistical methods are suitable or necessary for characterizing environmental surveys and determining

parameters. It is recommended to apply these methods to optimally analyze data and avoid misinterpretations.

CRediT authorship contribution statement

Tunga Salthammer: Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Anja Daniels:** Writing – original draft, Validation, Formal analysis, Data curation. **Wolfram Birmili:** Writing – original draft, Formal analysis, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are highly indebted to all children and adolescents and their families who participated in GerES V. We thank the Robert Koch Institute for the close cooperation with KiGGS Wave 2 and subsequent data sharing. GerES V was coordinated by a steering committee in the Department II 1 “Environmental Hygiene” at UBA. We thank Kantar Health Munich (now Oracle Life Sciences) for collecting the samples during fieldwork. GerES V field work and data processing received financial support from the German Federal Ministry for the Environment, Nature Conservation, and Nuclear Safety (BMU), Germany. The

Section “Indoor Hygiene, Health-related Environmental Impacts” at UBA was responsible for VVOC/VOC and aldehyde indoor air analyses. Plasticizer concentrations in house dust samples were determined by the Fraunhofer Institute for Process Engineering and Packaging (IVV) in Freising, Germany, under the UFOPLAN research program funded by BMU (grant No. FKZ 3715612040).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.indenv.2026.100167>.

Data availability

The data used in this study can be made available by the Robert Koch Institute (RKI) upon written request. Source: Robert Koch-Institut und Umweltbundesamt (2022). Das Umweltmodul zur Studie zur Gesundheit von Kindern und Jugendlichen in Deutschland (KiGGS Welle 2) — die Deutsche Umweltstudie zur Gesundheit von Kindern und Jugendlichen (GerES V). Scientific Use File 1. Version. <https://doi.org/10.7797/17-201417-1-1-1-umwelt>.

References

- [1] W.R. Ott, Environmental Statistics and Data Analysis, Lewis Publishers, Boca Raton, 1995, <http://dx.doi.org/10.1201/9780203756843>.
- [2] E. Limpert, W.A. Stahel, M. Abbt, Log-normal distributions across the sciences: Keys and clues, *Biosci.* 51 (2001) 341–352, [http://dx.doi.org/10.1641/0006-3568\(2001\)051\[0341:LNATSJ\]2.0.CO;2](http://dx.doi.org/10.1641/0006-3568(2001)051[0341:LNATSJ]2.0.CO;2).
- [3] F. Galton, The geometric mean, in vital and social sciences, *Proc. R. Soc. Lond.* 29 (1879) 365–367, <http://dx.doi.org/10.1098/rspl.1879.0060>.
- [4] D. McAllister, The law of the geometric mean, *Proc. R. Soc. Lond.* 29 (1879) 367–376, <http://dx.doi.org/10.1098/rspl.1879.0061>.
- [5] M. Eigen, R. Winkler, *Laws of the Game*, Princeton University Press, Princeton, NJ, 1981.
- [6] N.L. Johnson, S. Kotz, *Urn Models and their Application*, John Wiley & Sons, New York, NY, 1977.
- [7] W.R. Ott, D.T. Mage, A general purpose univariate probability model for environmental data analysis, *Comput. Oper. Res.* 3 (1976) 209–216, [http://dx.doi.org/10.1016/0305-0548\(76\)90029-0](http://dx.doi.org/10.1016/0305-0548(76)90029-0).
- [8] D.T. Mage, W.R. Ott, Refinements of the lognormal probability model for analysis of aerometric data, *J. Air Pollut. Control Assoc.* 28 (1978) 796–798, <http://dx.doi.org/10.1080/00022470.1978.10470662>.
- [9] P.G. Georgopoulos, J.H. Seinfeld, Statistical distributions of air pollutant concentrations, *Environ. Sci. Technol.* 16 (1982) 401A–416A, <http://dx.doi.org/10.1021/es00101a727>.
- [10] W.R. Ott, A physical explanation of the lognormality of pollutant concentrations, *J. Air Waste Manage. Assoc.* 40 (1990) 1378–1383, <http://dx.doi.org/10.1080/10473289.1990.10466789>.
- [11] C. Jia, J. D'Souza, S. Batterman, Distributions of personal VOC exposures: A population-based analysis, *Environ. Int.* 34 (2008) 922–931, <http://dx.doi.org/10.1016/j.envint.2008.02.002>.
- [12] R.E. Dodson, J.I. Levy, E.A. Houseman, J.D. Spengler, D.H. Bennett, Evaluating methods for predicting indoor residential volatile organic compound concentration distributions, *J. Expo. Sci. Environ. Epidemiol.* 19 (2009) 682–693, <http://dx.doi.org/10.1038/jes.2009.1>.
- [13] Z. Daraktchieva, J.C.H. Miles, N. McColl, Radon, the lognormal distribution and deviation from it, *J. Radiol. Prot.* 34 (2014) 183–190, <http://dx.doi.org/10.1088/0952-4746/34/1/183>.
- [14] E. Petermann, P. Bossew, J. Kemski, V. Gruber, N. Suhr, B. Hoffmann, Development of a high-resolution indoor radon map using a new machine learning-based probabilistic model and German radon survey data, *Environ. Health Perspect.* 132 (2024) <http://dx.doi.org/10.1289/ehp14171>.
- [15] Y. Bruinen de Bruin, K. Koistinen, S. Kephelopoulou, O. Geiss, S. Tirendi, D. Kotzias, Characterisation of urban inhalation exposures to benzene, formaldehyde and acetaldehyde in the European Union: Comparison of measured and modelled exposure data, *Environ. Sci. Pollut. Res.* 15 (2008) 417–430, <http://dx.doi.org/10.1007/s11356-008-0013-4>.
- [16] J.S. Park, K. Ikeda, Variations of formaldehyde and VOC levels during 3 years in new and older homes, *Indoor Air* 16 (2006) 129–135, <http://dx.doi.org/10.1111/j.1600-0668.2005.00408.x>.
- [17] W. Liu, J. Zhang, L. Zhang, B.J. Turpin, C.P. Weisel, M.T. Morandi, T.H. Stock, S. Colome, L.R. Korn, Estimating contributions of indoor and outdoor sources to indoor carbonyl concentrations in three urban areas of the United States, *Atmos. Environ.* 40 (2006) 2202–2214, <http://dx.doi.org/10.1016/j.atmosenv.2005.12.005>.
- [18] S. Langer, O. Ramalho, M. Derbez, J. Ribéron, S. Kirchner, C. Mandin, Indoor environmental quality in french dwellings and building characteristics, *Atmos. Environ.* 128 (2016) 82–91, <http://dx.doi.org/10.1016/j.atmosenv.2015.12.060>.
- [19] S. Yang, V. Perret, C. Hager Jörin, H. Niculita-Hirzel, J. Goyette Pernot, D. Licina, Volatile organic compounds in 169 energy-efficient dwellings in Switzerland, *Indoor Air* 30 (2020) 481–491, <http://dx.doi.org/10.1111/ina.12667>.
- [20] N. Liu, Z. Bu, W. Liu, H. Kan, Z. Zhao, F. Deng, C. Huang, B. Zhao, X. Zeng, Y. Sun, H. Qian, J. Mo, C. Sun, J. Guo, X. Zheng, L.B. Weschler, Y. Zhang, Indoor exposure levels and risk assessment of volatile organic compounds in residences, schools, and offices in China from 2000 to 2021: A systematic review, *Indoor Air* 32 (2022) e13091, <http://dx.doi.org/10.1111/ina.13091>.
- [21] A. Fernandez Lahore, R. Bethke, A. Daniels, K. Neumann, S. Ackermann, N. Schechner, K.-R. Brenske, E. Rucic, A. Murawski, M. Kolossa-Gehring, W. Birmili, Exposure of children and adolescents to volatile organic compounds in indoor air: Results from the German Environmental Survey 2014–2017 (GerES V), *Indoor Environ.* 2 (2025) 100082, <http://dx.doi.org/10.1016/j.indenv.2025.100082>.
- [22] W. Butte, B. Heinzow, Pollutants in house dust as indicators of indoor contamination, *Rev. Environ. Contam. Toxicol.* 175 (2002) 1–46.
- [23] L. Zhu, P. Hajeb, P. Fauser, K. Vorkamp, Endocrine disrupting chemicals in indoor dust: A review of temporal and spatial trends, and human exposure, *Sci. Total Environ.* 874 (2023) 162374, <http://dx.doi.org/10.1016/j.scitotenv.2023.162374>.
- [24] C. Schulz, A. Conrad, E. Rucic, G. Schwedler, L. Reiber, J. Peisker, M. Kolossa-Gehring, The German Environmental Survey for Children and Adolescents 2014–2017 (GerES V) – Study population, response rates and representativeness, *Int. J. Hyg. Environ. Health* 237 (2021) 113821, <http://dx.doi.org/10.1016/j.ijheh.2021.113821>.
- [25] W. Birmili, A. Daniels, R. Bethke, N. Schechner, G. Brasse, A. Conrad, M. Kolossa-Gehring, M. Debiak, J. Hurraß, E. Uhde, A. Omelan, T. Salthammer, Formaldehyde, aliphatic aldehydes (C2–C11), furfural, and benzaldehyde in the residential indoor air of children and adolescents during the German Environmental Survey 2014–2017 (GerES V), *Indoor Air* 32 (2022) e12927, <http://dx.doi.org/10.1111/ina.12927>.
- [26] M. Richter, F. Schühle, Experimental determination of 7-day uptake rates for diffusive sampling of 86 volatile and semi-volatile organic compounds relevant for indoor air monitoring and investigation on their sensitivity to exposure time and indoor climate, *Indoor Environ.* 2 (2025) 100095, <http://dx.doi.org/10.1016/j.indenv.2025.100095>.
- [27] R. Nagorka, W. Birmili, J. Schulze, J. Koschorreck, Diverging trends of plasticizers (phthalates and non-phthalates) in indoor and freshwater environments—why? *Env. Sci. Eur.* 34 (2022) 46, <http://dx.doi.org/10.1186/s12302-022-00620-4>.
- [28] A. für Innenraumrichtwerte, Bewertung von chemischen Innenraumluftverunreinigungen auf der Grundlage von Messergebnissen, *Bundesgesundheitsblatt* 68 (2025) 190–200, <http://dx.doi.org/10.1007/s00103-024-03999-y>.
- [29] T. Hesterberg, Bootstrap, *WIREs Comput. Stat.* 3 (2011) 497–526, <http://dx.doi.org/10.1002/wics.182>.
- [30] P.R. Bevington, D.K. Robinson, *Data Reduction and Error Analysis for the Physical Sciences*, McGraw Hill, New York, NY, 2003.
- [31] B.J.T. Morgan, *Elements of Simulation*, Chapman and Hall, London, 1984.
- [32] T. Salthammer, Calculation of kinetic parameters from chamber tests using nonlinear regression, *Atmos. Environ.* 30 (1996) 161–171, [http://dx.doi.org/10.1016/1352-2310\(95\)00055-4](http://dx.doi.org/10.1016/1352-2310(95)00055-4).
- [33] R.E. Walpole, R.H. Myers, S.L. Myers, K. Ye, *Probability and Statistics for Engineers and Scientists*, Pearson Education Ltd., Essex, UK, 2014.
- [34] D.V. O'Connor, D. Phillips, *Time-correlated Single Photon Counting*, Academic Press, London, UK, 1984.
- [35] T. Salthammer, Critical evaluation of approaches in setting indoor air quality guidelines and reference values, *Chemosphere* 82 (2011) 1507–1517, <http://dx.doi.org/10.1016/j.chemosphere.2010.11.023>.
- [36] B. Heinzow, H. Sagunski, Evaluation of indoor contaminants by means of reference and guide values: the German approach, in: T. Salthammer, E. Uhde (Eds.), *Organic Indoor Air Pollutants*, WILEY-VCH, Weinheim, 2009, pp. 189–211.
- [37] T. Salthammer, Y. Zhang, J. Mo, H.M. Koch, C.J. Weschler, Assessing human exposure to organic pollutants in the indoor environment, *Angew. Chem. Int. Ed.* 57 (2018) 12228–12263, <http://dx.doi.org/10.1002/anie.201711023>.
- [38] H. Solberg, Approved recommendation (1986) on the theory of reference values. Part 1. The concept of reference values, *Clin. Chim. Acta* 165 (1987) 111–118, [http://dx.doi.org/10.1016/0009-8981\(87\)90224-5](http://dx.doi.org/10.1016/0009-8981(87)90224-5).
- [39] H. Solberg, Approved recommendation (1987) on the theory of reference values. Part 5. Statistical treatment of collected reference values. Determination of reference limits, *Clin. Chim. Acta* 170 (1987) S13–S32, [http://dx.doi.org/10.1016/0009-8981\(87\)90151-3](http://dx.doi.org/10.1016/0009-8981(87)90151-3).
- [40] O.M. Poulson, E. Holst, J.M. Christensen, Calculation and application of coverage intervals for biological reference values, *Pure Appl. Chem.* 69 (1997) 1601–1612, <http://dx.doi.org/10.1351/pac199769071601>.

- [41] P.S. Horn, A.J. Pesce, Reference intervals: an update, *Clin. Chim. Acta* 334 (2003) 5–23, [http://dx.doi.org/10.1016/S0009-8981\(03\)00133-5](http://dx.doi.org/10.1016/S0009-8981(03)00133-5).
- [42] M. Hoopmann, A. Murawski, M. Schümann, T. Göen, P. Apel, N. Vogel, M. Kolossa-Gehring, C. Röhl, A revised concept for deriving reference values for internal exposures to chemical substances and its application to population-representative biomonitoring data in German children and adolescents 2014–2017 (GerES V), *Int. J. Hyg. Env. Health* 253 (2023) 114236, <http://dx.doi.org/10.1016/j.ijheh.2023.114236>.
- [43] A.H. Reed, R.J. Henry, W.B. Mason, Influence of statistical method used on the resulting estimate of normal range, *Clin. Chem.* 17 (1971) 275–284, <http://dx.doi.org/10.1093/clinchem/17.4.275>.
- [44] H. Fromme, M. Debiak, H. Sagunski, C. Röhl, M. Kraft, M. Kolossa-Gehring, The German approach to regulate indoor air contaminants, *Int. J. Hyg. Env. Health* 222 (2019) 347–354, <http://dx.doi.org/10.1016/j.ijheh.2018.12.012>.
- [45] B. Efron, R. Tibshirani, Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, *Stat. Sci.* 1 (1986) 54–75, <http://dx.doi.org/10.1214/ss/1177013815>.