Systematic Statistical Analysis of Microbial Data from Dilution Series

J. Andrés Christen & Albert E. Parker

Journal of Agricultural, Biological and Environmental Statistics

ISSN 1085-7117

JABES
DOI 10.1007/s13253-020-00397-0

Springer

Systematic Statistical Analysis of Microbial Data from Dilution Series

J. Andrés CHRISTEN and Albert E. PARKER

In microbial studies, samples are often treated under different experimental conditions and then tested for microbial survival. A technique, dating back to the 1880s, consists of diluting the samples several times and incubating each dilution to verify the existence of microbial colony-forming units or CFU's, seen by the naked eye. The main problem in the dilution series data analysis is the uncertainty quantification of the simple point estimate of the original number of CFU's in the sample (i.e., at dilution zero). Common approaches such as log-normal or Poisson models do not seem to handle well extreme cases with low or high counts, among other issues. We build a novel binomial model, based on the actual design of the experimental procedure including the dilution series. For repetitions, we construct a hierarchical model for experimental results from a single laboratory and in turn a higher hierarchy for inter-laboratory analyses. Results seem promising, with a systematic treatment of all data cases, including zeros, censored data, repetitions, intra- and inter-laboratory studies. Using a Bayesian approach, a robust and efficient MCMC method is used to analyze several real data sets.

Key Words: Dilution experiments; Binomial likelihood; Bayesian inference; Hierarchical models; MCMC.

1. INTRODUCTION

Dilution experiments are an important tool to detect the presence of microbes, even in very low concentrations, relying on basic microbiology techniques and relatively simple laboratory equipment. The microbes are first sampled from a natural environment, such as soil or drinking water, or from an engineered benchtop reactor system where they can be grown planktonically or in a biofilm before potentially being exposed to some treatment. Whatever the case, the sample containing microbes is transferred into a volume V_0 of liquid in tube 0, most commonly buffered water (Rice et al. 2017), but any appropriate liquid

J. A. Christen (✉)Centro de Investigación en Matemáticas (CIMAT-CONACYT), Jalisco S/N, Valenciana, 36023 Guanajuato, GTO, Mexico
(E-mail: *jac@cimat.mx*). A. E. Parker Department of Mathematical Sciences, Center for Biofilm Engineering (CBE), Montana State University, Bozeman, MT, USA
(E-mail: *parker@math.montana.edu*).

Journal of Agricultural, Biological, and Environmental Statistics
https://doi.org/10.1007/s13253-020-00397-0

Published online: 28 May 2020

medium may be used. It is then of interest to estimate the number of microbes in this tube 0, i.e., the *abundance* of microbes in the original sample. Analyzing the original sample directly may be possible using specialized equipment and more complex laboratory processes (e.g., a confocal scanning laser microscope Pitts and Stewart 2008; Parker et al. 2018a). Instead, subvolumes from tube 0 may be "spread-plated" or "drop-plated" (Herigstad and Hamilton 2001) onto growth media in a Petri dish to let bacteria grow until distinguishable by the naked eye (see Fig. 1c). When plating directly from samples that contain a high density of microbes, it is not possible to identify and count individual colonies, in which case it is necessary to dilute the original volume until only a few colonies can be counted after plating in a Petri dish. From these counts at some dilution(s), the microbial abundance is inferred in tube 0 and hence in the original sample. The process is called a *dilution series* and is described as follows.

To begin the dilution series, from the volume V_0 in tube 0 a subvolume V is taken which is diluted at a factor α, typically $\alpha = 10$, to form a new volume αV in a new tube at dilution $j = 1$. This process is repeated for $j = 1, \ldots, J-1$: a subvolume V is taken from the volume αV in tube j and then diluted by a factor α to form a new volume αV in a new tube at dilution $j+1$. A smaller volume u is then removed from each tube and then plated onto growth media in a Petri dish, and kept at optimal environment conditions (e.g., temperature and pH) to allow any viable microbes to grow and reproduce. Individual microbes or clumps of microbes will start forming colonies in the Petri dish, until, at a high enough dilution, they are distinguishable to the naked eye as dots called colony-forming units (CFUs).

Estimating microbial abundances from colonies dates at least as far back as the seminal work of Robert Koch in the 1880s (Prescott et al. 1996, p. 9). These estimates are expressed as CFUs rather than as the *number* of microbes because of a number of known limitations, the most obvious being the questionable assumption that each colony arises from an individual cell (Prescott et al. 1996, p.119); see also Cundell (2015) for a more recent discussion). Still, especially for a single microbial species isolated from a consortia in an environmental sample, or for a mono-culture grown in the laboratory, the CFU remains a useful quantitative measure for estimating microbial abundances. Dilution series for CFU counting are performed routinely in many government, academic and private laboratories for experimentation as well as for testing and public standard compliance (see, for example, Ben-David and Davidson 2014; FDA 2018, or a "dilution series" search in fda.gov). Indeed, CFU counting is a required international metric for assessing the efficacy of antimicrobial treatments in North America and Europe (Parker et al. 2018b).

In a dilution series, for some lower dilutions, too many CFUs may cluster and will be impossible to count and are reported as "too numerous to count" (TNTC). For higher dilutions, there will eventually be no CFUs (no microbial activity). Commonly, one dilution is then selected, namely dilution j, $0 \le j \le J-1$, for CFU counting, having a minimum number of distinguishable CFUs per plate or drop, and referred to as the *lowest countable dilution*. From this, the microbial abundance in the original sample is to be estimated. The crudest estimate is the (number of CFUs) $\times \alpha_0 \times \alpha^j \times \alpha_p$, where $\alpha_0 = V_0/V$ and $\alpha_p = V/u$ is the ratio of the dilution tube volume V over the volume plated, or drop volume, u. The usual formula used by practitioners is (number of CFUs) $\times V_0/u \times \alpha^j$, which is precisely

equal to the latter. See Fig. 1 for an illustration of the process, CFU counting and basic data analysis.

Instead of taking solely the lowest countable dilution, a possible variant is to weightedly average the CFUs across multiple dilutions (Hedges 2002; Maturin and Peeler 1998; Niemela 1983; Niemi and Niemela 2001; Parkhurst and Stern 1998) that is motivated by the Horvitz–Thompson estimator, popular in field ecology (Horvitz and Thompson 1952). Hamilton and Parker (2010) argue that the added information is minimal for common dilution experimental designs. Our investigation (presented here) leads to this same conclusion, which supports the microbiologist's conventional practice of using data from only the first countable dilution to estimate the microbial abundance in the original sample.

In most situations, the abundance of microbes in the original sample spans several orders of magnitude and therefore, it is common to estimate a *log abundance* that is the $\log_{10}$-transform of the CFU count in dilution 0. The estimated mean log abundance from statistical analyses can easily be normalized to a mean *log density* per unit of the volume or surface area of the original specimen, S_c, by simply subtracting by $\log_{10}(S_c)$.

The efficacy of an antimicrobial treatment is usually quantified by a *log reduction* (LR) which is the estimated mean log abundance of microbes that survived the treatment subtracted from the estimated mean log abundance of microbes in a concurrent control. The microbes in the control samples are subjected to the same conditions as the treated samples with the exception that the controls are subjected to a placebo treatment. Perhaps not surprisingly, the log abundances of the microbes in the control samples are typically much less variable than the microbes subjected to an antimicrobial (Parker et al. 2018b). Hence, the constant variance assumption of many statistical models is often violated when analyzing a data set that includes both control and treated samples.

There are two common approaches for analyzing CFU data. One uses a Poisson likelihood model of the counts (Haas et al. 1999), while the other uses a log-normal likelihood model of the counts (Hamilton et al. 2013). Both maximize the likelihood to provide microbial abundance estimates and quantify uncertainty. Both also can be extended to handle random effects (Zuur et al. 1999) due to samples being repeatedly collected from the same site, experiment and/or the same laboratory. Software is readily available to fit either of these types of models (see, e.g., Bates et al. 2015).

Haas et al. (1999, p. 213) argue that a Poisson maximum likelihood estimator (MLE) approach is preferred over the log-normal MLE. One reason is that the Poisson likelihood naturally deals with zero CFU counts. The obvious downside to the Poisson model is the requirement that the variance is equal to the mean. The control data that we present here clearly do not adhere to this restriction. The generalized Poisson and negative binomial are both extensions of the Poisson that allow for the variance to be independently estimated from the mean (Joe and Zhu 2005). Nonetheless, neither of these Poisson approaches allow one to directly analyze LRs.

The log-normal MLE approach overcomes the restriction that the mean equals the variance by including separate parameters for the mean and variance. To deal with the differing variability of microbes treated by antimicrobials versus controls, one can aggregate the log abundance estimates into LRs and then apply a normal model to the LRs (Hamilton et al. 2013). Unfortunately, this approach does not allow one to separately model or estimate the

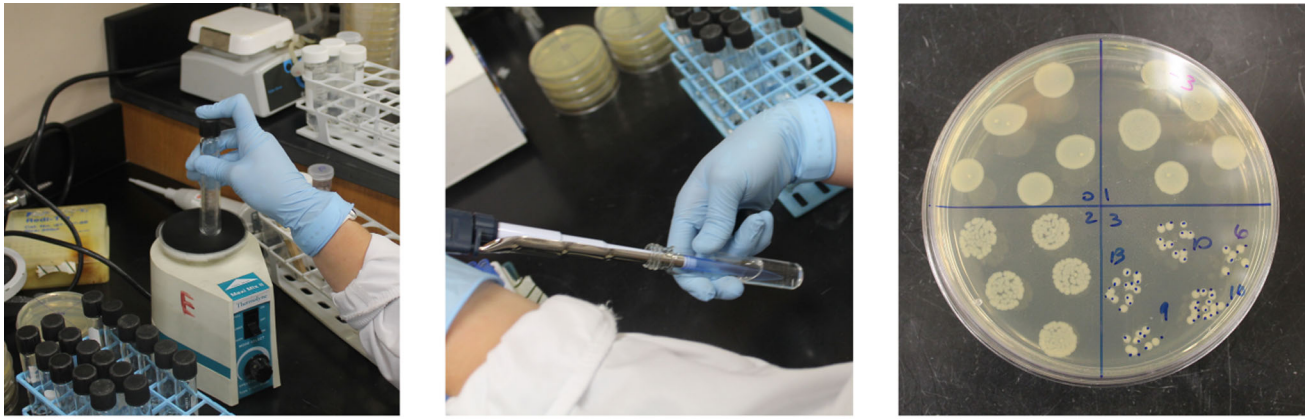

Figure 1. **a** Treated samples (here there are $K = 3$ repetitions) are transported and placed in dilution 0 tubes (top right) and from these subsequent dilutions are formed (bottom left) by adding buffered water and then homogenizing with an orbital shaker (center). **b** From each dilution, a volume is taken and drops are plated using an electronic micropipette. **c** Example of drops plated in Petri dish from dilutions 0, 1, 2 and 3. Dilution 3 (i.e., $j = 3$) is selected for CFU counting, and others are "too numerous to count" (TNTC). Counts for dilution 3 are 13, 10, 6, 9 and 16. As in the example in Sect. 3.1, $\alpha_0 = 1, \alpha = 10$ and $\alpha_p = 10^3$ leading to a simple mean estimator of 10.8×10^6. We use a binomial model to approach this estimation problem formally and use a Bayesian approach to quantify the uncertainty in estimates of CFU counts .

variance among counts on different plates at different dilutions. When presented with CFU data including a zero or TNTC, a common tactic when using a log-normal likelihood is to substitute in a small value (CFU = 0.5 or 1) for zero and to substitute in the largest count for TNTC (30 for the drop plate method, 300 for the spread plate method). Many published simulation studies show that as long as there are not too many censored data ($\dot{\leq}$15% of the total data set), the log-normal model has little bias and mean squared error when estimating the mean log abundance of organisms (see, e.g., Clarke 1998; Haas and Scheff 1990; Singh and Nocerino 2002; EPA 1996).

The approach that we present here has several advantages over the conventional approaches just described.

1. The design and assumptions of the dilution experiments lead directly to our binomial model in (1) and (2).

2. Any censored data are directly modeled in the binomial likelihood (i.e., no substitution rules are employed) resulting in zeros and TNTCs handled systematically by the same model.

3. Counts from multiple dilutions are directly incorporated into the model (i.e., they are not aggregated together before statistical analysis), which allows us to separate out the variance among counts on different plates and on different dilutions from other sources of variance that contribute to the variance of log abundances and LRs.

4. Our model accounts for clustering of the cells (with the miscount probability q) that violates the assumption that one microbe generates one CFU. This aspect of the model can also deal with miscounts by the technician who actually counts the CFUs.

5. Instead of summarizing the results from each experiment by a LR and then analyzing the LRs using a normal model, we provide an over-arching hierarchy for the analysis of LRs that has the explicit information regarding the CFUs and dilutions that led to the LR.

The paper is organized as follows: In Sect. 2 we build our binomial model from the description of the experimental design. Using a basic theorem, a simplification is obtained leading a straightforward likelihood. We also describe our Bayesian inference process and a hierarchical modeling strategy to analyze intra- and inter-laboratory data and the MCMC method used. Two examples are presented in Sect. 3 considering real data, and in Sect. 4, we present a discussion of the paper.

2. MODEL AND INFERENCE

As explained above, microbial samples are treated with, for example, a chemical agent at some concentration, or water temperature for a specified contact time, and the effect of this treatment is observed as the number of CFUs from surviving microbes using a dilution series. Commonly, several experiments may be conducted with the same treatment, and within each experiment, multiple repetitions or samples are also considered. To ease

analysis and notation, we concentrate on one experiment for a single treatment. We make a comment on the analysis of multiple treatments in Sect. 4.

Throughout, we will refer to the plate or drop from which CFUs were counted only as 'drop'; our model may consider in principle any such volume from which CFUs are counted simply by using a different division factor $\alpha_p = V/u$.

Let K be the number of repetitions of microbial samples to be analyzed in one experiment. Each sample, grown in an original surface area, volume, etc., S_c, is transported into a volume V_0 (typically buffered water) and carefully homogenized in tube 0 of each repetition $k = 1, 2, \ldots, K$. Some fraction $\alpha^{-1}\alpha_0^{-1}V_0$ of the volume V_0 in this first tube is taken to tube 1, diluted in $(1-\alpha^{-1})\alpha_0^{-1}V_0$ volume and carefully homogenized. This process is repeated to produce tubes $j = 1, \ldots, J-1$ for each repetition. Tubes 1 through $J-1$ will have $V = \alpha_0^{-1}V_0$ volume, and $\alpha_0 = V_0/V$ is the proportion of volume in tube 0 versus the common volume V in the rest of the dilution series tubes. In our examples $V_0 = V = 10\,\text{ml}$, $\alpha_0 = 1$ (intra-laboratory example studying heat treatments) and $V_0 = 40\,\text{ml}$, $V = 10\,\text{ml}$ and $\alpha_0 = 4$ (inter-laboratory example studying bleach treatments).

D 'drops' of volume u are then plated from each dilution tube. For example, for the drop plate method, a total of $D = 10$ drops, each with volume $u = 10\,\mu\text{l}$, are usually plated. For the spread plate method, $D = 2$ drops, each with volume $u = 0.1$ ml or $u = 1$ ml, is common. One or several Petri dishes may be used to grow CFUs from these drops, but we will not distinguish between different dishes from the same dilution tube. For $j = 1, \ldots, J-1$ the proportion plated is $\alpha_p = V/u$ and for $j = 0$, $\alpha_p\alpha_0 = V_0/u$.

Let N_j^k be the r.v. representing the number of CFUs in dilution j for repetition k and let n_j^k be any particular realization for it (we use the standard probabilistic notation of upper case being the r.v. and lower case a particular value for it, e.g., $P(Y \leq y|X = x)$, the probability of Y being less or equal to the particular value y conditional on $X = x$) . Let $Y_{j,i}^k$ be the CFU count in drop $i = 1, 2. \ldots, D$, for dilution j for repetition k. Our approach is to build a hierarchical model following the dilution process and the counting process just described. Due to homogenization, we can safely assume that N_j^k and $Y_{j,i}^k$ are binomial random variables. Namely

$$Y_{j,i}^k | N_j^k = n_j^k, q \quad \sim \text{Bi}(n_j^k, \alpha_p^{-1}\alpha_0^{-\delta_{0,j}}(1-q)) \tag{1}$$

$$N_j^k | N_{j-1}^k = n_{j-1}^k \quad \sim \text{Bi}(n_{j-1}^k, \alpha^{-1}\alpha_0^{-\delta_{1,j}}) \tag{2}$$

for $j = 0, 1, \ldots, J-1$, where $\delta_{i,j}$ is the Dirac function. Here, $1-q$ is included as an additional probability that each individual microbe in each drop actually does form a distinguishable colony and adds to the CFU count.

The proposed model above brings up an interesting philosophical point. This binomial model is simply describing the way that the experiment is conducted, as opposed to a Poisson or log-normal. As explained in the introduction, the above model is in fact describing the experimental design with no further assumptions in the statistical modeling than those already assumed by the experimenters conducting the dilution experiments. These assumptions are that: through careful homogenization, a drop from dilution 1 has a proportion $\alpha_p^{-1}\alpha^{-1}$ of CFUs from dilution 0 and dilution j has a proportion α^{-1} of CFUs from dilution

$j-1$. Moreover, only a small proportion q (maybe less that 5%) of CFUs fail to make countable colonies.

Moreover, use of the binomial permits direct estimation of the number of CFUs (N_j^k's) with no need for scaling and alleviates the main issues of the other models: over- and under-dispersion in the Poisson case and the use of substitution rules for 0's and TNTCs in the log-normal case.

To model all repetitions in a single experiment, and because it is common to consider log abundances when N_0^k is large, we take the usual approach of constructing a model linking all log abundances $\log_{10}(N_0^k+1)$, $k=1,2,\ldots,K$, as realizations from a population having a common mean E with some dispersion. This again reflects/models what is being done in the laboratory: The intention of performing repetitions is to try to estimate the (log) abundance of the experiment itself, and asses its variability by performing K repetitions under repeatable conditions. Accordingly, E is interpreted as the mean log abundance for a single experiment, which we will infer, and using a Bayesian approach, it will be taken as a r.v. and its posterior distribution estimated. However, before describing the details of the latter, we first comment on the following two points.

First, as opposed to usual practices, from the onset we add 1 when calculating the log abundance to properly define it when $N_0^k=0$, that is, when there are only zero CFUs as occurs when no microbes survive an antimicrobial treatment. Note that for $N_0^k \geq 10$ (as almost always occurs for control samples), $\log_{10}(N_0^k+1)$ is already nearly the same as $\log_{10}(N_0^k)$, so the interpretation of the log abundance defined as $\log_{10}(N_0^k+1)$ should be straightforward for these large values of N_0^k. When N_0^k is closer to 0, then a log abundance is less useful. Instead, one can examine directly N_0^k and its posterior for statistical inference. Still, Taylor's Theorem shows that $\log_{10}(N_0^k+1) \approx N_0^k$ for small N_0^k, which suggests defining the log abundance as $\log_{10}(N_0^k+1)$ even in this case. Moreover, this definition will permit us not only to be consistent in all cases, but also to discuss the limit of detection (LOD) when no counts are detected; see Sect. 2.2.

Second, it is common to normalize abundances to the volume or surface area of the original specimen, S_c, via N_0^k/S_c. We recommend applying this normalization to convert log abundances to log densities by $\log_{10}\left(N_0^k+1\right)-\log_{10}(S_c)=\log_{10}\left(\frac{N_0^k+1}{S_c}\right)$. This brings up an interesting point that is not well appreciated by microbiologists. Changing the units via S_c clearly changes the mean log density but leaves the variance unchanged (due to the multiplicative property of the log transform). For example, for biofilm samples, when changing the units from CFU/mm^2 to CFU/cm^2, S_c decreases by a factor of 100, so that the mean log density increases by 2 but the standard error of the sample mean (SEM) and the SD of the individual log densities remain unchanged. Hence, any frequentist hypothesis test of the population mean of $\log_{10}\left(N_0^k+1\right)$ that depends on a t-ratio of the sample mean to the SEM will always become "statistically significant" for a drastic enough change in units. This issue is mitigated when considering the log abundance of microbes compared to a concurrent control via a log reduction (LR), as occurs when assessing antimicrobial treatments. This is because the LR is unitless.

Returning to the experimental mean log abundance E, the first and nearly default modeling approach would be taking

$$\log_{10}\left(N_0^k+1\right) = E + \epsilon_k; \epsilon_k \sim N(0,\sigma) \tag{3}$$

or $\log_{10}\left(N_0^k+1\right) \sim N(E,\sigma)$. This model might indeed be appropriate for relatively small σ, to maintain $\log_{10}\left(N_0^k+1\right)$ positive, but the Gaussian model is certainly not well suited in general. To make a more robust modeling approach, first assume that

$$\mathbb{E}\left[\log_{10}\left(N_0^k+1\right)\middle|\ E=e\right] = e; \quad \text{for } k=1,2,\ldots,K \tag{4}$$

i.e., $\mathbb{E}\left[\log_{10}\left(N_0^k+1\right)\right] = E$. Second, we introduce A as a dispersion parameter, to stipulate the model

$$\log_{10}\left(N_0^k+1\right)\middle|\ E=e, A=a \ \sim \text{Ga}(a, ea^{-1}). \tag{5}$$

Here, we use the parametrization for the Gamma distribution $\text{Ga}(a,b)$ where b is the 'scale' parameter, and therefore, the expected value above is precisely e as required. Moreover, its standard deviation is $\frac{e}{\sqrt{a}}$ and its signal-to-noise ratio (i.e., mean over standard deviation or the inverse of the coefficient of variation) is $\sqrt{a}$, representing the unitless dispersion in the model which is specially well suited for positive r.v.'s such as $\log_{10}\left(N_0^k+1\right)$. Note also that the above gamma model correctly generalizes the Gaussian model (3) since for large A (e.g., $A>30$) the gamma distribution is already close to a Gaussian and $\log_{10}\left(N_0^k+1\right) \dot{\sim} N(E, \frac{E}{\sqrt{A}})$ and this is precisely the case when we have relatively small variances (low coefficient of variation), making the Gaussian model appropriate. Then simply, (5) generalizes the default Gaussian model in (3) to positive only values, only using a different, and perhaps better suited, parametrization.

The specification above in fact creates a hierarchical model and using a simple well-known result we may integrate out the N_j^k; $j=0,1,\ldots,J-1$ and the binomial model in (1) becomes

$$Y_{j,i}^k | N_0^k = n_0^k, q \ \sim \text{Bi}(n_0^k, \alpha^{-j}\alpha_p^{-1}\alpha_0^{-1}(1-q)). \tag{6}$$

See further details on this and other technical elements of our model in "Appendix A".

We use a Bayesian approach to make inferences for the parameters of interest by first stating prior distributions and performing an MCMC to sample from the posterior.

We require prior distributions for E, A, $N_0^1, N_0^2, \ldots, N_0^K$. We take a pragmatic approach in which the prior for N_0^k is a discrete uniform distribution from 0 to 10^M, $E \sim U(0,M)$ and $A \sim \text{Exp}(b)$, b is a scale parameter. 10^M is a maximum physical capacity of CFUs for the surface with surface area S_C or the volume with volume S_c, etc., put to treatment. In engineered reactor systems especially, experimentalists know the maximum microbial abundance in a sample (i.e., they know M). Indeed, they choose which dilutions to plate based on this knowledge. A is the shape parameter of the Gamma conditional distribution $f_{N_0^k|E,A}(n_0^k|e,a)$ in (5). A simple approach is to take the prior for A as exponential, resulting

in most of its mass from 0 to $2b$. In the drop plate examples below, we use $M = 10$ and $b = 500$. This prior parameter for A ($b = 500$) was calibrated such that each $\log_{10}\left(N_0^k + 1\right)$ had nearly the same prior as E, that is a $U(0, M)$ [i.e., a priori $f_{N_0^k|E,A}(n_0^k|e, a)$ is approximately $U(0, M)$ for all k; this may be done calibrating b with simulated samples from (5)].

We use the t-walk (Christen and Fox 2010) to produce a MCMC algorithm to sample from the resulting posterior distribution. The t-walk is a self-adjusting MCMC algorithm that requires the log posterior and two initial points. In the resulting MCMC in all examples, we typically generated chains of length 500,000 with an IAT (Geyer 1992) of 50 leading to an effective sample size of roughly 10,000. In our Python implementation the corresponding computations took 50 s on a 2.2 GHz processor. By simulating initial values for each N_0^k from its 'free' posterior (see "Appendix A") the burn-in resulted very short indeed in most cases. The initial value for E is taken as the mean of the $\log_{10}(N_0^k + 1)$s, and the initial value for A is taken by simulating from its prior distribution. Overall, the MCMC is very robust, working nearly unsupervised in all the examples we tested, including all those presented here.

2.1. GOODNESS OF FIT OF THE BINOMIAL MODEL

Our binomial model simply follows what is actually done in the laboratory, and therefore, we claim it models dilution series count data correctly. For example, it accounts for all extreme or censored data cases. However, there may be unaccounted sources of variability that could question the appropriateness of the binomial model. To support our claim that our approach is an overall better model for making the correct inferences, we here compare it to an alternative model. Obvious models to compare the binomial model with include the Poisson, negative binomial or generalized Poisson. The generalized Poisson or negative binomial are both Poisson mixtures (Joe and Zhu 2005) and would be good candidates for a comparison, but it is not clear how to model the dilution series, as we did in (1) and (2) with respect to the parameter of interest N_0^k (the abundance in the original sample) to provide a fair comparison with our model. Certainly alternative definitions can be attempted to use some Poisson mixture as a model for dilution data, but a straightforward generalization of our binomial approach is a beta-binomial distribution.

The beta-binomial distribution is the result of a mixture of a binomial and a beta distribution, and as such is a generalization of the binomial distribution. Namely, if $Y|S = s \sim \text{Bi}(n, s)$ and $S \sim \text{Beta}(\alpha, \beta)$, then $Y \sim \text{BetaBinomial}(n, \alpha, \beta)$. That is, this generalizes the binomial distribution by letting the success probability to be random, resulting in a variance that is independent of the mean, whereas for the binomial model, its expected value over variance is always > 1.

It is clear now how to extend our binomial model in (6), namely,

$$Y_{j,i}^k|N_0^k = n_0^k, S = s \sim \text{Bi}(n_0^k, s) \text{ and } S \sim \text{Beta}(s^*\lambda, (1 - s^*)\lambda),$$

where $s^* = \alpha^{-j}\alpha_p^{-1}\alpha_0^{-1}(1 - q)$. Here $E(S) = s^*$ as required, for any $\lambda > 0$. That is, instead of setting $Y_{j,i}^k|N_0^k = n_0^k \sim \text{Bi}(n_0^k, s^*)$ as in (6) we let s^* become a parameter, the

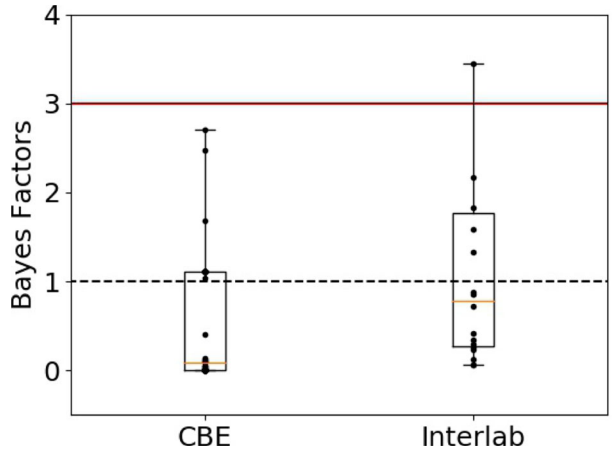

Figure 2. Boxplots of the posterior odds (Bayes Factors, BF) comparing the beta-binomial versus the binomial model, except for 3 cases with BF $>$ 4. From the 69 data sets analyzed, only 4 had a BF $>$ 3 and 43 had BF $<$ 1. These results fail to indicate clear evidence in favor of the beta-binomial model but instead seem to favor the binomial model .

r.v. S, with $E(S) = s^*$, and in this way generalize the binomial model to a beta-binomial. Here λ is an additional hyperparameter for the above beta (prior) distribution for S.

We use the Bayesian model comparison machinery, calculating the posterior odds of the beta-binomial versus the binomial model, namely, the Bayes factors (BF) of the beta-binomial in favor of the binomial model (see Kass and Raftery 1995, for example). In either case, the posterior probability for N_0^k is discrete and its normalization constant is calculated with a sum; this normalization constant is calculated for each of the binomial and the beta-binomial models. The posterior probability of each model is proportional to its normalization constant, and the BF is the ratio of these posterior probabilities or in fact the ratio of the normalization constants of each model.

We still need to fix λ to stipulate the Beta prior distribution for S in each case. If $s^*\lambda < 1$, the prior mode for S is at zero and would bias the beta-binomial model toward zero CFU counts in all cases. To avoid this we need $s^*\lambda \geq 1$, that is $\lambda \geq (s^*)^{-1}$. We have seen that not restricting $s^*\lambda \geq 1$ leads to a BF of practically zero in favor of the beta-binomial model, besides cases with zero CFU counts, as expected (results not shown). Since in most cases $(s^*)^{-1}$ is quite large, then setting $\lambda = (s^*)^{-1} + 1$ represents a neutral choice.

We performed the Bayesian model comparison on the data sets mentioned in Sect. 3. For the 51 CBE and 18 inter-laboratory individual dilution series count data, we found only 4 BF's above 3: one for the former, with a BF = 15 (tube 3 of experiment 70 °C, 10 min), and three for the latter data set, with BF's of 15, 5.5 (from tube 1 of laboratory 5, tube 1 of laboratory 6) and 3.45. Leaving out those BF's above 4, the rest of the BF's are plotted in the boxplots in Fig. 2. We are using the common recommendation that a BF above 3 provides positive evidence again the default model (Kass and Raftery 1995). These results do not provide strong evidence for the beta-binomial over the binomial model. Moreover, most of the BF's were below 1 which is evidence in favor of the binomial model over the alternative, more general, beta-binomial.

2.2. CENSORED DATA, ZERO COUNTS AND LEVEL OF DETECTION

As explained in the introduction, the usual practice when counting CFUs is that, besides the actual dilution selected for counting, the other $Y^k_{j,i}$s are not recorded. Therefore, the CFUs in the first uncounted dilutions could be considered as censored data since we know that drops below the selected dilution are TNTC, and also CFU counts in drops above the selected dilution could also be counted, even if all zero. However, this added difficulty in recoding and analysis does not seem to add any substantial information (Hamilton and Parker 2010), but this will depend on the particular design, mainly on the dilution parameters α and α_p. We have confirmed this using our model and the usual drop plate design using a simulation study (see "Appendix B"); we do not discuss this possibility any further.

Accordingly, let $j_k = 0, 1, \ldots, J-1$ be the dilution selected for counting (i.e., the lowest countable dilution) for repetition k and to ease notation let $Y^k_i = Y^k_{j_k,i}$, the actual CFU count in the ith drop. The selected dilution is part of our experimental design and is taken as the first dilution such that $Y^k_{j,i} \leq c$ ($c = 30$ when drop plating, $c = 300$ when spread plating). If zero CFUs appeared even in the first tube, we let $j_k = 0$ and $Y^k_i = Y^k_{0,i} = 0$. Dealing with zero counts, that is $Y^k_i = 0$ (i.e., no CFU counts even at dilution 0), represents no problem and may be dealt with consistently since (6) leads to a likelihood from which a posterior may be calculated; see Fig. 3a.

In the case where even at the highest dilution still the CFU count is above the threshold c, that is $Y^J_{j,i} > c$ then the CFU is recorded as TNTC, and we may treat this as right censored data. The likelihood in this case is $P[Y^J_{j,i} > c | N^k_0 = n^k_0, q]$, and this extreme case can also be dealt with, although with a higher computational burden; it involves calculating the binomial cdf in the likelihood. An example of this saturated data is presented in Fig. 3b, c. In the real data examples presented in Sects. 3.1 and 3.2, no saturated counts are present.

Regarding the miscount parameter q, note that for usual drop plate experimental designs $V = 10$ ml, $\alpha = 10, \alpha_p = 1000$, the success probability in the binomial model is (6) $\alpha^{-j_k}\alpha_p^{-1}(1-q)$ which will be quite similar to $\alpha^{-j_k}\alpha_p^{-1}$ for reasonable miss count probabilities $q \leq 0.1$. The effect of q will be barely noticeable. For other experimental settings, the effect of q could be more important, in which case q could be also considered a (nuisance) parameter to be included in the posterior, with a tight prior for q small. However, it would be advisable to make an experimental design to learn about the miscount parameter q, with a reference sample with a known microbial abundance and several repetitions. However, to our knowledge, these hypothetical experiments have not been conducted as yet. In the examples presented in Sect. 3 we simply fix $q = 0.05$; this should barely have an effect, as seen in Fig. 3a. Nonetheless, q is included for overall consistency of our approach.

Regarding the lower limit of detection (LOD, Currie 1968), that is, the minimum number of CFUs that can be detected in dilution 0 with the chosen experimental design, we may calculate L_K such that $P(10^E - 1 < L_K | Y^k_0 = 0, k = 1, \ldots K) < 0.95$, for various repetitions K. That is, setting all drop counts to zero, we define the lower LOD as the 95% upper quantile L_K of the corresponding posterior distribution for the microbial abundance $10^E - 1$. In Fig. 4 we present examples of these posteriors for the drop plate design ($\alpha = 10, \alpha_p = 1000$, $D = 10$ drops) with $K = 1, 3$ and 12 repetitions. The results are $L_1 = 110$, $L_3 = 50$ and $L_{12} = 30$. A substantial increase in the lower LOD is obtained when

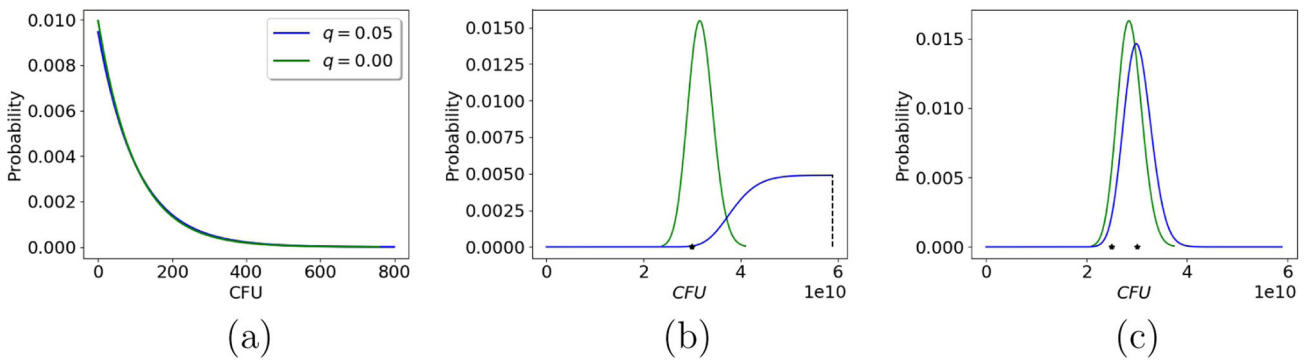

Figure 3. **a** Posterior distribution for the total CFUs in dilution 0, when zero CFUs are counted, $\alpha = 10$, $\alpha_p = 1000$ and $D = 10$ drops. A comparison is shown when the miscount probability is $q = 0.05$ or $q = 0.0$; little difference is observed in this experimental setting. **b** Example showing all TNTC at the highest dilution, ignoring the censorship setting $Y_i^k = c$ (green) and properly analyzing including censorship in the likelihood (blue). Note that in the latter, the posterior reaches a plateau, since *any* arbitrarily large value is equally likely, and is only truncated by the prior, i.e., a maximal feasible number of CFUs for the coupon surface. **c** Example of a hybrid case in which only three out of 10 drop counts where not TNTC, all close to $c = 30$, ignoring the censorship (green) and considering the censorship in the likelihood; already little differences are observed (Color figure online).

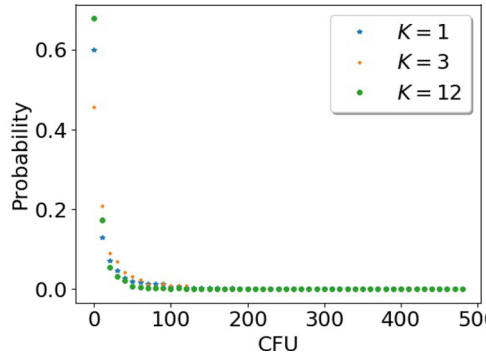

Figure 4. Posterior pmf for $10^E - 1$ (in 10 CFU bins), setting all drop counts to zero, with K repetitions. Drop plate design $\alpha = 10, \alpha_p = 1000$, $D = 10$ drops. Lower LODs, i.e., L_K such that $P(10^E - 1 < L_K | Y_0^k = 0, k = 1, \ldots K) < 0.95$, are $L_1 = 110$, $L_3 = 50$ and $L_{12} = 30$.

increasing from 1 to 3 repetitions, but the increase in the lower LOD is slower then onward, coinciding with the current practices of performing $K = 3$ repetitions per experiment. This same approach can be applied to similarly assess the upper LOD for any experimental design (a useful quantity to estimate when some drop counts are TNTC). We further comment on LODs in our approach in Sect. 4.

2.3. INTER-LABORATORY HIERARCHICAL MODEL ANALYSIS

If the treatment has also been repeatedly studied in L different laboratories, each laboratory will have an independent hierarchical E variable, denoted by $E_l; l = 1, 2, \ldots, L$. Analogously, we may define a global variable $\mathcal{E}$ and $\mathcal{A}$ where

$$E_l | \mathcal{E} = e, \mathcal{A} = a \ \sim \text{Ga}(a, ea^{-1}); l = 1, 2, \ldots, L \tag{7}$$

and again this generalizes the default Gaussian approach to positive values. This also describes what practitioners initially intend to do: infer an overall $\log_{10}(\text{CFU} + 1)$ across many laboratories and asses its variability. This constitutes an additional hierarchy, where now $\mathcal{E}$ represents the global mean for $\log_{10}(\text{CFU} + 1)$, for the experiment across laboratories. The t-walk can be used to generate a chain of posterior samples, as was the case for the hierarchical models within each laboratory. We use this approach in the inter-laboratory examples in Sect. 3.2.

3. EXAMPLES

All Python code and data from examples in Sects. 3.1 and 3.2 are available in the supplemental material.

3.1. INTRA-LABORATORY ANALYSIS

Data are taken from a series of experiments performed in the Center for Biofilm Engineering, Montana State University, MT, USA. Biofilms of *Sphingomonas paucimobilis* were

grown on a cylinder with surface area $S_c = 4.52$ cm^2 and then subjected to different temperatures for varying contact times. Each temperature and time combination represents a treatment, and each treatment was applied to $K = 3$ replicate biofilm samples, as described in Wahlen et al. (2016). By sonication the biofilm is harvested from the cylinders into $V_0 = 10$ ml of buffered water to form dilution $j = 0$. To begin a $\alpha = 10$-fold dilution series, next $\alpha^{-1}\alpha_0^{-1}V_0 = 1$ ml is taken from dilution 0 and then 9 ml of water is added to form dilution $j = 1$, etc., up to dilution $j = 6$. Ten drops of $u = 10\,\mu\text{l} = 10^{-3} \cdot 10\,\text{ml} = \alpha_p^{-1} \cdot V$ each are plated and placed at $36 \pm 2\,^\circ$C for 48 ± 2 h and in turn inspected for CFU formation. Therefore, we have $J = 7, \alpha_0 = 1, \alpha = 10, \alpha_p = 10^3$ and $D = 10$.

In Fig. 5 we present the results of our analysis of 4 treatments: room temperature for 15 min, 65 °C for 15 min, 70 °C for 10 min and 75 °C for 10 min. Also included are results from the classic simple analysis (using the log-normal approach) of calculating confidence intervals from a mean and a standard deviation on the estimated log abundances. The resulting intervals, in general, do not coincide with the variability in the posterior distributions and in some cases may result in confidence intervals that include negative values; see Fig. 5d.

A more extreme example is when most CFU counts are zero. This case is simply untreatable using a mean and standard deviation. For the 80 °C and 2-min treatment, two of the three repetitions had only zero counts (as mentioned in Sect. 2.2, we set $j_k = 0$ and all $y_i^k = 0, k = 1, 2$) and the third repetition had one drop with one CFU only at dilution 0, that is $j_3 = 0, y_1^3 = 1$ and $y_i^3 = 0, i = 2, 3, \ldots, D$. We show the corresponding posterior in Fig. 6b. Moreover, to appreciate the effect of the hierarchical model, we include the 'free' posterior distributions for the N_0^ks (see "Appendix A"), that is not considering the hierarchical model and each independent posterior for N_0^k based only on the likelihood based on (6); see Fig. 6a. Since for repetitions $k = 1, 2$ we only had zero counts there is a positive probability that $N_0^k = 0$, and results in that the mode of this free posterior is precisely at 0. However, since repetition $k = 3$ had one CFU, then this renders $N_0^k = 0$ logically impossible, and therefore, it has zero posterior probability at $N_0^k = 0$, see Fig. 6a.

The corresponding marginal posterior probabilities, now using the hierarchical model, seen in Fig. 6b, do not match completely with the former posteriors transformed to the $\log_{10}(\text{CFU} + 1)$ scale. This is indeed a result of the hierarchical model, and the shift in the individual marginal posteriors is a case of "borrowing strength" from one repetition to the other.

In many cases, it is desired to study the log abundance of a treatment with respect to its concurrent untreated control, that is the log reduction (LR). For example, it is common for antimicrobial products to make claims such as "kills 99.9%" of bacteria which corresponds to LR $= 3$. A considerable advantage of using a Bayesian approach is that the LR is analyzed explicitly through its constitutive parts (the controls and treated samples) which can open new possibilities for more comprehensive and goal-oriented analyses for dilution series experiments (see also the LR analysis in Sect. 3.2 and the 'activation probability' analysis discussed in Sect. 4). Namely, the same hierarchical model may be fitted to the control data obtaining a posterior sample for the hierarchical log abundance, which we call E_0 and inference regarding the LR is simply based on the posterior distribution of $E_0 - E$. Since we have MCMC posterior samples from both variables, obtaining a posterior

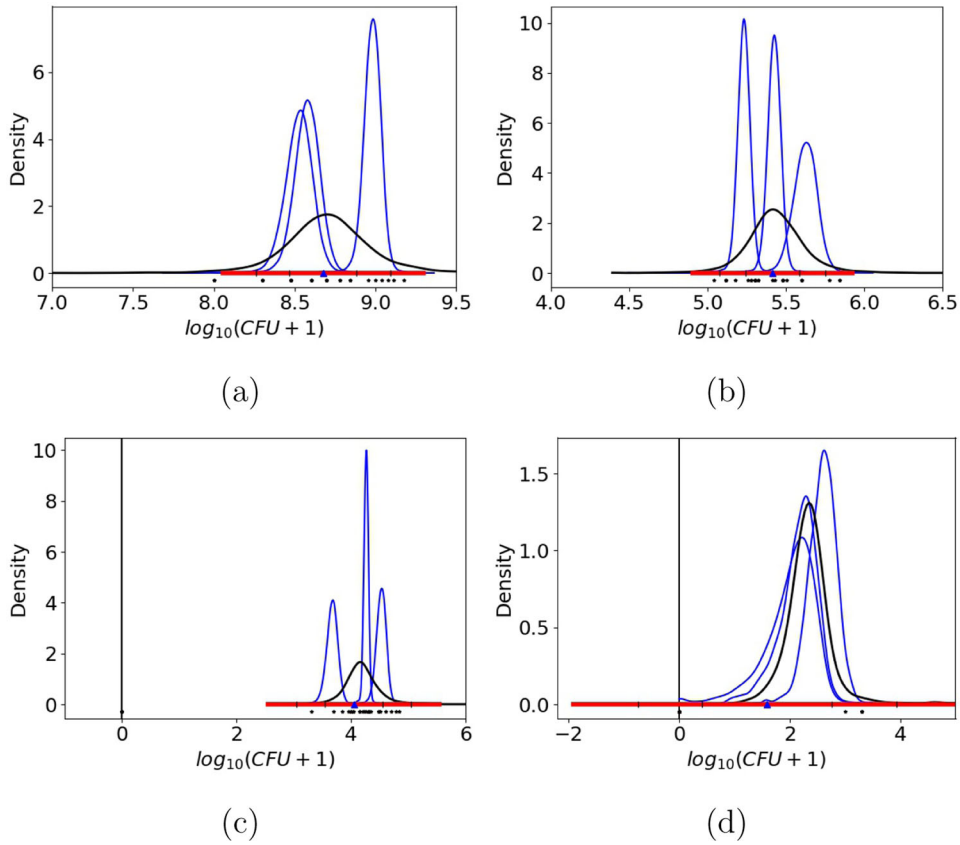

Figure 5. Posterior marginal distributions for E (thick black) and $\log_{10}(N_0^k + 1)$; $k = 1, 2, 3$ without hierarchical model (thin blue), along with the 3σ intervals and data, for: **a** Treatment room temperature for 15 min the 3σ interval seems over dispersed. **b** Treatment 65 °C for 15 min, the classic 3σ more or less coincides with the posterior variability. **c** Treatment 70 °C for 10 min, the 3σ interval looks clearly over dispersed. **d** Treatment 75 °C for 10 min the 3σ interval is wrong, covering negative values (Color figure online).

sample for $E_0 - E$ is immediate; see, for example, Fig. 6c. Moreover, we may calculate $P(LR > 3|Data)$ which in this case equals 0.9993, i.e., "with near certainty the 80 °C for 2 min treatment killed at least 99.9% of the biofilm".

3.2. INTER-LABORATORY ANALYSIS

In this example, we report on results from a multi-laboratory study of the so-called single tube method (ASTM 2013a,b), recently standardized by ASTM International (https://www.astm.org/), to test antimicrobial efficacy against biofilms of *Pseudomonas aeruginosa*. In this example, we focus on 3 laboratories from the inter-laboratory study of the single tube method, and a single experiment from each laboratory. In each experiment, we analyze $K = 3$ control samples and $K = 3$ samples treated with a low concentration of bleach (i.e., 123 ppm of sodium hypochlorite) against biofilms. The biofilms in the laboratories were harvested into a $V_0 = 40$ ml volume to form dilution

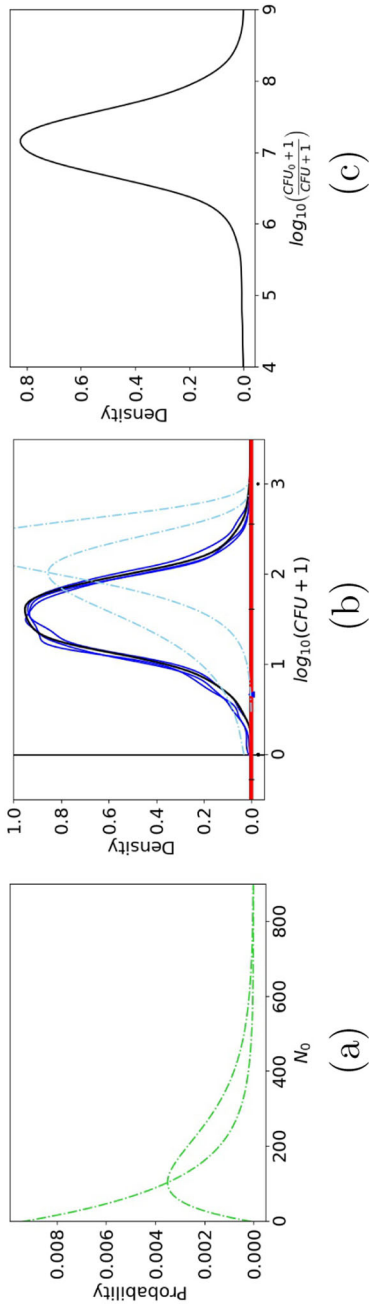

Figure 6. Results from treatment 80 °C for 2 min: **a** Individual ('free') posterior pmf's, without hierarchical model, for N_0^1 and N_0^2 (dashed, both identical) and N_0^3 (dashed and dotted) and **b** these same posteriors transformed to the $\log_{10}(\text{CFU}+1)$ scale as cdf's, the posterior marginal cdf for E (black) and for $\log_{10}(N_0^k+1)$ (blue solid lines) in the hierarchical model. **c** The log reduction of the experiment versus the control (room temperature); this is the posterior distribution of $E_0 - E$, where E_0 is the $\log_{10}(\text{CFU}+1)$ hierarchical parameter for the control experiment, which is in fact seen in Fig. 5a. Using this posterior distribution we calculate $P(\text{LR} > 3|\text{Data}) = 0.9993$ (Color figure online).

Table 1. Further investigating the reproducibility issues in our inter-laboratory example, according to our hierarchical posterior, the global log reduction $\mathcal{LR}$ is expected to be within 1.7 logs of the individual LR_ls for each laboratory (row 2). That is, the pooled hierarchical model seems to work well; however, its posterior distribution spans more that 5 orders of magnitude, as seen in Fig. 7c. This laboratory-to-laboratory inconsistency is also suggested in the individual posterior $P(LR_l > 3)$ (row 3), spanning from close to zero to $\frac{1}{2}$, while the inter-laboratory posterior LR is in fact $P[\mathcal{LR} > 3] = 0.0646$

Laboratories	$l = 1$	$l = 2$	$l = 3$
$E[\lvert\mathcal{LR} - LR_l\rvert]$	1.6443	1.4053	1.2654
$P(LR_l > 3)$	0.5348	0.0014	0.2257

0. As in the previous section, to begin a $\alpha = 10$-fold dilution series, $\alpha^{-1}\alpha_0^{-1}V_0 = 1$ ml is taken from dilution 0 and then 9 ml of water is added to form dilution 1, etc., up to dilution 6 (so there is a volume $V = 10$ ml in each dilution tube). Two drops of $u = 100\,\mu$l each are spread plated, and placed at $36 \pm 2\,^{\circ}$C for 26 ± 2 hr and in turn inspected for CFU formation. Therefore, we have $J = 7, \alpha_0 = 4, \alpha = 10$, $\alpha_p = 10/0.1 = 100$ and $D = 2$. Figure 7 presents the results of our analyses of the log densities ($\log_{10}(\text{CFU} + 1)$) for each of the control and bleach treatments, and also the LRs.

Unlike other fields of science where there is a demonstrable lack of reproducibility assessments, in the field of antimicrobial science, many organizations have pooled resources to quantifiably assess the reproducibility, across different laboratories, of methods that assess antimicrobial efficacy (Parker et al. 2018b). The reproducibility of test results of antimicrobials is of paramount importance for public health: regulators want to keep ineffective products out of the marketplace, and producers of highly effective products want to bring their products to market. All stakeholders want laboratory methods that accurately and reproducibly make decisions regarding which products are effective.

The reproducibility issue seems apparent in this example since the results from the LRs for this treatment, as seen in Fig. 7b, c span 4–5 orders of magnitude. In a classical setting, an analysis of variance (ANOVA) is performed on the dilution series data from multiple laboratories, but that inherits the problems of the log normal model already mentioned. In our setting, a formal reproducibility question may be put forward and then assessed with a posterior distribution. A first choice would be to compare the models for the mean log abundance E_l for each individual laboratory versus the full hierarchical model for the global mean log abundance $\mathcal{E}$, calculating the posterior distribution of each case, for example comparing the individual laboratory LR's $LR_l = E_l^0 - E_l$ versus the global LR $\mathcal{LR} = \mathcal{E}_0 - \mathcal{E}$; see Table 1 for a tentative discussion. This approach is one possibility, but a more in depth analysis is needed to address other reproducibility questions that stakeholders may phrase. The best approach is to establish the necessary posterior probabilities and, to provide guidance for the involved decisions to be taken, to maximize posterior expected utilities. Our Bayesian setting opens up these possibilities, but not without further effort.

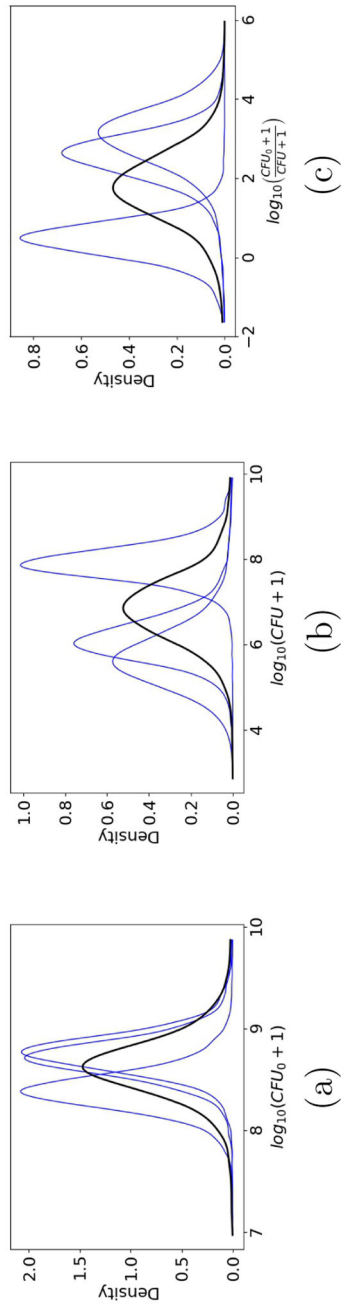

Figure 7. Inter-laboratory analysis of results for sodium hypochlorite treatment, individual E_ls for 3 laboratories not considering the hierarchical model (blue) and the global $\mathcal{E}$ (black) for the hierarchical model in 7 **a** control, **b** treatment and **c** LR (Color figure online).

4. DISCUSSION

A key aspect of our model is the use of a binomial likelihood ($\text{Bi}(N_0, \cdot)$) with the parameter of interest being N_0, the abundance of microbes in the original sample. We prefer the binomial likelihood because it models what the technicians actually do in the laboratory. This is in stark contrast to other common approaches (e.g., the log-normal and Poisson approaches) that provide only an approximation, and markedly different from the case where the statistician imposes an experimental design solely for convenience of the statistical analysis.

This work provides contributions to both the statistical and applied sciences. To the former, we show how to apply a Bayesian hierarchical model with a binomial likelihood to estimate and quantify uncertainty about microbial abundances (N_0) from dilution series experiments. The binomial likelihood has a rich history analyzing microbiological data in the case when, instead of CFUs, only a presence/absence response is measured from each tube in a dilution series from which N_0 can be estimated using MLE theory (McCrady 1915; Cochran 1950; Garthright 1993). Interesting, this MLE approach is called the "most probable number" technique, coined by McGrady in 1915 before MLE theory had been developed (Hamilton 2011). Our contribution is the first time to our knowledge that CFU data have been modeled with a binomial likelihood. Regarding our contribution to the applied sciences, we provide a sound alternative to the log-normal and Poisson modeling approaches that are commonly applied in the analysis of CFU data (Hamilton et al. 2013; Haas et al. 1999). We have shown that these common approaches have serious deficiencies when modeling CFUs. For example, in the Poisson case, the restriction on the variance does not hold for CFUs collected from control conditions (that tend to exhibit high abundances and low variability); and in the log-normal case, it is not possible to deal with 0's and TNTC's (without *ad hoc* substitution rules). The negative binomial or generalized Poisson models have been used to extend the Poisson model (Joe and Zhu 2005). Instead, we present a Bayesian approach with a binomial likelihood that allows us to directly estimate the abundance of microbes without a severe constraint on the variance like the negative binomial or generalized Poisson would do, and unlike these approaches, our approach can directly incorporate data from multiple dilutions, directly analyze log reductions from the CFUs recovered from control and treated samples that have different levels of variability, and account for miscounts (recently a shifted Poisson model has been suggested to also handle miscounts Ben-David and Davidson 2014). Furthermore, while there is readily available software for application of mixed effects models with either a log-normal or Poisson likelihood to assess reproducibility (see, e.g., Bates et al. 2015), we are not aware of similar extensions for the negative binomial or generalized Poisson approaches. We show how our approach may be applied for reproducibility assessments, by comparing the posterior for each individual laboratory to the posterior for the population of all laboratories. While this is an area of active research, the results we have presented seem promising.

A point not directly addressed in this paper is the analysis and comparison of multiple treatments, since here we concentrated on the novel Bayesian binomial approach using hierarchies for multiple repetitions and multiple laboratories for a single treatment. The usual objective in analyzing a series of treatments is to fit a regression model to test and/or

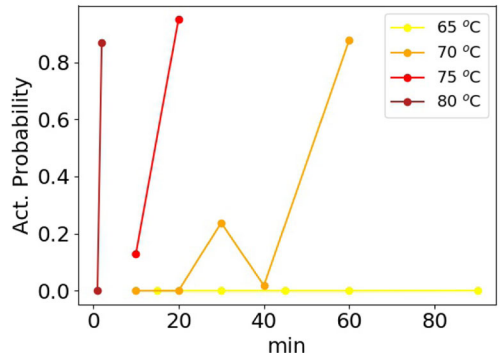

Figure 8. "Activation Probability" $P(E < e_h \mid \mathbf{Y} = \mathbf{y})$ with threshold $e_h = 2$, i.e., 100 CFUs, for the data described in Sect. 3.1. Coupons were treated at different temperatures and times; at 65 °C the log reduction was not achieved even over a 90 min exposure, whereas at 80 °C the Activation Probability is over 0.8 with an exposure time of only 2 min. There is a clear difference between 70 and 75 °C where similar results are obtained at 20 min at 75 °C and only after 60 min for 70 °C .

predict their effect (see, for example, Wahlen et al. 2016). One of the main benefits of performing a Bayesian inference is that we may ask as many questions of interest as we require regarding our parameter of interest, and these are answered with its corresponding posterior probability. A simple initial approach that exploits our Bayesian analysis would be to calculate the actual posterior probability of the desired result, and compare such posterior probabilities across treatments. For example, commonly in microbiological experiments we are interested in the number of surviving microbes after treatment. This translates in calculating an 'activation probability' of a treatment intended to kill microbial activity, i.e., $P(E < e_h \mid \mathbf{Y} = \mathbf{y})$ for e_h some (small) agreed threshold. If compared to a control E_0, then we could calculate the posterior probability of a log reduction above some threshold, that is $P(LR = E_0 - E > l_h \mid \mathbf{Y} = \mathbf{y}, \mathbf{Y}_0 = \mathbf{y}_0)$, etc., see Fig. 8 for an example. This approach lacks the formal predictive approach of fitting a regression model using covariates.

Although not trivial to do, a regression model may indeed be incorporated using a Bayesian approach, embedded in our hierarchical model and binomial likelihood, in a multi-experiment multi-laboratory scenario. These interesting possibilities are the focus of future research. In general, using a Bayesian approach opens the door to a formal uncertainty quantification (UQ) approach to the analysis of dilution series data, by exploiting the posterior distribution obtained and the ease in calculation of expectations or posterior probabilities given the MCMC sample obtained.

Instances of 0's and TNTCs relate directly to the lower and upper limits of detection (LOD) for the process used to generate CFUs (Currie 1968). The process includes how the dilution experiment was conducted such as the dilution factors (α_0 and α), which dilutions were plated (J), and the volume plated (u). The process also includes properties about the method used to harvest the bacteria from the sample (e.g., sonication, scraping, stomaching or a RODAC plate), the method used to disaggregate/homogenize the bacteria into single cells in the original volume V_0 and the environmental conditions used to incubate the bacteria after plating. In microbiology, it is common to refer to either $0.5/S_c$ or $1/S_c$ as the LOD of a CFU counting method. This is unfortunate since neither of these values necessarily

are associated with a high level of statistical confidence or probability regarding how many viable microbes survive in the sample. In our analyses (Fig. 4), we defined the LOD of the process used to generate CFUs as the lowest microbial abundance that can survive in the sample with probability 0.95 when there were no CFUs recovered from the sample. From the Bayesian analyses that we present, this calculation is straightforward: the LOD is simply the 95th percentile from the posterior distribution for $N_0|Y = 0$; see Figs. 3a and 4 . Our hierarchical model also easily and consistently can provide the LOD as a function of dilution experiment design (i.e., J, α_0, α and α_p) and the number of repetitions K by using the posterior for $10^E - 1$. For example, in Fig. 4, it is seen the expected result that the LOD decreases as the number of repetitions increases. The LOD for a particular experimental design can be calculated, beforehand, at minimal computational cost, to asses if the proposed design meets any LOD requirements.

Regarding the LOD or any other characteristic desired in an experiment, an intensive search can be conducted to optimize the design parameters in Bayesian analyses (Weaver et al. 2016; Huan and Marzouk 2014; Christen and Buck 1998). This, however, involves calculations of far higher costs, commonly needing parallel computing and other sophisticated software and numerical analysis resources. Given a design, parameters are simulated from the prior and in turn synthetic data sets from the model, to calculate the corresponding posterior for quantities of interest to assess the expected information gain in data from the design. This then needs to be repeated from a set of candidate designs to find an optimal design. We leave this interesting dilution experiment design possibility for future research.

ACKNOWLEDGEMENTS

JAC is partially funded by CONACYT CB-2016-01-284451, RDECOMM and ONRG grants. AEP is partially funded by NSF Grant 1516951. The authors gratefully acknowledge the industrial associates of the CBE, Lindsey Lorenz and Professor Emeritus Martin Hamilton.

[Received August 2019. Accepted May 2020.]

A. DETAILS ON THE HIERARCHICAL MODEL

Our hierarchical model may be summarized with the directed acyclical graph (DAG) depicted in Fig. 9a. As mentioned in the main text, using the following well-known result we may integrate out the N_j^k; $j = 0, 1, \ldots, J-1$. That is, if $Z \sim \text{Bi}(n, p_1)$ and $X|Z = z \sim \text{Bi}(z, p_2)$ then $X \sim \text{Bi}(n, p_1 p_2)$. Using it recursively (1) becomes for $j = 0, 1, \ldots, J-1$

$$Y_{j,i}^k | N_0^k = n_0^k, q \ \sim \text{Bi}(n_0^k, \alpha^{-j}\alpha_p^{-1}\alpha_0^{-1}(1-q)) \tag{8}$$

and the corresponding DAG is as shown in Fig. 9b. This substantiates Equation (6) provided earlier. Since considering $N_0^k = 0$ is important and may indeed happen, note that it must be the case that $P(Y_{j,i}^k = 0|N_0^k = 0, q) = 1$. Accordingly, we define $\text{Bi}(0, p)$ as the Dirac's delta at zero.

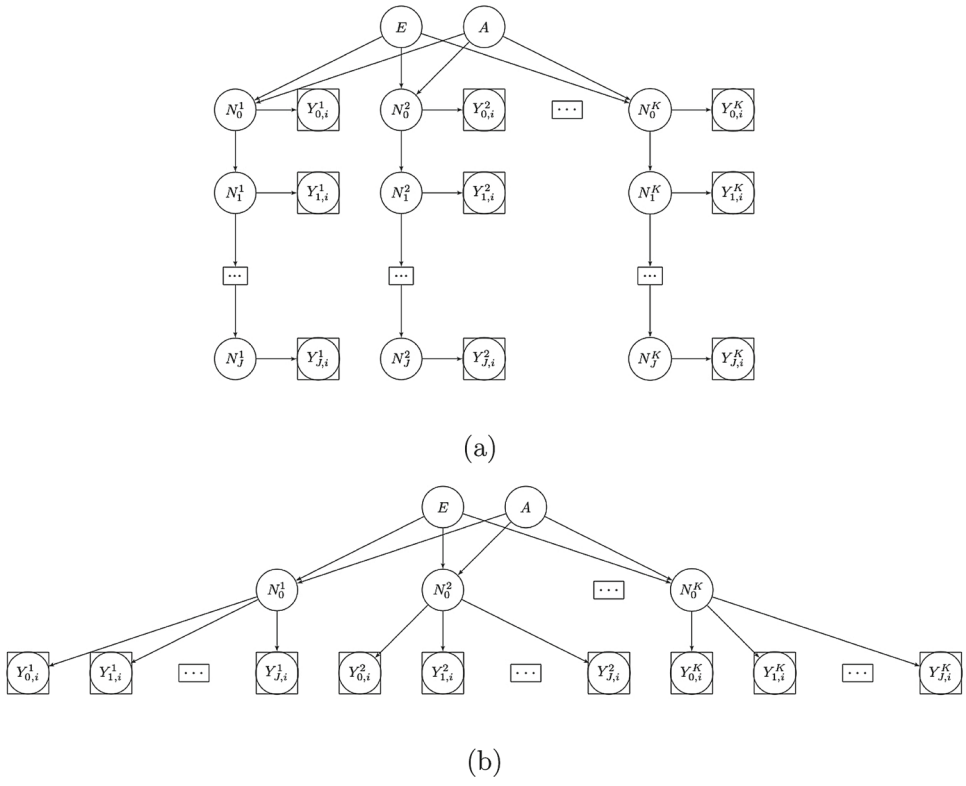

Figure 9. **a** Directed Acyclic Graph (DAG) representing our hierarchical model and **b** DAG representing our model once the N_j^ks have been integrated out, depending now only on the N_0^ks. Circle nodes are r.v.'s to be estimated, circle and square nodes are r.v.'s that are observables (the CFU counts) .

Note that, strictly speaking, $\log_{10}\left(N_0^k+1\right)$ is a discrete variable, and we are modeling it here as a continuous r.v. The binomial model may be changed to accept a non-integer "number of trials" N_0^k using gamma functions in the combination function, having as a particular case the conventional binomial pmf (this we do in our implementation providing no further details here). Then, for mathematical convenience $\log_{10}\left(N_0^k+1\right)$ is modeled as a continuous r.v., while still N_0^k is taken discrete.

Moreover, by ignoring the hierarchy involving E and A we may calculate the posterior distribution of each N_0^k independently. In this case, since it is a single discrete parameter, calculating its posterior pmf is straightforward, constructing a likelihood from (8). For illustration purposes and comparisons, this is done in Sect. 2.2 and in Fig. 6a. We call this the *free posterior* for N_0^k, the microbial abundance in the kth.

The full set of parameters are E, A, N_0^1, N_0^2, $\ldots$, N_0^K. Only binomial and Gamma densities are involved, and therefore, the corresponding likelihood is simple to write and calculate. The likelihood, indeed since repetitions are exchangeable, is

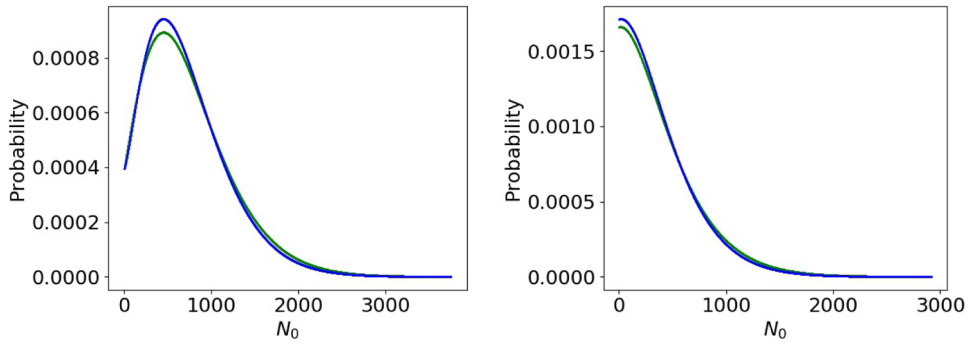

Figure 10. Expected posterior distribution when all dilution counts are considered (blue) and when only the first dilution counts are considered (green), for **a** drop plate design, with true value $N_0 = 500$, and **b** spread plate design, with true value $N_0 = 50$ (Color figure online).

$$f_{\mathbf{Y}|E,A,N_0^1,N_0^2,\ldots,N_0^K}(\mathbf{y}|e,a,n_0^1,n_0^2,\ldots,n_0^K) = \prod_{k=1}^{K}\left\{\prod_{i=1}^{D} f_{Y_i^k|N_0^k}(y_i^k|n_0^k)\right\} f_{N_0^k|E,A}(n_0^k|e,a) \tag{9}$$

where $f_{Y_i^k|N_0^k}(y_i^k|n_0^k)$ is the corresponding binomial pmf stated in (8) and $f_{N_0^k|E,A}(n_0^k|e,a)$ arises from (5), $\mathbf{Y}$ and $\mathbf{y}$ are the r.v.'s of observed CFU counts arranged in $K \times D$ matrices such that $\mathbf{Y} = (Y_i^k)$ and $\mathbf{y} = (y_i^k)$. $f_{N_0^k|E,A}(n_0^k|e,a)$ is established using the fact that if $\log_{10}(Z) \sim \text{Ga}(a,b)$, then $f_Z(z) = \frac{b^{-a}}{\Gamma(a)}(\log_{10}(z))^{a-1}y^{-(b\log(10))^{-1}-1}$.

B. MULTIDILUTION DATA ANALYSIS

We present results on whether it is worth analyzing CFU counts over all dilutions versus over only the first countable dilution. This question is difficult to answer in complete generality. We will then concentrate on a two common designs that correspond to data in Sects. 3.1 and 3.2 . Namely, $\alpha_0 = 1, \alpha = 10, \alpha_p = 1000, J = 6, c = 30$ for the drop plate data in Sect. 3.1 and $\alpha_0 = 4, \alpha = 10, \alpha_p = 100, J = 6, c = 300$ for the spread plate data in Sect. 3.2.

We study the extreme case when the first dilution can be counted, that is all counts are below the TNTC threshold c, and there is only one repetition ($K = 1$). We fix a true value for N_0^1 and simulate data at all dilutions using the binomial model in (8). Then, we calculate the discrete posterior pmf of N_0^1 and repeat the process averaging over all resulting posteriors over 120 simulated data sets. The process is done calculating the posterior when only the first dilution is used and when all dilution counts are considered in the calculation of the posterior. For the true value of the abundance N_0^1, we considered only a set of representative values for both designs, below a maximum to maintain expected simulated counts below c. Namely, $N_0^1 = 500$, 5,000, 10,000, 20,000, 30,000 for the drop plate data design and

$N_0^1 = 50, 500, 1{,}000, 2{,}000, 3{,}000$ for the spread plate data design. The posteriors with the most noticeable differences, although still quite low, were at $N_0^1 = 500$ and $N_0^1 = 50$, respectively; we present these posteriors in Fig. 10. In all other cases the differences in posteriors were even smaller (not shown), suggesting that the added difficulty in counting, recording and analyzing CFUs at all dilutions is not worth the expected information gain, and therefore, we recommend only to record and analyze the first countable dilution.

The reason that the posteriors are so similar can be seen in Fig. 3. The likelihood for TNTCs in dilution j becomes flat and provides no information for higher values than $2c\,\alpha^j\alpha_p$. If there are countable drops for dilution $j+1$, the likelihood for these data is placed at α-fold higher values well beyond $2c\,\alpha^j\alpha_p$ and therefore including the TNTCs likelihood in dilution j will not have any significant effect on the posterior, at least for the common case where $\alpha = 10$. Then including the censored likelihood terms for TNTC's will not add substantial information and does not justify the added computational cost.

REFERENCES

ASTM (2013a). An interlaboratory study was conducted by eight laboratories testing three samples of varying targeted results to establish a precision statement for test method e2871. ASTM International, West Conshohocken, PA, USA, https://www.astm.org/DATABASE.CART/RESEARCH_REPORTS/RR-E35-1008.htm.

— (2013b). Standard test method for evaluating disinfectant efficacy against pseudomonas aeruginosa biofilm grown in cdc biofilm reactor using single tube method. ASTM International, West Conshohocken, PA, USA, https://www.astm.org/Standards/E2871.htm.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1):1–48.

Ben-David, A. and Davidson, C. E. (2014). Estimation method for serial dilution experiments. Journal of Microbiological Methods, 107:214–221.

Christen, J. and Fox, C. (2010). A general purpose sampling algorithm for continuous distributions (the t-walk). Bayesian Analysis, 5(2):263–282.

Christen, J. A. and Buck, C. E. (1998). Sample selection in radiocarbon dating. Journal of the Royal Statistical Society: Series C (Applied Statistics), 47:543–557.

Clarke, J. (1998). Evaluation of censored data to allow statistical comparisons among very small samples with below detection limit observations. Environmental Science and Technology, 32:177–183.

Cochran, W. G. (1950). Estimation of bacterial densities by means of the "most probable number". Biometrics, 6:263–282.

Cundell, T. (2015). The limitations of the colony-forming unit in microbiology. European Pharmaceutical Review, 20(6):11–13.

Currie, L. (1968). Limits for qualitative detection and quantitative determination. Anal. Chem., 40:586.

EPA (1996). Guidance for data quality assessment. Office of Research and Development, US Enivironmental Protection Agency.

FDA (2018). BAM 4: Enumeration of *Escherichia coli* and the Coliform Bacteria. US Food and Drug Administration, https://www.fda.gov/food/laboratory-methods-food/bam-4-enumeration-escherichia-coli-and-coliform-bacteria.

Garthright, W. E. (1993). Bias of the logarithm of microbial density estimates from serial dilutions. Biometrical Journal, 35(3):299–314.

Geyer, C. (1992). Practical markov chain monte carlo. Statistical Science, 7(4):473–511.

Haas, C., Rose, J., and Gerba, C. (1999). Quantitative Microbial Risk Assessment.

Haas, C. N. and Scheff, P. A. (1990). Estimation of averages in truncated samples. Environental Science Technology, 24:912–919.

Hamilton, M., Hamilton, G., Goeres, D., and Parker, A. (2013). Guidelines for the statistical analysis of a collaborative study of a laboratory disinfectant product performance test method. JAOAC International, 96(5):1138–1151.

Hamilton, M. A. (2011). The P/N formula for the log reduction when using a semi-quantitative disinfectant test of type SQ1. Center for Biofilm Engineering.

Hamilton, M. A. and Parker, A. E. (2010). Enumerating viable cells by pooling counts for several dilutions. Center for Biofilm Engineering.

Hedges, A. J. (2002). Estimating the precision of serial dilutions and viable bacterial counts. International Journal of Food Microbiology, 76:207–214.

Herigstad, B. R. and Hamilton, M. (2001). How to optimize the drop plate method for enumerating bacteria. J Microbiological Methods, 44(2):121–129.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47:663–685.

Huan, X. and Marzouk, Y. M. (2014). Gradient-based stochastic optimization methods in Bayesian experimental design. International Journal for Uncertainty Quantification, 4(6):479–510.

Joe, H. and Zhu, R. a. (2005). Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. Biometrical Journal, 47(2):219–229.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. Journal of the American Statistical Association, 90(430):773–795.

Maturin, L. and Peeler, J. T. (1998). Chapter 3 – aerobic plate count, section: Conventional plate count method. In Council, B., editor, FDA Bacteriological Analytical Manual. US Food and Drug Administration.

McCrady, M. H. (1915). The numerical interpretation of fermentation-tube results. J. Infectious Disease, 17:183–212.

Niemela, S. (1983). Statistical evaluation of results from quantitative microbiological examinations. Nordic Committee on Food Analysis. Ord & Form AB, Uppsala.

Niemi, M. M. and Niemela, S. I. (2001). Measurement uncertainty in microbiological cultivation methods. Accreditation and Quality Assurance, 6:372–375.

Parker, A., Pitts, B., Lorenz, L., and Stewart, P. (2018a). Polynomial accelerated solutions to a large gaussian model for imaging biofilms: in theory and finite precision. Journal of the American Statistical Association, 113(524):1431–1442.

Parker, A. E., Hamilton, M. A., and Goeres, D. M. (2018b). Reproducibility of antimicrobial test methods. Scientific Reports, 8(1):12531.

Parkhurst, D. F. and Stern, D. A. (1998). Determining average concentrations of cryptosporidium and other pathogens in water. Environmental Science and Technology, 32:3424–34–3429.

Pitts, B. and Stewart, P. (2008). Confocal laser microscopy on biofilms: Successes and limitations. Microscopy Today, 16(4):18–21.

Prescott, L., Harley, J., and Kline, D. (1996). Microbiology. Wm. C. Brown Publishers, Dubuque, third edition edition.

Rice, E., Baird, R., and Eaton, A. (2017). Standard Methods for the Examination of Water and Wastewater. American Public Health Association, American Water Works Association, Water Environment Federation, 23rd edition edition.

Singh, A. and Nocerino, J. (2002). Robust estimation of mean and variance using environmental data sets with below detection limit observations. Chemometrics and intelligent laboratory systems, 60:69–86.

Wahlen, L., Parker, A., Walker, D., Pasmore, M., and Sturman., P. (2016). Predictive modeling for hot water inactivation of planktonic and biofilm-associated *Sphingomonas parapaucimobilis* to support hot water sanitization programs. Biofouling, 32(7):751–761.

Weaver, B. P., Williams, B. J., Anderson-Cook, C. M., and Higdon, D. M. (2016). Computational enhancements to Bayesian design of experiments using Gaussian processes. Bayesian Analysis, 11:191–213.

Zuur, A., Ieno, E., Walker, N., Savelliev, A., and Smith, G. (1999). Mixed Effects Models and Extensions in Ecology with R. Springer, New York