Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.

This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/authorsrights

Author's personal copy

Contents lists available at ScienceDirect

International Journal of Food Microbiology

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

Short communication

Estimating the correlation between concentrations of two species of bacteria with censored microbial testing data

Michael S. Williams *, Eric D. Ebel

Risk Assessment and Analytics Staff, Office of Public Health Science, Food Safety and Inspection Service, United States Department of Agriculture, 2150 Centre Ave, Building D., Fort Collins, CO 80526, United States

ARTICLE INFO

Article history:
Received 14 May 2013
Received in revised form 2 January 2014
Accepted 11 January 2014
Available online 17 January 2014

Keywords:
Campylobacter
Salmonella
Maximum likelihood estimation
Risk assessment

ABSTRACT

Indicator organisms, such as generic *Escherichia coli* (GEC) and coliforms, can be used to measure changes in microbial contamination during the production of food products. Large and consistent reductions in the concentration of these organisms demonstrates an effective and well-controlled production process. Nevertheless, it is unclear to what degree concentrations of indicator organisms are related to pathogenic organisms such as *Campylobacter* and *Salmonella* on a sample-by-sample basis. If a strong correlation exists between the concentrations of different organisms, then the monitoring of indicator organisms would be a cost-effective surrogate for the measurement of pathogenic organisms. Calculating the correlation between the concentrations of an indicator and pathogenic organism is complicated because microbial testing datasets typically contain a large proportion of censored observations (i.e., samples where the true concentration is not observable, with nondetects and samples that are only screen-test positive being examples). This study proposes a maximum likelihood estimator that can be used to estimate the correlation between the concentrations of indicator and pathogenic organisms. An example based on broiler chicken rinse samples demonstrates modest, but significant positive correlations between the concentration of the indicator organism GEC when compared to the concentration of both *Campylobacter* and *Salmonella*. A weak positive correlation was also observed between concentrations of *Campylobacter* and *Salmonella*, but it was not statistically significant.

Published by Elsevier B.V.

1. Introduction

Microbiological data are collected and tested for a broad range of applications in fields ranging from agronomy to zoology. It is common practice to assess the presence and concentration of multiple species of organisms in each sample. For example, the Food Safety and Inspection Service (FSIS) in the United States collects samples of poultry carcasses and measures the concentration of generic *Escherichia coli* (GEC), *Salmonella*, *Campylobacter* and coliforms (FSIS, 2008). The cost of testing for each of these species varies considerably, with the non-pathogenic GEC and coliform testing being less expensive and time consuming than testing for the pathogenic species such as *Salmonella, Listeria monocytogenes*, and *Campylobacter*. Species that are easily quantified and present in nearly all samples are referred to as indicator organisms. Testing for indicator organisms is advantageous because they are found in a much higher proportion of samples than pathogenic organisms.

Interest lies in determining whether the concentration of an indicator organism is positively correlated with concentrations of a target organism (Buchanan and Oni, 2012). If concentrations of GEC were sufficiently well correlated with concentrations of *Salmonella* or *Campylobacter*, then GEC concentrations would serve as reasonable indicators for concentrations of a pathogen. Using concentrations of indicator organisms, rather than concentrations of pathogenic organisms, would reduce sample size requirements and laboratory costs. A negative correlation between an indicator organism and a pathogen would also be of interest because it would suggest that competition between species was suppressing the population of pathogenic organisms. For example, a negative correlation would be expected between an effective probiotic and *Salmonella* (Van Immerseel et al., 2002; Gaggia et al., 2010).

Testing microbiological samples can be challenging because one sample can contain a single viable organism, while another sample can contain millions of organisms. This disparity is typically due to clustering of contaminants or the exponential growth rate of bacteria. The wide range of possible concentrations of contamination across samples often requires subdividing the sample and performing multiple tests on subsamples to detect and quantify different species. For organisms that occur infrequently, a large subsample is tested to determine if the organism is present in the sample. These tests are referred to as screening tests and they return a qualitative result of presence or absence of the target organism (+/−). Additional tests are performed on the remaining subsamples when the organism of interest was detected. These tests estimate the number of organisms in the sample.

Both the initial qualitative screening test and the subsequent test to quantify the number of organisms are subject to a minimum concentration at which detection and enumeration of the organism is possible.

* Corresponding author at: 2150 Centre Ave, Building D., Fort Collins, Colorado CO 80526, United States.
E-mail address: mike.williams@fsis.usda.gov (M.S. Williams).

0168-1605/$ – see front matter. Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.ijfoodmicro.2014.01.007

Author's personal copy

The terms limit of detection (denoted by L_{qual}) and limit of quantitation (denoted by L_{quan}) are used to denote the minimum number of organisms in the sample for which detection and quantitation are possible, respectively. Observations where the organism is not detected or quantitated are referred to as censored data, because the true concentration of the organism is below the limit of detection of the assay. Samples where the concentration is less than L_{qual} result in left-censored data, while a sample that generates a positive screening test result but that cannot be quantitated, results in an interval-censored datum (Helsel, 2005; Shorten et al., 2006; Pouillot and Delignette-Muller, 2010). Data also can be right-censored when the number of organisms in the sample is too large for accurate quantitation, but this phenomena occurs less frequently because the sample can easily be diluted until quantitation is possible. For this reason right-censored data will not be addressed in this study.

The test results of a single sample, collected during production, are rarely of interest. Instead, the goal of most analyses is to combine the test results from multiple samples to provide descriptive statistics for a population. The presence of censored observations implies that even simple statistics, such as the population mean and standard deviation, cannot be estimated without relying on a model to adjust for the influence of the unobserved data. Thus inference is model-based (Gregoire, 1998). A common approach is to fit the data to a parametric distribution and use the parameters of the fitted distribution to describe the characteristics of the population. The fitted distribution function can be used for quality control applications or for quantitative microbial food- or water-safety risk assessments. For risk assessment applications, not only are the parameter estimates of interest, the associated covariance matrix serves an important role in describing the uncertainty in the estimates, with the correlation between parameters playing a vital role in appropriate characterization of the exposure distribution (Haas, 1999; Wu and Tsang, 2004).

Correlation can be quantified using different statistics (Oakes, 1982; Lyles et al., 2001; Langdon et al., 2007; Newton and Rudel, 2007). For the purpose of assessing the utility of an indicator organism, Pearson's correlation coefficient is the most appropriate, because it measures the strength of the linear relationship between the log transformation of two normally distribututed random variables. This statistic, however, cannot be directly calculated using datasets where observations are censored without using ad hoc methods, such as substituting one half the L_{quan} for censored observations. The concern with the ad hoc methods is that they can cause bias in an estimator (Shorten et al., 2006). For example, one ad hoc method for estimating the correlation is to calculate the correlation coefficient using only samples that were successfully enumerated. Significance of the correlation can also be assessed using a number of methods (Hollander and Wolfe, 1973), but the validity of these methods is questionable, given that they ignore both the potential influence and variability of the censored observations.

This article presents a probabilistic framework for estimating the correlation between the concentrations of an indicator and pathogenic organisms for datasets containing left- or interval-censored observations. The discussion begins by defining a maximum likelihood estimator that provides an estimate of the correlation coefficient as well as parameters describing the individual distributions of each organism. An example based on the simultaneous testing of broiler chicken carcasses for GEC, *Campylobacter*, and *Salmonella* is provided.

2. Methods

Let X and Y be random variables describing the outcomes of laboratory tests for two different species of organism, where X will be assumed to be the indicator organism. Also define x_i and y_i as the data values for sample i. Assume a qualitative and quantitative test for each organism with limits of detection and limits of quantitation given by $L_{qual,x}$, $L_{quan,x}$, $L_{qual,y}$, and $L_{quan,x}$. The outcomes of the laboratory test results can take on different forms. For example, the results may be integer-valued data derived from colony counts per plate (Commeau et al., 2012; Williams and Ebel, 2012) or they can be estimates of the average concentration derived from the Most Probable Number technique (Cochran, 1950; Pouillot et al., 2013). Regardless of the form of the laboratory data, the goal of the analysis is to fit a single multivariate distribution to the data that describes average concentration of the indicator and pathogenic organisms and their correlation. The lognormal distribution is a commonly used distribution for describing microbial populations and fitting methods have been developed to address integer-valued data (Williams and Ebel, 2012) and Most Probable Number data (Pouillot et al., 2013). It is reasonable to fit integer count and estimated concentration data directly to a lognormal distribution (Busschaert et al., 2010), provided the proportion of samples whose concentration is below L_{qual} is not too large (Williams and Ebel, 2012) as is the case in this study.

Maximum likelihood estimation is the common approach for assessing the correlation between concentrations with censored data (Lyles et al., 2001; Newton and Rudel, 2007). When the log transformed test results from N samples are assumed to follow a multivariate normal distribution, as is common in quantitative microbiology (Stoline, 1991; Mitzenmacher, 2003), the parameter vector $\theta = (\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2, \rho_{XY})$ can be estimated by maximizing the multivariate likelihood function.

The contribution of each x or y value to the likelihood depends on whether the datum is left-, interval- or uncensored. To provide some insight into the form of the likelihood function it is helpful to first describe the likelihood function in a univariate application. In this setting, if X is successfully enumerated (uncensored), its contribution to the likelihood is $f_X(x)$. If X is left censored the cumulative density function, evaluated at the limit of detection $F_X(_{qual,x})$ is used because the only information provided by the sample is that the concentration is less than $L_{qual,x}$. If the X is interval censored, the contribution to the likelihood is $F_X(L_{quan,x}) - F_X(L_{qual,x})$, which represents the proportion of samples that were qualitatively positive, but returned a count of 0 during enumeration. The likelihood function for this application considers every combination of left-, interval- and uncensored results for X and Y in the bivariate normal distribution.

Extending the notational convention of Newton and Rudel (2007) and Hollander and Wolfe (1973) yields the likelihood function

$$\mathscr{L}(\theta) = \prod_{i=1}^{N} G_{X,Y}(x_i, y_i|\theta), \tag{1}$$

where the form of each term of the likelihood function varies according to the specific ranges of X and Y. The definition of each term, as a function of X and Y, is given by

$$\begin{aligned}
G_{X,Y}(x_i, y_i|\theta) &= f_{X,Y}(x_i, y_i); L_{quan,x} < x_i, L_{quan,y} < y_i \\
&= F_{X,Y}\left(L_{qual,x}, L_{qual,y}\right); x_i < L_{qual,x}, y_i < L_{qual,y} \\
&= F_{X,Y}\left(L_{quan,x}, L_{quan,y}\right) - F_{X,Y}\left(L_{qual,x}, L_{qual,y}\right); \\
&\qquad L_{qual,x} \le x_i < L_{quan,x}, L_{qual,x} \le y_i < L_{quan,y} \\
&= f_X(x_i) F_{Y|X}\left(L_{qual,y}|x_i\right); L_{quan,x} < x_i, y_i < L_{qual,y} \\
&= f_Y(y_i) F_{X|Y}\left(L_{qual,x}|y_i\right); x_i < L_{qual,x}, L_{quan,y} \le y_i \\
&= f_X(x_i)\left[F_{Y|X}\left(L_{quan,y}|x_i\right) - F_{Y|X}\left(L_{qual,y}|x_i\right)\right]; \\
&\qquad L_{quan,x} \le x_i, L_{qual,y} \le y_i < L_{quan,y} \\
&= f_Y(y_i)\left[F_{X|Y}\left(L_{quan,x}|y_i\right) - F_{X|Y}\left(L_{qual,x}|y_i\right)\right]; \\
&\qquad L_{qual,x} \le x_i < L_{quan,x}, L_{qual,y} \le y_i. \\
&= \int_{-\infty}^{L_{qual,x}} \int_{L_{quan,y}}^{L_{quan,y}} f_{X,Y}(x_i, y_i) dy dx; \\
&\qquad x_i < L_{qual,x}, L_{qual,y} \le y_i < L_{quan,y} \\
&= \int_{L_{qual,x}}^{L_{quan,x}} \int_{-\infty}^{L_{qual,y}} f_{X,Y}(x_i, y_i) dy dx; \\
&\qquad L_{qual,x} \le x_i < L_{quan,x}, y_i < L_{qual,y}.
\end{aligned}$$

Author's personal copy

The specific distribution functions in the likelihood are

$$\begin{aligned} f(x_i, y_i) &= \text{bivariate normal with parameter vector } \theta \\ f_X(x_i) &= \text{univariate normal with parameters } \mu_X, \sigma_X \\ f_Y(y_i) &= \text{univariate normal with parameters } \mu_Y, \sigma_Y \\ F_{Y|X}\left(L_{*,y}|x_i\right) &= \text{univariate normal CDF with parameters } \mu_{Y|x_i}, \sigma_{Y|X} \\ F_{X|Y}\left(L_{*,x}|y_i\right) &= \text{univariate normal CDF with parameters } \mu_{X|y_i}, \sigma_{X|Y} \\ F_{X,Y}\left(L_{*,x}, L_{*,y}\right) &= \text{bivariate normal CDF with parameters vector } \theta, \end{aligned}$$

and the marginal moments are given by

$$\begin{aligned} \mu_{Y|x_i} &= \mu_Y + (\rho_{XY}\sigma_Y/\sigma_X) * (x_i - \mu_X) \\ \mu_{X|y_i} &= \mu_X + (\rho_{XY}\sigma_X/\sigma_Y) * (y_i - \mu_Y) \\ \sigma_{Y|X} &= \sigma_Y\sqrt{1-\rho_{XY}^2} \\ \sigma_{X|Y} &= \sigma_X\sqrt{1-\rho_{XY}^2}. \end{aligned}$$

The formulation of this likelihood function assumes both X and Y have some interval-censored observations. In situations where no screening test is applied to either X or Y, $L_{qual} = L_{quan}$ and the interval censored terms can be ignored. This formulation is also not specific to the lognormal distribution, so alternative multivariate distributions can be used, provided there is biological support for the alternative multivariate distribution.

A nonlinear optimization routine can be used to estimate

$$\theta = \left(\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2, \sigma_{XY}\right).$$

Many optimization routines also provide a numerical approximation of the Hessian matrix, which can be inverted to provide the estimated variance–covariance matrix $\hat{\Sigma}$. Elements of this matrix provides estimates of $var[\hat{\mu}_*]$, $var[\hat{\sigma}_*^2]$, and $cov[\hat{\mu}_*, \hat{\sigma}_*^2]$, where $* = X,Y$. These estimates are used to describe the uncertainty in the parameter estimates.

Standard practice in microbiology is to express the mean and variance parameters describing microbial count data in $\log_{10}$, rather than $\log_e$, scale. The conversion between the two different scales is generally necessary for computation and is achieved by division of the $\hat{\mu}$ and $\hat{\sigma}$ parameters by $\log_e(10)$.

Ideally one would like to estimate the full parameter vector θ, yet maximizing the likelihood function for the full vector can sometimes be problematic. One option is to first fit the data for both variables to distributions independently and use the resulting parameter vectors $\left(\hat{\mu}_X, \hat{\sigma}_X^2\right)$ and $\left(\hat{\mu}_Y, \hat{\sigma}_Y^2\right)$ as starting values for fitting the full vector. Another approach is to hold the vectors $\left(\hat{\mu}_X, \hat{\sigma}_X^2\right)$ and $\left(\hat{\mu}_Y, \hat{\sigma}_Y^2\right)$ fixed and generate $\hat{\rho}_{XY}$ using a simplified likelihood function (Newton and Rudel, 2007).

3. Example

Interest lies in estimating the correlation between the concentrations of indicator and pathogenic organisms when the data contain left- and interval censored observations. This analysis examines the correlation between GEC, *Campylobacter* and *Salmonella* derived from broiler chicken rinse samples. Three test cases are presented, with the first two cases being an assessment of the correlation between the GEC indicator organism (denoted by X) and the two pathogenic organisms (Y). The third test case examines the correlation between *Campylobacter* and *Salmonella*. This example treats *Campylobacter* as the indicator organism, due to higher percent positive rate and concentration, when compared to *Salmonella*. These examples are of practical interest because previous studies have demonstrated mixed results regarding the significance of correlations between GEC, *Campylobacter* and *Salmonella* (Hue et al., 2011; Ghafir et al., 2008; Zhao et al., 2002).

Routines were written to calculate the value of the likelihood function $\mathscr{L}(\theta)$ using the R programming language (R Development Core Team, 2012). The nonlinear optimization function (*optim*) was used to generate the estimated parameter vector

$$\hat{\theta} = \left(\hat{\mu}_X, \hat{\sigma}_X^2, \hat{\mu}_Y, \hat{\sigma}_Y^2, \hat{\sigma}_{XY}\right).$$

Output from this routine also provides a numerical approximation for the variance–covariance matrix $\hat{\Sigma}$. The diagonal elements of $\hat{\Sigma}$ are used to generate the confidence intervals for $\hat{\rho}_{XY}$.

Convergence of the optimization routine was aided by accurate initial values for the μ and σ^2 parameters describing the univariate distribution of *Campylobacter* and *Salmonella*. These estimates were derived from the maximization of the univariate likelihood function for the lognormal distribution for each organism individually (Shorten et al., 2006).

4. Broiler chicken dataset

The examples are based on data from a large survey of broiler chicken production collected by FSIS in the United States. This example uses concentrations of GEC (X) and compares them to concentrations of *Campylobacter* and *Salmonella* (Y) found in 400 ml rinse samples collected at re-hang (FSIS, 2008). For this example, GEC was chosen over other indicator organisms because it is thought to be a good indicator of fecal contamination and more GEC samples were non-detects than Enterobacteriaceae or coliforms. The dataset consists of $N = 3{,}164$ observations. GEC and *Campylobacter* are typically found on the majority of poultry carcasses at re-hang, with the percentage of samples where the organism was found being 99.2 and 70.9%, respectively. Given this high percentage, a screening test was not used. The high concentrations of both GEC and *Campylobacter* at re-hang allow for enumeration using direct plating of small aliquot volumes, which were typically 0.1 *ml* of rinsate (or less if additional dilution was required). This aliquot size yields a limit of quantitation of $L_{quan,x} = 10cfu/ml$. For *Salmonella*, a screening test was performed and 45.8 percent of samples were test-positive. Screen-test positive samples were enumerated. Qualitative and quantitative limits of $L_{qual,y} = 0.01cfu/ml$ and $L_{quan,y} = 0.03cfu/ml$ were assumed, respectively.

The percentage of samples where both *Campylobacter* and *Salmonella* were detected was 33.1. The percentage of samples where GEC and either *Campylobacter* or *Salmonella* were detected was 70.5 and 45.6 percent, respectively.

For the sake of comparison, the correlations between the three species were approximated by calculating the Pearson's correlation coefficient using only the quantitated samples. The values were $\hat{\rho}^*_{GEC,Campy} = 0.271$, $\hat{\rho}^*_{GEC,Salm} = 0.159$, and $\hat{\rho}^*_{Campy,Salm} = 0.015$, with the * symbol indicating that these estimates were derived by ignoring the censored observations.

5. Results

The estimated parameter vectors for the three examples are given in Table 1. The estimated covariance terms are approximately an order of magnitude smaller than the $\hat{\sigma}^2$ values, so strong correlations are not expected. The parameter estimates $\hat{\mu}$ and $\hat{\sigma}^2$ demonstrate only slight shifts between the pairings of different organisms between the univariate and multivariate analyses. For example, the $\hat{\mu}_X$ and $\hat{\sigma}_X^2$ estimates for GEC are

Species pairing	$\hat{\mu}_X$	$\hat{\sigma}_X^2$	$\hat{\mu}_Y$	$\hat{\sigma}_Y^2$	$\hat{\sigma}_{XY}$
GEC, *Campylobacter*	2.78	0.65	2.10	2.53	0.122
GEC, *Salmonella*	2.78	0.65	−2.10	3.05	0.232
Campylobacter, *Salmonella*	2.10	2.55	−2.10	3.07	0.104

Author's personal copy

identical to 2 decimal places in both cases, whereas the $\hat{\sigma}_Y^2$ estimates for *Salmonella* are 3.05 and 3.07 for the data pairings (GEC, *Salmonella*) and (*Campylobacter*, *Salmonella*), respectively. Slight shifts in the parameter estimates are expected because the datasets differ for each species pairing and the maximum likelihood estimator is accounting for the correlation between the two species. This phenomenon was observed when a number of different initial values for the parameters were used and suggests that the maximum likelihood estimator is robust for this dataset.

The estimated correlation coefficients, as well as the associated 95% confidence intervals, are presented in Table 2. These estimates are calculated using two methods. The first is the Pearson's correlation coefficient (Hollander and Wolfe, 1973), which is calculated by ignoring the censored observations. These results are denoted by a * superscript. The remaining estimates are derived from the multivariate maximum likelihood estimator. Confidence intervals for $\hat{\rho}_{XY}$ were derived from the $\hat{\Sigma}$ matrix using the asymptotic normality assumption for the maximum likelihood estimator (Mood et al., 1973).

The results suggest a weak but significant correlation between GEC and the two pathogenic organisms, with the correlation being strongest for GEC and *Salmonella*. A weak, but positive correlation exists between *Campylobacter* and *Salmonella*, though the confidence interval does not support the assertion that the correlation is significant. This result also agrees with a previous study (Hue et al., 2011), where no positive correlation between concentration of these two pathogenic organisms was observed.

No consistent patterns exist between the $\hat{\rho}_{XY}^*$ and $\hat{\rho}_{XY}$ values, nor is their agreement in the relative width of the confidence intervals, even though the sample size for the multivariate maximum likelihood estimator is much larger. The agreement between $\hat{\rho}_{XY}^*$ and $\hat{\rho}_{XY}$ values is particularly poor for the data pairing of GEC and *Campylobacter*. This lack of agreement should serve as another cautionary example of problems associated with inferences derived from heavily censored data when the censored observations are ignored (Shorten et al., 2006). We assume that the results derived from the maximum likelihood estimator are more accurate, though such an assumption is still based on the appropriateness of the lognormal distribution for these data.

6. Conclusions

This study provides a framework for estimating the correlation between two different species of bacteria. The correlations observed in the broiler chicken dataset were not high, so the concentration of GEC would not be a particularly reliable indicator of concentrations of either *Campylobacter* or *Salmonella* in this application. Instead these results can be interpreted as indicating that carcasses with the highest levels of contamination at the beginning of the slaughter process have high levels of both pathogenic and nonpathogenic organisms. This information is useful because a slaughter facility could use sampling information on GEC to assess whether additional interventions would be necessary to achieve a finished product whose total microbial load was below a predetermined threshold.

Table 2
Estimated correlation coefficients and their confidence intervals between GEC, *Campylobacter*, and *Salmonella* derived from the FSIS broiler chicken baseline study (FSIS, 2008). Values denoted by the * symbol are calculated by ignoring the censored observations in the calculation of ρ. The ρ estimates and their confidence intervals derived from the maximum likelihood method are in the last two columns and are denoted by the ^ symbol.

Species Pairing	$\hat{\rho}_{XY}^*$	95% C. I. ρ_{XY}^*	$\hat{\rho}_{XY}$	95% C. I. $\hat{\rho}_{XY}$
GEC, *Campylobacter*	0.271	(0.233, 0.309)	0.095	(0.059,0.132)
GEC, *Salmonella*	0.159	(0.105, 0.212)	0.165	(0.125, 0.244)
Campylobacter, *Salmonella*	0.015	(−0.049, 0.079)	0.037	(−0.004,0.078)

Higher correlations are expected to exist between two points in the slaughter process and additional analyses with these data finds that the average log reduction in GEC and *Salmonella* concentrations between re-hang and post-chill are very similar, with both being slightly more than 2 log_{10} (FSIS, 2008). It is expected that the concentrations on individual poultry carcasses between re-hang and post-chill would be highly correlated. This is a natural consequence of the non-selective nature of antimicrobial agents and the simple effect of washing a carcass. This assertion is supported by the documented steady decline in the levels of both indicator organisms and pathogenic bacteria from repeated rinse samples (Lillard, 1988). Note, however, this is a fundamentally different problem than the one outlined here and estimation of the correlation would require paired samples of the same carcass at two points in the production process (e.g., re-hang and post-chill). These data, however, are generally not collected for reasons such as the reduction in organisms associated with each sampling event and the difficulty of locating previously sampled carcasses at the end of the production process.

Computer code, written in the R programming language (Development Core Team, 2012), is provided as supplementary material.

Appendix A. Supplementary computer code

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ijfoodmicro.2014.01.007.

References

Buchanan, R.L., Oni, R., 2012. Use of microbiological indicators for assessing hygiene controls for the manufacture of powdered infant formula. J. Food Prot. 75, 989–997.

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

Cochran, W.G., 1950. Estimation of bacterial densities by means of the most probable number. Biometrics 6, 105–116.

Commeau, N., Parent, E., Delignette-Muller, M.L., Cornu, M., 2012. Fitting a lognormal distribution to enumeration and absence/presence data. Int. J. Food Microbiol. 155, 146–152.

FSIS, 2008. The Nationwide Baseline Data Collection Program: Young Chicken Survey. Food Safetey and Inspection Service. United States Department of Agriculture, Washington, DC (URL: http://www.fsis.usda.gov/PDF/Baseline_Data_Young_Chicken_2007-2008.pdf).

Gaggia, F., Mattarelli, P., Biavati, B., 2010. Probiotics and prebiotics in animal feeding for safe food production. Int. J. Food Microbiol. 141, S15–S28 (Supplement).

Ghafir, Y., China, B., Dierick, K., DeZutter, L., Daube, G., 2008. Hygiene indicator microorganisms for selected pathogens on beef, pork, and poultry meats in Belgium. J. Food Prot. 71, 35–45.

Gregoire, T.G., 1998. Design-based and model-based inference in survey sampling: appreciating the difference. Can. J. For. Res. 28, 1429–1447.

Haas, C.N., 1999. On modeling correlated random variables in risk assessment. Risk Anal. 19, 1205–1214.

Helsel, D.R., 2005. Nondetects and Data Analysis. Wiley Inter-science, Hoboken, New Jersey.

Hollander, M., Wolfe, D.A., 1973. Nonparametric Statistical Methods. John Wiley & Sons, New York, New York.

Hue, O., Allain, V., Laisney, M.A., LeBouquin, S., Lalande, F., Petetin, I., Rouxel, S., Quesne, S., Gloaguen, P.Y., Picherot, M., Santolini, J., Bougeard, S., Salvat, G., Chemaly, M., 2011. *Campylobacter* contamination of broiler caeca and carcasses at the slaughterhouse and correlation with *Salmonella* contamination. Food Microbiol. 28, 862–868.

Langdon, M.J., Taylor, C.C., West, R.M., 2007. Classification of type 1-censored bivariate data. Comput. Stat. Data Anal. 51, 4562–4576.

Lillard, H.S., 1988. Comparison of sampling methods and implications for bacterial decontamination of poultry carcasses by rinsing. J. Food Prot. 51, 405–408.

Lyles, R.H., Fa, D., Chuachoowong, R., 2001. Correlation coefficient estimation involving a left censored laboratory assay variable. Stat. Med. 20, 2921–2933.

Mitzenmacher, M., 2003. A brief history of generative models for power law and lognormal distributions. Internet Math. 1, 226–251.

Mood, A.M., Graybill, F.A., Boes, D.C., 1973. Introduction to the Theory of Statistics, 3rd edition. McGraw-Hill, New York, New York.

Newton, E., Rudel, R., 2007. Estimating correlation with multiply censored data arising from the adjustment of singly censored data. Environ. Sci. Technol. 41, 221–228.

Oakes, D.A., 1982. A concordance test for independence in the presence of censoring. Biometrics 38, 451–455.

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

Author's personal copy

Pouillot, R., Hoelzer, K., Chen, Y., Dennis, S., 2013. Estimating probability distributions of bacterial concentrations in food based on data generated using the most probable number (MPN) method for use in a risk assessment. Food Control 29, 350–357.

R Development Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. 3-900051-07-0 (URL: http://www.R-project.org).

Shorten, P.R., Pleasants, A.B., Soboleva, T.K., 2006. Estimation of microbial growth using population measurements subject to a detection limit. Int. J. Food Microbiol. 108, 369–375.

Stoline, M.R., 1991. An examination of the lognormal and Box and Cox family of transformations in fitting environmental data. Environmentrics 2, 85–106.

Van Immerseel, F., Cauwerts, K., Devriese, L.A., Haesebrouck, F., Ducatelle, R., 2002. Feed additives to control *Salmonella* in poultry. Worlds Poult. Sci. J. 58, 501–514.

Williams, M.S., Ebel, E.D., 2012. Methods for fitting the Poisson-lognormal distribution to microbial testing data. Food Control 27, 73–80.

Wu, F.C., Tsang, Y.P., 2004. Second-order monte carlo uncertainty/variability analysis using correlated model parameters: application to salmonid embryo survival risk assessment. Ecol. Mod. 177, 393–414.

Zhao, T., Doyle, M.P., Fedorka-Cray, P.J., Zhao, P., Ladely, S., 2002. Occurrence of *Salmonella enterica* serotype Typhimurium DT104A in retail ground beef. J. Food Prot. 65, 403–407.