International Journal of Food Microbiology 155 (2012) 146–152

Contents lists available at SciVerse ScienceDirect

International Journal of Food Microbiology

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

Fitting a lognormal distribution to enumeration and absence/presence data

Natalie Commeau [a,b,c,*], Eric Parent [a], Marie-Laure Delignette-Muller [d,e], Marie Cornu [b,1]

[a] UMR 518 AgroParisTech-INRA MIA, 16 rue Claude Bernard 75005 Paris, France
[b] Laboratoire de sécurité des aliments, ANSES, 23 av du Général de Gaulle, 94706 Maisons-Alfort, France
[c] AgroParisTech ENGREF, 19 avenue du Maine, F 75732 Paris, France
[d] Université Lyon 1, CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, France
[e] VetAgro Sup Campus Vétérinaire de Lyon, F-69280 Marcy l'Etoile, France

ARTICLE INFO

Article history:
Received 27 May 2011
Received in revised form 12 December 2011
Accepted 29 January 2012
Available online 3 February 2012

Keywords:
Microbial contamination assessment
Limit of detection
Limit of quantification
Maximum likelihood estimation
EM algorithm

ABSTRACT

To fit a lognormal distribution to a complex set of microbial data, including detection data (e.g. presence or absence in 25 g) and enumeration data (e.g. 30 cfu/g), we compared two models: a model called $\mathcal{M}_{CLD}$ based on data expressed as concentrations (in cfu/g) or censored concentrations (e.g. <10 cfu/g, or >1 cfu/25 g) versus a model called $\mathcal{M}_{RD}$ that directly uses raw data (presence/absence in test portions, and plate colony counts). We used these two models to simulated data sets, under standard conditions (limit of detection (LOD) = 1 cfu/25 g; limit of quantification (LOQ) = 10 cfu/g) and used a maximum likelihood estimation method (directly for the model $\mathcal{M}_{CLD}$ and via the Expectation–Maximisation (EM) algorithm for the model $\mathcal{M}_{RD}$. The comparison suggests that in most cases estimates provided by the proposed model $\mathcal{M}_{RD}$ are similar to those obtained by model $\mathcal{M}_{CLD}$ accounting for censorship. Nevertheless, in some cases, the proposed model $\mathcal{M}_{RD}$ leads to less biased and more precise estimates than model $\mathcal{M}_{CLD}$.

1. Introduction

Detection and enumeration of contaminating microorganisms are regularly collected nationally and internationally for quantitative risk assessments of food products (Beaufort et al., 2007; Uyttendaele et al., 2009; Habib et al., 2008; Burfoot et al., 2010). However, in many cases, particularly when the target microorganism of interest is a pathogenic species, available data are limited because analyses are labour-intensive, and they are poorly informative due to a high proportion of non-detects (i.e. when a detection has been performed and the result is absence) and zero counts (i.e. an enumeration was performed and zero colony was enumerated).

How can this type of complex data set be analysed to properly assess microbial contamination? Assuming that concentrations (e.g. in cfu/g) in the contaminated products follow a lognormal distribution, it was frequent in the 1990s and early 2000s to calculate a log mean concentration based only on enumeration results (i.e. ignoring non-detects; "log" refers to the logarithm to base 10), see Abadias et al. (2006). Zero counts were either ignored, substituted with the limit of quantification (LOQ) or a fraction of the LOQ. Since then, this practice has been criticised for the bias it introduces (Helsel, 2006; Lorimer and Kiermeier, 2007) and the maximum likelihood estimation (MLE) approach combined with statistical censoring has been introduced to fit a single lognormal distribution to the full data set which can then be applied to all food products from which the samples have been taken (Lorimer and Kiermeier, 2007; Busschaert et al., 2010; Pouillot and Delignette-Muller, 2010). In statistics, the term "censoring" is used when an observation has not been quantified, but is known (or presumed) to exceed or to be less than a threshold value. In left-censoring, a data point is presumed to be below the threshold value, whereas in right-censoring the data point is above the threshold value. Thus, for MLE fits as proposed by Lorimer and Kiermeier (2007); Busschaert et al. (2010); Pouillot and Delignette-Muller (2010), positive enumeration data are expressed as concentrations (in cfu/g), and all other data are expressed as censored concentrations: a zero count is considered as below the LOQ (left censoring), presence is considered as above the limit of detection (LOD) (right censoring) and absence is considered as a concentration below the LOD (left censoring). The LOD is the minimal theoretical concentration required for a test portion so that the result of detection is "presence".

These concentration-like data, i.e. the results expressed as concentrations (in cfu/g) for positive enumeration results, and as censored concentrations in all other cases, may actually convey less information in statistical analysis than directly using raw data generated in the laboratory, i.e. either colony counts per plate or binary presence/absence data. A few methods are currently available for assessing the

* Corresponding author at: UMR 518 INRA-MIA, 16 rue Claude Bernard 75005 Paris, France.
E-mail address: natalie.commeau@agroparistech.fr (N. Commeau).
[1] Present address: Institut de Radioprotection et de Sûreté Nucléaire (IRSN), PPRENV, SERIS, LM2E, Cadarache, France.

0168-1605/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijfoodmicro.2012.01.023

distribution of bacterial concentrations from raw data Commeau et al. (in press), but they are rarely used in practice.

The objective of this work was to determine in which conditions the use of concentration-like data leads to less accurate estimation of the parameters of the lognormal distribution than the use of raw data. Models and estimation methods are illustrated in this paper with simulated data sets to study the statistical properties of the estimates. Realistic parameters have been chosen with regard to microbiological contamination in the food industry. After presenting the models and estimation methods, we suggest some criteria for comparing the methods' performances.

2. Materials and methods

2.1. How should raw data be converted into concentration-like data?

This section describes the notations used below to denote the current ways in which microbiology laboratories convert raw data into concentration-like data. To illustrate our notations, numerical values have been taken from the methods for *Listeria monocytogenes* (ISO11290-1, 1996; ISO11290-2, 1998), but can be applied to other target microorganisms. For detection analysis, if result x (coded as 0 = absence and 1 = presence) is presence, it means that the target has grown during the enrichment step then it is likely that, before the enrichment step, there was at least one cell in the test portion of mass $M = 25$g, and so the concentration is presumed to be greater than the LOD $= 1/M = 0.04$ cfu/g. If the result is absence, the concentration is presumed to be below the LOD.

Regarding enumeration, (ISO11290-2, 1998), once a test portion of $M' = 10$ g has been taken, it is diluted in a solution so that the final volume is $V_d = 100$ ml from which a volume $V_r = 1$ ml is removed and pour-plated onto a Petri dish containing the appropriate medium. Finally, after 24 h or 48 h of incubation, y colonies of the target [microorganism] are counted. If $y \neq 0$, then these colonies are assumed to come from the multiplication of y cells initially present in V_r. The number of cells in the volume V_d is considered to be proportional to y and is equal to $y/\frac{V_r}{V_d} = y/0.01$ cfu. Because all these cells come from the test portion of mass M' g, the concentration of the target in the test portion is $y/\left(M'\frac{V_r}{V_d}\right) = y/0.1$ cfu/g. If $y = 0$, then it is considered that the number of cells in the test portion is less than 1 colony in a volume V_r, thus the concentration is below the LOQ$= 1/\left(M'\frac{V_r}{V_d}\right) = 10$ cfu/g. Note that in this case study, colonies are not confirmed.

For enumeration, if n dilutions are made, the conversion formula becomes $y/\left(M'\prod_{i=1}^{n}\frac{V_{ri}}{V_{i}d}\right)$, where V_{di} is the volume of the ith dilution V_{ri} is the volume removed at the ith dilution. This case is not considered in the article to simplify notations.

2.2. Models to estimate the contamination distribution

Fitting a lognormal distribution to a data set, obtained as described in the previous section, relies on assumptions, which can be conceptualized under a model.

The most simplistic models, ignoring non detects and zero counts, or substituting them with an arbitrary value, are not considered in this study because they are not appropriate models to estimate the parameters of contamination properly (Helsel (2006); Shorten et al. (2008); Busschaert et al. (2010)).

2.2.1. The recommended current practice

It is possible to take all the available data into account, censored and non-censored. The approach is described in Lorimer and Kiermeier (2007); Busschaert et al. (2010); Pouillot and Delignette-Muller (2010)). A concentration datum c follows a lognormal distribution:

$$c \sim \mathcal{LN}(\mu, \sigma).$$

In this case, a detection datum can be expressed as a censored concentration (or concentration-like datum): less than the LOD (absence) or greater than the LOD (presence). An enumeration concentration-like value is either the value of c, with $c \geq LOQ$ or the information $c < LOQ$ when censored. This model is denoted $\mathcal{M}_{CLD}$, as data in this model are concentration-like ones.

2.2.2. A physically based approach

As mentioned above, data expressed as concentrations, concentration-like data, are obtained from raw data, i.e. colony counts for enumeration (y) or presence/absence for detection (x). We therefore designed a hierarchical model based on raw data and that mimics the "real" assay as it is actually carried out in laboratories. Enumeration data follow a Poisson distribution (P), detection data follow a Bernoulli distribution (Ber) and the unknown concentration data λ follow a lognormal distribution (LN):

$$\begin{aligned} y &\sim \mathcal{P}\left(\lambda M' \frac{V_r}{V_d}\right) \\ x &\sim \mathcal{B}er(1 - exp(-\lambda M)) \\ \lambda &\sim \mathcal{LN}(\mu, \sigma). \end{aligned}$$

The term $1 - exp(-\lambda M)$ is the probability of obtaining at least one colony when cells follow a Poisson distribution with parameter λM. Conditionally to the concentration λ, detection result x and enumeration result y are independent. We call this hierarchical model $\mathcal{M}_{RD}$, as it links raw data to the parameters μ and σ. It corresponds to the model A-test portion in Commeau et al. (in press). If ignoring detection results, this model is known as the Poisson LogNormal model (ILSI, 2010; Reinders et al., 2003).

To use this model, it is necessary to have the protocol, or at least the LOD and the LOQ. If only the concentration-like data are given, it is not always possible to get the LOQ so there can be a loss of information when converting raw data into concentration-like data.

2.3. Parameters estimation

For the most simplistic models as well as for model $\mathcal{M}_{CLD}$, maximum likelihood can be used analytically to estimate parameters μ and σ, for example with the package fitdistrplus described in Pouillot and Delignette-Muller (2010). The likelihood of this model is equal to:

$$\begin{aligned} L(C|\mu,\sigma) &= \prod_{i=1}^{i=N_{abs}} F(C_i \leq LOD) \prod_{j=1}^{j=N_{pres}} F\left(C_j > LOD\right) \prod_{k=1}^{k=N_0} F(C_k < LOQ) \prod_{l=1}^{l=N_{\neq 0}} F(C_l = c_l) \\ &= \prod_{i=1}^{i=N_{abs}} \Phi\left(\frac{logLOD-\mu}{\sigma}\right) \prod_{j=1}^{j=N_{pres}} \left(1-\Phi\left(\frac{logLOD-\mu}{\sigma}\right)\right) \\ &\times \prod_{k=1}^{k=N_0} \Phi\left(\frac{logLOQ-\mu}{\sigma}\right) \prod_{l=1}^{l=N_{\neq 0}} \frac{1}{c_l\sqrt{2\pi\sigma^2}} exp\left(\frac{-(logc_l-\mu)^2}{2\sigma^2}\right), \end{aligned}$$

where N_{abs} is the number of absence results, N_{pres} the number of presence results, N_0 the number of enumeration results with no colony and $N_{\neq 0}$ the number of enumeration result with at least one colony present in a data set. The cumulative distribution function (cdf) of the log-normal distribution is denoted F and the standard normal

cdf is denoted Φ. The random variables are denoted C_i, C_j, C_k and C_l and the values taken by C_l are noted c_l.

The likelihood cannot be analytically calculated for model $\mathcal{M}_{CLD}$. The likelihood of stochastic variables X and Y conditionally to λ is equal to:

$$L(X,Y|\lambda) = \prod_{i=1}^{i=N_{abs}} \mathbb{P}(X_i=0) \prod_{j=1}^{j=N_{pres}} \mathbb{P}\left(X_j=1\right) \prod_{k=1}^{k=N_0} \mathbb{P}(Y_k=0) \prod_{l=1}^{l=N_{\neq 0}} \mathbb{P}(Y_l=y_l)$$

$$= exp(-N_{abs}\lambda M)(1-exp(\lambda M))^{N_{pres}} exp\left(-N_0\lambda M' \frac{V_r}{V_d}\right)$$

$$\times \prod_{l=1}^{l=N_{\neq 0}} exp\left(-\lambda M' \frac{V_r}{V_d}\right)\left(\lambda M' \frac{V_r}{V_d}\right)_l^y \frac{1}{y_l!},$$

where X_i, X_j, Y_k and Y_l are stochastic variables and y_l are the values taken by Y_l. The likelihood of X and Y conditionally to μ and σ is equal to $L(X, Y|\lambda)$ integrated on λ:

$$L(X,Y|\mu,\sigma) = \int L(X,Y|\lambda)[\lambda]d\lambda$$

$$= \int L(X,Y|\lambda) \frac{1}{\lambda\sqrt{2\pi\sigma^2}} exp\left(\frac{-(log\lambda-\mu)^2}{2\sigma^2}\right)d\lambda.$$

Bayesian inference has been proposed by Commeau et al. (in press). When writing the likelihood, the latent variable λ appears. To find the maximum of the likelihood, the Expectation Maximisation (EM) algorithm is used, which is an iterative method for finding maximum likelihood estimates of parameters, where the model depends on unobserved latent variables. The EM iteration consists of two processes, the E-step, which computes the expectation of the log-likelihood evaluated using the current estimate for the parameters, and an M-step, which consists in finding the values of parameters maximizing the expected log-likelihood found on the E-step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E-step. The EM algorithm was first described in Dempster et al. (1977). In this work, the E-step was performed using importance sampling (Robert and Casella, 2004) and the M-step was performed using a simple gradient search in a bidimensional space. This algorithm is written in R (R Development Core Team, 2011) and is available upon request from the first author. The algorithm is easy to code because the derivative of $Q(\theta, \theta') = \int ln[y, \lambda|\theta][\lambda|\theta', y]d\lambda$ is easy to calculate.

2.4. Simulating detection and enumeration protocols

When carrying out detection and enumeration protocols, several measurement errors can occur: on mass M and volumes V_d and V_r in the enumeration protocol and in the detection protocol for mass M'. All these errors were described as follows, assuming plausible coefficients of variation (0.1% for test portion mass and 2% for dilution volume and masses) commonly found in standard laboratory working conditions.

- Enumeration errors
  - Test portion mass : $m' \sim \mathcal{N}(M', 0.001M')$
  - Dilution volume: $v_d \sim \mathcal{N}(V_d, 0.02V_d)$
  - Removed volume: $v_r \sim \mathcal{N}(V_r, 0.02V_r)$
- Detection errors
  - Test portion mass : $m \sim \mathcal{N}(M, 0.02M)$

Here, upper-case letters for volume and mass refer to the intended masses and volumes (as written in the protocol) and the lower-case letters refer to the mass and the volume actually simulated to model what is obtained in a laboratory. The standard deviations of the normal distributions are expressed in terms of the relative accuracy of the device used. We considered that there is no error in the detection and enumeration results. For example, if six colonies are counted on a Petri dish that means that there were six cells in the pour-plated volume V_r.

To model enumeration result y, a test portion mass m', a dilution volume v_d and a removed volume v_r were simulated from their distributions and then an enumeration result y was simulated from the Poisson distribution with parameter $\lambda m' \frac{V_r}{V_d}$, where λ is the concentration of the test portion and $\lambda \sim \mathcal{N}(\mu, \sigma)$. Finally, the raw datum y was converted into a concentration c_y: $c_y = \frac{y}{M' V_r/V_d} = y \times LOQ$, if $y \neq 0$ and $c_y < \frac{1}{M' V_r/V_d} = LOQ$, if $y = 0$.

Detection result x was simulated from a Bernoulli distribution with parameter $1 - exp(-\lambda m)$, with $\lambda \sim \mathcal{N}(\mu, \sigma)$ and m simulated from its distribution as described above. The result was converted into concentration using the following formula: if $x = 1$, then $c_x \geq \frac{1}{M} = LOD$, else, $c_x < LOD$.

The intended values of masses and volumes were used in these conversion formulae for both enumeration and detection to mimic real laboratory [assay] conditions. For simulation purposes, values of the mean log concentration μ range from -3 to 4 log cfu/g with an increment of 0.5 log cfu/g between two values. For the standard deviation of the log concentration σ, values 0.7, 1 and 1.5 log cfu/g were chosen.

2.5. Simulating a sampling plan

For each parameter value, 1000 data sets were simulated. In the simulations, we mimicked the strategy applied in the experimental plan of Beaufort et al. (2007) or Uyttendaele et al. (2009). In these articles, detection analyses were first performed. Each time the result for one test portion was presence, another test portion (somehow linked to the positive test portion) was enumerated. In these simulations, first detection data were simulated and then as many enumeration data as positive detection results were simulated. Therefore, each data set had 100 detection results and from 0 to 100 enumeration data. The putative link between the positive test portion and the consecutively enumerated test portion is not explicitly taken into account, neither in the simulations, nor in the estimations. This link is not considered because the food is solid and microorganisms do not move inside it. It is therefore possible that a test portion taken from a food sample has a colony on it (the detection result is then 'presence', so the concentration-like datum is right-censored and above the LOD) and that another test portion removed from the same sample has no colony at all so the enumeration result is 'no colony' and is left-censored (below the LOQ). As a consequence, we do not consider that the concentration is between the LOD and the LOQ. We consider instead that we have two different data, one right-censored with a concentration above the LOD while the second is left-censored with a concentration below the LOQ. This approach is different from the one chosen in Busschaert et al. (2010) in which a presence detection result and a 0-colony enumeration result on two different test portions from the same sample are considered as one interval-censored datum (between the LOQ and the LOD).

The intended values of volumes and masses were: $V_d = 100$ ml, $V_r = 1$ ml, $M = 25$ g and $M' = 10$ g. According to these numerical values, the LOD was equal to 0.04 cfu/g and the LOQ, 10 cfu/g. These values are commonly used, e.g. for *L. monocytogenes* (ISO11290-1, 1996; ISO11290-2, 1998).

2.6. Bias and variance in the estimates of mean and variance

To assess the quality of each estimate, its bias and variance were calculated. Bias is the difference between the mean of the estimate and the true value of the parameter. For instance, for μ, the formula is $E(\hat{\mu} - \mu)$, where $\hat{\mu}$ is the result of an estimation of μ for a given data set. Biases for μ and for σ were calculated for each data set and for each model.

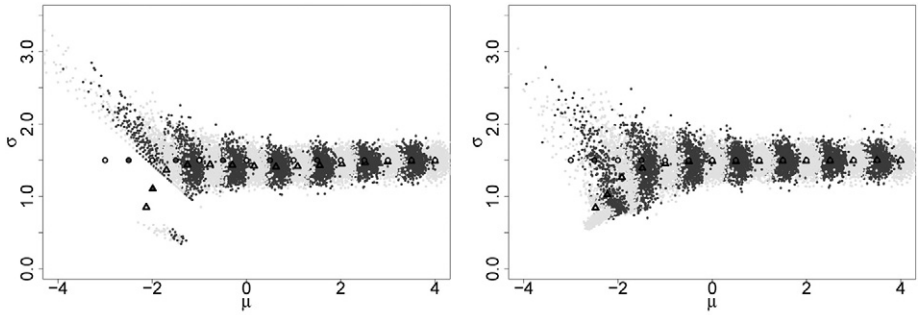

Fig. 1. Joint distribution of inference on μ (the mean of the log concentration in log cfu/g) and σ (the standard deviation of the log concentration in log cfu/g). Each shade of grey (light or dark grey) corresponds to a value of μ, from −3 to 4 with an increment step of 0.5. Each point represents an inference of μ and σ. Filled circles are the true values of μ and σ. Filled triangles are the mean of the estimates for each value of μ. For each value of the parameter pair (μ, σ), 100 data sets were simulated. Each data set contains 100 detection results and as many enumeration results as presence detection data. Information on estimate bias can be seen as the distance between the squares and the triangles of the same color.

The sum of the bias and the variance is equal to the mean square error (MSE). For instance, for μ, it is equal to $MSE(\hat{\mu}) = (E(\hat{\mu})-\mu))^2 + V(\hat{\mu}) = E\left((\hat{\mu}-\mu)^2\right)$. The MSE is also equal to the expected value of the squared error loss. The smaller the MSE are, the better the inferences are.

For the second parameter, the MSE was used on $ln(\hat{\sigma})$: $MSE(ln(\hat{\sigma})) = E\left((ln(\hat{\sigma})-ln\sigma)^2\right) = (E(ln(\hat{\sigma})-ln\sigma))^2 + V(ln(\hat{\sigma}))$, where "ln" is the natural logarithm. The logarithm was taken because σ is always positive. Moreover, it is the order of magnitude that is important. Therefore, taking the logarithm of σ instead of σ is useful for penalising estimations of σ that are too low.

2.7. Estimating the 97.5th percentile

In addition to assessing μ and σ, the 97.5th percentile on the total population was studied for the two models.

3. Results

3.1. Results of the two inference methods

Fig. 1 shows the estimates of μ and of σ when $\sigma = 1.5$ log cfu/g. The simulated data set inferred a higher concentration and was less dispersed than the real data set. The pattern was quite similar for $\sigma = 0.7$ and $\sigma = 1$ (results not shown), although $\hat{\mu}$ and $\hat{\sigma}$ were less dispersed. For low values of μ, there was a correlation between $\hat{\mu}$ and $\hat{\sigma}$. Information on estimate bias can be seen in Fig. 1. This bias decreased when μ increased and was virtually nil when μ was sufficiently high (the exact value depended on the model and on σ; for instance, for $\sigma = 1.5$ log cfu/g and for model $\mathcal{M}_{CLD}$, the bias was nil for $\mu > 2$ log cfu/g). The higher the value of σ was, the higher the bias for the same value of μ was (results not shown). The model with the smallest bias was model $\mathcal{M}_{RD}$. Overall, models $\mathcal{M}_{CLD}$ and $\mathcal{M}_{RD}$ led to overestimation of low μ values. Overestimation was worse for model $\mathcal{M}_{CLD}$ than for model $\mathcal{M}_{RD}$. Standard deviation σ was underestimated with both models and on average, the estimates of σ with model $\mathcal{M}_{CLD}$ were higher than the estimate using model $\mathcal{M}_{RD}$, except for $\sigma = 0.7$ log cfu/g.

For a given value of σ, prevalence in detection and percentage of non-zero counts in enumeration results decreased when μ decreased. This is expected since, when μ decreases, the concentration decreases as well; therefore, there are more and more zero-count enumeration results and absence-of-detection results.

Sometimes, for $\mu = -3$ log cfu/g, all 100 simulated detection data points were "absence". Under the assumed sampling plan, this means that no enumeration was simulated. In that case, estimations were not made for either of the models.

3.2. Bias and variance of the estimates μ and σ

Fig. 2 shows the results of MSE for μ. The MSE($\hat{\mu}$) for $\sigma = 1.5$ log cfu/g are not shown because they are very close to the MSE($\hat{\mu}$) when $\sigma = 1$ log cfu/g. When $\mu \geq 1$ log cfu/g, models $\mathcal{M}_{CLD}$ and $\mathcal{M}_{RD}$ gave MSE($\hat{\mu}$) almost equal to 0. When μ was greater than 0.5 log cfu/g, the MSE for $ln\hat{\sigma}$, whatever the value of σ, were almost equal in both models and very close to zero (not shown).

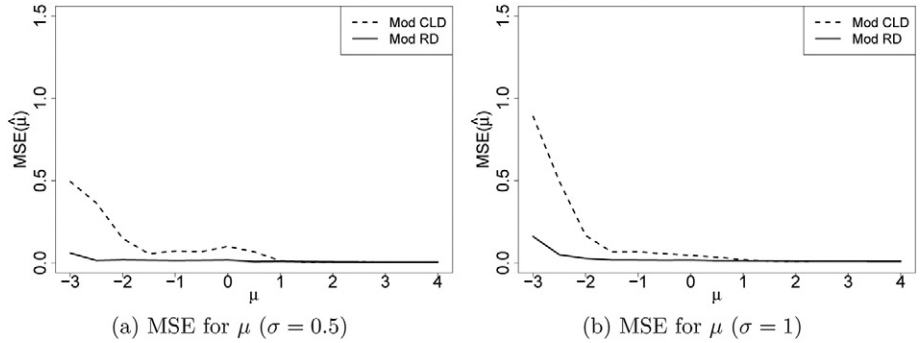

Fig. 2. MSE($\hat{\mu}$) (in (log cfu/g)^2) for the two models and for $\sigma = 0.5$ log cfu/g (a) and $\sigma = 1$ log cfu/g (b). Formula is as follows: $MSE(\hat{\mu}) = E\left((\hat{\mu}-\mu)^2\right)$, where $\hat{\mu}$ is the estimate of μ. The parameter μ ranges from −3 to 4 log cfu/g with an increment of 0.5 log cfu/g between two values.

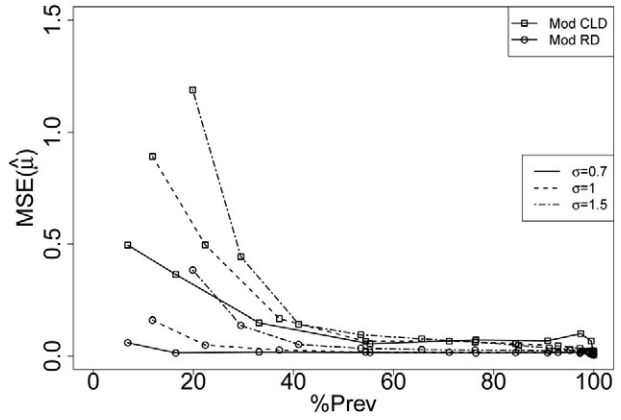

Fig. 3. MSEs for the four models and for μ with respect to prevalence. Each point couple (line, point) represents a model. Data sets were simulated with a log mean μ which varies from −3 to 4 log cfu/g and a log standard deviation σ which is equal to 0.7 (plain line), 1 (—) and 1.5 (-.-.-) log cfu/g.

When using real data sets, the true values of μ and σ are of course unknown. Available information is the percentage of presence in detection results (i.e. observed prevalence). For this reason, we also represented the MSE as a function of this percentage (Fig. 3). When the prevalence was equal to 100%, models $\mathcal{M}_{CLD}$ and $\mathcal{M}_{RD}$ were similar (see Appendix A). Note that in such a case the simplistic models (ignoring non detects and zero counts, or replacing them with arbitrary values) also provide similar estimates. When prevalence decreased from 100% to zero, the MSE of μ for models $\mathcal{M}_{CLD}$ and $\mathcal{M}_{RD}$ remained low, whatever the value of σ. Nevertheless, the MSE for these two models increased when prevalence was below 40%, especially for model $\mathcal{M}_{CLD}$. The MSE of μ for model $\mathcal{M}_{CLD}$ reached 0.25 $(\log \text{cfu/g})^2$ only when the prevalence was near 40%. The same level was attained for a prevalence of 25% with model $\mathcal{M}_{RD}$. When there was more than 85% of absence, the MSE for these two models was higher than 0.4 $(\log \text{cfu/g})^2$, at least for $\sigma = 1.5$ log cfu/g.

3.3. Estimating the 97.5th percentile of the total distribution

When studying microbial risk, it is interesting to examine the high percentiles of pathogen concentration, especially if data are collected at the moment of consumption or at the end of shelf life. In fact, high concentrations are the ones that have the greatest impact on the exposure (Pouillot et al., 2009). For each couple (μ, σ), 1000 values of the 97.5th percentile were estimated. Fig. 4 represents the 950th value of the 97.5th percentile estimated with both models. This value is higher for model $\mathcal{M}_{CLD}$ than for model $\mathcal{M}_{RD}$. The real 97.5th percentile was always lower than the ones estimated with both models (the estimation is fail-safe). The estimation using model $\mathcal{M}_{RD}$ was closer to the real value of the percentile as shown in Fig. 4.

4. Discussion

Under the assumption that "true" contamination follows a lognormal distribution, pathogen contamination can be properly assessed from a combination of enumeration and detection results, using methods accounting for zero counts and non-detects. These methods include the model $\mathcal{M}_{CLD}$ with an MLE approach that considers censored data and model $\mathcal{M}_{RD}$ that applies an EM algorithm to raw data. The main objective of this article was to demonstrate the added value of accounting for raw data, i.e. using $\mathcal{M}_{RD}$, instead of $\mathcal{M}_{CLD}$, which is based on concentration-like data.

As illustrated by each figure, model $\mathcal{M}_{RD}$ provides slightly more precise and accurate estimations of parameters μ and σ of the lognormal contamination (and, consequently, of a high percentile) than those of model $\mathcal{M}_{CLD}$. The main advantage of the model using the MLE technique directly is that it is very easy to use: for example, using R software with the fitdistrplus package. For model $\mathcal{M}_{RD}$, to our knowledge, no package exists in R to easily apply the EM algorithm specially devoted to this model structure. A function was written in R for this study and is available upon request from the first author. Regarding computational burden, the MLE algorithm runs three times faster than the EM algorithm in model $\mathcal{M}_{RD}$.

In addition to the likelihood-based frequentist methods discussed in this study, a Markov–Chain Monte Carlo (MCMC) algorithm can also be applied to the so-called models $\mathcal{M}_{CLD}$ and $\mathcal{M}_{RD}$. This type of Bayesian analysis has recently been published (Busschaert et al., 2011; Gonzales-Barron and Butler, 2011b; Commeau et al., in press). Codes are very easy to write for software such as OpenBugs (Thomas et al., 2006), but this programme takes longer to run than the EM algorithm. These MCMC analyses were not included in the present comparison, due to this computational burden. Nevertheless, they are promising for regression analyses, i.e. to consider additional information, e.g. between- and within-batch variability (Gonzales-Barron and Butler, 2011a; Commeau et al., in press), the company from which food samples are taken, or the food preparation method (Busschaert et al., 2011). Finally, in Bayesian analysis, additional information can be obtained through expertise and used as priors.

We excluded simplistic models from our comparison. Ignoring censored values or replacing enumeration censored data by the LOD is a source of bias as previously stated (Helsel, 2006; Lorimer and Kiermeier, 2007; Shorten et al., 2008) for both μ and σ estimators: the estimated mean is greater than the real mean, whereas the estimated standard deviation is less than the true one. This result is obvious given that the fitting procedure only applies to a subpopulation of the results, the "positive" subpopulation.

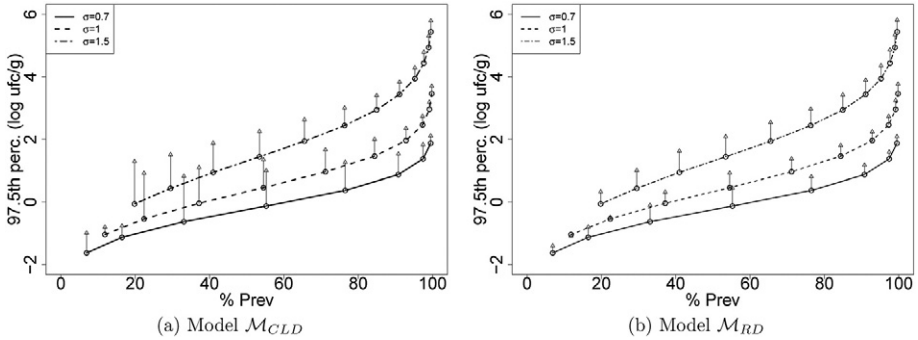

Fig. 4. 97.5th percentile for the total population (in log cfu/g), depending on the percentage of absence results. Open circles are the real values of the percentiles and triangles are the 950th value of the assessed percentiles for each simulated data set. Vertical lines connect the real percentile to its corresponding mean estimated percentile. Data sets were simulated with a log mean μ which varies from −3 to 4 log cfu/g and a log standard deviation σ which is equal to 0.7 (plain line), 1 (—) and 1.5 (-.-.-) log cfu/g.

In some cases, the data sets were so uninformative that even the "best" method could not extract any information. In the sampling strategy simulated here, enumeration was performed only as many times as there were presence data. When observed prevalence is low, this leads to few enumeration results, most (or all) of them being zero counts. For some simulated data sets of this study, there were no presence results at all and all estimates were then either impossible or aberrant. This is an extreme case but can occur in real situations, for instance for samples naturally low in contamination, when the LOD is high. For instance, Cornu et al. (2005) show that the 99.9th confidence interval for prevalence (percentage of presence in 25 g) of *L. monocytogenes* in French pork rillettes at the end of production is between 0.10% and 0.50%. In this case, presence is rare and positive enumeration results even rarer. Turning to more informative protocols (by lowering both LOQ and LOD so as to get less zero count or absence results, see for example Gnanou Besse et al. (2004)) would be a promising way to estimate the concentration parameters better, but this would increase the cost of microbiological analyses. To assess the distribution of a microbial population in a food product at a given level of accuracy, it may be necessary to analyse samples with low LOD and LOQ. A cost–benefit analysis should be performed to determine whether it is worth lowering the LOD and LOQ. Data obtained with various protocols (and therefore various LOD and/or LOQ values) can be combined in complex data sets. Methods based on models $\mathcal{M}_{CLD}$ and $\mathcal{M}_{RD}$ can be easily used to analyse a complex data set with various LODs and LOQs, as illustrated by Busschaert et al. (2010) for model $\mathcal{M}_{CLD}$.

Assessing accurately and precisely a statistical distribution characterising microbial counts is essential in numerous applications of food microbiology, e.g. for risk assessments, for characterising contamination of the hazard at the initial step of a modelled pathway in an exposure assessment, or in risk management, for use in application risk-based metrics, such as Performance Objectives (PO) and Food Safety Objectives (FSO). Nevertheless, this estimation is blurred by diverse sources of uncertainty, resulting from measurement errors in the raw data, possible loss of information from raw data to the data set used in the fitting, sampling uncertainty and the intrinsic properties of the fitting method. Given that routine data sets in food microbiology are usually obtained from low numbers of samples and analytical methods have poor precision, targeting perfectly accurate and precise estimations seems unattainable. For this study, realistic targets were chosen, relative to the current practise used routinely in laboratories. Thus, it was considered here that the bias in the estimation of a 97.5th percentile should be less than twice the between-laboratory standard deviation for the enumeration of *L. monocytogenes*, as assessed by Augustin and Carlier (2006) at 0.28 log cfu/g. Then, in practice, the 97.5th percentile is considered in this study to be "correct" as long as the difference between the real value of the 97.5th percentile and 95% of the estimates of the 97.5th percentile is below 0.5 log cfu/g.

According to this decision rule applied to the estimation of the 97.5th percentile, and taking all other results and considerations discussed above into account, we make the following recommendations to fit a lognormal distribution properly to a data set obtained under the assumptions of the simulations:

1. If prevalence is greater than 85% (i.e. less than 15% of absence), all estimates of the 97.5th percentile are "correct". MSE values for the estimation of the log mean using $\mathcal{M}_{CLD}$ and $\mathcal{M}_{RD}$ are very low. model $\mathcal{M}_{CLD}$ However, the model $\mathcal{M}_{CLD}$ is preferred because it is easy to run;
2. If prevalence is between 25% and 85%, MSE values on the estimation of the log mean are greater for model $\mathcal{M}_{CLD}$ than for model $\mathcal{M}_{RD}$, and estimates of the 97.5th percentile are closer to the real value for the model $\mathcal{M}_{CLD}$. Our recommendation is to use only model $\mathcal{M}_{RD}$.
3. Below prevalence of 25%, models $\mathcal{M}_{CLD}$ and $\mathcal{M}_{RD}$ both give poor results, with high MSE values on the estimation of the log mean, and many "incorrect" estimates of the 97.5th percentile. This is easily explained by the poor informativeness of the sample, under the conditions tested here (data set of 100 detection, then less than 25 "presence", and less than 25 enumerations performed, most or all of them being zero counts). Our best recommendation would be to analyse additional samples, if possible with lower LOQ and LOD. Nevertheless, in this situation, if a statistical analysis has to be done, model $\mathcal{M}_{RD}$ gives slightly better estimates than the other model.

All conclusions developed here are contingent on the way data sets were simulated. For instance, errors in the test portion mass, in the first suspension volume and the volume pour-plated on Petri dishes were considered. However, the model used to simulate data does not consider any errors in colony counting. In reality, this type of error can occur, especially when there are a high number of colonies on plates. If this type of error affects only high numbers of colonies, the inference on μ and σ is practically unchanged. If there are errors in counts of small numbers of colonies, the parameter estimates are more biased (result not shown).

Another key feature is the simulated sampling procedure. In this article, it mimics a practise described in Beaufort et al. (2007): first, detection is performed on a test portion; then, if the result is positive, another test portion is taken for enumeration analysis. As a consequence, enumeration is carried out as many times as there are positive detection results. This does not provide much information when μ is low as there is only a small number of enumeration results. If there were as many enumeration analyses as detection analyses, the inference would be slightly better for low values of μ. For instance, if each data set had 100 detection data points and 100 enumeration data points, then for $\sigma = 1.5$ log cfu/g, both the bias and variance of the estimates are lower when $\mu \leq -1.5$ log cfu/g for models $\mathcal{M}_{CLD}$ and $\mathcal{M}_{RD}$ (results not shown).

Here it was assumed that the second test portion (for enumeration) is independent of the first one (for detection), even if enumeration is performed only when detection is positive. Another assumption was made by Busschaert et al. (2010), who assumed that the results are completely correlated: if an enumeration result leads to 0 colony, detection and enumeration results are considered to be one interval-censored concentration datum between the LOD and the LOQ. In this paper, it is considered that there are two data, one above the LOD and the other below the LOQ. Thus, it is possible that a test portion has a lot of *L. monocytogenes* on it (for e.g. a colony) so that the detection result is "presence" whereas a test portion next to the first has no *L. monocytogenes* so the enumeration result is 0. The real concentration of both test portions is then high above the LOQ. This may happen because the food is solid. In reality, it is likely that a correlation exists between the results of two test portions from the same unit of food but it is neither 0 (our assumption), neither 1 (the assumption of Busschaert et al. (2010)). The ideal case is when both detection and enumeration can be performed on the same test portion, the only reliable way to derive interval-censoring. In Commeau et al. (in press), we present additional Bayesian methods to account for variability between test portions within a food unit (and other levels of variability as well).

An assumption for both models $\mathcal{M}_{CLD}$ and $\mathcal{M}_{RD}$ is that the microbial concentration in a test portion is at least below the LOD but both models can handle pathogen free test portions. As the "concentration × test portion mass = number of *L. monocytogenes*", it is possible that a test portion has no *L. monocytogenes* while the concentration of the food sample is 0.01 cfu/g. In that case, if test portions have a mass of 10 g, there will be on average one test portion in ten contaminated with one cell. Another way to model 0 colony test portion is to consider a three-parameter model. The third parameter is the proportion of uncontaminated food in the product considered. This kind of model is presented in Gonzales-Barron et al. (2010)

with various distributions such as the Poisson and the negative binomial distributions.

The LOD and LOQ are fixed through the protocol. However, it is possible that a contaminated test portion with a concentration above the LOD (resp. above the LOQ) gets an absence result (resp. gets a 0 colony result). For instance, if a test portion of 25 g is contaminated with concentration equal to 0.1 cfu/g, the probability that a detection analysis leads to the result 'absence' is $exp(-0.1\times25)=0.08$. Model $\mathcal{M}_{CLD}$ cannot take into account this situation whereas model $\mathcal{M}_{RD}$ can: if a detection result $?y=1$ cfu, the corresponding unobserved log concentration λ can be below the LOQ.

Taking into account both detection and enumeration data is very important because often few data are available to assess a foodborne pathogen contamination. Models presented in this work can handle with both kinds of data. Model $\mathcal{M}_{RD}$ together with the EM algorithm is a way to estimate the mean and the standard deviation of a contamination without being time consuming while being more accurate than the model $\mathcal{M}_{CLD}$.

Appendix A

When there is no zero count, both models $\mathcal{M}_{CLD}$ and $\mathcal{M}_{RD}$ have the same expectation and similar variances. The expectation and the variance of model $\mathcal{M}_{CLD}$ are as follows:

$$E(c)=e^{\mu'+(\sigma')^2/2},$$

where $\mu'=\mu ln(10)$ and $\sigma'=\sigma ln(10)$, and

$$V(c)=e^{2\mu'+(\sigma')^2}\left(e^{(\sigma')^2}-1\right).$$

For model $\mathcal{M}_{RD}$, the expectation of $K=\frac{y}{LOQ}$ is as follows:

$$\begin{aligned}E(K)&=\frac{E(y)}{LOQ}\\&=\frac{1}{LOQ}E(E(y|\lambda))\\&=E(\lambda)=e^{\mu'+(\sigma')^2/2},\end{aligned}$$

thus $E(c)=E(K)$. The variance of K is equal to the following:

$$\begin{aligned}V(K)&=\frac{V(y)}{(LOQ)^2}\\&=\frac{1}{(LOQ)^2}(E(V(y|\lambda))+V(E(y|\lambda)))\\&=\frac{1}{(LOQ)^2}(E(\lambda LOQ)+V(\lambda LOQ))\\&=\frac{1}{(LOQ)^2}\left(LOQE(\lambda)+(LOQ)^2V(\lambda)\right)\\&=\frac{e^{\mu'+(\sigma')^2/2}}{LOQ}+e^{2\mu'+(\sigma')^2}\left(e^{(\sigma')^2}-1\right)\\&=\frac{e^{\mu'+(\sigma')^2/2}}{LOQ}+V(c).\end{aligned}$$

For a given value of σ, when μ is sufficiently high so that the probability of having zero counts is negligible, the term $\frac{e^{\mu'+(\sigma')^2/2}}{LOQ}$ is much smaller than $V(c)$. In fact, the limit of $V(K)/V(c)$, as μ approaches infinity is unity. In this case, models $\mathcal{M}_{CLD}$ and $\mathcal{M}_{RD}$ are similar because their first two moments are almost equal.

References

Abadias, M., Canamas, T., Asension, A., Anguera, M., Vinas, I., 2006. Microbial quality of commercial "Golden Delicious" apples throughout production and shelf-life in Lleida (Catalonia, Spain). International Journal of Food Microbiology 108, 404–409.

Augustin, J.C., Carlier, V., 2006. Lessons from the organization of a proficiency testing program in food microbiology by interlaboratory comparison: analytical methods in use, impact of methods on bacterial counts and measurement uncertainty of bacterial counts. Food Microbiology 23, 1.

Beaufort, A., Rudelle, S., Gnanou-Besse, N., Toquin, M.T., Kerouanton, A., Bergis, H., Salvat, G., Cornu, M., 2007. Prevalence and growth of *Listeria monocytogenes* in naturally contaminated cold-smoked salmon. Letters in Applied Microbiology 44 (4), 406–411.

Burfoot, D., Archer, J., Horváth, E., Hooper, G., Allen, V., Hutchison, M., Harrison, D., 2010. Fate of *Salmonella* spp. on broiler carcasses before and after cutting and/or deboning. Technical Report. Campden BRI and University of Bristol. http://www.efsa.europa.eu/en/scdocs/doc/45e.pdf, last accessed on May 11, 2011.

Busschaert, P., Geeraerd, A., Uyttendaele, M., Van Impe, J., 2010. Estimating distributions out of qualitative and (semi)quantitative microbiological contamination data for use in risk assessment. International Journal of Food Microbiology 138, 260–269.

Busschaert, P., Geeraerd, A., Uyttendaele, M., Van Impe, J., 2011. Hierarchical Bayesian analysis of censored microbiological contamination data for use in risk assessment and mitigation. Food Microbiology 28, 712–719.

Commeau, N., Cornu, M., Albert, I., Denis, J.-B., Parent, E., in press. Listeria in food, risk assessment accounting for between and within batch variability: a Bayesian modeling. Risk Analysis.

Cornu, M., Damerdji, Y., Beaufort, A., 2005. Evaluation de l'exposition à *L. monocytogenes* : exemple du calcul de la fréquence d'exposition liée aux rillettes. AFSSA – Bulletin Épidémiologique 4–5.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B: Methodological 39, 1–38.

Gnanou Besse, N., Audinet, N., Beaufort, A., Colin, P., Cornu, M., Lombard, B., 2004. A contribution to the improvement of *Listeria monocytogenes* enumeration in cold-smoked salmon. International Journal of Food Microbiology 91, 119–127.

Gonzales-Barron, U., Butler, F., 2011a. Characterisation of within-batch and between-batch variability in microbial counts in foods using poisson-gamma and poisson-lognormal regression models. Food Control 22, 1268–1278.

Gonzales-Barron, U., Butler, F., 2011b. A comparison between the discrete poisson-gamma and poisson-lognormal distributions to characterise microbial counts in foods. Food Control 22, 1279–1286.

Gonzales-Barron, U., Kerr, M., Sheridan, J.J., Butler, F., 2010. Count data distributions and their zero-modified equivalents as a framework for modelling microbial data with a relatively high occurrence of zero counts. International Journal of Food Microbiology 136, 268–277.

Habib, I., Sampers, I., Uyttendaele, M., Berkvens, D., De Zutter, L., 2008. Baseline data from a Belgium-wide survey of *Campylobacter* species contamination in chicken meat preparations and considerations for a reliable monitoring program. Applied and Environmental Microbiology 74, 5483–5489.

Helsel, D., 2006. Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. Chemosphere 65, 2434–2439.

ILSI, 2010. Impact of microbial distribution on food safety. , pp. 1–68.

ISO11290-1, 1996. Horizontal method for the detection and enumeration of *Listeria monocytogenes*. Part 1: detection method. .

ISO11290-2, 1998. Horizontal method for the detection and enumeration of *Listeria monocytogenes*. Part 2: enumeration method. .

Lorimer, M., Kiermeier, A., 2007. Analysing microbiological data: Tobit or not Tobit? International Journal of Food Microbiology 116, 313–318.

Pouillot, R., Delignette-Muller, M.L., 2010. Evaluating variability and uncertainty separately in microbial quantitative risk assessment using two R packages. International Journal of Food Microbiology 142, 330–340.

Pouillot, R., Goullet, V., Delignette-Muller, M.L., Mahé, A., Cornu, M., 2009. Quantitative risk assessment of *Listeria monocytogenes* in French cold smoked salmon: II. Risk characterization. Risk Analysis 29, 809–819.

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 3-900051-07-0.

Reinders, R., De Jonge, R., Evers, E., 2003. A statistical method to determine whether micro-organisms are randomly distributed in a food matrix, applied to coliforms and *Escherichia coli* O157 in minced beef. Food Microbiology 20, 297–303.

Robert, C., Casella, G., 2004. Monte-Carlo statistical methods. Springer. 645pp.

Shorten, P., Pleasants, A., Soboleva, T., 2008. Estimation of microbiological growth using population measurements subject to a detection limit. International Journal of Food Microbiology 108, 369–375.

Thomas, A., O'Hara, B., Ligges, U., Sturtz, S., 2006. Making BUGS open. R News 6, 12–17.

Uyttendaele, M., Busschaert, P., Valero, A., Geeraerd, A., Vermeulen, A., Jacxens, L., Goh, K., de Loy, A., Van Impe, J., Devlieghere, F., 2009. Prevalence and challenge tests of *Listeria monocytogenes* in Belgian produced and retailed mayonnaise-based delisalads, cooked meats products and smoked fish between 2005 and 2007. International Journal of Food Microbiology 133, 94–104.