Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

Estimating true mean value and standard deviation of data censored by limit of detection

Timotej Verbovšek
University of Ljubljana

Gregor Šega
University of Ljubljana

Abstract

Values below the limit of detection (LOD) are a problem in several fields of science, and there are numerous approaches for replacing the missing data. We present a new mathematical solution for maximum likelihood estimation that allows us to estimate the true values of the mean and standard deviation for normal and lognormal distributions and is significantly faster than previous implementations. We provide the implementation in R, Mathematica, and Excel.

Keywords: limit of detection (LOD), substitution value, Maximum Likelihood Estimation (MLE), exact calculation.

1. Introduction

Truncated samples pose problems when statistical analyses of data are performed, as laboratories can quantify only the values above the limit-of-detection (LOD) threshold. Several methods have been proposed to replace the unknown values below the LOD with values that ensure that errors are minimized when the statistical analyses are performed. The earliest studies were performed by Cohen (1949), Cohen (1950), and several comparisons of the substitution methods and reviews of each method were performed in various fields of mathematics, statistics, earth sciences, and medicine (Beal (2001), Gilliom and Helsel (2010), Hewett and Ganser (2007), Hornung and Reed (1990), Lambert, Peterson, and Terpenning (1991), Senn, Holford, and Hockey (2012),Succop, Clark, Chen, and Galke (2004)), with the most relevant studies performed by Helsel (Helsel (1990), Helsel (2005), Helsel (2006), Helsel (2011)).

The simplest methods rely on the substitution of missing values with some arbitrary number, and more advanced methods rely on predictions based on maximum likelihood estimation (MLE) or other statistical approaches. The latter are recommended; however, simple substitution is still performed and even suggested, for example, by the US Environmental Protection

Agency (EPA (2006)) for datasets with $< 15\%$ missing values. Simple substitution can be performed using some fraction of the LOD values (calculated as LOD $\times$ substitution value factor), and the common values or approaches for the replacement of the missing values are the following (Beal (2001), Senn *et al.* (2012)):

1. a constant of zero (substitution value factor equal to 0) is used
2. the value of the LOD itself (factor = 1) is used
3. some fraction of the LOD (usually LOD/2 or LOD/$\sqrt{2}$, factor = 1/2 or $1/\sqrt{2}$) is used
4. values are simply ignored and are not included in the analysis ("No Data" replacement)
5. other statistical approaches are used to replace the values (extrapolation, regression)

Substitution values also differ in different fields of study; Helsel (2006) noted that in water chemistry, the most common substitution value is LOD/2, and in air chemistry, the most common value is LOD/$\sqrt{2}$. Approaches for the substitution of multiple detection limits have been also used (Lee and Helsel (2007)).

As shown above, several approaches have been presented to find the "correct" substitution value (substitution value factor = C); however, none has introduced an exact mathematical solution to find such a value. A comparison of different substitution methods was performed by Verbovšek (Verbovšek (2011)) for the normal and lognormal (Perkins, Cutter, and Cleveland (1990)) distributions, which are commonly fitted to data. The results of the comparison "indicated that the best substitution method is by LOD $\times 1/\sqrt{2}$, as it produced the smallest errors." However, the results are highly dependent on the baseline data, so the conclusion is not definitive. Helsel (Helsel (2011)) suggested using MLE and derived the system of nonlinear equations, which should be "solved by iterative approximation using the Newton-Raphson method." The method is numerically exhausting and time-consuming and exceeds the knowledge of statistics and mathematics of a typical researcher who deals with LOD-censored data. It should be also emphasized that in many cases, the goal of the calculations is not to obtain the substituted values themselves (they just replace the missing values) but to estimate the basic statistics of distributions – the mean (μ) and the standard deviation (σ).

The main result of our method is a simple-to-use mathematically derived solution based on MLE that can calculate both the mean value and standard deviation of the observed normal/lognormal distribution. Our approach reduces the problem to solving only one equation, which can be done by any statistics computer tool (even a simple program such as MS Excel).

Our method also derives the substitution value factor (C), which is used for the replacement of missing data below the LOD. However, if we need to estimate the mean value, it turns out that the substitution factor C is not the same as the factor used to estimate the standard deviation. Thus, if one needs to estimate the mean and standard deviation of the underlying data, such a substitution should not be applied, and the parameters obtained by our procedure are the only correct estimates, because the MLE-derived substitution value relies on statistical parameters. Therefore, we would like to emphasize that we do not recommend the simple substitution methods, as these have long been described as poor methods for computing descriptive statistics (Gilliom and Helsel (2010)). We are, however, aware that many users

find the MLE method (or other high-level statistical methods) too advanced to use; therefore, our approach simplifies the MLE method for practical use.

As our approach is available to anyone performing an analysis (we provide the formulae for the Mathematica and R software and a template for the MS Excel software), we provide two calculation methods. The first approach is preferred for calculating the mean and/or standard deviation of the data. The second approach is for calculating the substitution value factor (C) from a model sample, which is then used to "simply" replace the missing values for the other sets of data. Therefore, with our work, we do not present just "another number" that should be used instead of the LOD, LOD $\times 1/2$, etc.; instead, we present a mathematically simple solution that estimates both the mean and standard deviation for the specific data and can be used to determine the factor C, which depends on the underlying data.

2. Methods

Our method is based on Maximum Likelihood Estimation. This approach uses three pieces of information to perform computations: (a) numerical values above reporting limits, (b) the proportion of data below each reporting limit, and (c) the mathematical formula for an assumed distribution Helsel (2011). The most crucial consideration for MLE is how well data fit the assumed distribution. A major problem with MLE is that for small data sets there is often insufficient information to determine whether the assumed distribution is correct or not, and so whether parameters are estimated reliably. MLE has been shown to perform poorly for data sets with less than 25–50 observations (Helsel (2011)) however simulations of real normally distributed data show that the method produces satisfactory results even for small samples. For larger data sets, MLE is an efficient way to estimate parameters, given that the chosen distribution is correct.

The proposed method is based on Maximum Likelihood Estimation. The setup is as follows: we have a sample of n values $X_1, X_2, \ldots, X_n$, but some of them are below the LOD threshold and cannot be measured. We aim to determine the mean value of all the values in the sample.

Suppose X_i are normally distributed random variables with mean μ and variance σ^2 (which are unknown to us) and we know the LOD. Our measured sample then consists of k measurements $x_1, x_2, \ldots, x_k$ and we know that another $n-k$ values are below LOD. We try to evaluate Maximum likelihood statistics of μ and σ. In order to do that we look at the likelihood of our sample:

$$L((x_1, x_2, \ldots, x_k), n-k, \mathrm{LOD}; \mu, \sigma) = \Phi\left(\frac{\mathrm{LOD}-\mu}{\sigma}\right)^{n-k} \prod_{i=1}^{k} (f((x_i-\mu)/\sigma)/\sigma)\,,$$

where $\Phi(x)$ is the cumulative distribution function of a standard normal variable and $f(x)$ is the probability density function of a standard normal variable. In order to maximize the function we must solve the following two equations:

$$\sigma(n-k)\frac{1}{\Phi\left(\frac{\mathrm{LOD}-\mu}{\sigma}\right)} f\left(\frac{\mathrm{LOD}-\mu}{\sigma}\right) = \sum_{i=1}^{k}(x_i-\mu) \tag{1}$$

and

$$(n-k)\frac{1}{\Phi\left(\frac{\mathrm{LOD}-\mu}{\sigma}\right)} f\left(\frac{\mathrm{LOD}-\mu}{\sigma}\right)(-(\mathrm{LOD}-\mu))\sigma - k\sigma^2 + \sum_{i=1}^{k}(x_i-\mu)^2 = 0\,. \tag{2}$$

It turns out that solving this system can be numericaly unstable, it usually is time consuming, and, most importantly, it is too complex to perform for an average user of LOD-censored measurements, who is most commonly a non-statistician.

Somewhat unexpected is that the system can be reduced to one equation which is easier to solve. It follows that μ can be computed from the value of σ:

$$\mu(k\mathrm{LOD} - \sum_{i=1}^{k} x_i) = k\sigma^2 + \mathrm{LOD}\sum_{i=1}^{k} x_i - \sum_{i=1}^{k} x_i^2\,. \tag{3}$$

The system reduces to one equation:

$$\sigma(n-k)\frac{1}{\Phi\left(\frac{k(\mathrm{LOD}^2-\sigma^2)+\sum_{i=1}^{k}x_i^2-2\mathrm{LOD}\sum_{i=1}^{k}x_i}{\sigma(k\mathrm{LOD}-\sum_{i=1}^{k}x_i)}\right)} f\left(\frac{k(\mathrm{LOD}^2-\sigma^2)+\sum_{i=1}^{k}x_i^2-2\mathrm{LOD}\sum_{i=1}^{k}x_i}{\sigma(k\mathrm{LOD}-\sum_{i=1}^{k}x_i)}\right)$$

$$= \sum_{i=1}^{k} x_i - k\frac{k\sigma^2 + \mathrm{LOD}\sum_{i=1}^{k}x_i - \sum_{i=1}^{k}x_i^2}{k\mathrm{LOD}-\sum_{i=1}^{k}x_i}\,.$$

This equation can be solved even with Excel, which is commonly used in fields where data below LOD can be found. The solution of the equation gives the estimate for σ, from which we can derive also the estimate for μ using (3).

3. Models and software

We provide the implementation in R, Mathematica, and Excel.

MS Excel.

Usage of `Lod_auto_compute_lognormal.xlsm`:

The file is a notebook in MS Excel. In order to run it properly the following must be prepared:

1. Macros must be enabled (it can be done when the file opens by pressing "Enable content")
2. Excel Add-on Solver is enabled (go to File > Options, click Add-Ins, and then in the Manage box, select Excel Add-ins. Click Go. In the Add-Ins available box, select the Solver Add-in check box, and then click OK.)

The file has two sheets, Normal (for normal distribution) and Lognormal (for lognormal distribution)

Normal: The data (measurements) should be entered (copied) into column B (yellow background, thick frame). Note that if there is already data in the B column, it should be first deleted (in case the new data has less entries). The cell F4 holds the number of the missing data, one must enter the number (again the cell has yellow background, thick frame). LOD

(level of detection) is entered into the cell F8. The values of σ and μ are computed by pressing the button "Calculate!". The initial value of σ can be (prior to Computing) entered in the cell F10. The value is used by internal procedures of MS Visual Basic so some initial values result in Solver returning the error message. In this case another initial value (higher or lower) can be used. For nearly all valid data the button "Reset sigma" writes an appropriate initial value into F10 which produces the correct final result. So when the Solver finds a solution one must choose "Keep Solver solution" and "OK", this writes the estimated value of σ into the cell F10. Estimate of μ is in the cell F15. If the value is not valid, there is ERROR! written next to it. In this case "Calculate!" should be used again (possibly with the changed initial value of σ, cell F10). Also the theoretical value of C, corresponding to this set of data, is calculated and shown in the cell F17.

Lognormal: Everything as with normal, with the addition of estimations of average and standard deviation (because they are not equal to the parameters μ and σ).

Mathematica.

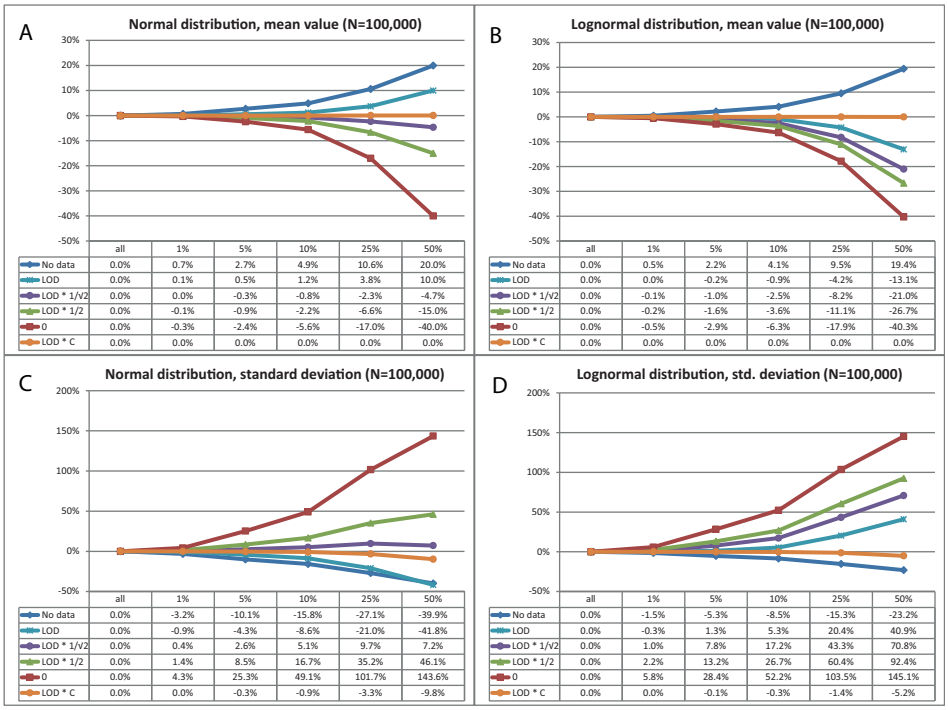
We provide the function `LODEstimateOfSigma[n, k, LOD, sumx, sumx2]` that returns the σ of the normal distribution. The parameters of the function are: n is the size of the sample, k is the number of known values (that are larger thamn LOD), LOD is the level of detection, $sumx$ is the sum of known values and $sumx2$ is the sum of squares of known values. The parameter μ of the normal distribution is calculated from σ using the formula derived by us. Examples are provided in Mathematica notebook `lod_normal.nb`.

R.

The same function is available in R and the example is provided in `lodR.r`.

4. Results and discussion

We have tested the proposed method on statistical datasets of normally and lognormally distributed data, which are two of the most commonly applied distributions for analysis data. We provide an Excel file with $N = 100,000$ (`LOD_Excel-100000.xlsx`) for testing the method, and two other files (`LOD_Excel-100.xlsx` and `LOD_Excel-36.xlsx`) with $N = 100$ and $N = 36$ as user templates (all three datasets are also used in Figure 2). In the tested dataset, five spreadsheet columns correspond to five degrees of truncation (missing data): 1%, 5%, 10%, 25% and 50%. We have tested six scenarios of substitution value factors (1, 0, 1/2, $1/\sqrt{2}$, C–our proposed calculated value, and No Data). For each of 30 possible outcomes (6 scenarios $\times$ 5 truncation values) of truncated and substituted sets, we calculated the mean value and standard deviation and compared them to the true values of the untruncated dataset. Figure 1 presents the comparison of all six scenarios depending on the degree of truncation for normal (Figure 1A and 1B) and lognormal distributions (Figure 1C and 1D). The comparison is quantified by plotting the error of the method vs. the degree of truncation. As the degree of truncation increases from 1% to 50%, the error increases rapidly for both distributions, up to the values of -40.0% and -40.3% for replacement with the zero value (Figures 1A and 1B). If No Data is used, the error is comparably large. The commonly used replacement methods with LOD $\times 1/2$ and LOD $\times 1/\sqrt{2}$ produce results that are approximately twice as good (error of -15.0% for LOD $\times 1/2$ and -4.7% for LOD $\times 1/\sqrt{2}$; the mean value and normal distribution), although the errors are still unacceptable for large degrees of truncation (larger than 5%). However, our suggested method gives the exact calculation and has zero error for



A	all	1%	5%	10%	25%	50%
No data	0.0%	0.7%	2.7%	4.9%	10.6%	20.0%
LOD	0.0%	0.1%	0.5%	1.2%	3.8%	10.0%
LOD * 1/√2	0.0%	0.0%	-0.3%	-0.8%	-2.3%	-4.7%
LOD * 1/2	0.0%	-0.1%	-0.9%	-2.2%	-6.6%	-15.0%
0	0.0%	-0.3%	-2.4%	-5.6%	-17.0%	-40.0%
LOD * C	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

B	all	1%	5%	10%	25%	50%
No data	0.0%	0.5%	2.2%	4.1%	9.5%	19.4%
LOD	0.0%	0.0%	-0.2%	-0.9%	-4.2%	-13.1%
LOD * 1/√2	0.0%	-0.1%	-1.0%	-2.5%	-8.2%	-21.0%
LOD * 1/2	0.0%	-0.2%	-1.6%	-3.6%	-11.1%	-26.7%
0	0.0%	-0.5%	-2.9%	-6.3%	-17.9%	-40.3%
LOD * C	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

C	all	1%	5%	10%	25%	50%
No data	0.0%	-3.2%	-10.1%	-15.8%	-27.1%	-39.9%
LOD	0.0%	-0.9%	-4.3%	-8.6%	-21.0%	-41.8%
LOD * 1/√2	0.0%	0.4%	2.6%	5.1%	9.7%	7.2%
LOD * 1/2	0.0%	1.4%	8.5%	16.7%	35.2%	46.1%
0	0.0%	4.3%	25.3%	49.1%	101.7%	143.6%
LOD * C	0.0%	0.0%	-0.3%	-0.9%	-3.3%	-9.8%

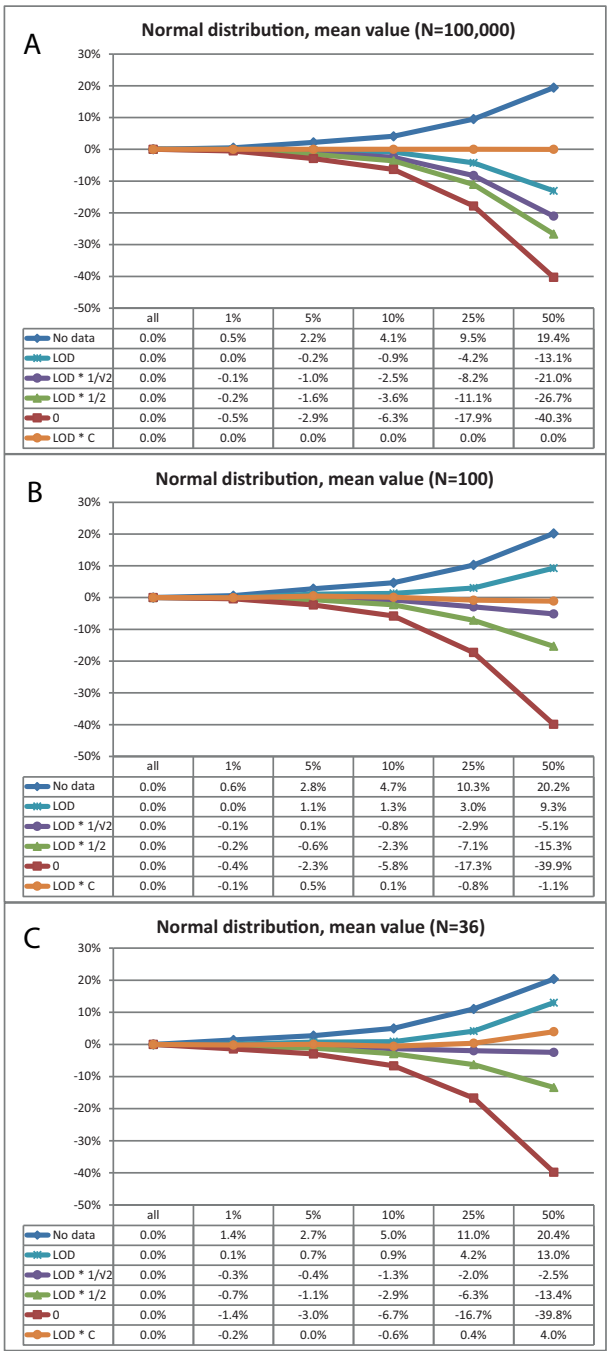
D	all	1%	5%	10%	25%	50%
No data	0.0%	-1.5%	-5.3%	-8.5%	-15.3%	-23.2%
LOD	0.0%	-0.3%	1.3%	5.3%	20.4%	40.9%
LOD * 1/√2	0.0%	1.0%	7.8%	17.2%	43.3%	70.8%
LOD * 1/2	0.0%	2.2%	13.2%	26.7%	60.4%	92.4%
0	0.0%	5.8%	28.4%	52.2%	103.5%	145.1%
LOD * C	0.0%	0.0%	-0.1%	-0.3%	-1.4%	-5.2%

Figure 1: Comparison of errors for the mean values, normal distribution (Figure 2A) and lognormal distribution (Figure 2B) and for the standard deviations, normal distribution (Figure 2C) and lognormal distribution (Figure 2D).

the calculation of the mean (Figures 1A and 1B). One should note that the error values are exactly zero, which confirms the correctness of the calculation and the replacement of the mean value with LOD $\times C$.

One should note that the calculation of the standard deviation shows some error (maximum values of -9.8% and -5.2% for both distributions, in the case of 50% truncation; see Figures 1C and 1D). However, the non-zero error values appear only if one uses the same replacement value of LOD $\times C$ for the estimation of both the mean and standard deviation. In the case of such a substitution, one of these two parameters (in this case, the mean value) is restrained to have zero error (Figures 1A and 1C), and the other always has a non-zero error (Figures 1B and 1D). Nevertheless, it is possible to estimate both the mean and standard deviation values correctly with our proposed calculation, provided in the file `LOD_auto_compute.xlsm`. If the data from Figure 1 are entered into this Excel template (or R or Mathematica), both the mean and standard deviation values are calculated with errors very close to zero, confirming the correctness of our method. The errors are progressively smaller as the amount of data grows. The process of inputting the data into the Excel file is explained in the file `readme.txt`.

The influence of the sample size is presented in Figure 2. It is obvious that the error is virtually equal to zero with a very large sample number of $N = 100,000$ (Figure 2A), but the method performs well even with small samples. With $N = 100$ measurements (Figure 2B), the average of all errors is -0.3%, smaller than that of any of the other replacement methods. Figure 2C shows that even with the smallest samples ($N = 36$), the error is negligible; even with 50% of the samples below the LOD, the result only varies by 4%, and our method again



A	all	1%	5%	10%	25%	50%
No data	0.0%	0.5%	2.2%	4.1%	9.5%	19.4%
LOD	0.0%	0.0%	-0.2%	-0.9%	-4.2%	-13.1%
LOD * 1/√2	0.0%	-0.1%	-1.0%	-2.5%	-8.2%	-21.0%
LOD * 1/2	0.0%	-0.2%	-1.6%	-3.6%	-11.1%	-26.7%
0	0.0%	-0.5%	-2.9%	-6.3%	-17.9%	-40.3%
LOD * C	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

B	all	1%	5%	10%	25%	50%
No data	0.0%	0.6%	2.8%	4.7%	10.3%	20.2%
LOD	0.0%	0.0%	1.1%	1.3%	3.0%	9.3%
LOD * 1/√2	0.0%	-0.1%	0.1%	-0.8%	-2.9%	-5.1%
LOD * 1/2	0.0%	-0.2%	-0.6%	-2.3%	-7.1%	-15.3%
0	0.0%	-0.4%	-2.3%	-5.8%	-17.3%	-39.9%
LOD * C	0.0%	-0.1%	0.5%	0.1%	-0.8%	-1.1%

C	all	1%	5%	10%	25%	50%
No data	0.0%	1.4%	2.7%	5.0%	11.0%	20.4%
LOD	0.0%	0.1%	0.7%	0.9%	4.2%	13.0%
LOD * 1/√2	0.0%	-0.3%	-0.4%	-1.3%	-2.0%	-2.5%
LOD * 1/2	0.0%	-0.7%	-1.1%	-2.9%	-6.3%	-13.4%
0	0.0%	-1.4%	-3.0%	-6.7%	-16.7%	-39.8%
LOD * C	0.0%	-0.2%	0.0%	-0.6%	0.4%	4.0%

Figure 2: Comparison of errors for the mean values, normal distribution, with sample size of 100.000 (Figure 3A), 100 (Figure 3B) and 36 (Figure 3C).

has the lowest average error (0.7%). The true error of our suggested method is therefore equal to zero, but it deviates slightly from this due to a low number of samples.

Computational details

The results in this paper were obtained using MS Excel 2016, Mathematica 11.3 and R 3.4.1. R itself and all packages used are available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/.

Conclusions

Our newly proposed mathematical solution for MLE therefore allows us to estimate the true values of the mean and standard deviation for normal and lognormal distributions, and it is significantly faster than previous implementations; we encourage users to use this approach instead of simple substitution methods.

References

Beal SL (2001). "Ways to Fit a PK Model with Some Data Below the Quantification Limit." *Journal of Pharmacokinetics and Pharmacodynamics*, **28**(5), 481–504. doi:10.1023/A:1012299115260. URL https://doi.org/10.1023/A:1012299115260.

Cohen AC (1949). "On Estimating the Mean and Standard Deviation of Truncated Normal Distributions." *Journal of the American Statistical Association*, **44**(248), 518–525. doi:10.1080/01621459.1949.10483324. https://www.tandfonline.com/doi/pdf/10.1080/01621459.1949.10483324, URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1949.10483324.

Cohen AC (1950). "Estimating the Mean and Variance of Normal Populations from Singly Truncated and Doubly Truncated Samples." *Annals of Mathematical Statistics*, **21**(4), 557–569. doi:10.1016/0304-4076(86)90002-3. https://www.jstor.org/stable/2236606, URL https://projecteuclid.org/euclid.aoms/1177729751.

EPA (2006). *Data Quality Assessment: Statistical Methods for Practitioners. (EPA QA/G-9S)*. United States Environmental Protection Agency. URL https://www.epa.gov/quality/guidance-data-quality-assessment.

Gilliom RJ, Helsel DR (2010). "Estimation of Distributional Parameters for Censored Trace Level Water Quality Data: 1. Estimation Techniques." *Water Resources Research*, **22**(2), 135–146. doi:10.1029/WR022i002p00135. https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/WR022i002p00135, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/WR022i002p00135.

Helsel DR (1990). "Less than obvious - statistical treatment of data below the detection limit." *Environmental Science & Technology*, **24**(12), 1766–1774. doi:10.1021/es00082a001. https://doi.org/10.1021/es00082a001, URL https://doi.org/10.1021/es00082a001.

Helsel DR (2005). "More Than Obvious: Better Methods for Interpreting Nondetect Data." *Environmental Science & Technology*, **39**(20), 419A–423A. doi:10.1021/es053368a.

PMID: 16295833, https://doi.org/10.1021/es053368a, URL https://doi.org/10.1021/es053368a.

Helsel DR (2006). "Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it." *Chemosphere*, **65**(11), 2434 – 2439. ISSN 0045-6535. doi:https://doi.org/10.1016/j.chemosphere.2006.04.051. Environmental Chemistry, URL http://www.sciencedirect.com/science/article/pii/S0045653506005157.

Helsel DR (2011). *Statistics for Censored Environmental Data Using Minitab®and R*. 2nd edition. John Wiley & Sons, Ltd, Hoboken, New Jersey. ISBN 9781118162729. doi:10.1002/9781118162729. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9781118162729.

Hewett P, Ganser GH (2007). "A Comparison of Several Methods for Analyzing Censored Data." *The Annals of Occupational Hygiene*, **51**(7), 611–632. ISSN 0003-4878. doi:10.1093/annhyg/mem045. https://academic.oup.com/annweh/article-pdf/51/7/611/332745/mem045.pdf, URL https://doi.org/10.1093/annhyg/mem045.

Hornung RW, Reed LD (1990). "Estimation of Average Concentration in the Presence of Nondetectable Values." *Applied Occupational and Environmental Hygiene*, **5**(1), 46–51. doi:10.1080/1047322X.1990.10389587. https://doi.org/10.1080/1047322X.1990.10389587, URL https://doi.org/10.1080/1047322X.1990.10389587.

Lambert D, Peterson B, Terpenning I (1991). "Nondetects, Detection Limits, and the Probability of Detection." *Journal of the American Statistical Association*, **86**(414), 266–277. doi:10.1080/01621459.1991.10475030. https://www.tandfonline.com/doi/pdf/10.1080/01621459.1991.10475030, URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1991.10475030.

Lee L, Helsel D (2007). "Statistical analysis of water-quality data containing multiple detection limits II: S-language software for nonparametric distribution modeling and hypothesis testing." *Computers & Geosciences*, **33**(5), 696 – 704. ISSN 0098-3004. doi:https://doi.org/10.1016/j.cageo.2006.09.006. URL http://www.sciencedirect.com/science/article/pii/S0098300406002056.

Perkins JL, Cutter GN, Cleveland MS (1990). "Estimating the Mean, Variance, and Confidence Limits from Censored (<Limit of Detection), Lognormally-Distributed Exposure Data." *American Industrial Hygiene Association Journal*, **51**(8), 416–419. doi:10.1080/15298669091369871. https://doi.org/10.1080/15298669091369871, URL https://doi.org/10.1080/15298669091369871.

Senn S, Holford N, Hockey H (2012). "The ghosts of departed quantities: approaches to dealing with observations below the limit of quantitation." *Statistics in Medicine*, **31**(30), 4280–4295. doi:10.1002/sim.5515. https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.5515, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5515.

Succop P, Clark S, Chen M, Galke W (2004). "Imputation of Data Values That are Less Than a Detection Limit." *Journal of occupational and environmental hygiene*, **1**, 436–41. doi:10.1080/15459620490462797.

Verbovšek T (2011). "A comparison of parameters below the limit of detection in geochemical analyses by substitution methods." *RMZ - Materials and Geoenvironment*, **58**(4), 393–404.

Affiliation:

Timotej Verbovšek
University of Ljubljana
Faculty of Natural Sciences and Engineering
Department of Geology
Aškerčeva ulica 12
SI-1000 Ljubljana, Slovenia
E-mail: timotej.verbovsek@ntf.uni-lj.si

Gregor Šega
University of Ljubljana
Faculty of Mathematics and Physics
Department of Mathematics
Jadranska ulica 19
SI-1000 Ljubljana, Slovenia
E-mail: gregor.sega@fmf.uni-lj.si

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd