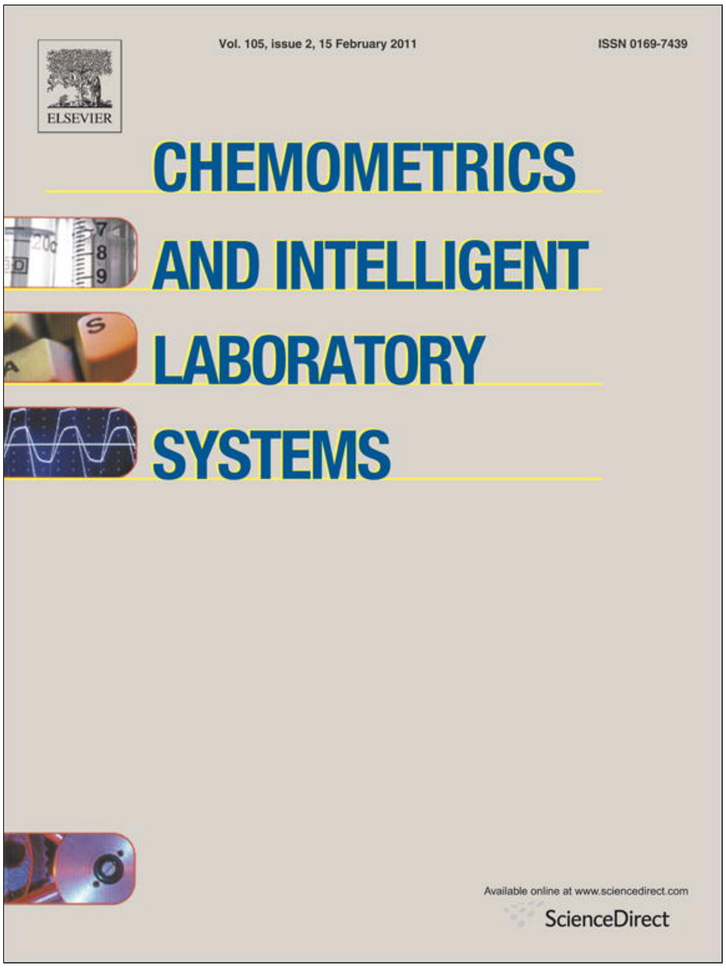

Vol. 105, issue 2, 15 February 2011
ISSN 0169-7439
ELSEVIER
CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS
Available online at www.sciencedirect.com
ScienceDirect

Chemometrics and Intelligent Laboratory Systems 105 (2011) 188–194

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Limit of detection determination for censored samples

Andrew L. Rukhin [*], Daniel V. Samarov

Statistical Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD 20899-0001, United States

ARTICLE INFO

Article history:
Received 6 October 2010
Received in revised form 10 December 2010
Accepted 4 January 2011
Available online 11 January 2011

Keywords:
Confidence bounds
Critical level
Decision limit
Heteroscedasticity
Maximum likelihood estimator
Tolerance limits
Truncation parameter

ABSTRACT

The problem of setting the limit of detection is considered for censored samples and heterogeneous errors. After formal definitions of the critical level and of the method detection limit, we obtain simplified maximum likelihood-type estimators. The estimation problem of the truncation parameter and its uncertainty are reviewed. Upper confidence bounds for the limit of detection are derived and some simulation results are given.

Published by Elsevier B.V.

1. Setting of the problem: Critical level, decision limit, and limit of detection

This work was motivated by the problem of Limit of Detection (LOD) determination in trace explosive particle detectors. To solve the problem in the classical setting of continuous variables (and only this case is considered here), a homogeneous linear regression model relating the mass x to the values of instrument response $Y(x)$ is usually employed [1]. In this set-up the limit of detection is defined as the mass value x_c with a given quantile of $Y(x_c)$-distribution.

This setting is extended to allow heterogeneous and/or censored samples. Namely, the following model for the readings $Y(x)$ as a function of mass x was suggested by the data obtained in the Surface and Microanalysis Science Division at the National Institute of Standards and Technology (NIST):

$$Y(x)=\begin{cases}0, & W(x)<h;\\ W(x) & W(x)\geq h.\end{cases} \tag{1}$$

Here, $W(x)=a+bx+$noise, is the response with the mean $a+bx$ and the variance σ^2, which (for sufficiently small positive x) follows a linear regression model with Gaussian errors. An unknown non-negative threshold parameter h is such that positive readings are possible if and only if the underlying normally distributed amplitude $W(x)$ is above this terminus h.

When $x=0$, i.e., in the sample of blanks, the error variance may be different, say, σ_0^2. The intercept a interpretable as the background noise is the mean of the blank sample, only uncensored part of which is observed.

Thus our model has five unknown parameters a,b,h σ_0, and σ, such that

$$Pr(Y(x)=0)=\Phi\left(\frac{h-a-bx}{\sigma}\right),\quad x>0,$$
$$Pr(Y(0)=0)=\Phi\left(\frac{h-a}{\sigma_0}\right).$$

Here Φ is the distribution function of standard normal distribution. (The notation list is provided in Appendix A.) While all these parameters are of interest, to estimate the limit of detection, a more convenient parametrization is through b,h σ_0, σ, and $z_\star=(h-a)/\sigma_0$, so that $\Phi(z_\star)=Pr(Y(0)=0)$. Possibility of censored samples and heterogeneous errors as well as the following definitions are in agreement with the existing methodology [2].

The first task is to set an alarm threshold *critical level* (LC) such that the probability of false positive readings is a given number α (say, $\alpha=0.05$), i.e.,

$$Pr(Y(0)\leq LC)=1-\alpha. \tag{2}$$

So LC is the $(1-\alpha)$-th quantile of the $Y(0)$-distribution corresponding to the blank sample, but only provided that it exceeds h, which means that $Pr(Y(0)=0)<1-\alpha$.

[*] Corresponding author. Tel.: +1 301 975 2951; fax: +1 301 975 3144.
E-mail addresses: andrew.rukhin@nist.gov (A.L. Rukhin), daniel.samarov@nist.gov (D.V. Samarov).

0169-7439/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.chemolab.2011.01.001

Let $z_{1-\alpha}$ denote the percentile of standard normal distribution, $P(Z>z_{1-\alpha})=\alpha$. Since $h=a+z_{\star}\sigma_0$,

$$\mathrm{LC} = \max(a + z_{1-\alpha}\sigma_0, h) = a + \max(z_{1-\alpha}, z_{\star})\sigma_0.$$

Practical importance of the critical level is that it can be used as a benchmark when designing a LOD experiment after observing the blank sample. Indeed it is desirable to have future observations exceeding LC. However, as was noted by the referee, sometimes the instrument parameters are set so as to balance false positive and false negatives vis a vis the LOD.

The mass value x_c, at which only a small percentage 100β% of the amplitude readings is below LC, by definition represents the Limit of Detection (LOD) (also known as the Method Detection Limit),

$$Pr(Y(x_c)>LC) = 1-\beta.$$

Thus the probability of false negative readings when $x \geq x_c$ cannot exceed the given number β. Since

$$\Phi\left(\frac{LC-a-bx_c}{\sigma}\right) = \beta,$$

x_c satisfies the formula,

$$a + bx_c = LC + z_{1-\beta}\sigma = a + \max(z_{1-\alpha}, z_{\star})\sigma_0 + z_{1-\beta}\sigma.$$

Therefore, if $y_c = LC + z_{1-\beta}\sigma$ is the so-called *decision limit*, the LOD can be determined from any of the following identities,

$$x_c = \frac{y_c-a}{b} = \frac{LC + z_{1-\beta}\sigma-a}{b} = \frac{\max(z_{1-\alpha}, z_{\star})\sigma_0 + z_{1-\beta}\sigma}{b}.$$

It is clear that LOD involves four parameters $z_{\star}, \sigma_0, \sigma$, and b, as well as two error probabilities α and β. These two probabilities are chosen by the instrument developer to keep the rates of false positives and false negatives small. The error probability α enters the definition of LC as in Eq. (2) corresponding to the blank sample only. However, the LOD depends on both α and β. Fig. 1 illustrates graphically the LC, the LOD and the decision limit.

In Section 2 we discuss the estimation of LC which involves that of a, h and σ_0. Explicit, simplified maximum likelihood-type estimators are suggested there. Section 3 deals with the LOD estimation, and upper confidence bounds are discussed in Section 4. These mathematical results were used as a basis for e-metrology tool at a public server, www.limitsofdetection.com/lod_astm.htm as discussed in Section 5. Some simulation results are provided in Section 6. All mathematical derivations are collected in Appendix A, in particular the estimation problem of the truncation parameter h and its uncertainty is reviewed there.

2. Blank sample: LC estimation

Let the available data be $x_i, y_i = (y_1^{(i)}, \ldots, y_{n_i}^{(i)}), i = 0, 1, \ldots, N$, n_i denoting the sample sizes (in ion mobility spectrometry applications, usually $n_0=30, n_i \equiv 15$ for $i \geq 1$.) Here $x_0 < x_1 < \cdots < x_N$ are the different mass values, so that y_i is the corresponding vector of n_i responses. In most cases $x_0=0$, but there are situations in which x_0 is a small positive mass.

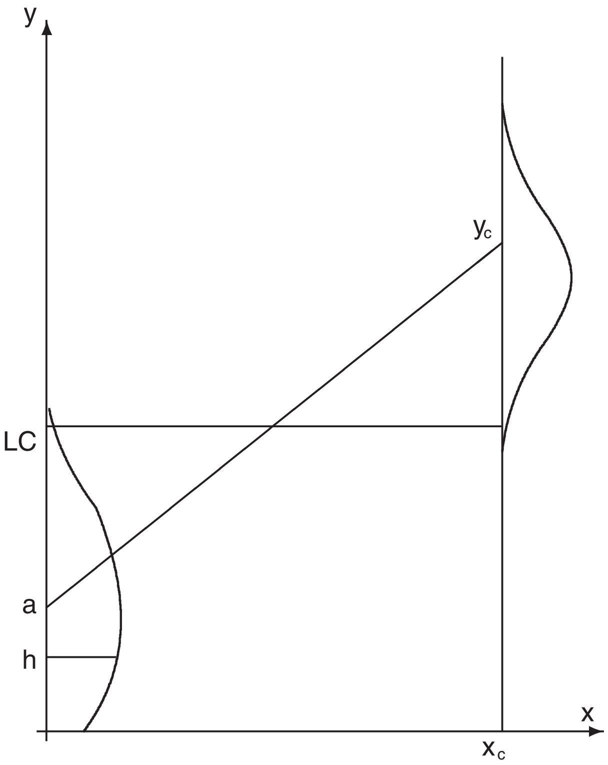

Fig. 1. The graphical interpretation of LC, x_c and y_c. The data points below h are unobservable.

Assuming that $x_0=0$ and that the number of all positive responses in the blank sample, m_0, is positive, the maximum likelihood estimator of the background level a on the basis of the blank sample maximizes

$$\Phi^{\nu_0}\left(\frac{h-a}{\sigma_0}\right) \prod_{j:y_j^{(0)}>0} \frac{1}{\sigma_0}\phi\left(\frac{y_j^{(0)}-a}{\sigma_0}\right), \quad h \leq \min_{j:y_j^{(0)}>0} y_j^{(0)}. \tag{3}$$

Here ϕ is the density of standard normal distribution, and ν_0 denotes the number of zero readings in the sample of blanks, $\nu_0 = n_0 - m_0$.

The maximum likelihood estimator of the terminus h on the basis of the blank sample is

$$\hat{h} = \min_{j:y_j^{(0)}>0} y_j^{(0)}. \tag{4}$$

With h replaced by this estimate, the problem becomes that of finding the values of a and σ_0 on the basis of a left censored normal sample in which m_0 observations are fully measured while ν_0 data points are censored at $\hat{h}$.

This is a classical statistical problem which appears in biostatistics, and life-time data analysis. See for example [3] where the numerical solution of likelihood equations is performed via tabulation of an auxiliary function of $\sum_{j=1}^{m_0}\left(y_j^{(0)}-\hat{h}\right)^2 / \left(\bar{y}^{(0)}-\hat{h}\right)^2$ and of m_0/n_0. Chemistry applications are discussed in [4]. Environmental applications are in [5,7].

A more advanced method implemented in *R*-software is also available in [6] where maximum likelihood estimators are developed for the log-normal setting. As there is no explicit form solution, the parameter estimates require a numerical optimization procedure with potential convergence problems. Our simpler approximate solution has a closed form given below.

To estimate $z_\star$ we use the fact that $\Phi(z_\star) \approx \nu_0/n_0$ and employ a Bayesian method with noninformative Jeffrey's prior by approximating $\Phi(z_\star)$ via $(\nu_0+0.5)/(n_0+1)$ [8]. The reason for this approximation commonly used in categorical data analysis is that it is applicable when $\nu_0=0$ or when $\nu_0=n_0$. This suggestion leads to the statistic,

$$\hat{z}_\star = \Phi^{-1}\left(\frac{\nu_0+0.5}{n_0+1}\right),$$

This formula along with Eq. (18) in Appendix A gives the explicit estimator of σ_0,

$$\hat{\sigma}_0 = \frac{\hat{z}_\star\left(\bar{y}_0 - \hat{h}\right)}{2} + \sqrt{\left(1+\frac{\hat{z}_\star^2}{4}\right)\left(\bar{y}_0 - \hat{h}\right)^2 + s_0^2}. \tag{5}$$

Here, $\bar{y}_0 = \sum_{j:y_j^{(0)}>0} y_j^{(0)} / m_0$, is the mean of all positive responses in the blank sample, $s_0^2 = m_0^{-1}\sum_{j:y_j^{(0)}>0}\left(y_j^{(0)} - \bar{y}_0\right)^2$. An associated estimate of the intercept a follows from the definition of $z_\star$,

$$\hat{a} = \hat{h} - \hat{z}_\star\hat{\sigma}_0. \tag{6}$$

Since $\hat{h}$ overestimates h, an almost unbiased estimator of h

$$\tilde{h} = 2\hat{h} - \min_{j:y_j^{(0)}>\hat{h}} y_j^{(0)}, \tag{7}$$

can be employed [11].

The estimator $\tilde{h}$ leads to a better estimate of σ_0,

$$\tilde{\sigma}_0 = \frac{\hat{z}_\star\left(\bar{y}_0 - \tilde{h}\right)}{2} + \sqrt{\left(1+\frac{\hat{z}_\star^2}{4}\right)\left(\bar{y}_0 - \tilde{h}\right)^2 + s_0^2}, \tag{8}$$

and of the background level a,

$$\tilde{a} = \tilde{h} - \hat{z}_\star\tilde{\sigma}_0. \tag{9}$$

The corresponding estimator of LC $-$ a has the form,

$$\widetilde{LC-a} = \max\left(z_{1-\alpha}\tilde{\sigma}_0, \tilde{h} - \tilde{a}\right) = \max(z_{1-\alpha}, \hat{z}_\star)\tilde{\sigma}_0.$$

If $m_0 \geq 2$, the upper confidence bound, $\bar{\sigma}_0$, on σ_0 can be taken to be $\bar{\sigma}_0 = \hat{\sigma}_0\sqrt{(m_0-1)/\chi_\gamma^2(m_0-1)}$. Here $\chi_\gamma^2(\nu)$ is the γ-th percentile of the $\chi^2(\nu)$-distribution. Then $a + z_{1-\alpha}\bar{\sigma}_0$ is an upper confidence bound on the $(1-\alpha)$-th percentile of the censored normal $Y(0)$-sample. One of the ways to determine it directly, is to apply available methods for confidence estimation of lognormal distribution parameters as described in [9, Sec 5.2.2c]. See also [10].

If $x_0 > 0$, then estimating σ_0 is not feasible, but we can put

$$z_\star = (h - a - bx_0)/\sigma,$$

so that $LC - a = bx_0 + z_\star\sigma$, and

$$x_c = \frac{LC - a + z_{1-\beta}\sigma}{b} = x_0 + \frac{\left(z_\star + z_{1-\beta}\right)\sigma}{b}.$$

Observe that $\hat{z}_\star = \Phi^{-1}((\nu_0+0.5)/(n_0+1))$ still can be used as an estimator of $z_\star$. Thus, in all situations, the LOD determination reduces to estimation of parameters b and σ. This problem is considered in Section 3.

3. LOD estimator and its uncertainty

If $x_0 = 0$, we can treat the centered data from non-blank samples, $\tilde{y}_j^{(i)} = y_j^{(i)} - \tilde{a}, i = 1, \ldots, N,$ as satisfying the homogeneous linear model with zero intercept, $\tilde{y}_j^{(i)} = bx_i + \epsilon_{ij}$, and constant variance $\mathrm{Var}(\epsilon_{ij}) \equiv \sigma^2, i = 1, \ldots, N, j = 1, \ldots, n_i$.

Since b cannot be estimated on a sample of blanks alone, estimation of LOD demands at least one sample with positive readings for a positive mass. If the i-th sample has $m_i (m_i \leq n_i)$ positive responses, it is required that $m_1 \geq 2$, i.e., there must be at least two positive readings for the smallest positive mass x_1 (otherwise this sample is discarded and a sample for a larger mass is drawn.)

The estimators of b and σ can be determined in a straightforward way if the positive mass samples have only a small percentage of zeros. Namely, let

$$d = \frac{1}{\sqrt{\sum_i m_i x_i^2}}, \tag{10}$$

$$\hat{b} = d^2\sum_i m_i x_i \bar{y}_i, \tag{11}$$

and

$$\hat{\sigma}^2 = \frac{1}{\nu}\sum_{i,j}\left(\tilde{y}_j^{(i)} - \hat{b}x_i\right)^2. \tag{12}$$

Here $\bar{y}_i = \sum_j \tilde{y}_j^{(i)} / m_i$ is the mean of all positive responses in the i-th sample, and $\nu = \sum_{i=1}^N m_i - 1$.

The approximate distributions of the estimators b and σ are supposed to be:

$$\hat{b} \sim b + d\sigma Z,$$

where

$$\hat{\sigma}^2 \sim \sigma^2\frac{\chi^2(\nu)}{\nu},$$

with independent standard normal Z and χ^2-distributed random variable $\chi^2(\nu)$ (ν denotes the degrees of freedom.) Then an upper $(1-\gamma/2)$-confidence bound on σ can be found, $\bar{\sigma}^2 = \nu\hat{\sigma}^2/\chi_{0.5\gamma}^2(\nu)$. Thus, $P\left(\sigma^2 < \bar{\sigma}^2\right) = 1 - \gamma$.

A natural point estimate of the LOD is

$$\widehat{\mathrm{LOD}} = \frac{\max(z_{1-\alpha}, \hat{z}_\star)\tilde{\sigma}_0 + z_{1-\beta}\hat{\sigma}}{\hat{b}}. \tag{13}$$

To adjust for its bias, one can use the estimator

$$\widetilde{\mathrm{LOD}} = \widehat{\mathrm{LOD}}\left[1 - \frac{\hat{\sigma}^2}{\hat{b}^2 d^2}\right]_+$$

[12].

To find the approximate uncertainty, i.e., the square root of the mean squared error approximable by $b^{-2}\left[\mathrm{Var}(\max(z_\alpha, \hat{z}_\star)\tilde{\sigma}_0) + \mathrm{Var}(z_\beta\hat{\sigma}) + \mathrm{Var}\left(\hat{b}\right)\mathrm{LOD}^2\right]$, the propagation of error formula gives,

$$\widetilde{\mathrm{Var}}\left(\widehat{\mathrm{LOD}}\right) = \frac{1}{\hat{b}^2}\left[\left[\max(z_{1-\alpha}, \hat{z}_\star)\right]^2\frac{\tilde{\sigma}_0^2}{m_0} + z_{1-\beta}^2\frac{\hat{\sigma}^2}{\nu} + d^2\hat{\sigma}^2\widetilde{\mathrm{LOD}}^2\right]. \tag{14}$$

4. Confidence and tolerance bounds on LOD

An upper $(1-\gamma)$ confidence limit H for the LOD satisfies the condition,

$$Pr\,(\mathrm{LOD} \le H) \ge 1-\gamma.$$

It makes sense to seek this confidence bound to be of the form

$$H = \frac{\max(z_{1-\alpha}, \bar{z}_\star)\bar{\sigma}_0 + z_{1-\beta}\bar{\sigma}}{\bar{b}}, \qquad (15)$$

where $\bar{z}_\star$ is an upper bound on $z_\star$ and $\bar{b} = \hat{b} - c\hat{\sigma}, c > 0$, is interpretable as a lower confidence bound for b. To specify H, we must choose the constant c.

Let $T(\nu)$ be a random variable with t-distribution, ν degrees of freedom, and let $t_{1-\gamma/2}(\nu)$ be the $(1-\gamma/2)$-percentile of this distribution, $Pr(T(\nu) \le t_{1-\gamma/2}(\nu)) = 1-\gamma/2$. Then with $c = dt_{1-\gamma/2}(\nu)$ one has

$$Pr\left(b \ge \bar{b}\right) = Pr\left(\hat{b} - b \le c\hat{\sigma}\right)$$

$$= Pr\left(dZ \le c\sqrt{\frac{\chi^2(\nu)}{\nu}}\right) = Pr\left(T(\nu) \le \frac{c}{d}\right) = 1-\gamma/2.$$

If $\bar{z}_\star, \bar{\sigma}_0$ and $\bar{\sigma}$ are such that

$$Pr\left(\max(z_{1-\alpha}, \bar{z}_\star)\bar{\sigma}_0 + z_{1-\beta}\bar{\sigma} \ge \max(z_{1-\alpha}, z_\star)\sigma_0 + z_{1-\beta}\sigma\right) \ge 1-\gamma/2,$$

the Bonferroni inequality [12] shows that indeed H is an upper $(1-\gamma)$ confidence bound on LOD.

Of course this method can work only if $\bar{b} > 0$, i.e., if

$$\frac{\hat{b}}{\hat{\sigma}} > c = dt_{1-\gamma/2}(\nu) = \frac{t_{1-\gamma/2}(\nu)}{\sqrt{\sum_i m_i x_i^2}}. \qquad (16)$$

When the estimators are based on just one sample, $(N=1)$, a heuristic interpretation of Eq. (16) is that the estimated variation coefficient $\hat{\sigma}/\left(\hat{b}x\right)$ of $Y(x)$ is small enough.The upper tolerance bound U for LOD gives a region which with probability $1-\gamma$ guarantees that the future $Y(\mathrm{LOD})$ observation will be inside it with a large probability $1-P$,

$$Pr\,(Pr\,(Y(LOD) \le a + bU) \ge 1-P) \ge 1-\gamma,$$

or

$$Pr(bU - bx_c \ge z_{1-P}\sigma) \ge 1-\gamma.$$

In other words, in the case of the tolerance bound U for LOD, we expect that $U \ge x_c + z_{1-P}\sigma/b$, while for the confidence bound H a less stringent condition, $H \ge x_c$, must hold.

The Bonferroni inequality shows that with $\bar{b}$, as defined earlier, a conservative choice for U is

$$U = \frac{\max(z_{1-\alpha}, \bar{z}_\star)\bar{\sigma}_0 + \left(z_{1-\beta} + z_{1-P}\right)\bar{\sigma}}{\bar{b}}. \qquad (17)$$

This upper limit has the form of the Lieberman–Miller tolerance bound [12].

As an alternative approach for deriving these bounds, notice that for the given value of the decision limit y_c, the ratio $\left(a + \hat{b}x_c - y_c\right)/\left(dx_c\hat{\sigma}\right)$ has an approximate t-distribution with ν degrees of freedom. Assuming that Eq. (16) holds, set this ratio equal to $t_{1-\gamma/2}(\nu)$ to get

$$Pr\left(x_c \le \frac{y_c - a}{\hat{b} - t_{1-\gamma/2}(\nu)d\hat{\sigma}}\right) \ge 1-\gamma/2.$$

By replacing $y_c - a$ by its upper $(1-\gamma/2)$-confidence bound $\max(z_{1-\alpha}, \bar{z}_\star)\bar{\sigma}_0 + z_{1-\beta}\bar{\sigma}$, a $(1-\gamma)$ confidence bound for LOD obtains,

$$\frac{\max(z_{1-\alpha}, \bar{z}_\star)\bar{\sigma}_0 + z_{1-\beta}\bar{\sigma}}{\hat{b} - dt_{1-\gamma/2}(\nu)\hat{\sigma},}$$

and this bound coincides with Eq. (15).

5. Implementation

The methodology presented in the previous sections was used as a basis for e-metrology tool at NIST and also at a public server, www.limitsofdetection.com/lod_astm.htm. It was intended to facilitate the calculation of the limit of detection based on definitions and assumptions which were used to formulate the proposed ASTM standard method WK 19817. The goal is to provide the users, manufacturers and testers of detectors with a practical tool to determine LOD for an analyte of choice (notably, explosives or drugs-of-abuse). An analysis of process blanks, as described in Section 2, is performed to determine the background level a, Eq. (6), and variation σ_0, Eq. (5), leading to LC. Then reference swipes can be prepared and analyzed starting at low analyte levels and working up to a level that elicits responses above LC which allows LOD determination.

The output includes the linear least squares estimator of the slope b, Eq. (11), the estimator of the standard error σ, Eq. (12), LOD estimator, Eq. (13), uncertainty of this estimator, Eq. (14), and the upper confidence limit on LOD, Eq. (15).

6. Simulation results

In this section we present the results from two simulation studies. The first study compares the performance of the simplified likelihood-type estimators (5) and (6) (which we refer to as sMLE) against the traditional MLE based approach in the normal and log-normal distribution settings. The second study is a validation of the upper bound on the LOD, H shown in Eq. (11).

The goal of comparing the sMLE against the MLE in the normal and log-normal settings is to see how our approach compares to the current "state of art". Also we wanted to illustrate that under a simple construction which does not require any type of numerical optimization, one is able to get comparable results. Specifically, because the MLE and LMLE do not have a closed form solution for LOD, an iterative method is needed to estimate the model parameters. Our explicit estimator is intuitively appealing without risking a convergence failure.

In the first study we focus on the sample of blanks, i.e. when $x=0$, so that $W(0) \sim N(a, \sigma_0)$, with population parameters $\sigma_0 = 30$, $a = 40$ and $h \in [20, 50]$. Note that the values of the parameters used in this study reflect quantities which have been observed in practice. Using these parameters the data are sampled from Eq. (1) with $x=0$.

In order to assess the performance of the sMLE and MLE approaches we use the Mean Squared Error (MSE) of their respective estimates of the parameters a and σ_0. The MSE for each parameter is calculated as follows: for a given h, we draw $N = 1000$ random samples of size $M = 15$ or 30 from Eq. (1). For each sample drawn a and σ_0 are estimated using the sMLE and MLE procedures. Finally the MSEs are calculated by taking the average of the sum of squares difference between the estimated and population parameters.

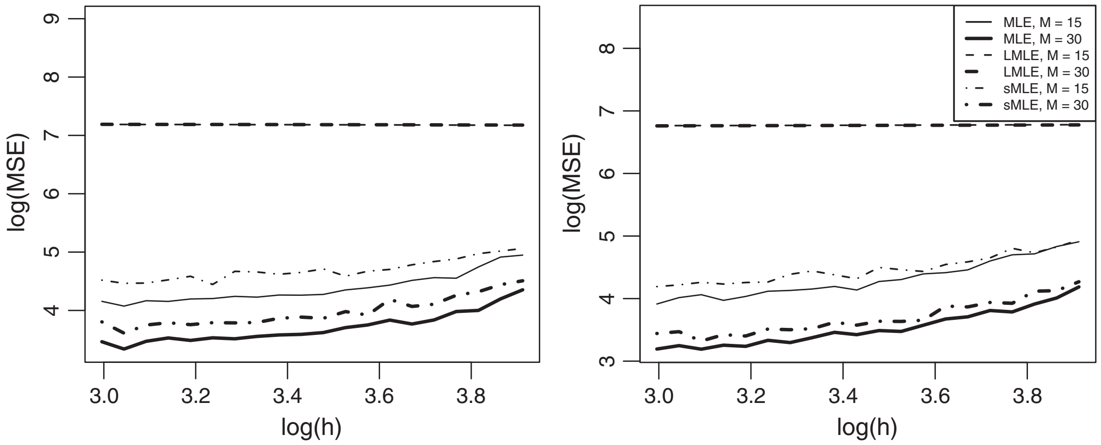

Fig. 2. The results of the MLE and sMLE based procedures for normal data. The x-axis corresponds to the log of the threshold parameter h, and the y-axis the log of the MSE. The left plot shows the results for the intercept a and the plot on the right the standard deviation parameter σ_0. In both plots the solid, long-dashed and short-dashed lines correspond to the MSEs using the MLE, MLE based on the assumption of log-normality and sMLE approaches respectively for a sample size of $M=15$. Similarly the bold lines display the results for a sample size of $M=30$.

The results in the normal setting are shown in Fig. 2. In this figure the x-axis corresponds to the log of the threshold parameter h, and the y-axis the log of the MSE. The plot on the left displays the results for the intercept term a and the plot on the right the standard deviation σ_0. In both plots the solid, long-dashed and short-dashed lines correspond to the $log(\mathrm{MSE})$ at different values of $log(h)$ using the MLE, MLE based on the assumption of log-normality [6] and sMLE approaches respectively for a sample size of $M=15$. Similarly the bold lines display the results for a sample size of $M=30$. Note, the parameters estimated using MLE and MLE based on log-normality were calculated using the R package sand [6]. The sMLE functionality has also been implemented in R. Code for MLE and sMLE as well as the simulation studies is available upon request.

As expected for larger M we observe lower MSEs. The poor performance of the log-normal based MLE is also to be expected. Of particular interest is the comparable performance of the MLE and sMLE procedures.

Fig. 3 shows the results in the log-normal setting, i.e., when $W(0)\sim LN(a,\sigma_0)$, where LN denotes the log-normal distribution. The layout in this figure is the same as in Fig. 2 except that now the solid, long-dashed and short-dashed lines correspond to the MLE based on the log-normal, the sMLE under the assumption of normality and the sMLE under the assumption of log-normality for a sample size of $M=15$. Again the bold lines correspond to a sample size of 30.

The performance of the two approaches is fairly close when log-normality is taken into consideration. Note, we do not show the results for sMLE under the assumption of normality as the log MSEs in this setting are an order of magnitude larger (in the 300s).

In the second study we look to verify the bound H on the LOD. The same population parameters are used in this analysis as in the first study with a few changes/additions; first we let $x\in\{0,1,2,3\}$, next we define the slope $b=100$ and non-blank standard deviation $\sigma=30$, finally we set the significance level to be $\gamma=0.05$. For each h we draw 1000 random samples of size M from $W(x)\sim N(a+bx,\sigma(x))$, where $\sigma(0)=\sigma_0$ and $\sigma(x)=\sigma$ for all $x>0$. The observed data are then sampled from Eq. (1). In order the assess the accuracy of the upper bound we calculate H from each random sampling and count the number of times the LOD$>H$. The results from this study are shown in Fig. 4.

Both plots in Fig. 4 show the density estimate of H based on 1000 simulations of size 15 (solid line) and 30 (dashed line) sampled from Eq. (1). With the exception of a few outlying values of H the

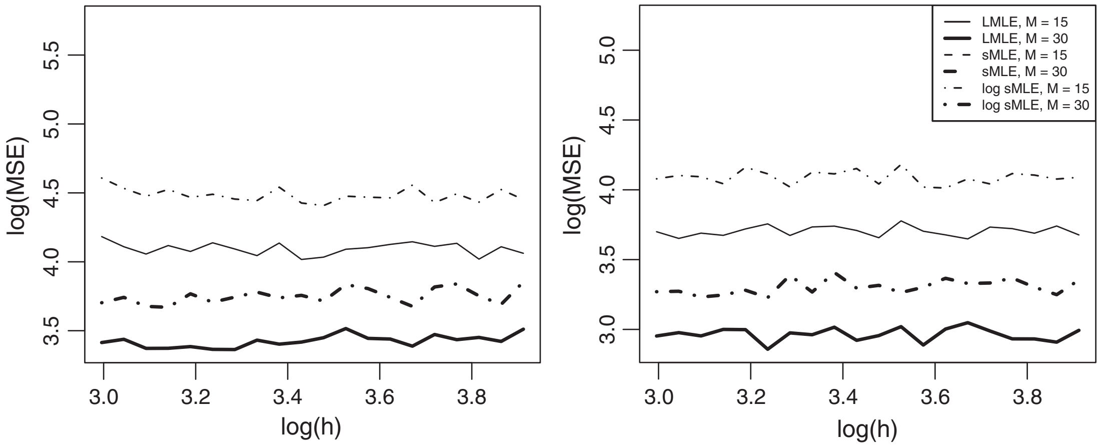

Fig. 3. The results for the MLE and sMLE based procedures for log-normal data. The layout of this figure is the same as in Fig. 2 except that now the solid, long-dashed and short-dashed lines correspond to the MLE based on the log-normal, the sMLE under the assumption of normality and the sMLE under the assumption of log-normality for a sample size of $M=15$. Again the bold lines correspond to a sample size of 30.

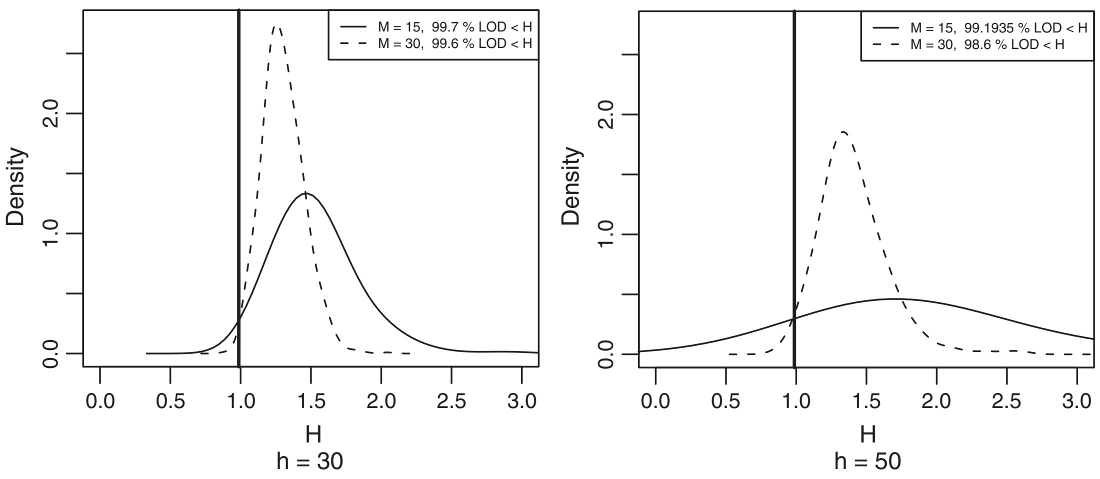

Fig. 4. Results from the simulation study of the upper bound H of the LOD. Both plots show the density estimate of H based simulations of size 15 (solid line) and 30 (dashed line) sampled from Eq. (1). The bold vertical line is the actual LOD. The left and right plots show, respectively the results given a threshold of $h = 30$ and 50. The legend in top right of each plot displays the percent of the time the LOD greater than the estimated H.

densities appear to be fairly Gaussian. Furthermore for larger (smaller) M the estimated densities appear more (less) dispersed. The bold vertical line is the actual LOD. The left and right plots show, respectively the results given a threshold of $h = 30$ and 50. Note that larger values of h also contribute to the dispersion of the estimated density. The legend in top right of each plot shows the percent of the time the LOD was greater than the H. As expected in all cases $Pr\ (LOD \le H) \ge 0.95$.

Acknowledgment

The authors are grateful to M. Verkouteren for invaluable advice and information he provided during this work.

Appendix A

A.1. Notation summary

We provide here the list of main characteristics studied in the main body of this paper.

Table 1
Parameters and their estimators (including upper confidence bounds).

Parameter	Estimator	Definition
a	$\hat{a}, \tilde{a}$	Intercept of response W
b	$\hat{b}, \overline{b}$	Slope of response W
σ_0^2	$\hat{\sigma}_0^2, \tilde{\sigma}_0^2, \overline{\sigma}_0^2$	Error variance, blank sample
σ^2	$\hat{\sigma}^2, \overline{\sigma}^2$	Error variance, non-blank sample
h	$\hat{h}, \tilde{h}$	Truncation parameter
$z_\star = (h-a)/\sigma_0$	$\overline{z}_\star$	Percentile of $Y(0)$ distribution
$\mathrm{LOD} = x_c$ $= (\max(z_{1-\alpha}, z_\star)\sigma_0 + z_{1-\beta}\sigma)/b$	$\widehat{\mathrm{LOD}}, \widetilde{\mathrm{LOD}}$, H, U	Limit of detection
$\mathrm{LC} = \max(a + z_{1-\alpha}\sigma_0, h)$	$\widetilde{\mathrm{LC}}$	Critical level
$y_c = LC + z_{1-\beta}\sigma$		Decision limit

Table 2
Error probabilities.

Error probability	Meaning
α	Defines LC (false positive)
β	Defines LOD (false negative)
γ	Defines confidence limit H for LOD
P	Defines tolerance limit U for LOD

Table 3
Other notation.

Notation	Meaning
$Y(x)$	Observable positive response at mass x
$W(x)$	Unobservable Gaussian response at mass x
N	Total number of samples (including blanks)
n_0	Sample size of blanks
m_0	Number of positive readings in the blank sample
$\nu_0 = n_0 - m_0$	Number of zero readings in the blank sample
$\overline{y}_0$	The sample mean of positive responses in the blank sample
s_0^2	The sample variance of positive responses in the blank sample
$n_i, i \ge 1$	Size of i-th non-blank sample
$m_i, i \ge 1$	Number of positive readings in i-th non-blank sample
$\nu = \sum_{i=1}^{N} m_i - 1$	Degrees of freedom
$d = 1/\sqrt{\sum_i m_i x_i^2}$	Defines confidence limit H for LOD
ϕ	Standard normal density
Φ	Normal cumulative distribution function
z_α	Normal percentile of order $\alpha, \Phi(z_\alpha) = \alpha$

A.2. Estimators of the slope and intercept

By differentiating Eq. (3) in a and σ_0, one obtains simultaneous equations for the maximum likelihood estimators $\hat{a}$ and $\hat{\sigma}_0$,

$$\frac{\nu_0 \varphi\left(\frac{\hat{h}-\hat{a}}{\hat{\sigma}_0}\right)}{\Phi\left(\frac{\hat{h}-\hat{a}}{\hat{\sigma}_0}\right)} = \frac{1}{\hat{\sigma}_0} \sum_{j=1}^{m_0} \left(y_j^{(0)} - \hat{a}\right),$$

$$\frac{\nu_0\left(\hat{h}-\hat{a}\right) \varphi\left(\frac{\hat{h}-\hat{a}}{\hat{\sigma}_0}\right)}{\Phi\left(\frac{\hat{h}-\hat{a}}{\sigma_0}\right)} = \frac{1}{\hat{\sigma}_0} \sum_{j=1}^{m_0} \left(y_j^{(0)} - \hat{a}\right)^2 - m_0 \hat{\sigma}_0.$$

It follows that

$$\left(\hat{h}-\hat{a}\right) \sum_{j=1}^{m_0} \left(y_j^{(0)} - \hat{a}\right) = \sum_{j=1}^{m_0} \left(y_j^{(0)} - \hat{a}\right)^2 - m_0 \hat{\sigma}_0^2,$$

so that

$$\hat{\sigma}_0^2 = \frac{1}{m_0}\sum_{j=1}^{m_0}\left(y_j^{(0)}-\hat{a}\right)^2-\left(\bar{y}^{(0)}-\hat{a}\right)\left(\hat{h}-\hat{a}\right) = s_0^2+\left(\bar{y}^{(0)}-\hat{a}\right)\left(\bar{y}^{(0)}-\hat{h}\right). \tag{18}$$

This formula is used in Section 2 to obtain the standard deviation estimate $\tilde{\sigma}_0$ in Eq. (5).

A.3. Truncation parameter estimators

We present here approximate formulas for the distribution of $\hat{h}$. This estimator admits the representation,

$$\hat{h} = a+\sigma_0 \min_{j:Z_j^{(0)}>(h-a)/\sigma_0} Z_j^{(0)}$$

$$= h+\sigma_0 \min_{j:Z_j^{(0)}>(h-a)/\sigma_0}\left[Z_j^{(0)}-\frac{h-a}{\sigma_0}\right] = h+\sigma_0 V,$$

where for $t=(h-a)/\sigma_0$,

$$V = \min_{j:Z_j^{(0)}>t}\left(Z_j^{(0)}-t\right),$$

and $Z_j^{(0)}, j=1,\ldots,n=n_0$, are independent standard normal random variables. Let $Z_{(1)}^{(0)} \le Z_{(2)}^{(0)} \le \cdots \le Z_{(n)}^{(0)}$ denote the corresponding order statistics. Then for $v \ge 0$,

$$Pr\left(\min_{j:Z_{(j)}^{(0)}>t} Z_{(j)}^{(0)} > t+v\right) = \sum_{j=0}^{n-1} Pr\left(Z_{(j)}^{(0)}<t,\ Z_{(j+1)}^{(0)}>t+v\right)$$

$$= \sum_{j=0}^{n-1}\binom{n}{j}[\Phi(t)]^j[1-\Phi(v+t)]^{n-j} = [1-\Phi(v+t)+\Phi(t)]^n-[\Phi(t)]^n.$$

Thus,

$$Pr(V>v) = [1-\Phi(v+t)+\Phi(t)]^n-[\Phi(t)]^n.$$

According to the asymptotic theory of extreme order statistics [13] if $b_n = \sqrt{2\log n}$, $t \approx 0.5\log(4\pi\log n)/b_n - b_n$, then

$$Pr(b_n V > v) \approx 1-e^{-e^{-v}}, \quad v>0.$$

In other words, for large n, $\sqrt{2\log n}\,V$ is distributed like $\max(0,G)$ with G having a Gumbel distribution.

It follows that

$$E\left(\hat{h}-h\right)^2 \approx \frac{\sigma_0^2}{2\log n} E\max\left(0,G^2\right) = \frac{1.1267\sigma_0^2}{2\log n}.$$

The standard deviation of $\tilde{h}$ can be estimated by the spacing,

$$\min_{j:y_j^{(0)}>\hat{h}} y_j^{(0)}-\hat{h} = \hat{h}-\tilde{h}.$$

As in [11, sec 3], an approximate upper $(1-\delta)$ confidence bound on h is

$$\bar{h} = \left(1-\frac{\delta}{1-\delta}\right)\hat{h}+\frac{\delta}{1-\delta}\tilde{h}.$$

References

[1] A. Hubeaux, G. Vos, Decision and detection limits for linear calibration curves, Anal. Chem. 42 (1970) 849–855.

[2] L.A. Currie, Detection and quantification limits: basic concepts, international harmonization, and outstanding (“low-level”) issues, Appl. Rad. Isot. 61 (2004) 145–149.

[3] A.C. Cohen, Truncated and Censored Samples: Theory and Applications, M. Dekker, New York, 1991.

[4] J.Y. Tsay, I.-W. Chen, H.-W. Maxon, L. Heminger, A statistical method for determining normal ranges from laboratory data including values below the minimum detectable value, Clin. Chem. 25 (1979) 2011–2014.

[5] R.D. Gibbons, D.E. Colman, Statistical Methods for Detection and Quantification of Environmental Contamination, Wiley, New York, 2001.

[6] E.L. Frome, J.P. Watkins, Statistical Analysis of Data with Non-detectable Values, Technical Report ORNL/TM-2004/146, Oak Ridge National laboratory, TN, 2004.

[7] D.R. Helsel, Nondetects and Data Analysis: Statistics for Censored Environmental Data, Wiley, New York, 2005.

[8] E.L. Lehmann, G. Casella, Theory of Point Estimation, second ed, Springer, New York, 1998.

[9] J.F. Lawless, Statistical Models and Methods for Lifetime Data, second ed, Wiley, New York, 2003.

[10] P. Wild, R. Hordan, A. Leplay, R. Vincent, Confidence intervals from probabilities of exceeding threshold limits with censored log-normal data, Environmetrics 7 (1995) 247–259.

[11] D.S. Robson, J.H. Whitlock, Estimation of truncation point, Biometrika 51 (1964) 33–39.

[12] R. Miller, Simultaneous Statistical Inference, second ed, Springer, New York, 1981.

[13] H. David, H.N. Nagaraja, Order Statistics, third ed, Wiley, New York, 2003.